



María Claudia González Rátiva
Jorge Antonio Mejía Escobar
Universidad de Antioquia (Colombia)

Resumen: Este artículo presenta los resultados de una investigación sobre la frecuencia fonemática en dos corpus del español colombiano: literatura colombiana escrita y trova colombiana (canto popular repentista). Con base en la aplicación del programa *Cratilo*® se establece la frecuencia léxica y la frecuencia fonemática de los corpus. El análisis de las restricciones lingüísticas sobre la frecuencia fonemática permite realizar *compensaciones* a los datos para trabajar principalmente sobre el léxico de contenido. Los resultados de la frecuencia fonemática del español colombiano se contrastan en la discusión con otros trabajos sobre el tema con el fin de presentar una reflexión en torno al sistema fonológico de la lengua.

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

1. Marco teórico

Este estudio se enmarca 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 un instrumento (el programa *Cratilo*®²) que nos permite manejar y cualificar almacenes masivos de información.

Para Rojo (2002: 4), 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”. 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 gramática³. Nuestro propósito en este trabajo ha sido evitar que la ceguera predomine, al tener la posibilidad de generar condiciones para la reorganización del

¹ Consideraciones iniciales en torno a este trabajo se presentaron en el XXIV Congreso Nacional de Lingüística, Literatura y Semiótica, Valledupar, Colombia, 26 de noviembre de 2006.

² *Cratilo*® es un programa computacional elaborado en la Universidad de Antioquia para crear concordancias universales de textos. Fue creado por J.A. Mejía, F.J. Álvarez y J.A. Sánchez y financiado por el CODI (Comité de Investigaciones) de la Universidad de Antioquia. Permite además etiquetar, con un dispositivo abierto para un total de 100 categorías y permite hacer consultas a la concordancia generada, que funciona como base de datos. Procesa textos escritos en los caracteres del inglés, francés, alemán, italiano, portugués, latín y español. El corpus de las quince obras 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, por John Fredy Zapata Morales.

³ 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 2000 palabras (Rojo 2008); y, por ejemplo, el *Corpus230*, consiste en 344.619 frases, 235.891 unidades léxicas y alrededor de 15 millones de palabras, cuya fuente es la Web.

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: 20). Esto nos permitió, por ejemplo, realizar *compensaciones* para ver si la dominancia de algunos fonemas se debe a formas gramaticales de alta frecuencia en los corpus. De ese modo aspiramos a tener y presentar una comprensión mejor del fenómeno estudiado: la frecuencia fonemática, que es la base para el establecimiento de las tendencias del sistema fónico y para su aplicación en disciplinas asociadas a las tecnologías del habla y a la enseñanza y aprendizaje de la lengua.

1.1. Los fonemas del español de uso en Colombia

La fonología, basada en las posibilidades articulatorias humanas, establece el dominio fónico de una lengua. Del conjunto fonémico expresado por la Asociación Fonética Internacional, la variedad colombiana de español hace uso de 23 categorías, como se muestra en la Tabla 2.

Consonantes										Vocales			
	Zona Bilabial		Labio-dental		Dental	Alveolar	Alveopalatal	Palatal	Velar		Anterior	Central	Posterior
Modo													
Oclusivo	p	b			t	d					i		u
Fricativo		f				s					e		o
Africado							ks						
Nasal		m					n					a	
Líquido lateral							l						
Líquido vibrantes							r						

Tabla 2. Fonemas de uso del español hablado en Colombia

La norma colombiana asume como fonema el sonido africado, alveopalatal sordo /ks/, 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 extendido fenómeno del yeísmo, proyectado desde los trabajos basados en el ALEC, Atlas Lingüístico Etnográfico de Colombia (Flórez 1978) cuya solución para buena parte del área andina del país es /j/, consonante fricativa palatal sonora.

1.2. 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 *et al.* 1975).

Autores	Año del estudio	Corpus	Nº de fonemas	Variedad
Zipf y Rogers	1939	Escrito	5 000	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	62 980	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/Cuéltara	2004	Oral	---- (300 000 frases)	mexicana

Tabla 1. Tipo de corpus y número de fonemas analizados

En la discusión de este trabajo 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.

