FRECUENCIA FONEMÁTICA DEL ESPAÑOL DE COLOMBIA*

PHONEMIC FREQUENCY OF COLOMBIAN SPANISH

*María Claudia González Rátiva***
Universidad de Antioquia, Colombia

*Jorge Antonio Mejía Escobar****
Universidad de Antioquia, Colombia

Artículo de investigación. Recibido 01-06-2011, aceptado 30-08-2011

* Una síntesis de este trabajo se presentó en el XVI Congreso Internacional de la Asociación de Lingüística y Filología de la América Latina (ALFAL), realizado entre el 6 y el 9 de junio de 2011, en Alcalá de Henares (Madrid). El resumen y las tablas allí presentadas son las mismas de este trabajo.

** claudiagonzalez@comunicaciones.udea.net.co

*** jamejia@udea.edu.co

Resumen

Este artículo presenta resultados de una investigación sobre la frecuencia de fonemas del español colombiano, con base en la aplicación del programa Cratilo® a dos corpus, literatura y trova, en esta variedad de habla. Un algoritmo fonético que estableció el uso de fonemas en el corpus y la realización de compensaciones basadas en la frecuencia léxica permitió obtener la frecuencia fonemática del español de Colombia, lo que complementa estrategias de la lingüística del corpus y la lingüística tradicional. Los resultados de frecuencia fonemática se contrastan con algunos otros estudios sobre otras variedades hispánicas, con el fin de generar una reflexión sobre el sistema fonológico del español.

Palabras clave: *frecuencia fonemática, lingüística del corpus, español, programa Cratilo®, lengua española.*

PHONEMIC FREQUENCY OF COLOMBIAN SPANISH

Abstract

This article presents results of a research project on the frequency of phonemes of Colombian Spanish, on the basis of the application of the Cratilo® software to two corpora, literature and "trova" (improvised popular song) of this speech variety. A phonetic algorithm establishing the use of phonemes in the corpus and compensations on lexical frequency made it possible to obtain the frequency of phonemes in Colombian Spanish, thus using strategies from corpus linguistics and traditional linguistics. Phonemic frequency findings were compared to results reported in other Hispanic speech varieties, in order to provide a reflection on the Spanish phonological system.

Keywords: *phonemic frequency, corpus linguistics, Spanish, Cratilo® program, Spanish language.*

Universidad Nacional de Colombia, Facultad de Ciencias Humanas, Departamento de Lingüística

Introducción

LA DESCRIPCIÓN DE la base articulatoria de una lengua es parte importante de su caracterización, su historia y los elementos fundamentales para su enseñanza. A pesar de ello, hay escasez de trabajos al respecto (Gil, 2007, p. 254). Un aspecto importante en el marco de la disposición articulatoria lo constituye el establecimiento de la frecuencia de fonemas[1] de la lengua o variedad de habla. La aplicación de los resultados ofrecidos por un estudio sobre frecuencia de fonemas incluye, entre otros, el campo de las tecnologías del habla y la enseñanza de la lengua (Pérez, 2003, p. 1). En este sentido, el objetivo de este artículo es presentar los resultados de un trabajo de investigación realizado sobre la frecuencia de fonemas del español de Colombia con el fin de determinar la tendencia de nuestra variedad de habla en la constitución de la cualidad fonética de uso general de la lengua.

Antecedentes

Sobre la frecuencia fonemática existen trabajos clásicos, anteriores a los elaborados con base en corpus electrónicos. La Tabla 1 sintetiza la información de algunos de los estudios realizados sobre la frecuencia de fonemas en lengua española (Moreno Sandoval, Torre, Curto & De la Torre, 2006; Pérez, 2003; Alcina & Blecua, 1975).

La Tabla 1 muestra que los estudios sobre frecuencia de fonemas han ido incrementando el número de fonemas y, en los últimos años, se han centrado en corpus orales. Además, se destaca que no se encuentra en la bibliografía un estudio sobre el español de Colombia. En este sentido, los resultados de este trabajo son un aporte a la comprensión del uso del español en nuestra región.

Algunos de estos trabajos explicitan los mecanismos manuales o computarizados de transcripción fonológica. Sin embargo, se da por sentado que la ortografía del español es fundamentalmente fonológica, ya que, en comparación con otras lenguas, hay bastante adecuación en la relación fonema-grafema. Como este trabajo incluye tanto corpus oral como escrito, en el apartado metodológico se explicará el algoritmo mediante el cual se convirtió la información ortográfica en información fonológica.

---

1 Según Gil (2007, p. 204), la frecuencia de fonemas es el recuento estadístico de la aparición de estos en una lengua. Un estudio sobre frecuencia de fonemas "develaría de qué orden es la contribución que la ocurrencia de cada segmento individual o de sus diversas combinaciones hacen a la conformación general del castellano".

Tabla 1. Tipo de corpus y número de fonemas analizados

Autores	Año del estudio	Corpus	N.º de fonemas	Variedad de habla española
Zipf y Rogers	1939	Escrito	5000	ibérica
Tomás Navarro	1946	Escrito	20.000	ibérica
Alarcos Llorach	1961	Escrito	- (25 cartas)	ibérica
Pierre Delattre	1965	Oral	-	ibérica
Lloyd y Schnitzer	1967	Escrito (listado)	- (252.404 sílabas)	ibérica
Guirao y Borzone	1972	Escrito y oral	252.404	argentina
Quilis y Esgueva	1980	Oral (conversación)	160.000	ibérica
Rojo	1991	Escrito	3'650.000 (ARTHUS)	ibérica e hispanoamericana
Guirao y García Jurado	1993	Oral	- 74.460 sílabas	argentina
Listerri y Mariño	1993	Oral	100.000	ibérica
Pérez	2003	Oral (lectura)	75.269	chilena
Pineda/Cuétara	2004	Oral	- (300.000 frases)	mexicana

Los resultados estadísticos de los estudios de la Tabla 1 se presentan en la discusión con el fin de incluir en el análisis los resultados sobre la variedad colombiana. Se establecerá un cuadro comparativo con los resultados de algunos de estos estudios para mirar el grado de similitud o diferencia en las estadísticas, según las variedades de habla española de estas investigaciones.

Una precisión conceptual

Consideramos el fonema como cada una de las unidades fonológicas que componen cada una de las palabras de una lengua (Ashby & Maidment, 2005, p. 138). Son, además representantes subyacentes de los respectivos alófonos o versiones del mismo sonido (Ashby & Maidment, 2005, p. 139). A través de la pronunciación, los hablantes internalizan el componente fonológico, el sistema de reglas y representaciones (Ashby & Maidment, 2005, p. 138). De esta manera, cada palabra hablada en la superficie contiene las unidades sonoras contrastivas del componente.

Tabla 2. Fonemas de uso del español hablado en Colombia

Consonantes							
Zona / **Modo**	**Bilabial**	**Labio-dental**	**Dental**	**Alveolar**	**Alveo-palatal**	**Palatal**	**Velar**
Oclusivo	p b		t d				k g
Fricativo		f		s		ʝ	x
Africado					k͡s	tʃ	
Nasal	m			n		ɲ	
Líquido lateral				l			
líquido vibrante				ɾ r			

Vocales			
	Anterior	**Central**	**Posterior**
Alta	i		u
Media	e		o
Baja		a	

Nota: **Los fonemas consonánticos sordos aparecen al lado izquierdo de la casilla; los sonoros, a la derecha.**

En general, la fonología recurre a las posibilidades articulatorias humanas para establecer el componente fonológico de una lengua. Los parámetros articulatorios tradicionales son suficientes para establecer la clasificación de los fonemas del español de uso en Colombia, presentado en la Tabla 2. Utilizamos la última notación revisada del Alfabeto Fonético Internacional, del año 2005[2] (Internacional Phonetic Asociation, 2011).

Se asume como fonema de uso en Colombia el sonido africado, alveopalatal sordo /t͡ʃ/, con pronunciación, escritura y aceptación social plena, y que, además, es considerado variante de prestigio en algunas comunidades urbanas del país (observaciones en estudio). También se toma en cuenta el tan extendido fenómeno del yeísmo, proyectado desde los trabajos dialectológicos basados en el Atlas Lingüístico Etnográfico de Colombia (ALEC) (Flórez, 1978), cuya solución para buena parte del área andina del país es /ʝ/, consonante fricativa palatal sonora.

Metodología

Con el fin de establecer la frecuencia de fonemas del español colombiano, se adecuan dos corpus transliterados mediante un algoritmo fonético y se generan concordancias entre símbolo y fonema en un programa de barrido léxico. De esta manera, se obtienen 5'682.417 muestras de fonemas.

