Habla culta de Caracas 1973-2011
Un subcorpus de propósito especial para el estudio diacrónico del habla caraqueña

Krístel Guirado
Universidad Central de Venezuela, Universidad de Zaragoza

The term *Corpus Reengineering* is proposed to refer to the process of reconfiguration of speech samples for reuse in various scopes (Guirado 2014 y 2015). The process was developed with two corpora of Spanish spoken in Caracas, aiming at the construction a *special purpose corpus* for diachronic studies: *Corpus del habla culta de Caracas* 1968-77 (cf. Rosenblat & Bentivoglio 1979) and *Corpus sociolingüístico de Caracas PRESEEA 2004-10* (cf. Bentivoglio & Malaver 2014). The methodology included the following steps: i. describe the structuring of the original corpus; ii. evaluate the inoperative and stable aspects of each architecture; and, iii. create the new design and estimate its representativeness. Such methodology produced a new corpus for the real time study of linguistic phenomena in a specific speech community (cultured speakers): *Habla culta de Caracas 1973-2011. Corpus diacrónico*. It is concluded that *Corpus Reengineering* is a rewarding practice for Corpus Linguistics that provides useful and interesting products for the linguistic community.

Keywords: *Corpus Reengineering*, special purpose corpus, corpus linguistics, Spanish spoken in Caracas, diachronic study.

1. Los corpus electrónicos y su reingeniería

La creación de corpus específicos a partir del rediseño radical y la reconcepción de uno o varios corpus responde a la necesidad de incrementar y diversificar los ámbitos de análisis de los fenómenos del lenguaje. El término *Reingeniería de Corpus* se usa para designar las tareas relacionadas con la reconfiguración de materiales de habla (orales y escritos) recopilados y estructurados en diversas bases y cuerpos de datos cuando su productividad se torna vulnerable (cf. Guirado 2014 y 2015).

CHIMERA. Romance Corpora and Linguistic Studies 5.1 (2018), 111-116. ISSN 2386-2629

De acuerdo con el criterio de representatividad, los corpus pueden ser objeto de diversas tipologías. Atkins, Clear & Ostler (1992) diferencian entre *corpus* y *subcorpus*. Para Sinclair (1996) y Torruella y Llisterri (1999), un subcorpus es cualquier porción seleccionada de un corpus mayor. Pearson propone un tipo de corpus que no se puede incluir dentro de las clasificaciones previas, el *corpus de propósito especial* (*especial purpose corpus*): "a corpus whose composition is determined by the precise purpose for which it is to be used" (Pearson 1998: 48). En la presente comunicación, se muestra el proceso para la producción de un corpus de propósito especial a partir de la reingeniería de dos corpus orales para el estudio del español hablado en Caracas.

2. Los corpus objeto de reingeniería

2.1 *Corpus del Habla Culta de Caracas 1968-77* (CHCC 68-77)

En 1966, Caracas se suma al *Proyecto para el estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica* (cf. Lope Blanch 1987). Entre los años 1968 y 1977, se grabaron unas 240 conversaciones, con la participación de 320 hablantes, todos caraqueños cultos, residenciados en la ciudad e hijos de padres caraqueños.

Según la metodología preestablecida en el denominado *Proyecto del Habla Culta* (cf. Rabanales 1992), se tomaron en cuenta solo dos variables sociales para distribuir las grabaciones: edad y sexo. A cada grupo etario le corresponde un porcentaje de grabación del total de encuestas: i) 25 a 35 años (30%); ii) 36 a 55 años (45%); y, iii) más de 56 años (25%). Para cada rango, se procuró una distribución equitativa (50%) entre mujeres y hombres.

2.2 *Corpus sociolingüístico de Caracas PRESEEA 2004-10* (PRESEEA-CSC 04-10)

El PRESEEA-CSC 04-10 forma parte del *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (cf. Moreno Fernández 1997). El desarrollo del proyecto en Caracas está coordinado por las investigadoras Paola Bentivoglio e Irania Malaver.