2. Metodología

Para obtener datos estadísticos sobre el español usado en Colombia, tomamos dos corpus o muestras de lengua de trabajos elaborados con *Cratilo*®: 15 novelas colombianas del siglo XX y 15 sesiones finales del festival de la trova de Medellín entre los años 1977-1999. En el corpus de lengua escrita aparecen obras de gran renombre en la literatura colombiana, como *Cien años de soledad*, *La otra raya del tigre* e *Ilona llega con la lluvia*. También hay obras como *La marquesa de Yolombó*, que en algunos fragmentos aspira a reflejar de manera escrita el español hablado. De otra parte, el corpus oral está tomado del Festival de la trova⁴, evento que se realiza anualmente en Medellín, durante la Feria de las flores. Contiene la transliteración de quince sesiones finales entre 1977 y 1999. Así, la muestra oral y la escrita, el registro normalizado y el autóctono, 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 realizaciones de fonemas; el corpus oral tiene 75.421 realizaciones de fonemas. En total, el corpus para realizar el estudio de frecuencia de fonemas en el español colombiano alcanza las 5'682.417 realizaciones de fonemas.

El programa utilizado para el análisis del corpus fue el *Cratilo*®. La aplicación de este programa permite un seguimiento cuantitativo y cualitativo de datos lingüísticos. Cuantitativo, porque el procesamiento produce estadísticas de uso de todas las formas gráficas; cualitativo, porque podemos mirar en formato de concordancia todos los datos que responden a un formato de consulta para verificar que no se produzcan distorsiones o que no intervengan factores que no habían sido tenidos en cuenta inicialmente.

2.1. Adaptación fonema-grafema en el corpus para el análisis con el *Cratilo*®

Una vez establecidas las unidades fonológicas, se procedió a adaptar el corpus elegido, transliterado ortográficamente, a un corpus que corresponda al principio fonémico (Mosterín 1981: 34) un fonema -un símbolo- un grafema, en este caso un fonema -un carácter-. Con el fin de aprovechar los recursos ofrecidos por el programa *Cratilo*® se diseñó un algoritmo fonético para la adecuación del corpus al principio fonémico. Se creó así lo que llamamos en nuestra investigación el “corpus del lector”: cada uno de los fonemas de uso del español colombiano está representado en el corpus para análisis con un solo carácter gráfico.

3. Resultados

Cratilo® ofrece en primera instancia el cálculo de frecuencias presentado en la Tabla 3. Teniendo en cuenta la gran diferencia numérica entre los dos corpus, podemos afirmar que, en general, las frecuencias conservan una gran afinidad y estabilidad. Esto nos lleva a pensar que cada uno de los corpus es una muestra equilibrada y *rentable* en cuanto a su representatividad y significatividad; cada corpus es fuente aceptable de datos en este estudio.

⁴ El Festival de la trova es una competencia cultural de canto de coplas, 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. Como es un enfrentamiento verbal entre dos personas, estos pueden ser de diversa procedencia, pero generalmente son en su mayoría antioqueños y caldenses, zonas estas que se han caracterizado por cultivar esta manifestación cultural.

Fonema	Frecuencia corpus trova	Porcentaje	Fonema	Frecuencia corpus literario	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%
r	3.856	5,11%	r	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%	f	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%
j	879	1,17%	x	41.071	0,73%
r	600	0,80%	r	40.571	0,72%
s	518	0,69%	j	37.280	0,66%
f	363	0,48%	f	34.634	0,62%
ŋ	266	0,35%	ŋ	14.175	0,25%
ɲ	162	0,21%	ɲ	12.113	0,22%
ks	34	0,05%	ks	5.331	0,11%

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

Se destaca en este primer resultado 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%. Al revisar la frecuencia léxica de este corpus oral, observamos que entre los primeros lugares se encuentran las preposiciones “de” y “en”, el artículo definido “el” y el pronombre relativo “que”. 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.). Esto podría favorecer el aumento de la frecuencia del fonema /e/. Además, que en la transliteración del corpus oral se encuentre ya elidida una palabra de alta frecuencia como “para” y aparezca con su forma dialectal “pa”, puede restar la frecuencia de la vocal /a/ en esta muestra.