Algoritmo fonético para el tratamiento fonológico de corpus transliterados

Teniendo en cuenta la relativa cercanía entre las unidades del componente fonológico y los grafemas que se utilizan para la escritura de la lengua, se diseñó un algoritmo fonético o la secuencia programática de órdenes para la adecuación del corpus al principio fonémico (Mosterin, 1981, p. 34): un fonema - un símbolo - un grafema, en este caso un fonema - un carácter.

La pauta general que se siguió para la programación fue el reemplazo de caracteres, en un orden preciso, que permitiera aprovechar el código alfabético latino, como se observa en el Anexo 3. De esta manera, las poligrafías de los fonemas y las polifonías de los grafemas (Mosterin, 1981, p. 187) del español[3] se equilibran y

2 La Asociación Fonética Internacional (2011) presenta, en varios de sus enlaces, el sonido africado palatal tal como aquí se presenta, sin barra de ligadura, debido a la posibilidad de unión estrecha de sus símbolos componentes. Contrariamente, no aparece el africado alveopalatal, por lo que se requiere tal diacrítico.

3 Ejemplo de poligrafía de fonema: /s/ > s, z, c, "sol, zapato, circo"; y de polifonía de grafema: g > /gáto/, /xénte/.

cada uno de los fonemas de uso del español colombiano está representado en el corpus para análisis de sus frecuencias con un solo carácter gráfico. Se creó así lo que llamamos en esta investigación el corpus del lector, un corpus adaptado a un *software* de barrido léxico que pasa de ser una transliteración ortográfica, a un corpus de correspondencias fonológicas.

Corpus[4]

Los textos o corpus de lengua corresponden a 15 novelas colombianas del siglo XX y a 15 sesiones finales del Festival de la Trova de la ciudad de Medellín entre los años 1977-1999.

En el corpus escrito aparecen 14 obras de gran renombre en la historia de la literatura colombiana, como *Cien años de soledad*, *La otra raya del tigre* e *Ilona llega con la lluvia*. El listado de obras que integran este corpus escrito se presenta como Anexo 1. Las obras fueron escritas por trece autores colombianos de distintas regiones[5]. Es interesante anotar que, dentro de la muestra escrita, aparecen obras como *La marquesa de Yolombó*, que en algunos fragmentos aspira a reflejar de manera escrita el español hablado. Ante obras de este tipo, los hablantes generan procesos de identidad, ya que constituyen variantes regionales de lengua.

El texto oral, transliterado ortográficamente, está representado por la *trova*[6], competencia cultural de poesía popular cantada o canto repentista con acompaña-

4 Los textos transliterados ortográficamente son parte de la base de datos del programa computarizado Cratilo®. El corpus de obras literarias fue preparado en principio para hacer estudios sobre el contenido de la novela del siglo XX en Colombia. El corpus de trova fue preparado para una investigación de la Maestría en Literatura Colombiana de la Universidad de Antioquia, hecha por John Fredy Zapata Morales, quien sustentó su trabajo en 2005 y ahora es profesor de la Universidad Industrial de Santander, en Colombia.

5 Uno bogotano, uno santandereano, uno vallecaucano, uno caldense, cuatro costeños y cinco antioqueños.

6 Para Zapata, "La trova antioqueña es una forma de poesía oral, improvisada que, a manera de contienda, se difundió por gran parte de la región cafetera colombiana de influencia paisa, durante el periodo de la Colonización antioqueña y fue traída —a causa de los procesos migratorios vividos a mediados del siglo XX en Colombia— a la ciudad de Medellín; su forma más difundida es la que se conoce con el nombre de trova sencilla —una cuarteta octosílaba rimada en los versos pares, acompañada con tiple, en la cual dos trovadores —o troveros— improvisan alternadamente sobre un tema que puede ser libre o impuesto por un jurado cuando se canta en festivales o a iniciativa del público o del trovador cuando se canta en otras circunstancias como fiestas, actos cívicos, ceremonias, y demás celebraciones sociales" (2010, p. 132).

miento (Bermúdez, 2006, p. 92), en la cual dos "trovadores" improvisan rimas sobre temas que deben enlazar y acompañar con música de tiple. La trova, enmarcada en una estructura de verso, hace parte del habla espontánea y de manifestaciones culturales propias del ámbito nacional (Zapata, 2010).

Este corpus oral está tomado del Festival de la Trova, evento que se realiza anualmente en Medellín, durante la Feria de las Flores. Contiene la transliteración de quince sesiones finales entre 1977 y 1999. Como es una competencia a modo de enfrentamiento, siempre ocurre entre dos personas que pueden ser de diversa procedencia, pero, en su mayoría, son de la región antioqueña o del antiguo Caldas, zona que se ha caracterizado por cultivar esta manifestación cultural. Las sesiones que componen este corpus se registran en el Anexo 2.

De esta manera, la muestra oral y la escrita se complementan y los datos estadísticos que surjan de estos dos corpus darán cuenta de sus semejanzas y diferencias. El corpus escrito alcanza un total de 5'606.996 muestras de fonemas; el corpus oral tiene 75.421 muestras de fonemas. En total, el corpus seleccionado para realizar el estudio de frecuencia de fonemas en el español colombiano alcanza las 5'682.417 ocurrencias de fonemas.

El programa o software Cratilo®[7]

El programa utilizado para el análisis del corpus fue Cratilo®. Es un programa computarizado que permite etiquetar con un dispositivo abierto para un total de cien categorías y, además, facilita la realización de consultas a la concordancia generada, que funciona como base de datos. La aplicación de este programa permite un seguimiento cuantitativo y cualitativo de datos lingüísticos. Cuantitativo, porque el procesamiento produce estadísticas de uso, en nuestro caso en términos de porcentajes de frecuencias[8], de todas las formas gráficas; cualitativo, porque podemos mirar en formato de concordancia todos los datos que responden a un formato

7 Cratilo® es producto de una investigación realizada en la Universidad de Antioquia para la creación de un software de concordancias universales de textos. Fue desarrollado por J. A. Mejía, F. J. Álvarez y J. A. Sánchez y financiado por el Comité de Investigaciones (CODI) de la Universidad de Antioquia. Procesa textos escritos en los caracteres del inglés, francés, alemán, italiano, portugués, latín y español.

8 Los estudios sobre frecuencia de fonemas se han presentado en su mayoría en porcentajes de frecuencias de uso, lo que los hace comparables. Estudios como los de Rojo (1991) y Pérez (2003) no incluyen más procedimientos estadísticos.

de consulta para verificar que no se produzcan distorsiones o que no intervengan factores que no habían sido tenidos en cuenta inicialmente.

El uso de este programa como recurso de análisis lingüístico nos sitúa en un punto intermedio entre lingüística del corpus y la lingüística tradicional o intuitiva. La primera se dedica al análisis estadístico de grandes cantidades de datos, mientras que la lingüística intuitiva se apoya en un análisis cualitativo, basado principalmente en la percepción del usuario. Nuestra perspectiva está en una línea media porque partimos de la percepción cualitativa y la confrontamos con los datos proporcionados por el programa Cratilo®, que nos permite manejar y cualificar almacenes masivos de información.

Para Rojo, la recolección de materiales diferencia la lingüística del corpus de otras líneas de análisis del lenguaje, ya que en ella el lingüista "lanza una búsqueda sobre un conjunto de textos y recibe, como resultado de una búsqueda 'ciega' llevada a cabo por la máquina, todos los casos —a veces en cantidades realmente aplastantes— que responden formalmente a lo que ha solicitado" (2002, p. 4). Esta metodología trae como ventaja el poder considerar todos los datos pertinentes del corpus y analizar el comportamiento de un fenómeno de lengua, además de tener a la mano su factor estadístico, datos que tienen una importante cabida en las explicaciones de la gramática[9].

Nuestro propósito en este trabajo ha sido evitar que la ceguera predomine, al tener la posibilidad de generar condiciones para la reorganización del cuerpo base de trabajo con fundamentos gramaticales; es decir, reconstruir el universo estudiado a través de una mediación entre la teoría y la realidad, evitando así la "dependencia empírica" (Caravedo, 1999, p. 20). Esto nos permite, por ejemplo, revisar la relación entre las frecuencias y algunas categorías gramaticales para mirar su influencia en los resultados porcentuales. De ese modo, aspiramos a tener y presentar una mejor comprensión del fenómeno estudiado: la frecuencia de fonemas, que es la base para el establecimiento de las tendencias o hábitos articulatorios de la variedad de habla.

