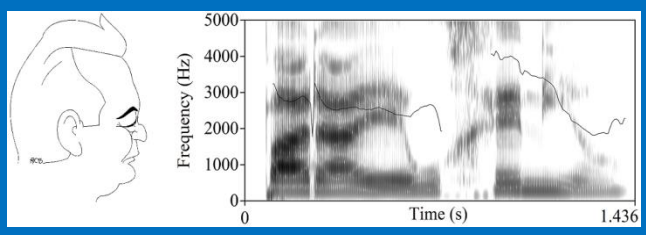
Corpus de habla espontánea para el estudio de la entonación

Francisco José Cantero Serena



Fernández Planas, A. Ma. (ed.) (2016): *53 reflexiones sobre aspectos de la fonética y otros temas de lingüística*, Barcelona, págs. 151-160.
ISBN: 978-84-608-9830-6.

Corpus de habla espontánea para el estudio de la entonación

Francisco J. Cantero Serena
Universitat de Barcelona
cantero@ub.edu

Para Eugenio Martínez Celdrán, maestro de una generación de fonetistas, cuyas enseñanzas se expanden fecundas por muchos caminos.

1. INTRODUCCIÓN

El estudio de la lengua oral ha estado mediatizado, tradicionalmente, por la transcripción escrita de los textos orales. Curiosamente, para obtener esos textos orales también se ha empleado la lengua escrita como soporte y, de hecho, buena parte de la investigación experimental en fonética, todavía hoy, está basada en fragmentos de habla de laboratorio, inducida por la lectura de textos escritos.

Frente a los corpus de habla de laboratorio, los corpus de habla espontánea constituyen ejemplos de habla genuina, no inducida, ni preparada, ni condicionada por los investigadores; sino conjuntos de enunciados extraídos de conversaciones reales, en las que los informantes ni siquiera saben que su discurso está siendo objeto de estudio. El sentido de elaborar corpus de habla espontánea es, justamente, acercarse al estudio del habla real, es decir, de aquello que hacen los hablantes cuando se relacionan: de la comunicación oral genuina.

La mayoría de los corpus orales empleados en lingüística (*Speech Corpora*) está focalizada en el desarrollo de aplicaciones tecnológicas (como el reconocimiento automático), por lo que se recurre frecuentemente a formatos de laboratorio: frases leídas, palabras aisladas, dígitos encadenados, etc[1].

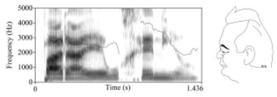
Solo algunos corpus se centran en la comunicación oral genuina (*Spoken Language Corpora*), y estos constituyen un avance metodológico esencial en el estudio de la lengua oral, porque ofrecen fragmentos de habla real en su contexto, producidos por hablantes reales en su contexto real[2].

Frente al término «habla coloquial», empleado frecuentemente, preferimos el de «habla espontánea» para referirnos a su caracterización metodológica: no es habla preparada, ni

[1] En Llisterri et al. (2005) se ofrece una lista muy completa de este tipo de corpus en español.

[2] Son ejemplos bien conocidos de este tipo de corpus: el *Corpus de Català Contemporani de la Universitat de Barcelona (CCCUB)* para el catalán, el *Santa Barbara Corpus of Spoken American English* para el inglés, o el *IFA Spoken Language Corpus* para el holandés.

 ISBN 978-84-608-9830-6



leída, ni es habla profesional; «coloquial» se refiere a un registro de habla y en habla «espontánea» también pueden aparecer otros registros[3].

Los corpus que presentamos aquí han sido elaborados bajo los auspicios del *Laboratorio de Fonética Aplicada (LFA)* de la Universidad de Barcelona (www.ub.edu/lfa). Todos ellos son corpus orales en el sentido de los *Spoken Language Corpora*, y los constituyen diversos conjuntos de enunciados extraídos de conversaciones genuinas, de los que ofrecemos no una mera transcripción textual, sino un análisis melódico pormenorizado, junto con el achivo sonoro.