Una segunda observación recae en el fonema /d/: en el corpus oral 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”, “tiple ‘e cedro”. La elisión podría favorecer este descenso. Por el contrario, la frecuencia de uso del fonema fricativo palatal sonoro /j/ en el corpus oral hace que aparezca este fonema en un lugar más alto del listado que en el corpus escrito. Esta diferencia puede deberse al reiterado uso del pronombre de primera persona “yo” (379 apariciones), tan característico de las formas del repentismo como en el uso del enunciado: “yo le digo a usté”.

Teniendo en cuenta que las diferencias de orden de aparición de frecuencia entre el corpus oral y el escrito pueden estar asociadas a formas gramaticales, se asumen como restricción lingüística el uso de las palabras de conexión sintáctica o del “tejido conectivo”: aquellos lexemas gramaticales de gran recurrencia en las lenguas analíticas, como preposiciones, conjunciones, artículos y determinantes. Efectivamente, el léxico de mayor frecuencia está conformado por formas gramaticales que no son la base de la significación lingüística de los enunciados. Se determina, entonces, observar el comportamiento de frecuencia fonemática, elidiendo del corpus el tejido conectivo.

3.1. Compensación del corpus

Los resultados del corpus compensado aparecen en las Tablas 4 y 5.

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%	f	6,05%
f	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%
j	1,17%	r	1,04%
r	0,80%	x	0,90%
x	0,69%	ɲ	0,87%
ɲ	0,48%	f	0,63%
ʃ	0,35%	ʃ	0,46%
ɲ	0,21%	ɲ	0,28%
ks	0,05%	ks	0,06%

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%	f	6,88%
f	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%
j	0,66%	j	0,84%
f	0,62%	f	0,78%
ʃ	0,25%	ʃ	0,32%
ɲ	0,22%	ɲ	0,27%
ks	0,11%	ks	0,14%

Como en el rango de frecuencia léxica las primeras 21 formas correspondieron en su mayoría a léxico gramatical, se procedió a retirarlo de cada corpus. En los dos corpus, hubo coincidencia de 16 de esas 21 formas léxicas; las cinco palabras restantes alcanzaron frecuencias similares a las de tejido conectivo. Al suprimirse las formas léxicas de tejido conectivo, la compensación del corpus permitirá establecer si la frecuencia fonemática depende de la frecuencia léxica del tejido conectivo o de la frecuencia léxica de las formas de contenido en el uso de la lengua.

Esta compensación permite corroborar que efectivamente el uso de palabras funcionales como “que”, “el” y “de” inciden en el resultado de rango de frecuencias fonemáticas del corpus oral, ya que sin ellas el fonema /e/ desciende al segundo lugar, aun con la compensación realizada sobre el fonema /a/ en palabras como “la”, “a” y “pa”, que tienen también una alta frecuencia de aparición. Cabe destacar que la distancia porcentual entre los dos fonemas se amplía en la compensación, lo que afirma la adhesión del fonema /a/ al léxico de contenido de la lengua, como el más frecuente en el español de uso de Colombia. En relación con las consonantes, 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 los artículos definidos “los” y “las”, es la causa de su alta aparición. De igual forma sucede con el fonema /l/. Podría pensarse que su ubicación frecuencial está marcada por el uso del artículo definido y por el dativo pronominal (le, les, lo, la) ya que 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” inciden en la ubicación del rango general de frecuencia fonemática, ya que, una vez compensado, decrece en casi 1%. Por último, una vez compensado el fonema /j/ de la palabra “yo”, observamos cómo se presenta una diferencia pequeña pero importante, dada la baja frecuencia de este fonema. Se puede afirmar, entonces, que en el corpus oral de la trova, el rango alcanzado por este fonema, en primera instancia, depende del uso del pronombre de primera persona.