9 Desde la década de 1960 se han realizado trabajos lingüísticos con base en corpus electrónicos, que, a medida que avanza la tecnología, van creciendo en número de formas y muestras. Así, el *Brown Corpus* contenía 500 muestras de unas 2.000 palabras (Rojo, 2008) y, por ejemplo, el Corpus 230 consiste en 344.619 frases, 235.891 unidades léxicas y alrededor de 15 millones de palabras, cuya fuente es la web.

Resultados

El programa Cratilo® ofrece en primera instancia el cálculo de frecuencias presentado en la Tabla 3.

Tabla 3. Frecuencia de fonemas en dos corpus de español colombiano

Fonema	Frecuencia corpus trova	Porcentaje	Fonema	Frecuencia corpus Literaio	Porcentaje
e	10.868	14,41%	a	768.449	13,71%
a	9.925	13,16%	e	753.505	13,44%
o	8.223	10,90%	o	538.730	9,61%
s	6.221	8,25%	s	519.082	9,26%
i	5.318	7,05%	i	405.301	7,23%
n	5.153	6,83%	n	395.988	7,06%
ɾ	3.856	5,11%	ɾ	327.330	5,84%
l	3.577	4,74%	l	287.685	5,13%
t	3.427	4,54%	d	285.426	5,09%
k	3.323	4,41%	t	226.999	4,05%
d	2.854	3,78%	k	212.496	3,79%
m	2.527	3,35%	u	179.151	3,20%
u	2.193	2,91%	m	166.812	2,98%
b	2.155	2,86%	b	160.543	2,86%
p	2.090	2,77%	p	140.749	2,51%
g	889	1,18%	g	52.575	0,94%
ɟ	879	1,17%	x	41.071	0,73%
r	600	0,80%	r	40.571	0,72%
x	518	0,69%	ɟ	37.280	0,66%
f	363	0,48%	f	34.634	0,62%
tʃ	266	0,35%	tʃ	14.175	0,25%
ɲ	162	0,21%	ɲ	12.113	0,22%
k͡s	34	0,05%	k͡s	6.331	0,11%

Se destaca de la Tabla 3 que, aun teniendo en cuenta la gran diferencia numérica entre los dos corpus, el orden de frecuencia de fonemas coincide en 14 de los 23 fonemas, como lo muestran los segmentos resaltados de la tabla. Podemos afirmar que, en general, las frecuencias conservan una gran afinidad y estabilidad en los dos corpus. Esto nos lleva a plantear cada uno de los corpus como una muestra completa y equilibrada y que cualquiera de ellos es rentable para el análisis propuesto. Cada corpus se asume como fuente aceptable de datos en el estudio que realizamos.

Este primer resultado de frecuencias sobre los corpus separados muestra la variación que sufren los fonemas vocálicos /a/ y /e/ en los dos primeros lugares del rango de frecuencias. En el corpus oral, hay primacía de la vocal /e/ y la diferencia con la /a/ es de más del 1%; en el corpus escrito, hay un leve predominio de la /a/ sobre la /e/, que solo llega al 0,27%.

Dado que el programa Cratilo® nos permite, en primer lugar, realizar estadísticas sobre el léxico de la lengua, se revisó en el corpus qué características léxicas podrían condicionar los primeros resultados. Si se revisa el orden de frecuencia léxica del corpus oral, se observa que en los cinco primeros lugares hay tres palabras con /e/ (*que*, *de* y *el*) cuya recurrencia (1709/75.421) afecta la frecuencia de este fonema vocálico. También debemos considerar el uso reiterado de algunas formas verbales muy usadas en los versos repentistas, como la primera persona del pretérito indefinido (*llegué, bailé*, etc.), que aumentan la frecuencia del fonema /e/.

Otro aspecto que repercute en la aparición en segundo lugar del fonema /a/ es que en la transliteración del corpus oral se encuentre ya elidida una palabra de alta frecuencia léxica como *para* y aparezca con su forma apocopada *pa'*, lo que viene a disminuir la frecuencia de la vocal /a/ en esta muestra.

Otra observación recae en el fonema /d/: la Tabla 3 muestra que en el corpus oral /d/ aparece en un lugar más bajo de la lista de frecuencias que en el corpus escrito. Una revisión sobre la transliteración del corpus oral da cuenta de una elisión del fonema /d/ en posición de final de palabra y, en algunos casos, de la preposición *de*, como en los ejemplos *usté* y *tiple 'e cedro*. La elisión bien podría ser la causa de su ubicación más baja en el listado de frecuencias del corpus oral. Al contrario, el fonema fricativo palatal sonoro /ʝ/ aparece en el corpus oral en un lugar más alto del orden que en el corpus escrito. Esta diferencia está relacionada con el reiterado uso del pronombre de primera persona *yo* (379/75.421), tan característico de las formas del repentismo como en el uso del enunciado: "yo le digo a usté".

En relación con el corpus escrito, también entre las cinco primeras palabras del orden de frecuencia se encuentran *de*, *que* y *el* (136.547/5'606.966). La forma *la*

ocupa el segundo lugar en el orden de frecuencia léxica en ambos corpus y la forma *a* ocupa el séptimo en corpus oral y el sexto en el escrito. Los porcentajes internos de estos lugares son similares para los dos corpus: en el corpus oral, *la* tiene una recurrencia de 658/75 421, que equivale a 0,38%; en el corpus escrito, *la* tiene una recurrencia de 43.356/5'606.996, que equivale a 0,34%. En este sentido, se reitera que los dos corpus se comportan como muestras rentables y sus resultados de frecuencia son comparables.

Compensación del corpus

El alto porcentaje de frecuencia léxica de las primeras 30 formas de los dos corpus de análisis nos lleva a indagar qué tanto condicionamiento tienen esas cifras en el orden y la frecuencia de fonemas de estos corpus. Este porcentaje de frecuencias comienza a reducir la distancia porcentual entre forma y forma entre los lugares 25 a 30 del orden de frecuencias. Dentro de estas 25 formas, 15 palabras se repiten en los dos corpus. De manera coincidente, resultan ser palabras funcionales (Smith & Witten, 1993, p. 5)[10] las palabras de conexión sintáctica o del "tejido conectivo"; aquellas palabras enlace de alta frecuencia en las lenguas analíticas, como preposiciones, determinantes y conjunciones.

Debido a su frecuencia alta de aparición en los corpus, se eliminaron de los dos corpus en estudio 21 de esas palabras funcionales y otras que están en el límite entre la función estructural o funcional y el contenido referencial, como algunos adverbios. De esta manera, se compensa o se filtra el corpus de análisis con el fin de comprobar si el léxico de mayor frecuencia condiciona la frecuencia fonemática de la muestra. El léxico de cada corpus que sirvió de base para la compensación puede observarse en el Anexo 4.

Resultados comparativos de la compensación de corpus

Los resultados de frecuencia de fonemas, a partir del corpus compensado y sin compensar aparecen en las Tablas 4 y 5.

10 "[…]function words are exemplified by prepositions, articles, auxiliary verbs, pronouns, and such –words whose principal role is more syntactic than semantic […] serve primarily to clarify relationships between the more meaning-laden elements of linguistic expression […] Compares to other vocabulary items, function words demonstrate high frequency usage" (Smith & Witten,1993, p. 5).

Tabla 4. Corpus oral. Comparación entre muestra sin y con compensación

Fonema	Porcentaje de frecuencia	Fonema	Porcentaje de frecuencia compensada
e	14,41%	a	14,57%
a	13,16%	e	12,28%
o	10,90%	o	11,78%
s	8,25%	i	7,53%
i	7,05%	n	7,29%
n	6,83%	ɾ	6,05%
ɾ	5,11%	t	5,42%
l	4,74%	s	5,17%
t	4,54%	d	4,13%
k	4,41%	b	3,75%
d	3,78%	k	3,68%
m	3,35%	m	3,56%
u	2,91%	l	3,42%
b	2,86%	u	2,82%
p	2,77%	p	2,77%
g	1,18%	g	1,55%
ʝ	1,17%	r	1,04%
r	0,80%	x	0,90%
x	0,69%	j	0,87%
f	0,48%	f	0,63%
tʃ	0,35%	tʃ	0,46%
ɲ	0,21%	ɲ	0,28%
k͡s	0,05%	k͡s	6,06%

La Tabla 4 permite afirmar que, una vez compensado el corpus oral a través de la eliminación de 21 formas de alta frecuencia, entre ellas las formas funcionales *que*, *el* y *de*, hay una relación entre la frecuencia léxica y la frecuencia fonemática en esta muestra de habla, ya que el orden /e/ /a/ se invierte en el corpus compensado, aun habiendo eliminado también formas como *a*, *la* y *pa*, que tienen también una alta frecuencia de aparición. Cabe destacar que la distancia porcentual entre los dos fonemas se amplía en el corpus compensado, lo que afirma la mayor adhesión del fonema /a/ al léxico de contenido de la lengua, como el más usado en el español de Colombia.