2. CRITERIOS PARA LA ELABORACIÓN DE LOS CORPUS

Elaboramos los corpus a partir de la grabación de programas de televisión (v. Cantero, 2002) con participación de público en directo, generalmente anónimo (pero del cual se obtienen numerosos datos personales, como la edad, la procedencia, la profesión o el perfil sociolingüístico). Optamos por las grabaciones televisivas, frente a otros medios, porque en ellas podemos contar con numeroso público no profesional: en radio, por ejemplo, normalmente nos encontramos con locutores profesionales y tertulianos «profesionalizados». Además, en televisión se suceden las entrevistas a pie de calle, los debates y, especialmente, los concursos abiertos al público y los *reality-show*.

Del corpus de grabaciones extraemos los enunciados a analizar, según una serie de criterios que permitan determinar su calidad y pertinencia:

Sobre los informantes:

1. Despreciamos los enunciados emitidos por locutores profesionales (actores, presentadores o periodistas) y por todos aquellos participantes que siguen un guión previo o que leen: de este modo, nos aseguramos que los enunciados que constituyen el corpus están producidos por hablantes genuinamente espontáneos.
2. Despreciamos, igualmente, los participantes de programas televisivos con un *casting* previo (que, a menudo, se alarga varios meses, durante los cuales el concursante, de hecho, se profesionaliza como «locutor» o «actor»).
3. Comprobamos que el informante no tiene ningún defecto de habla y es hablante nativo.

Sobre los enunciados:

1. Seleccionamos solo enunciados significativos, producidos con naturalidad en el curso de una conversación real.
2. No seleccionamos más de 10 enunciados de un mismo informante.
3. Comprobamos que no hay solapamiento de voces, ni ruidos o música de fondo que impidan el análisis.

[3] En español, el corpus de referencia en habla coloquial es el corpus VAL.ES.CO (v. Briz, 1995 y 2002).

El etiquetaje de cada enunciado incluye su transcripción, su localización en la grabación, los datos identificativos del programa televisivo (incluyendo hora y día de emisión), la descripción del contexto conversacional en que se produjo el enunciado y los datos del informante.

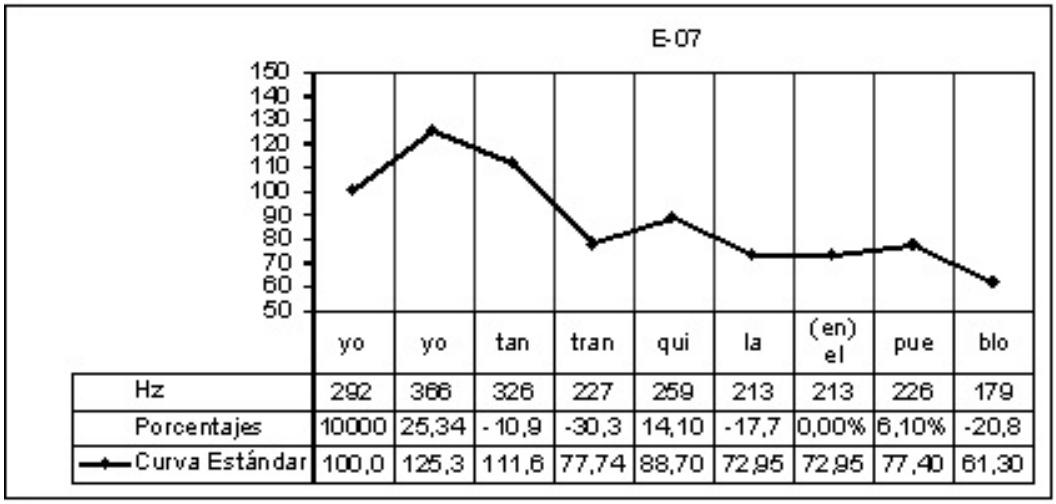
3. CORPUS DEL ESPAÑOL PENINSULAR

Contamos con diversos corpus de habla espontánea del español. Un primer corpus peninsular, en el que no se distingue la procedencia de los informantes, y dos conjuntos de corpus específicos, por comunidades autónomas (v. Ballesteros et al., 2010):

1. *Corpus Alfonso* (1999): se trata de un corpus de 6 horas de grabación, con un total de 90 enunciados, producidos por 37 informantes (15 hombres y 22 mujeres).