De otra parte, 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 palabras del tejido conectivo del discurso. 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 los artículos definidos femeninos y las preposiciones “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 marcas de plural de los artículos compensados, perdió su posición ante el fonema /i/ que también fue compensado, especialmente mediante la conjunción “y”. Se sigue afianzando la idea de que el lugar del rango adquirido por /s/ está ligado fuertemente a la marca del plural del léxico de contenido. Igualmente sucede con el fonema /l/; podemos seguir relacionando su dependencia de uso con el artículo definido y sus formas para el registro de su frecuencia. Además, aunque se mantiene en un rango estable, /d/ desciende ante su opuesto sordo /t/, y con un porcentaje inferior; la preposición “de” en su forma simple y en la contracción “del” parece tener incidencia en el cambio. /k/ desciende en el rango, abajo de los labiales /m/ y /b/. 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 asociado 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 /ks/ y su diferencia frecuencial entre los dos corpus. Parece ser evidente que hay muy poco léxico disponible de uso común con “x” 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 /ks/ en el uso escrito.

La restricción lingüística y la compensación del corpus permitieron ratificar la relación de algunos fenómenos gramaticales y la frecuencia fonemática. Debido a la significatividad de la compensación, la Tabla 6 presenta los resultados generales sobre la frecuencia de fonemas en estas muestras de español colombiano a partir de los datos compensados y el rango frecuencial fonemático que establece el promedio entre los dos corpus.

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%
r	6,05%	r	6,88%	r 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%
j	0,87%	j	0,84%	j 0,85%
f	0,63%	f	0,78%	f 0,71%
ʃ	0,46%	ʃ	0,32%	ʃ 0,59%
ɲ	0,28%	ɲ	0,27%	ɲ 0,28%
ks	0,06%	ks	0,14%	ks 0,10%

Tabla 6. Distribución de frecuencias de fonemas en dos corpus del español colombiano⁵.

La compensación comparada de los dos corpus vuelve a mostrar que las zonas estables son muy amplias y sus diferencias mínimas, a excepción del fonema /s/. Los dos corpus se manifiestan de manera muy homogénea, lo que reitera su validez y representatividad. Así, el promedio presentado da cuenta del rango porcentual de frecuencias de los fonemas del español en el uso colombiano.

Con base en esta clasificación, se desprenden las siguientes afirmaciones sobre el sistema fonémico en este corpus de español colombiano: 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

⁵ Para establecer los promedios, dada la particularidad de cada corpus, se toma cada uno como totalidad, independiente del número de fonemas, pues si se ponderara con base en este parámetro se estaría asumiendo que son un continuo, mientras en realidad son cualitativamente diferentes como hechos de lengua.

ellos con más de 7% de frecuencia y entre ellos suman el 52.67% del total de realizaciones. Con menos del 1% de aparición hay 7 fonemas (/r/, /x/, /j/, /f/, /tʃ/, /ɲ/, /ks/). Estos no alcanzan el 5% del total de la frecuencia y 5 de ellos tienen en común ser articulaciones “complejas” (a excepción de /x/ y /f/, que comparten la maticidad): africadas, palatales, o múltiples.

4. Discusión. La frecuencia de fonemas en lengua española

El análisis hecho nos permite ahora realizar una presentación contrastiva con algunos de los trabajos en el área (Tabla 7) especialmente con aquellos que han tomado corpus de distintas variedades de uso del español, para observar tendencias del sistema fónico en general, y particularidades del uso de la lengua en su variedad colombiana.

Podemos observar que el porcentaje de uso de las vocales, en este trabajo como en los tomados para el contraste, 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 los casos del uso de las preposiciones que la incluyen. 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 abiertas y las cerradas /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%), un poco menos de la mitad, una vez compensada con su aspecto funcional. Esto resalta aún más este aspecto gramatical.