En relación con las consonantes de la Tabla 4, con una distancia porcentual mayor del 3% entre corpus sin y con compensación, se ubica en un lugar mucho menor el fonema /s/. La compensación tiende a mostrar que el uso de /s/ como morfema de plural presente en las formas *los* y *las* es la causa de su alta aparición. Otro tanto sucede con el fonema /l/. Podría pensarse que su ubicación frecuencial está marcada por el uso de las formas *le*, *les*, *lo* y *la*, ya que, efectivamente, entre el corpus oral compensado y sin compensar la distancia porcentual de este fonema decrece en 1,32%. Otro tanto lo constituye el fonema /k/. Palabras como *que*, *con*, *porque* y *aquí* afectan su ubicación en el rango general de frecuencia de fonemas, ya que, una vez compensado, decrece en casi 1%.

De otra parte, en la Tabla 5, la compensación del corpus escrito nos permite verificar que muy buena parte del peso de /e/ en el rango de frecuencia depende de su uso en varias palabras del tejido conectivo de esta muestra. Su alta compensación, sin embargo, no influye para que pierda su segundo lugar como fonema más usado en este corpus. De manera similar, con la compensación de /a/, especialmente producida por la eliminación de las formas *a* y *para*, el fonema /a/ no alcanza a cambiar su primacía en el rango de frecuencia fonemática. Esto simplemente reafirma que el uso de /a/ tiene una fuerte presencia en el léxico de contenido de la lengua.

En relación con el consonantismo del corpus escrito, el fonema /s/, presente en las formas compensadas *los* y *las*, como marca de plural, perdió su posición ante el fonema /i/, que también fue compensado, especialmente mediante la conjunción *y*. Sin embargo, /s/ permanece como uno de los fonemas de mayor uso, lo que podría estar relacionado con su función como marca del plural del léxico de contenido.

El caso contrario sucede con el fonema /l/. En la Tabla 5 se registra un fuerte descenso en su orden al compensar las formas funcionales asociadas a este fonema, relacionando así su dependencia de uso con formas como *la*, *el*, *los*, *las*, *al* y *le*.

En este mismo sentido, aunque se mantiene en un rango estable, /d/ desciende ante /t/, y con un porcentaje inferior; la preposición *de*, en su forma simple y en la forma *del*, causa este cambio.

Tabla 5. Corpus escrito. Comparación entre muestra sin y con compensación

Fonema	Porcentaje de frecuencia	Fonema	Porcentaje de frecuencia compensada
a	13,71%	a	14,20%
e	13,44%	e	11,94%
o	9,61%	o	10,63%
s	9,26%	i	8,11%
i	7,23%	n	7,03%
n	7,06%	ɾ	6,88%
ɾ	5,84%	s	6,50%
l	5,13%	t	5,11%
d	5,09%	d	4,68%
t	4,05%	m	3,63%
k	3,79%	b	3,51%
u	3,20%	k	3,43%
m	2,98%	l	3,15%
b	2,86%	u	3,08%
p	2,51%	p	2,75%
g	0,94%	g	1,18%
x	0,73%	x	0,92%
r	0,72%	r	0,91%
ɟ	0,66%	ɟ	0,84%
f	0,62%	f	0,78%
tʃ	0,25%	tʃ	0,32%
ɲ	0,22%	ɲ	0,27%
k͡s	0,11%	k͡s	0,14%

La Tabla 5 también muestra un descenso en el rango de /k/. Si se tiene en cuenta que la forma gráfica *que* ocupó en el corpus escrito el tercer lugar de la frecuencia léxica, es muy probable que su eliminación en el proceso de compensación esté firmemente asociada con esta posición más baja en el rango frecuencial fonemático.

Cabe anotar, como parte final de este apartado, el último lugar ocupado por el fonema /k͡s/ y su diferencia frecuencial entre los dos corpus. Parece evidente que hay muy poco léxico disponible de uso común con este fonema y por ello su escasa frecuencia en el corpus oral y un breve aumento de palabras que hacen uso de la grafía "x" y su correspondencia fónica /k͡s/ en el uso de la lengua escrita.

La compensación del corpus permite ratificar la relación de algunas palabras funcionales y la frecuencia de fonemas, principalmente en lo que se refiere a aquellas que incluyen los fonemas /e/ /l/ /s/ y /k/.

Frecuencia fonemática del español de Colombia

Dada la representatividad y homogeneidad de los corpus, y la relevancia del tratamiento compensatorio de los datos, presentamos en la Tabla 6 los resultados generales sobre la frecuencia de fonemas en estas muestras compensadas de español colombiano y el rango frecuencial fonemático que establece el promedio entre los dos corpus, el oral y el escrito.

La compensación comparada de los dos corpus vuelve a mostrar que las zonas estables resaltadas en la Tabla 6 son muy amplias y sus diferencias son mínimas, a excepción del fonema /s/, principalmente, ya que la distancia porcentual en ese caso supera el 1%. Los dos corpus se manifiestan de manera muy homogénea, lo que reitera la validez de haberlos usado como corpus representativos. Así, el promedio presentado en la columna de la derecha da cuenta del rango porcentual de frecuencias de los fonemas del español en el uso colombiano.

La Tabla 6 permite afirmar, sobre la base del corpus estudiado, que, en el componente fonológico del español colombiano son válidas las siguientes aseveraciones:

- El fonema con mayor frecuencia es /a/.
- Los fonemas vocálicos en orden decreciente son /aeoiu/.
- Los cinco primeros lugares en frecuencia los ocupan /aeoin/, cada uno de ellos con más de 7% de frecuencia y entre ellos suman el 52,67% del total de realizaciones.
- Con menos del 1% de aparición se encuentran 7 fonemas (/r/, /x/, /ʝ/, /f/, /ɲ/, /t͡ʃ/, /k͡s/). Estos fonemas no alcanzan el 5% del total de la frecuencia: entre ellos tres fricativos (/x/, /ʝ/, /f/) y cinco que son articulaciones complejas, no

simples (Thomas, Diamante, Bouquiaux & Cloarec-Heiss, 1985) bien sea por la multiplicidad (Lewis, 2004), /r/, la africación, /ʧ/, /k͡s/, o su amplio contacto articulatorio /ʝ/, /ɲ/.

Tabla 6. Distribución de frecuencias de fonemas en dos corpus del español colombiano[11]

Fonema	Porcentaje de frecuencia compensada en corpus oral	Fonema	Porcentaje de frecuencia compensada en corpus escrito	Promedio en orden descendente	
a	14,57%	a	14,20%	a	14,38%
e	12,28%	e	11,94%	e	12,11%
o	11,78%	o	10,63%	o	11,20%
i	7,53%	i	8,11%	i	7,82%
n	7,29%	n	7,03%	n	7,16%
ɾ	6,05%	ɾ	6,88%	ɾ	6,47%
t	5,42%	s	6,50%	s	5,84%
s	5,17%	t	5,11%	t	5,26%
d	4,13%	d	4,68%	d	4,41%
b	3,75%	m	3,63%	m	3,63%
k	3,68%	b	3,51%	b	3,60%
m	3,56%	k	3,43%	k	3,55%
l	3,42%	l	3,15%	l	3,28%
u	2,82%	u	3,08%	u	2,95%
p	2,77%	p	2,75%	p	2,76%
g	1,55%	g	1,18%	g	1,36%
r	1,04%	x	0,92%	r	0,98%
x	0,90%	r	0,91%	x	0,91%
ʝ	0,87%	ʝ	0,84%	ʝ	0,85%
f	0,63%	f	0,78%	f	0,71%
ʧ	0,46%	ʧ	0,32%	ʧ	0,39%
ɲ	0,28%	ɲ	0,27%	ɲ	0,28%
k͡s	6,06%	k͡s	0,14%	k͡s	0,10%

11 Para el establecimiento de los promedios, dada la particularidad de cada corpus, se toma cada uno como totalidad, independientemente del número de fonemas que contiene, pues si se ponderara en relación con el número de fonemas, se estaría asumiendo que son un continuo, mientras en realidad son cualitativamente diferentes como fenómenos de lengua.

Discusión

Consideramos el orden frecuencial de la última columna de la Tabla 6 como el resultado del proceso investigativo sobre la frecuencia de fonemas en un corpus amplio en su mayoría de léxico de contenido del español de uso en Colombia. En este sentido, esta tabla representa la muestra colombiana, que nos permite ahora realizar un análisis comparativo con algunos de los trabajos en el área, especialmente con aquellos que han tomado corpus de distintas variedades de uso de la lengua española, para observar tendencias del componente fonológico en general y particularidades del uso de la lengua en su variante colombiana. La Tabla 7 sintetiza esta información e incluye los resultados de este estudio.