2. *Corpus Ballesteros* (2011): 5 corpus de habla espontánea, de las comunidades de Asturias, Navarra, Euskadi, Castilla León y Madrid. Se recogió un total de 58 horas de grabación, de las que se extrajeron 1.000 enunciados producidos por un total de 302 informantes (179 hombres y 123 mujeres).

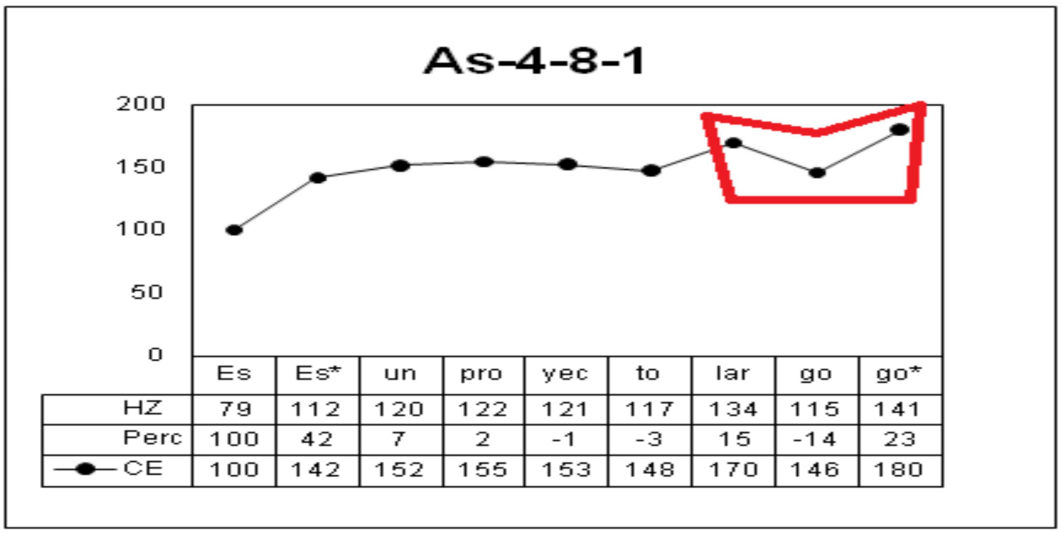
3. *Corpus Mateo* (2013): 5 corpus de habla espontánea, de las comunidades de Andalucía, Canarias, Castilla La Mancha, Extremadura y Murcia. Se recogió un total de 309 horas de grabación, de las que se extrajeron 1.851 enunciados producidos por 475 informantes (274 hombres y 201 mujeres).



	yo	yo	tan	tran	qui	la	(en) el	pue	blo
Hz	292	366	326	227	259	213	213	226	179
Porcentajes	10000	25,34	-10,9	-30,3	14,10	-17,7	0,00%	6,10%	-20,8
Curva Estándar	100,0	125,3	111,6	77,74	88,70	72,95	72,95	77,40	61,30