Autores	Zapf y Rogers (1999)	Alarcos Llorach (1965)	Guitao y Borzone (1972)	Quilis/ Esquivela (1980)	Rojas (1991)	Guitao/ García Jurado (1993)	Listerri/ Marín (1993)	Pérez (2003)	Pineda et al. (2004)	Mejía/ González 2008
variedad	ibérica	ibérica	argentina	ibérica	ibérica y americana	argentina	ibérica	chilena	mexicana	colombiana
Fonemas										
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%	9,83%	9,98%	9,35%	10,75%	10,37%	9,28%	9,31%	11,20%
u	1,76%	2,10%	3,08%	3,33%	3,13%	2,79%	2,33%	3,03%	3,01%	2,93%
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%	2,39%	2,68%	2,60%	2,38%	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%
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%
ʃ	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%
ʎ	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%
ks	-----	-----	-----	-----	-----	-----	-----	-----	-----	0,10%
ɲ	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%
r	5,90%	7,00%	5,58%	5,19%	2,49%	5,40%	4,25%	6,19%	5,70%	6,47%
ɾ	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%	4,23%	5,05%	5,43%	5,28%
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%

Tabla 7. Comparación de la distribución de porcentaje de frecuencias de fonemas del español⁶.

⁶ Para la elaboración de esta 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, y Ivonne López, (2004). Se incluyeron en esta Tabla 7 los fonemas fricativo interdental sordo, /θ/, y lateral palatal, /ɲ/, que se

Una diferencia significativa se presenta en el fonema /l/, ya que a diferencia de los otros estudios, en los cuales oscila en un rango entre 3,88% (Argentina) y 5,43% (México), en el caso colombiano, compensado por su dependencia frecuencial respecto de las formas del artículo definido y de los pronombres de dativo, bajó a un promedio de 3.28% de uso.

Una diferencia un tanto amplia 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/, /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 /r/, 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 /tʃ/, /ɲ/, /j/, /ks/, ninguno de los cuales llega a 0.50%, siendo /ɲ/ el que ocupa el último lugar en siete de los 10 trabajos en contraste. La variedad mexicana es la única que presenta /j/ como el menos frecuente y la colombiana, la única que presenta /ks/, 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.

Zipf/ Rogers (1939) ibérica	Alarcos Llorach (1965) ibérica	Guirao/ Bozzano (1972) argentina	Quilis/ Eguavea (1980) ibérica	Rojo (1991) ibérica/ americana	Guirao/ García Jurado (1993) argentina	Listern/ Mairío (1993) ibérica	Pérez (2003) chilena	Pineda/ Cuatara (2004) mexicana	Mejía/ González 2008 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
r	r	i	r	l	i	i	i	i	r
l	l	r-r	t	d	r-r	r	r-r	r-r	s
d	t	t	d	t	t	r-r	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	u	b	l	d	k	k	b
b	b	u	m	p	m	m	u	u	k
m	m	m	n	m	b	b	m	m	l
p	u	p	p	r	u	p	p	p	u
j	p	b	b	n	p	u	b	b	p
u	θ	g	r	θ	g	θ	g	g	g
θ	g	f	θ	g	j	g	f	f	r
r	f	x	g	x	x	x	x	x	x
g	x	j	x	f	f	k	j	r	j
f	r	r	f	r	ʃ	f	r	j	f
k	k	ʃ	j	ʃ	r	ʃ	ʃ	ʃ	ʃ
x	j	p	k	j	p	r	p	p	p
p	ʃ		ʃ	p		j			ks
ʃ	p		p						

Tabla 8. Comparación del rango de fonemas del español.

En términos generales, la Tabla 8 corrobora lo afirmado por autores como Quilis (1993: 79) sobre la base articulatoria del español o de sus hábitos fonéticos: la preponderancia del uso de fonemas de articulación anterior, vocales anteriores y consonantes alveolares y dentales. Hay que anotar que esta

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 incidencia en la distribución frecuencial ni en el rango de fonemas totales. En esos estudios, el total del porcentaje de frecuencia de las consonantes obviamente 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%).

anterioridad no se incluye la labialidad; los fonemas labiales, /p, b, m, f, u/ aparecen entre el grupo de frecuencias de uso medio, a excepción de /o/.

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 /n/. 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%.