En la Tabla 7 podemos observar que el porcentaje de uso de las vocales, en este trabajo como en los tomados para la comparación, es muy cercano a la mitad del total del corpus. En concordancia, las consonantes superan el 52% del total de frecuencia. Los dos fonemas vocálicos más usados son /a/ y /e/, que intercambian su posición en el rango, aparentemente sin relación alguna a la variedad de lengua. Sin embargo, se observa que en los corpus ibéricos /a/ supera a /e/, o están más cercanos, o equiparan su frecuencia. Para la variedad chilena, Pérez (2003) advierte que, en su corpus, el 5% de las ocurrencias correspondió a la preposición *a*. En nuestro caso, compensado el corpus, el resultado plantea la preponderancia de la vocal /a/, aún sin las palabras funcionales que incluyen /a/.

La relación de esos dos fonemas vocálicos con las demás vocales es de una distribución gradual en su orden de aparición. La serie vocálica abierta se ve separada solo en las variedades chilena y mexicana; en la variedad colombiana aparece en el orden /aeo/.

Es el fonema consonántico /s/ el que inicia el rango consonántico y el distanciamiento entre las vocales +silábicas y ±silábicas /i, u/, en ese orden, en todos los trabajos, a excepción de este estudio, en el cual la serie vocálica se interrumpe con la aparición en el rango del fonema /n/. Si bien podemos relacionar el descenso de /s/ en el rango con la compensación de /s/ como marca de plural, es interesante anotar que, en los demás corpus, su frecuencia supera el 8%, que, sumado a la frecuencia de la consonante interdental /θ/ en las variedades ibéricas, de todas maneras no supera la mexicana (10,34%), lo que contrasta en muy buena medida con la colombiana (5,84%), que representa un poco menos de la mitad, una vez compensada con su aspecto funcional. Esto resalta aún más este aspecto gramatical.

Tabla 7. Comparación de la distribución de porcentaje de frecuencias de fonemas del español[12]

Autores	Zipf y Rogers	Alarcos Llorach	Guirao y Burzone	Quilis/ Esgueva	Rojo	Guirao/ García Jurado	Listerri/ Mariño	Pérez	Pineda/ Cuétara	González/ Mejía
Año	1939	1965	1972	1980	1991	1993	1993	2003	2005	2011
Variedad	ibérica	ibérica	argentina	ibérica	ib. e hispa-nam	argentina	ibérica	chilena	mexicana	colombiana
i	4,20%	8,60%	7,27%	7,38%	7,51%	6,59%	6,89%	7,46%	8,50%	7,82%
e	12,20%	12,60%	14,51%	14,67%	13,46%	14,99%	13,73%	14,13%	13,53%	12,11%
a	14,06%	13,70%	12,45%	12,19%	13,46%	13,27%	13,43%	12,31%	12,07%	14,38%
o	9,32%	10,30%	9,85%	9,98%	9,55%	10,75%	10,37%	9,28%	9,31%	11,20%
u	1,76%	2,10%	3,08%	3,33%	3,15%	2,79%	2,33%	3,05%	3,01%	2,95%
TOTAL	**41,54%**	**47,30%**	**47,16%**	**47,55%**	**47,13%**	**48,39%**	**46,75%**	**46,23%**	**46,42%**	**48,46%**
p	2,92%	2,10%	2,76%	2,77%	2,59%	2,68%	2,60%	2,58%	2,57%	2,76%
t	4,46%	4,60%	4,92%	4,53%	4,31%	4,48%	4,63%	4,92%	4,66%	5,26%

12 Para la elaboración de la tabla 7 se tomaron los datos encontrados en algunos de los estudios mencionados y, especialmente, en el estudio de Luis A. Pineda, Luis Villaseñor Pineda, Javier Cuétara, Hayde Castellanos, e Ivonne López (2004). Se incluyeron en la tabla 7 los fonemas fricativo interdental sordo, /θ/, y lateral palatal, /ʎ/, que se toman en cuenta en estudios de otras variedades. Sin embargo, no se tiene en cuenta la frecuencia de los archifonemas que aparecen estudiados en los trabajos de Quilis y de Alarcos. Dichos porcentajes de frecuencia son muy bajos y probablemente no tienen repercusión en la distribución frecuencial ni en el rango de fonemas totales. En esos estudios, el total del porcentaje de frecuencia de las consonantes se ve afectado por esta disminución; en el caso de Alarcos, en 4,30% (para un total de 52,70%); en el caso de Quilis en 5,22% (para un total de 52,43%); en el caso de Rojo 11,42% (para un total de 52,87%).

k	3,84%	3,80%	4,37%	3,98%	3,81%	4,31%	4,04%	3,94%	3,90%	3,55%	
b	3,26%	2,50%	2,45%	2,37%	2,65%	3,08%	2,92%	1,92%	2,11%	3,60%	
d	5,06%	4,00%	4,16%	4,24%	4,72%	4,00%	3,96%	4,84%	5,57%	4,41%	
g	1,02%	1,00%	0,94%	0,94%	0,87%	1,11%	0,90%	0,94%	0,82%	1,36%	
f	0,72%	1,00%	0,67%	0,55%	0,68%	0,53%	0,51%	0,75%	0,80%	0,71%	
θ	1,74%	1,70%		1,45%	1,69%		1,53%				
s	8,12%	8,00%	9,72%	8,32%	7,55%	9,39%	8,28%	9,61%	10,34%	5,84%	
ʝ	2,40%	0,40%	0,55%	0,41%	0,21%	0,72%	0,19%	0,69%	0,32%	0,85%	
x	0,58%	0,70%	0,65%	0,57%	0,73%	0,70%	0,63%	0,74%	0,70%	0,91%	
k͡s										0,10%	
ʧ	0,30%	0,40%	0,33%	0,37%	0,27%	0,40%	0,40%	0,32%	0,15%	0,39%	
m	2,98%	2,50%	3,04%	3,06%	2,56%	3,17%	3,63%	2,62%	2,77%	3,63%	
n	5,94%	2,70%	7,67%	2,78%	2,39%	7,14%	7,48%	7,78%	7,09%	7,16%	
ɲ	0,36%	0,20%	0,28%	0,25%	0,19%	0,24%	0,27%	0,24%	0,13%	0,28%	
ɾ	5,90%	7,00%	5,58%	5,19%	2,49%	5,40%	4,25%	6,19%	5,70%	6,47%	
r	1,04%	0,60%	0,50%	1,93%	0,38%	0,39%	0,40%	0,64%	0,62%	0,98%	
l	5,20%	4,70%	4,25%	4,23%	5,12%	3,88%	4,25%	5,05%	5,43%	3,28%	
ʎ	0,60%	0,50%		0,38%			0,54%				
TOTAL	**56,44%**	**48,40%**	**52,84%**	**48,32%**	**43,21%**	**51,62%**	**51,41%**	**53,77%**	**51,12%**	**51,54%**	

En relación con el fonema /l/, los datos muestran una notoria diferencia entre la muestra colombiana y las demás, ya que en los otros estudios hay un rango de oscilación entre 3,88% (Argentina) y 5,43% (México), mientras que en el caso colombiano hay un 3,28% de uso, lo que corrobora el peso de la filtración de las ocho palabras funcionales que incluyen /l/.

Otra diferencia importante se presenta en el fonema /n/, debido a que en algunos de los estudios se agregó la categoría /N/, archifonema que representa la desfonologización de los fonemas /n/ y /m/ en posición de coda silábica. Como ya se había advertido, para efectos de comparación se hicieron algunos ajustes en las tablas presentadas por los autores en sus trabajos, como en el caso de /R/, cuya frecuencia resultante solo puede corresponder al fonema /ɾ/, ya que la posición de coda silábica desde el punto de vista fonológico es restrictiva a este fonema. En ese caso, se sumaron los porcentajes de archifonema y fonema. Para /N/, no se pudo realizar esta operación, ya que la coda silábica puede estar ocupada por cualquiera de los fonemas /m/ y /n/; igual que en los otros archifonemas /BDG/. Por ello, en los estudios que incluyeron /N/ aparece /n/ con una frecuencia porcentual baja en relación con los otros rangos.

Es interesante anotar que la posición del fonema menos frecuente en lengua española varía entre las articulaciones dorsales /ʧ, ɲ, ʝ, k͡s/, todos los sonidos palatales, ninguno de los cuales llega a 0,50%, en donde /ʃ/ es el que ocupa el último lugar en siete de los 10 trabajos en comparación. La variedad mexicana es la única que presenta /ʃ/ como el menos frecuente y la colombiana, la única que presenta /k͡s/, con los dos porcentajes frecuenciales más bajos de todo el universo presentado (0,13% y 0,10%, respectivamente).