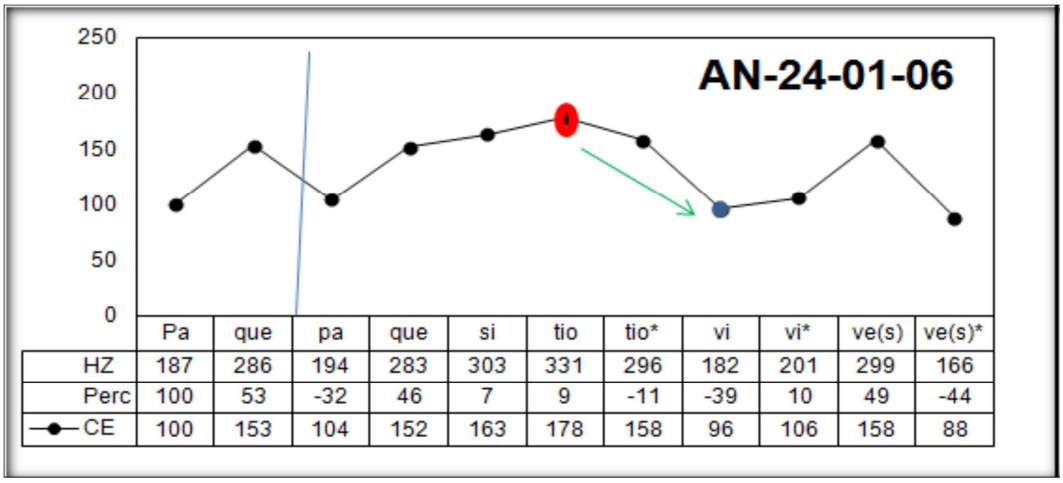
Figura 1. *Gráfico del enunciado* Yo tan tranquila (en) el pueblo, *del Corpus Alfonso.*

 2016 ISBN 978-84-608-9830-6



	Es	Es*	un	pro	yec	to	lar	go	go*
HZ	79	112	120	122	121	117	134	115	141
Perc	100	42	7	2	-1	-3	15	-14	23
CE	100	142	152	155	153	148	170	146	180

Figura 2. *Gráfico del enunciado* Es un proyecto largo, *del Corpus Ballesteros (Asturias).*



	Pa	que	pa	que	si	tio	tio*	vi	vi*	ve(s)	ve(s)*
HZ	187	286	194	283	303	331	296	182	201	299	166
Perc	100	53	-32	46	7	9	-11	-39	10	49	-44
CE	100	153	104	152	163	178	158	96	106	158	88

Figura 3. *Gráfico del enunciado* ¿Pa(ra) qué sitio vives?, *del Corpus Mateo (Andalucía).*

Los corpus de las variedades del español peninsular (Ballesteros + Mateo) cuentan, en total, con: 367 horas de grabación, 2851 enunciados y 777 informantes (453 hombres y 324 mujeres). Dejamos para más adelante la elaboración de los corpus específicos de aquellas comunidades en las que el español convive con otras lenguas románicas (Catalunya, Balears, Aragón y Galicia).

En conjunto, nuestro corpus de habla espontánea, que se ha ido enriqueciendo a lo largo de los últimos años, cuenta con un total de:

1. 373 horas de grabación
2. 2941 enunciados
3. 814 informantes (468 hombres y 346 mujeres)

En él se basaron nuestros primeros trabajos sobre la entonación lingüística del habla espontánea, como Cantero et al. (2002 y 2005) y, sobre todo, Cantero y Font-Rotchés (2007), Ballesteros (2011) y Mateo (2013) –centrados, estos últimos, en la descripción de la entonación prelingüística.