La Tabla 8 muestra una división acorde con el siguiente fundamento. Desde un punto de vista combinatorio, relacionado con las posibilidades de funcionamiento en la sílaba hispánica, se encuentran tres grupos de fonemas. En un primer grupo, resaltado en amarillo, y con índices de porcentaje frecuencial alrededor y sobre el 5%, están los fonemas que pueden aparecer en el habla española en cualquier posición silábica (a excepción de /r/), especialmente en función núcleo-silábica. 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 r/ históricamente han actuado en otras lenguas con dicha función; /krk/ “cuello” en checo, /litl/ “pequeño” en inglés, /abn/ “tener” en alemán, y fonéticamente en español coloquial se encuentran casos como [ntoes] por “entonces”, que permiten ver la potencialidad núcleo-silábica del fonema. 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 alargamiento⁷.

En segunda instancia, y resaltado en azul, tenemos un grupo dominado por las 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 pueden funcionar como ataque y coda silábicas (a excepción de /u/), su descenso en el 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.

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 /j/, /ks/, /tʃ/, /ɲ/ 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: 166), 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 puede imprimirle rasgos característicos y de identidad a una lengua.

5. Conclusiones

En relación con la metodología: la semejanza hallada en corpus distintos en procedencia y número recalca la homogeneidad de estructura lingüística de ellos para análisis el lingüístico. Con corpus medianamente robustos se generan resultados similares y de igual significatividad que con corpus de gran robustez. En la misma línea, la creación y aplicación del algoritmo fonético, con base en el principio fonémico, permitió un acercamiento más real a la lengua de uso y al sistema fonológico, ya que el conteo

⁷ /r/ puede conseguir continuidad en el habla, no a través de la vibración, sino por medio del fenómeno de fricativización.

estadístico se hizo sobre la base de fonemas y no de grafemas. Con este método se establece una pauta que puede mejorar futuros análisis de corpus escritos y una base de contraste entre varios corpus. Además, el análisis de la restricción lingüística y el procedimiento de compensación de los lexemas gramaticales de alta frecuencia léxica corroboraron su incidencia en la frecuencia fonemática. Así, se estableció la frecuencia fonemática sobre el léxico de contenido de la lengua.

En el español de uso colombiano el fonema con mayor frecuencia es /a/. Los fonemas vocálicos en orden decreciente de frecuencia son /aeoiu/. Los cinco fonemas más usados son /aeoin/, cada uno 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/, /tʃ/, /ɲ/, /ks/), que no alcanzan el 5% del total de la frecuencia. Cinco de ellos tienen en común ser articulaciones "complejas" (a excepción de /x/ y /f/, que comparten la maticidad): africadas, palatales, o múltiples.

Al comparar este trabajo con otros del ámbito hispánico puede afirmarse que /a e/ son los fonemas más usados, sin que se encuentre una relación con alguna causa específica. Las diferencias frecuenciales entre los demás resultados y el nuestro en los casos de los fonemas /l/ y /s/ reafirman su uso preponderante en los aspectos gramaticales anotados: uso del artículo y marca de plural, respectivamente. El fonema menos frecuente en lengua española varía entre /tʃ, ɲ, ʝ, ks/, siendo /ɲ/ el que ocupa el último lugar en siete de los trabajos analizados.

El análisis sobre los rangos fonemáticos en comparación determinó que hay en español una marcada tendencia al de uso de fonemas anteriores, de articulación anterior, vocales anteriores y consonantes alveolares y dentales, con exclusión de los fonemas labiales, /p, b, m, f, u/. Los fonemas sonoros son los más usados en la lengua; sin embargo, el fonema consonántico más usado es el fricativo alveolar sordo /s/ y el menos usado es el nasal palatal sonoro /ɲ/.

El rango de uso de fonemas de la lengua española tiene estrecha relación con la posibilidad de funcionamiento en la estructura silábica, ya que se encontraron tres grupos: los más usados, que son los que pueden aparecer en cualquier posición de la sílaba y son núcleo-silábicos o potencialmente nucleares. Los fonemas medianamente usados son los oclusivos y los labiales, cuya funcionalidad silábica permite su aparición en ataque y coda (a excepción de /u/), siendo esta última posición (implosiva, débil) causante de procesos históricos que reafirman la tendencia del español hacia la sílaba abierta. Entre los menos usados están los fonemas que en la lengua no aparecen, o en raras ocasiones se presentan como coda silábica.