El orden descendente de la frecuencia de fonemas, como rango fonemático, tal como se presenta en la Tabla 8, nos acercará a otra reflexión, surgida de esta comparación: las tendencias fonológicas del sistema en uso y el establecimiento, sobre este recuento estadístico, de la base articulatoria de la lengua española.

Si se toma como referencia el cuadro articulatorio presentado en páginas anteriores, es difícil encontrar una relación directa entre la frecuencia y las categorías allí presentadas. De igual manera, es poco evidente una correlación entre este rango frecuencial y clasificaciones de tipo acústico o perceptual.

En términos generales, la Tabla 8 corrobora lo afirmado por autores como Quilis (1993, p. 9) sobre la base articulatoria del español, sus hábitos fonéticos o la disposición articulatoria (Gil, 2007, p. 224): la preponderancia del uso de fonemas pronunciables en una zona articulatoria central-anterior, vocales anteriores y consonantes alveolares y dentales. Le siguen en frecuencias de uso medio las

consonantes anteriores labiales y dentales, las consonantes posteriores (toda la serie oclusiva oral) y la vocal posterior alta.

La preferencia del español por los elementos sonoros (16 de 24) se complementa con la primacía de uso de los fonemas sonoros (las vocales alcanzan casi el 50% de frecuencia), pero contrasta con que el fonema consonántico más usado sea sordo, /s/ y el menos usado, en general, sea sonoro /ɲ/. No obstante, los fonemas sonoros, por lo menos en nuestro estudio, alcanzan el 80,48% de uso, mientras los fonemas sordos llegan al 19,52%.

Más allá de la interpretación articularia anotada, la Tabla 8 permite explicar, de acuerdo con los espacios resaltados en tres grupos diferentes, la disposición de aparición de los fonemas del español desde un punto de vista combinatorio, relacionado con las posibilidades de funcionamiento del fonema en la agrupación silábica hispánica.

En un primer grupo, resaltado en amarillo, y con índices de porcentaje alrededor y sobre el 5%, están los fonemas que aparecen en el habla española en cualquier posición silábica (a excepción de /ɾ/), especialmente en función nuclear, de núcleo-silábico o silábicas. Esta funcionalidad favorece enormemente su rango de frecuencia de uso en el sistema y brinda mayores posibilidades de aumento en la frecuencia porcentual de uso.

El sistema fonológico hispánico ha relegado la función silábica a las vocales, aunque las consonantes /s, n, l, ɾ / históricamente han actuado en otras lenguas con dicha función; /kɾk/ 'cuello' en checo, /lit**l**/ 'pequeño' en inglés, /ab**n**/ 'tener' en alemán, y fonéticamente en español coloquial se encuentran casos como [**n**toes] por "entonces", que permiten ver por lo menos en este fonema una potencialidad como sonido núcleo-silábico. Asociados con el carácter de silabicidad, están los rasgos de sonoridad y de cantidad; al ser un fonema silábico, o potencialmente silábico en la lengua, está también entre los más audibles y susceptibles de alargamiento[13].

En segunda instancia, y resaltado en azul, tenemos un grupo dominado por las consonantes oclusivas (las orales y la nasal /m/) y los fonemas labiales /p,b,m,u/, cuyos porcentajes de uso están entre el 1,5% y el 5%, y aunque funcionan como ataque y coda silábicas —a excepción de /u/—, su descenso de rango se relaciona más con su restricción al funcionar como coda silábica. Cuando estos fonemas aparecen en el habla en posición de coda (implosivos), se presentan fenómenos como la neutralización, el debilitamiento, la vocalización o la pérdida, lo que refuerza la tendencia de la lengua hacia la sílaba abierta, terminada en vocal.

13 /r/ puede conseguir continuidad en el habla, no a través de la vibración, sino por medio del fenómeno de fricativización.

Tabla 8. Comparación del rango de fonemas del español

Zipf y Rogers (1939) ibérica	Alarcos Llorach (1965) ibérica	Guirao y Burzone (1972) argentina	Quilis/ Esgueva (1980) ibérica	Rojo (1991) ibérica/ americana	Guirao/ García Jurado (1993) argentina	Listerri/ Mariño (1993) ibérica	Pérez (2003) chilena	Pineda/ Cuétara (2005) mexicana	González/ Mejía (2011) colombiana
a	a	e	e	e	e	e	e	e	a
e	e	a	a	a	a	a	a	a	e
o	o	o	o	o	o	o	s	s	o
s	i	s	s	s	s	s	o	o	i
n	s	n	i	i	n	n	n	i	n
ɾ	ɾ	i	ɾ	l	i	i	i	n	ɾ
l	l	ɾ	t	d	ɾ	t	ɾ	ɾ	s
d	t	t	d	t	t	ɾ	l	d	t
t	d	k	l	k	k	l	t	l	d
i	k	l	k	u	d	k	d	t	m
k	n	d	o	b	l	d	k	k	b
b	b	u	m	p	m	m	u	u	k

m	p	ɟ	u	θ	r	g	f	ɣ	x	ʃ	ʄ
m	u	p	θ	g	f	x	r	ɣ	ɟ	ʄ	ʃ
m	p	b	g	f	x	ɟ	r	ʄ	ʃ		
n	p	b	r	θ	g	x	f	ɟ	ɣ	ʄ	ʃ
m	ɾ	n	θ	g	x	f	r	ʄ	ɟ	ʃ	
b	u	p	g	ɟ	x	f	ʄ	r	ʃ		
b	p	u	θ	g	x	ɣ	f	ʄ	r	ʃ	ɟ
m	p	b	g	f	x	ɟ	r	ʄ	ʃ		
m	p	b	g	f	x	r	ɟ	ʄ	ʃ		
l	u	p	g	r	x	ɟ	f	ʄ	ʃ	k͡s	

La presencia de algunos de estos fonemas en posiciones de coda silábica ha generado históricamente en nuestra lengua los llamados *grupos consonánticos cultos* (pt, kt) en palabras como "aptitud", "octavo", etc., que se han establecido como uno de los rasgos que caracterizan dialectalmente a las variedades conservadoras en su pronunciación. En este sentido, el español colombiano, en especial los dialectos de la zona andina del país, han sido relacionados con una mayor conservación de patrones fonológicos (Quesada, 2001).

El tercer grupo de fonemas en el rango de frecuencias, señalado en verde, con los porcentajes más bajos del sistema, está conformado por aquellos fonemas que históricamente no aparecen con función de coda silábica o su aparición es extremadamente escasa. Hablamos del fonema vibrante múltiple /r/, de los fonemas palatales /ɟ/, /k͡s/, /ʧ/, /ʃ/ y de las fricativas /f/ y /x/.

La posición consonántica posnuclear va en contravía de la tendencia hispánica a la sílaba abierta, tradicional e históricamente descrita para la lengua. Además, un rasgo que podría unir estos fonemas es su carácter articulatorio complejo. La africación, la multiplicidad, las posiciones articulatorias de amplio contacto no están entre las principales preferencias de selección de las lenguas (Crystal, 1999, p. 166; Thomas et ál., 1985), muy seguramente por su dificultad articulatoria, que exige fisiológica y energéticamente al hablante. Sin embargo, su inclusión en cualquier sistema lingüístico permite lograr un equilibrio acústico y le imprime rasgos característicos y de identidad a una lengua.

Conclusiones

A partir del estudio de la frecuencia de fonemas en dos corpus del español de uso colombiano, a través del uso del programa Cratilo®, podemos llegar a las siguientes conclusiones:

En relación con el tratamiento y manejo del corpus, pudimos advertir que un resultado estadístico similar, sobre corpus distintos en procedencia y número, recalca la homogeneidad de estructura lingüística de los corpus o materiales para análisis lingüístico: con corpus medianamente robustos se generan resultados similares y de igual rentabilidad que con corpus de gran robustez.

La creación y aplicación del algoritmo fonético, a partir de la aplicación del principio fonémico, permitió un acercamiento más real a la lengua de uso y al componente fonológico de la lengua, ya que el conteo estadístico se hizo sobre la base de fonemas y no de grafemas. Con este método se establece una pauta que mejorará próximos análisis de corpus de material escrito y la base de contraste entre varios corpus.