La muestra de Caracas está formada por 108 entrevistas hechas a informantes nacidos en la ciudad, hijos de padres caraqueños. El corpus se dividió en tres grupos generacionales (1: 20 a 34 años; 2: 35 a 54; y 3: 55 o más años); tres grados de instrucción (1: primaria; 2: secundaria; 3: superior); y sexo.

3. El estudio diacrónico del habla culta caraqueña

Usar los materiales del proyecto PRESEEA para crear un corpus diacrónico del habla culta de Caracas representó una economía de recursos y tiempo, pero trajo algunos inconvenientes metodológicos.

Primero, los criterios de selección de los hablantes difieren en ambos proyectos, de modo que no todos los entrevistados del nivel de instrucción universitaria de PRESEEA-CSC 04-10 completan las características requeridas en las bases del proyecto *Norma Culta* para ser considerados hablantes de esta modalidad[1]. Segundo, resulta difícil establecer criterios precisos de representatividad de la muestra, ya que no se dispone de datos estadísticos precisos sobre la realidad para calcular la proporción ideal de entrevistados.

Para salvar este escollo, se decidió partir de la noción de representatividad asociada a los corpus de propósito especial: "el criterio de representatividad debe restringirse a la del dominio de estudio específico para el que son creados" (Pérez Hernández & Moreno Ortiz 2009: 76). De esta forma, se pudo optar por determinar una cuota de afijación de tres hablantes por casilla[2]. Con este número logré, por una parte, ajustar la posibilidad de ubicar hablantes "cultos" en PRESEEA-CSC 04-10 y, por otra, establecer una proporción relativamente representativa de la población culta caraqueña en ambos períodos[3].

Asimismo, se consideró el criterio de longitud discursiva de las entrevistas para la distribución del nuevo corpus. Para ello, se reajustó la duración de las mismas a los porcentajes establecidos para los grupos etarios en el proyecto original (30%: 25 a 35 años; 45%: 36 a 55 años; y, 25%: 56 años en adelante); de

[1] Se considera informante culto: "1) aquel que tuviera estudios universitarios completos, 2) que conociera a lo menos una lengua extranjera, 3) que hubiera realizado lecturas relevantes y 4) que, en lo posible, hubiera viajado fuera del país" (Rabanales 1992: 258).

[2] Moreno Fernández 2005 señala que "normalmente cada cuota es representada por entre tres y cinco informantes" (312).

[3] Como se mencionó antes, no se dispuso de una cifra exacta de la totalidad de hablantes "cultos" caraqueños; sin embargo, se obtuvieron algunos indicadores con los cuales se pudo hacer algunas estimaciones indirectas. Censo de 2006: 314.291 habitantes trabajan en ámbitos profesionales y de categoría directiva en el Distrito Capital; asimismo, el promedio aproximado de escolaridad de la población mayor de 25 años es de 9 años, es decir solo una minoría de encuestados había completado los estudios universitarios. Censo de 2011: de 1.026.706 habitantes caraqueños que respondieron los datos educativos, apenas 84.091 tenían estudios universitarios, y solo 67.549 los completaron (5 a 6 años) (Fuente: http://www.ine.gov.ve).

modo que, en lugar de más hablantes para el grupo 2, se le adjudicó mayor tiempo[4].

El nuevo corpus se denomina *Habla culta de Caracas 1973-2011. Corpus diacrónico* (HCC/CD 73-11), para hacer referencia a su propósito especial y al período de grabación que abarcan las encuestas compiladas (el año de grabación más distante de las primeras encuestas y el año más próximo de las últimas, respectivamente, cf. tabla 2, *infra*). Los 36 hablantes cuyas grabaciones constituyen el corpus están distribuidos como se presume por sexo, tres grupos etarios (1:25 a 35 años; 2:36 a 55 años; y, 3:56 años en adelante), y dos períodos de grabación (HCC 73 y HCC 11). De acuerdo con los porcentajes establecidos, la duración en minutos de cada entrevista por grupo etario quedaría como sigue: 25 a 35 años: 30 min.; 36 a 55 años: 45 min; y, 56 años en adelante: 25 min. En la tabla 1, se puede observar la distribución del nuevo corpus:

Tabla 1. Distribución en hablantes y minutos del HCC/CD 73-11

		GRUPO ETARIO							
		Grupo 1 20 a 35 años		Grupo 2 36 a 55 años		Grupo 3 56 años o más		Total	
Período de grabación		hombres	mujeres	hombres	mujeres	hombres	mujeres	∑	*Min.*
HCC-1973	*n*	3	3	3	3	3	3	18	600
	min.	90	90	135	135	75	75		
HCC-2011	*n*	3	3	3	3	3	3	18	600
	min.	90	90	135	135	75	75		
Total		12		12		12		**36**	1200
		30% =360		45% =540		25% = 300			

En la submuestra correspondiente al período grabado entre 1973 y 1975, se incluyeron 14 entrevistas publicadas en Rosenblat & Bentivoglio (1979: 11-233). Adicionalmente, se escogieron del corpus matriz cuatro encuestas inducidas de un solo informante grabadas entre estos años –se digitalizaron y corrigieron– para completar las 18 requeridas en el primer período (un hombre y una mujer del grupo 1 y un hombre y una mujer del grupo 2). Para los materiales del segundo período se seleccionaron 18 transcripciones de hablantes pertenecientes al grupo 3 de instrucción del corpus PRESEEA-CSC 04-10 (Bentivoglio & Malaver 2014).

[4] Se parte del criterio de que depende de la segunda generación "en su mayor parte, el prestigio y expansión de la norma culta que ella misma contribuye a formar" (Rabanales 1992: 264-5). La reasignación de los porcentajes de acuerdo con la longitud discursiva tiene su origen en la distribución que Rosenblat hace para la publicación de las entrevistas de un solo informante (Rosenblat & Bentivoglio 1979: 10).

Las entrevistas seleccionadas se postcodificaron con una nomenclatura que describe la reconfiguración de los materiales. La identificación de las encuestas del HCC-CD 73-11 se puede observar en el tabla 2:

Tabla 2. Postcodificación y distribución de encuestas del HCC/CD 73-11

	GRUPO ETARIO					
	Grupo 1 20 a 35 años		Grupo 2 36 a 55 años		Grupo 3 56 años o más	
Período de grabación	hombres	mujeres	hombres	mujeres	hombres	mujeres
HCC-1973	CAH1A.73	CAM1A.74	CAH2A.73	CAM2A.75	CAH3A.73	CAM3A.73
	CAH1B.73	CAM1B.74	CAH2B.73	CAM2B.75	CAH3B.73	CAM3B.74
	CAH1C.73	CAM1C.73	CAH2C.73	CAM2C.75	CAH3C.73	CAM3C.73
HCC-2011	CAH1A.05	CAM1A.04	CAH2A.04	CAM2A.05	CAH3A.04	CAM3A.05
	CAH1B.07	CAM1B.04	CAH2B.07	CAM2B.05	CAH3B.05	CAM3B.09
	CAH1C.09	CAM1C.04	CAH2C.08	CAM2C.08	CAH3C.11	CAM3C.09

El nuevo código resume los datos en ciudad (CA), sexo (H/M), grupo etario (1/2/3), distinción del hablante (A/B/C) y año de grabación (73-75 y 04-11). El corpus completo suma un total de 20 horas de grabación. Queda pendiente la tarea de calcular el número de palabras.

4. Comentarios finales

Melis, Flores y Bogart 2003 consideran que durante el siglo pasado tuvo lugar un tercer período evolutivo en la historia del español. A partir de esta afirmación, Pons Bordería 2014 propone tratar el siglo XX "como espacio relevante para el estudio de la evolución diacrónica", en virtud de la trascendencia de algunos rasgos de su "historia externa" (1001). En concreto, el autor plantea estudiar: "*microdiacronías*: pequeños periodos de tiempo de especial relevancia en la datación de un cambio lingüístico" (1007).