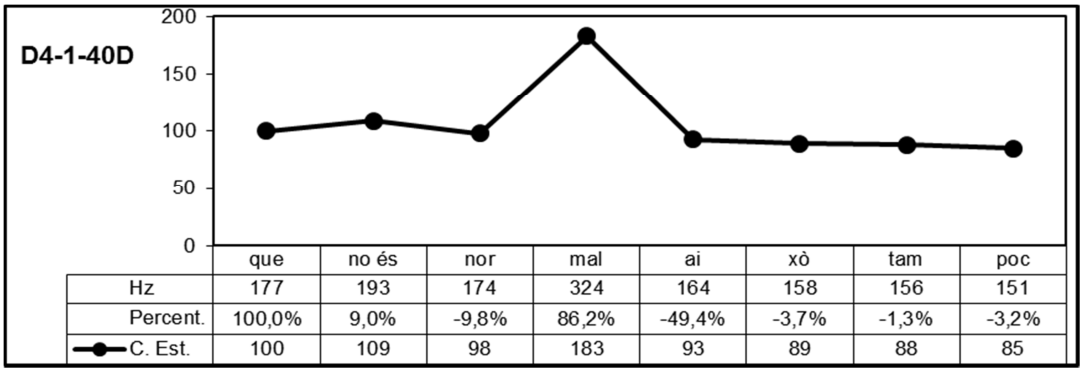
Sobre este corpus, también se vienen realizando diversos análisis acústicos centrados en la pronunciación del español, sobre el vocalismo (Alfonso, 2010) y sobre los sonidos laterales (Andrés, 2014), vibrantes (Ortiz, 2012 y 2014) y aproximantes (Sola, 2011, 2014a y 2014b).

4. OTROS CORPUS DE HABLA ESPONTÁNEA

Con los mismos criterios, se elaboró un corpus de habla espontánea en catalán central, que sirvió de base para la descripción de la entonación lingüística del catalán (Font-Rotchés, 2007). El *Corpus Font-Rotchés* consta de:

1. 47 horas de grabación
2. 580 enunciados
3. 160 informantes (98 hombres y 67 mujeres)

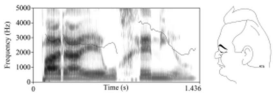
Este corpus sigue ampliándose en la actualidad, con el objeto de reflejar las distintas variedades del catalán, lo que permitirá caracterizar los distintos perfiles melódicos de la lengua. El corpus ha servido también para la descripción de la entonación de (des)cortesía en catalán (Devís y Cantero, 2014) y para la descripción del vocalismo del catalán (Rius-Escudé, 2015).



	que	no és	nor	mal	ai	xò	tam	poc
Hz	177	193	174	324	164	158	156	151
Percent.	100,0%	9,0%	-9,8%	86,2%	-49,4%	-3,7%	-1,3%	-3,2%
C. Est.	100	109	98	183	93	89	88	85

Figura 4. *Gráfico del enunciado* Que no és normal això tampoc, *del Corpus Font-Rotchés*.

 2016 ISBN 978-84-608-9830-6



Siguiendo el mismo protocolo de elaboración, contamos también con corpus de habla espontánea en alemán y en portugués:

1. *Corpus multisistémico de partículas modales del alemán* (Torregrosa, 2011).
2. *Corpus de portugués de Brasil* (Araújo, 2014; Cantero y Font-Rotchés, 2013; Mendes, 2013)

No siguen el mismo protocolo, en cambio, otros corpus de habla (espontánea o semiespontánea) que, por sus características, no pueden elaborarse sin ninguna participación del investigador: se trata de nuestros corpus de ELE, cuyo objetivo es describir el perfil melódico de los hablantes de español como lengua extranjera (v. Cantero y Devís, 2011).

En estos casos, en los que no es posible contar con programas televisivos, los informantes son seleccionados siguiendo criterios de homogeneidad y grabados directamente por el investigador, en situación conversacional (desfocalizando, en todo momento, el objeto de estudio):

1. *Español hablado por hablantes taiwaneses* (Liu y Cantero, 2002; Liu, 2005)
2. *Español hablado por brasileños* (Fonseca, 2013)
3. *Español hablado por húngaros* (Baditzné Pálvölgyi, 2012)
4. *Español hablado por suecos* (Martorell, 2014)
5. *Español hablado por norteamericanos* (Muñoz, 2014)

En definitiva, contamos con un amplio conjunto de corpus de habla espontánea para el estudio de la entonación, susceptible también de ser utilizado para el análisis acústico segmental. Nuestra intención es ir poniéndolo a disposición de la comunidad científica, paulatinamente, a partir de 2016 en nuestra web http://www.ub.edu/lfa/

5. REFERENCIAS BIBLIOGRÁFICAS

ALFONSO, R. (2010): El vocalismo del español en habla espontánea, tesis doctoral, Universitat de Barcelona.