La compensación, consistente en la elisión de los fonemas asociados con las primeras 21 formas gráficas de los dos corpus —coincidentes en su mayoría con formas funcionales o del tejido conectivo— afectó la frecuencia fonemática; una vez realizada, la frecuencia de fonemas homogeneizó aún más las diferencias entre los corpus. Así, se establecieron algunas relaciones entre condiciones gramaticales de la lengua que contribuyen en la determinación de dicha frecuencia. La compensación, entonces, permite determinar esa frecuencia sobre el léxico de contenido de la lengua, aquel que ejerce mayor influencia en la semántica de los enunciados del discurso.

Con base en el análisis del corpus tomado para este estudio, se determinó que en el español de uso colombiano: el fonema con mayor frecuencia es /a/; el orden de frecuencia de las vocales es /aeoiu/; los cinco primeros lugares en frecuencia los ocupan /aeoin/, cada uno de ellos con más de 7% de frecuencia y entre ellos suman el 52,67% del total de realizaciones; y con menos del 1 % de aparición se encuentran siete fonemas (/ɾ/, /x/, /ʝ/, /f/, /ʧ/, /ɲ/, /k͡s/), cinco de ellos de articulación compleja.

Sobre la discusión de los resultados de este trabajo comparados con otros estudios del ámbito hispánico, se corrobora que /a/ y /e/ son los fonemas que ocupan los dos primeros lugares en frecuencia, sin que se encuentre una relación directa con alguna causa específica. Las diferencias de frecuencia entre los demás resultados y este trabajo en los casos de los fonemas /l/ y /s/ reafirman su uso preponderante en palabras funcionales anotadas a lo largo de este estudio.

Al realizar un análisis sobre los rangos fonemáticos en comparación se encontró que en español hay una marcada tendencia al uso de fonemas de pronunciación en zona articulatoria centro-anterior. Los fonemas sonoros son los más usados en la lengua; el fonema consonántico más usado es el fricativo alveolar sordo /s/ y el menos usado es el nasal palatal sonoro /ɲ/.

Al relacionar la frecuencia de fonemas con la estructuración silábica de la lengua, se determinó que hay en tres grupos de fonemas, a saber: los más usados (5% o más), que son aquellos que aparecen en cualquier posición de la sílaba y son núcleo-silábicos o potencialmente nucleares; aquellos fonemas medianamente usados, los oclusivos y los labiales (1,5% - 5%), cuya funcionalidad silábica permite su aparición en ataque y coda —a excepción de /u/—, siendo esta última posición —implosiva, débil— causante de procesos históricos fonético-fonológicos que reafirman la tendencia del español hacia la sílaba abierta; y, con porcentajes de frecuencia muy bajos, los fonemas que en la lengua no aparecen o en raras ocasiones se presentan como coda silábica.

Finalmente, es importante anotar que este tipo de estudios permite mencionar algunos alcances y sugerencias de aplicaciones o nuevos trabajos alrededor de sus conclusiones. Así, además de su contribución al estudio de la variedad de habla, el aporte a su reconocimiento cultural y las posibilidades de servir como material de apoyo en su enseñanza, podemos dar ejemplos de cómo las contribuciones en aspectos básicos de todo corpus pueden "(1) Servir a la investigación básica en la descripción de la lengua [y] (2) Prestar servicio en aplicaciones tecnológicas concretas" (Llisterri & Poch, 1994). Específicamente, el análisis de las formas lingüísticas del corpus y su compensación permite establecer, en parte, aquello que para la comunicación es relevante en términos de materia de expresión y qué no. Esto podría tener alcance en las tecnologías de la comunicación. En el mismo sentido, la información fonológica básica generada por bases de datos y estadísticas de frecuencia, como las aquí presentadas, podrían servir de soporte para el proceso de la clasificación de la fase de reconocimiento de patrones y el establecimiento de modelos de unidades lingüísticas en las que se basa una interfaz de comunicación para acceder a la información solicitada por un usuario en áreas como el reconocimiento del habla, la síntesis de voz y otras disciplinas relacionadas con las tecnologías del habla.

Referencias

Alarcos Llorach, E. (1965). *Fonología española*. Madrid: Gredos.

Alcina, J. & Blecua, J. M. (1975). Frecuencia de fonemas en español. En *Gramática española* (pp. 430-435). Barcelona: Ariel.

Ashby, M. & Maidment, J. (2005). *Introducing phonetic science*. Cambridge: Cambridge University Press.

Bermúdez, E. (2006). Del humor y del amor: Música de parranda y música de despecho en Colombia (1). *Cátedra de Artes*, *3*, 81-108. Consultado el 26 de mayo de 2010 en http://www.ebermudezcursos.unal.edu.co/humoryamor1.pdf

Caravedo, R. (1999). *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español.* Salamanca: Ediciones Universidad de Salamanca.

Crystal, D. (1994). *Enciclopedia del lenguaje de la Universidad de Cambridge*. Madrid: Taurus.

Flórez, L. (1978). Algunas formas de pronunciar muchos colombianos el español [separata]. *Thesaurus, Boletín del Instituto Caro y Cuervo*, XXXIII.

Gil Fernández, J. (2007). *Fonética para profesores de español: de la teoría a la práctica.* Madrid: Arco Libros.

Guirao, M. & García Jurado, M. A. (1993). *Estudio estadístico del español*. Buenos Aires: Consejo Nacional de Investigaciones Científicas y Técnicas.

Guirao, M. & Borzone de Manrique, A. (1972). Fonemas, sílabas y palabras del español de Buenos Aires. *Filología*, *XVI*, 135-165.

Internacional Phonetic Asociation. (2011). *Alphabet*. Consultado el 26 de mayo de 2010 en http://www.langsci.ucl.ac.uk/ipa/ipachart.html

Lewis, A. M. (2004). Coarticulatory effects on spanish trill production. En A. Agwuele, W. Warren & S.-H. Park (Eds.), *Proceedings of the 2003 Texas Linguistics Society Conference* (pp. 116-127). Consultado el 26 de mayo de 2010 en http://www.lingref.com/cpp/tls/2003/paper1073.pdf

Llisterri, J. & Mariño, J. B. (1993). *Spanish adaptation of SAMPA and automatic phonetic transcription. SAM-A/UPC/001/VI. ESPRIT project 6819 (SAM-A Speech Technology Assessment in Multilingual Applications).* Consultado el 17 de junio de 2010 en http://liceu.uab.cat/~joaquim/publicacions/SAMPA_Spanish_93.pdf

Llisterri, J. & Poch, D. (1994). Proyecto de una base de datos acústicos de la lengua española. En *Actas del Congreso de la Lengua Española, Sevilla, 7-10 octubre de 1992*. Madrid: Instituto Cervantes. Consultado el 17 de junio de 2010 en http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_llisterripoch.htm

Moreno Sandoval, A., Torre Toledano, D., Curto, N., & De la Torre, R. (2006). Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito. En L. Buera, E. Lleida, A. Miguel & A. Ortega (Eds.), *IV Jornadas en Tecnología del Habla. Zaragoza. Del 8 al 10 de Noviembre de 2006* (77-81). Zaragoza: Universidad de Zaragoza. Consultado el 17 de septiembre de 2010 en http://jth2006.unizar.es/finals/4jth_116.pdf

Mosterin, J. (1981). *La ortografía fonémica del español*. Madrid: Alianza Universidad.

Pérez, H. E. (2003). Frecuencia de fonemas. *e-rthabla*, *1*. Consultado el 10 de junio de 2010 en http://lorien.die.upm.es/~lapiz/e-rthabla/numeros/N1/N1_A4.pdf

Pineda, L. A., Villaseñor Pineda, L., Cuétara, J., Castellanos, H., & López, I. (2004). DIMEX100: A New Phonetic and Speech Corpus for Mexican Spanish. En C. Lemaître, C. A. Reyes & J. A. González (Eds.), *Iberamia*, 2004, LNAI 3315, 974-983, [en línea]. Consultado el 17 de junio de 2010 en http://leibniz.iimas.unam.mx/~luis/DIME/publicaciones/papers/DIMEX100-LNAI3315.pdf

Quesada, M. Á. (2001). La fonética del español americano en pugna: dialectos radicales y conservadores en lucha por la supremacía. Ponencia presentada en el *Segundo Congreso Internacional de la Lengua Española, Valladolid*. Consultado el 9 de

septiembre de 2010 en http://congresosdelalengua.es/valladolid/ponencias/unidad_diversidad_del_espanol/2_el_espanol_de_america/quesada_m.htm

Quilis Morales, A. & Esgueva Martínez, M. A. (1980). Frecuencia de fonemas en el español hablado. *LEA: Lingüística española actual*, *2*(1), 1-25.

Quilis, A. (1993). *Tratado de Fonología y Fonética Españolas*. Madrid: Gredos.

Rojo, G. (1991). Frecuencia de fonemas en español actual. En *Homenaxe ó Profesor Constantino García* (pp. 451-467). Santiago de Compostela: Universidad de Santiago de Compostela.