Finalmente, además de contribuir al estudio de la variedad de habla, a su reconocimiento cultural y a la selección de material de apoyo para su enseñanza, se espera realizar un aporte a las tecnologías de la comunicación, dado que el análisis de las restricciones lingüísticas del corpus y su compensación permite establecer en parte qué es relevante para la comunicación en términos de materia de expresión y qué no. En el mismo sentido, la información fonológica apoyada en 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 para 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 relacionadas con las tecnologías del habla.

Referencias bibliográficas

- Alarcos Llorach, Emilio. 1965. *Fonología española*. Madrid, Gredos.
- Alcina, J. y Blecua, José Manuel. 1975. Frecuencia de fonemas en español, en *Gramática española*. Barcelona, Ariel: 430-435.
- Caravedo, Rocío. 1999. *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español*. Salamanca, Ediciones Universidad de Salamanca.
- Crystal, David. 1994. *Enciclopedia del lenguaje*. Madrid, Taurus.
- Flórez, Luis. 1978. Sobre algunas formas de pronunciar muchos colombianos el español. *Thesaurus*, Boletín del Instituto Caro y Cuervo, T. XXXIII, separata.

- Guirao, Miguelina y García Jurado, María Amalia. 1993. *Estudio estadístico del español*. Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas.
- Guirao, Miguelina y Borzone de Manrique, Ana. 1972. Fonemas, sílabas y palabras del español de Buenos Aires. *Filología*, XVI: 135-165.
- Llisterri, J. y Mariño, J. 1993. *Spanish adaptation of SAMPA and automatic phonetic transcription*. [en línea]. http://liceu.uab.cat/~joaquim/publicacions/SAMPA_Spanish_93.pdf
- Moreno Sandoval, Antonio; Torre Toledano, Doroteo; Curto, Natalia; y De la Torre, Raúl. 2006. Inventario de frecuencias fonémicas y silábicas del castellano espontáneo y escrito. *IV Jornadas en Tecnología del Habla, Zaragoza* [en línea]. Disponible en http://jth2006.unizar.es/finals/4jth_116.pdf
- Mosterin, Jesús. 1981. *La ortografía fonémica del español*. Madrid, Alianza Universidad.
- Pérez, Hernán Emilio. 2003. Frecuencia de fonemas. *e-rthabla*, 1, Abril 2003, [en línea]. Disponible en http://orien.die.upm.es/~lapiz/e-rthabla/numeros/N1/N1_A4.pdf
- Pineda, L. A.; Villaseñor P., L.; Cuétara, J.; Castellanos, H.; e I. López. 2004. DIMEx100: A New Phonetic and Speech Corpus for Mexican Spanish. En C. Lemaître, C.A. Reyes, and J.A. Gonzalez (Eds.), *Iberamia*, 2004, LNAI 3315: 974-983, [en línea]. Disponible en <http://leibniz.iimas.unam.mx/~luis/DIME/publicaciones/papers/DIMEx100-LNAI3315.pdf>
- Quilis Morales, Antonio y Esgueva Martínez, M. A. 1980. Frecuencia de fonemas en el español hablado. *LEA: Lingüística española actual*, II-I: 1-25.
- Quilis, Antonio. 1993. *Tratado de Fonología y Fonética Españolas*. Madrid, Gredos.
- Rojo, Guillermo. 2008. Lingüística de corpus y lingüística del español. XV Congreso de la ALFAL, Montevideo, agosto 18-21 de 2008. Disponible en http://www.lehrbuch-online.de/downloads/spanische_sprachwissenschaft/rojo.pdf
- _____. 2002. Sobre la lingüística basada en el análisis del corpus. *Hizkuntza-corpusak*, 24/25. Disponible en: www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54A.pdf
- _____. 1991. Frecuencia de fonemas en español actual. *Homenaxe ó Profesor Constantino García*, Universidad de Santiago de Compostela: 451-467.
- 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 Expérimentale*, 15 : 111-147.