ANDRÉS EDO, B. (2014): «Análisis acústico de los sonidos laterales en el habla espontánea del español», *Phonica*, 9-10, pp. 13-20. http://revistes.ub.edu/index.php/phonica/article/view/10944 [18/11/2016].

ARAÚJO, M. L. (2014): *Entonação das interrogativas e das declarativas do português brasileiro falado em Minas Gerais: Modelos para o Ensino de Línguas.* Dissertação de Mestrado, Universidade de Brasília.

BADITZNÉ PÁLVÖLGYI, K. (2012): *Spanish Intonation for Hungarian learners: yes/no questions.* Biblioteca Phonica, 15. http://www.publicacions.ub.edu/revistes/phonica-biblioteca/15_Kata.pdf [18/11/2016].

BALLESTEROS, M. (2011): *La entonación del español del norte.* tesis doctoral, Universitat de Barcelona.

BALLESTEROS, M.; M. MATEO y F. J. CANTERO (2010): «Corpus oral para el análisis melódico de las variedades del español», en P. Cano López, S. Cortiñas Ansoar, B. Dieste Quiroga. I. Fernández López y L. Zas Varela (eds): *Actas del XXXIX Simposio Internacional de la SEL*, Santiago de Compostela, Universidad de Santiago de Compostela, p..66.

BRIZ, A. (coord.) (1995): *La conversación coloquial (Materiales para su estudio)*, anejo XVI de la Revista Cuadernos de Filología, Universidad de Valencia.

BRIZ, A. y GRUPO VAL.ES.CO. (2002): *Corpus de conversaciones coloquiales*, Anejo de la Revista *Oralia*, Madrid, Arco-Libros.

CANTERO, F. J. (2002): *Teoría y análisis de la entonación*, Barcelona, Edicions de la UB.

CANTERO, F. J.; M. A. DE ARAÚJO; Y. H. LIU; Y. K. WU y A. ZANATTA (2002): «Patrones melódicos de la entonación interrogativa del español en habla espontánea», en J. Díaz García (ed.): *Actas del II Congreso de Fonética Experimental*, Sevilla, Universidad de Sevilla, pp. 118-123.

CANTERO, F. J.; R. ALFONSO; M. BARTOLÍ; A. CORRALES y M. VIDAL (2005): «Rasgos melódicos de énfasis en español», *Phonica*, vol. 1.
http://revistes.ub.edu/index.php/phonica/article/view/5571 [18/11/2016].

CANTERO, F. J. y E. DEVÍS (2011): «Análisis melódico de la interlengua», en A. Hidalgo, Y. Congosto y M. Quilis (eds.): *El estudio de la prosodia en España en el siglo XXI: perspectivas y ámbitos.* anejo nº 75 de *Quaderns de Filologia*, Universitat de València, pp. 285-299.

CANTERO, F. J. y D. FONT-ROTCHÉS (2007): «Entonación del español peninsular en habla espontánea: patrones melódicos y márgenes de dispersión», *Moenia*, 13, pp.69-92.

CANTERO, F. J. y D. FONT-ROTCHÉS (2009): «Protocolo para el análisis melódico del habla», *Estudios de Fonética Experimental*, XVIII, pp. 17-32.

CANTERO, F. J. y D. FONT-ROTCHÉS (2013): «The intonation of absolute questions of Brazilian Portuguese», *Linguistics and Literature Studies*, 1(3), pp. 148-149.

CANTERO, F. J. y Y.-H- LIU (2002): «La entonación prelingüística del español hablado por taiwaneses: establecimiento de un corpus», en J. Díaz García (ed.): *Actas del III Congreso de Fonética Experimental*, Sevilla, Publicaciones de la Universidad de Sevilla, pp. 238-242.