El presente trabajo pone en evidencia que la intervención en la arquitectura original de un corpus requiere de una revisión fundamental para generar cambios que permitan ofrecer un producto útil y rentable para la comunidad de usuarios. En este caso, la reingeniería del HCC/CD 73-11 no estuvo exenta de los mismos dilemas metodológicos propios de la LC, especialmente en lo concerniente a los criterios de representatividad y tamaño. No obstante, un tratamiento adecuado de los metadatos de construcción permitió zanjar los obstáculos de representatividad y presentar un corpus que no arriesga la proyección de algunas tendencias de uso de la variedad caraqueña en el tiempo.

El HCC/CD 73-11 constituye una muestra útil para el estudio exploratorio de tendencias y, en muchos casos, verdaderas fuentes de datos para el análisis

exhaustivo de fenómenos del lenguaje asociados a factores socio-culturales específicos.

Referencias bibliográficas

Atkins, S., J. Clear & Ostler N. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.

Bentivoglio, P. & Malaver I. 2014. Corpus sociolingüístico *Preseea* Caracas 2004-2010. CD-rom. Caracas: Fondo Editorial de Humanidades y Educación, Universidad Central de Venezuela.

Pons Bordería, S. 2004. *Conceptos y aplicaciones de la Teoría de la Relevancia*. Madrid: Arco/Libros.

Guirado, K. 2014. Corpus Diacrónico del Habla de Caracas 1987/2013. *Boletín de Lingüística* 41/42: 19-44.

Guirado, K. 2015. Marcadores discursivos de Caracas. En A. Valencia & A. Vigueras (coords.): *Más sobre marcadores hispánicos: Usos de España y América en el corpus de estudios de la norma culta*. México D.F.: UNAM, 69-122.

Lope Blanch, J. M. 1987. El Estudio Coordinado de la Norma Culta de las Principales Ciudades de Lengua Española. *Actas del VII Congreso. Asociación de Lingüística y Filología de América Latina (ALFAL)* I. Santo Domingo: ALFAL, 163-67.

Melis, C., M. Flores & Bogard S. 2003. La historia del español: propuesta de un tercer período evolutivo, Nueva Revista de Filología Hispánica, 51: 1-56.

Moreno Fernández, F. 1997. Metodología del «Proyecto para el Estudio Sociolingüístico del Español de España y América». *Trabajos de sociolingüística hispánica*. Alcalá de Henares: Universidad de Alcalá, 137-161.

Moreno Fernández, F. 2005. *Principios de sociolingüística y sociología del lenguaje*. Barcelona: Ariel.

Pearson, J. 1998. *Terms in Context. Studies in Corpus Linguistics 1*. Amsterdam/Philadelphia: John Benjamins.

Pérez Hernández, C. & Ortiz A. M. 2009. Lingüística Computacional y Lingüística de Corpus. Potencialidades para la investigación textual. En N. Rodríguez Ortega (dir.). *Teoría y literatura artística en la sociedad digital: construcción y aplicabilidad de colecciones textuales informatizadas*. Gijón: TREA, 67-96.

Rabanales, A. 1992. Fundamentos teóricos y pragmáticos del «Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades del mundo hispánico». *Boletín de Filología* (XXXIII): 251-72.

Rosenblat, Á. (dir.) & Bentivoglio P. (ed.). 1979. *El habla culta de Caracas. Materiales para su estudio*. Caracas: Facultad de Humanidades y Educación, Universidad Central de Venezuela.

Sinclair, J. M. 1996. Preliminary recommendations on Corpus Typology. EAG-TCWG-CTYP/P. En *EAGLES*. <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>.

Torruella, J. & Llisterri J. 1999. Diseño de corpus textuales y orales. En J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (eds). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Milenio, 45-77. En línea: <http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf>.