Rojo, G. (2002). Sobre la lingüística basada en el análisis del corpus. *Hizkuntza-corpusak. Oraina eta geroa*, 24/25. Consultado el 27 de mayo de 2010 en www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54a.pdf

Rojo, G. (2008). Lingüística de corpus y lingüística del español. Ponencia plenaria presentada en el XV Congreso de la Asociación de Lingüística y Filología de América Latina, Montevideo. Consultado el 10 de junio de 2010 en http://www.lehrbuch-online.de/downloads/spanische_sprachwissenschaft/rojo.pdf

Smith, T. C. & Witten, I. H. (1993). Language inference from function words. *Working Paper Series*, *3*, 25.

Thomas, J. M.-C., Diamante, E., Bouquiaux, L. & Cloarec-Heiss, F. (1985). *Iniciación a la fonética. Fonética articulatoria y fonética distintiva*. Madrid: Gredos.

Zapata Morales, J. F. (2010). Oralidad y escritura en la trova antioqueña. *Lingüística y Literatura*, *57*, 131-145.

Zipf, G. K. & Rogers, F. M. (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*, 111-147.

Anexo 1. Obras literarias del corpus escrito

Título	Autor	Año de publicación	Editorial
La Marquesa de Yolombó	Tomás Carrasquilla	1928	Instituto Caro y Cuervo
Viaje a pie	Fernando González	1929	Le livre libre
El libro de los viajes y de las presencias	Fernando González	1959	Alberto Aguirre, Ed.
Una mujer de cuatro en conducta	Jaime Sanín Echeverri	1961	Universidad de Antioquia
Respirando el verano	Héctor Rojas Herazo	1962	Tercer Mundo / Antares
La casa grande	Álvaro Cepeda Samudio	1962	Mito
Cien años de soledad	Gabriel García Márquez	1967	Sudamericana
La otra raya del tigre	Pedro Gómez Valderrama	1977	Siglo XXI
Reptil en el tiempo	María Helena Uribe Echavarría	1986	Molino de papel
En diciembre llegaban las brisas	Marvel Moreno	1987	Plaza y Janés
La virgen de los sicarios	Fernando Vallejo	1994	Alfaguara
Empresas y tribulaciones de Maqroll el Gaviero; *La nieve del almirante*; *Ilona llega con la lluvia*; *Un bel morir*; *La última escala del Tramp Steamer*; *Amirbar*; *Abdul Bashur, soñador de navíos*; *Tríptico de mar y tierra.*	Álvaro Mutis	1995	Alfaguara
Que viva la música	Andrés Caicedo	1997	Instituto Colombiano de Cultura
Pensamientos de guerra	Orlando Mejía Rivera	1999	Ministerio de Cultura

Anexo 2. Sesiones del Festival de la Trova que generan el corpus oral

Festival nacional de la trova, Medellín	Trovador que inicia	Trovador que contesta	Tema
III, 1977	Mario Mosquera "Mario Tierra"	"Tista" Ortega	Libre
IV, 1978	Jorge Carrasquilla.	Luis Fernando Macías.	Libre
V, 1979 (Semifinal)	Leonel Guillén	Albeiro Jaramillo "Ladrillo"	Libre
V, 1979	Hernán Darío Hincapié	Jorge Mario Correa	Libre
V, 1979	Rodrigo Mejía "El Bobo de Caldas"	Hernán Darío Hincapié	Libre
VIII, 1982	Albeiro Jaramillo "Ladrillo"	Rodrigo Mejía "El bobo de Caldas"	Libre
VIII, 1982	Rodrigo Mejía "El Bobo de Caldas"	Albeiro Jaramillo "Ladrillo"	Embarazo
IX, 1983	Jorge Carrasquilla	Leonardo Díaz "El tráfico"	Espantos
IX, 1983	John Jairo Pérez Ortiz	Alejandro Echavarría "El Negro"	Las fiestas de los pueblos
IX, 1983	John Jairo Pérez	Jorge Carrasquilla	Libre
X, 1984	Saulo García Gómez "Gelatina"	César Augusto Betancur "Pucheros"	Libre
X, 1984	Saulo García Gómez "Gelatina"	Jorge Carrasquilla	Libre
XI, 1985	Leonardo Díaz "El Tráfico"	Germán Darío Carvajal "Minisicuí"	Libre
XI, 1985	Jaime Hernández "Paso 'e reina"	Germán Darío Carvajal "Minisicuí"	Libre
XI, 1985	Jorge Carrasquilla	Jaime Hernández "Paso 'e reina"	Libre
XI, 1985	Leonardo Díaz "El Tráfico"	Jaime Hernández "Paso 'e reina"	Libre
XII, 1987	César Augusto Betancur "Pucheros"	John Jairo Pérez Ortiz	Los negocios
XIII, 1988	Gustavo Hernán Aristizábal	Germán Darío Carvajal "Minisicuí"	Una casa campesina
XIII, 1988	Gustavo Hernán Aristizábal	Carlos Mario García "El Tigre"	Pedidos al Niño Dios
XIII, 1988	Carlos Mario García "El Tigre"	Gustavo Aristizábal	La vejez
XIV, 1989	Martín Bedoya "Rajaleña"	Saulo García "Gelatina"	3 vírgenes en la tierra

XVI, 1991	Aicardo Martínez "El Cura"	Raúl Mario Castaño "Crispeta"	Comidas callejeras
XVI, 1991	Julio César Arcila "El Cachetón"	Raúl Mario Castaño "Crispeta"	Antioquia
XVI, 1991	Aicardo Martínez "El Cura"	Raúl Castaño "Crispeta"	Va por ti Medellín
XVI, 1991	Raúl Mario Castaño "Crispeta"	Julio César Arcila "El Cachetón"	Libre
XVI, 1991	Raúl Mario Castaño "Crispeta"	Carlos Mario García "El Tigre"	Libre
XVII, 1992	Diego López "Dieguito"	Aicardo Martínez "El Cura"	Libre
XVII, 1992	Roberto Arturo Palacio "Toto"	Julio César Arcila "El Cachetón"	Libre
XVII, 1992	Julio César Arcila "El Cachetón"	Aicardo Martínez "El Cura"	Chanchullo
XIX, 1995	Raúl Mario Castaño "Crispeta"	Aicardo Martínez "El Cura"	Libre
XIX, 1995	Jorge Eduardo Agudelo "Mazamorra"	Raúl Mario Castaño "Crispeta"	Libre
XX, 1997	Ramiro Gómez "Tutti-Frutti"	Roberto Arturo Palacio "Toto"	Libre
XX, 1997	Ramiro Gómez "Tutti-Frutti"	Diego Alberto López "Dieguito"	Libre
XX, 1997	Ramiro Gómez "Tutti-Frutti"	Diego Alberto López "Dieguito"	Libre
XXI, 1999	Gilberto Díaz "Pategrillo"	Rolando García	Libre
XXI, 1999	Juan Fernando Londoño "Frisoles"	Gilberto Díaz "Pategrillo"	Libre
XXI, 1999	Nicolás Escobar "Cantarrana"	Edwin Giraldo "Radioloco"	Libre
XXI, 1999	Rolando García	Edwin Giraldo "Radioloco"	Libre
XXI, 1999	Nicolás Escobar "Cantarrana"	Rolando García	Libre

Anexo 3. Pautas de conversión del corpus escrito al "corpus del lector"

Signos de puntuación por espacios	ge por je	cu por ku
Grafemas mayúsculos por minúsculos	gi por ji	ch por z
Vocales tildadas por vocales sin tilde	gue por ge	c por k
ü por u	gui por gi	rr por c
z por s	que por ke	*espacio* + r por *espacio* + c
v por b	qui por ki	y + *espacio* por i + *espacio*
ce por se	ca por ka	ll por y
ci por si	co por ko	*según las palabras,* w por b ó w por u

Anexo 4. Listado de palabras que fueron compensadas de los dos corpus

Corpus oral		Corpus escrito	
Forma gráfica	**Frecuencia**	**Forma gráfica**	**Frecuencia**
que	800	de	66.655
la	658	la	43.356
y	548	que	39.558
de	474	y	37.037
el	435	el	30.334
a	413	a	29.744
no	373	en	29.648
lo	343	los	18.309
un	271	con	14.340
en	254	un	13.550
le	194	las	12.803
con	188	no	12.395
porque	153	por	10.734
si	144	del	10.669
los	132	una	10.362
una	126	lo	8.915
por	123	para	7.654
pa’	122	al	7.620
pero	98	le	7.535
aquí	93	como	6.249
muy	92	más	5.406