DEVÍS, E. y F. J. CANTERO (2014): «The intonation of mitigating politeness in Catalan», *Journal of Politeness Research*, 10, 1, pp. 127-149.

FONSECA DE OLIVEIRA, A. (2013): *Caracterización de la entonación del español hablado por brasileños.* tesis doctoral, Universitat de Barcelona.

FONT-ROTCHÉS, D. (2007): *L'entonació del català.* Barcelona. Publicacions de l'Abadia de Montserrat.

LIU, Y-H (2005): *La entonación del español hablado por taiwaneses*, Biblioteca Phonica, 2.
http://www.publicacions.ub.edu/revistes/phonica-biblioteca/esp_taiw/esp_taiw.pdf [18/11/2016].

LLISTERRI, J.; C. DE LA MOTA, M. J. MACHUCA, A. RÍOS y M. RIERA (2005): «Corpus orales para el desarrollo de las tecnologías del habla en español», *Oralia*, 8, pp. 289-325.

MARTORELL, L. (2014): «Aproximació als trets melòdics de les interrogatives de l'espanyol parlat per suecs», *Phonica*, 9-10.
http://revistes.ub.edu/index.php/phonica/article/view/10968 [18/11/2016].

MATEO RUIZ, M. (2013): *La entonación del español meridional*, tesis doctoral, Universitat de Barcelona.

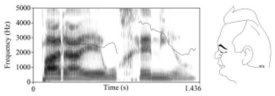
MENDES, S. R. (2013): *A entonação no processo de ensino-aprendizagem de PLE. Proposta didática para o ensino de modelos de entonação interrogativa do português do Brasil- Estado de São Paulo.* Dissertação de Mestrado em Linguística Aplicada, Universidade de Brasília.

MUÑOZ, A. (2014): «Aproximación al perfil melódico de la interlengua de anglófonos que hablan español como segunda lengua», *Phonica*, 9-10, pp. 115-122.
http://revistes.ub.edu/index.php/phonica/article/view/10970 [18/11/2016].

ORTIZ DE PINEDO, N. (2012): «Las vibrantes del español en habla espontánea», *Phonica*, 8. pp. 44-67.
http://revistes.ub.edu/index.php/phonica/article/view/10189 [18/11/2016].

ORTIZ DE PINEDO, N. (2014): «Análisis acústico de la vibrantes del español en habla espontánea», *Phonica*, *9-10*, pp. 21-32.
http://revistes.ub.edu/index.php/phonica/article/view/10958 [18/11/2016].

RIUS-ESCUDÉ, A. (2015): *Les vocals del català central en parla espontània*, tesi doctoral, Universitat de Barcelona.



SOLA PRADO, A. (2011): «Las aproximantes [β, δ, γ] del español en habla espontánea», *Phonica*, *7*, pp. 118-140.
http://revistes.ub.edu/index.php/phonica/article/view/5609 [18/11/2016].

SOLA PRADO, A. (2014a): «Estudio sobre las aproximantes [ß, ∂, γ] del español en habla espontánea», *Phonica*, 9-10, pp. 41-46.
http://revistes.ub.edu/index.php/phonica/article/view/10960 [18/11/2016].

SOLA PRADO, A. (2014b): «Caracterización acústica de las aproximantes [ß, ∂, γ] del español en habla espontánea», en Y. Congosto, M. L Montero Curiel y A. Salvador (eds): *Fonética Experimental, Educación Superior e Investigación*, vol. *I*, Madrid, Arco-Libros, pp. 436-464.

TORREGROSA AZOR, J. (2011): *Análisis multisistémico de las partículas modales del alemán*. Biblioteca Phonica, 13-14, pp. 135-147.
http://revistes.ub.edu/index.php/phonica/article/view/10973 [18/11/2016].