Corpus Diacrónico del Habla de Caracas 1987/2013

Krístel Guirado
kristelguirado@gmail.com
Universidad Central de Venezuela

Resumen

En el presente artículo, se describe el proceso de *reingeniería de corpus* (Guirado 2013) aplicado a dos corpus sincrónicos del habla caraqueña, con el objetivo de crear un nuevo corpus para el estudio diacrónico de esta variedad. La metodología incluye: i. describir la estructuración de los corpus originales; ii. evaluar los aspectos inoperantes y los estables de cada arquitectura; y, iii. crear el rediseño y estimar su representatividad. Se obtiene así el *Corpus Diacrónico del Habla de Caracas 1987/2013* (CDHC'87/13). Los artículos incluidos en este volumen tienen en común que usan como muestra para el estudio del fenómeno que analizan parte o la totalidad de este subcorpus, razón por la cual han remitido a la consulta del presente artículo en sus metodologías.

Palabras clave: *reingeniería de corpus*, habla de Caracas, análisis diacrónico, Lingüística de Corpus.

Abstract

In this paper I describe the process of *corpus reengineering* (Guirado 2013) applied to two synchronic corpuses of Caracas speech. The purpose is to obtain a new corpus, *Corpus Diacrónico del Habla de Caracas 1987/2013* (CDHC'87/13), for the diachronic study of this speech variety. The method consists of: i. describing the structure of the two original corpuses; ii. evaluating the aspects which are inoperative and those which are stable in every structure; and, iii. creating a new design and estimating its representativity. The articles in this volume use part or all of this new corpus for analyzing the phenomenon under study. Thus, their sections on method and corpus description make reference to this article, so as to avoid the repetition of information.

Keywords: *corpus reengineering*, Caracas speech, diachronic analysis, Corpus Linguistics.

Recibido el 9 de septiembre de 2014. Aceptado el 17 de diciembre de 2014.

Introducción: Los corpus electrónicos y la reingeniería de los mismos

Los conceptos de Francis (1992 [1979, 1982]),[1] Sinclair (1996)[2] y McEnery (2003)[3] son tal vez los más citados en la historia de la Lingüística de Corpus (LC), seguramente porque los tres expresan claramente el objetivo principal de un corpus y los parámetros fundamentales de su construcción: una agrupación de textos, diseñada de forma representativa, para el estudio de la lengua. En principio, el desarrollo de nuevas tecnologías no tiene ninguna relación con la noción de corpus; no obstante, el adelanto que implica el uso de la informática en el perfeccionamiento y desarrollo de bases de datos ha obligado a los autores a incluir las propiedades que facilitan el análisis informatizado de los corpus como criterio fundamental en su descripción (Sánchez 1995; McEnery y Wilson 1996; Santalla 2005; Sinclair 2005; McEnery, Xiao y Tono 2006). Parodi (2008) nos ofrece una de las definiciones más completa, la cual incluye origen, propósito, composición, formato, representatividad y extensión de un corpus:

> un conjunto amplio de textos digitales de naturaleza específica y que cuenta con una organización predeterminada en torno a categorías identificables para la descripción y análisis de una variedad de lengua. Este conjunto de textos debe mostrar, de preferencia, accesibilidad desde entornos computacionales y visibilidad de modo que se posibilite su uso en diversas investigaciones con el fin de asegurar acumulación de conocimientos e integración de la investigación de una lengua particular o en comparación con otra. También debe cumplir con **aportar detalles relevantes acerca de su recolección y procedencia**. De modo más específico, **se espera se almacene en conjunto con otros corpus diversos con el fin que se permita su comparación e, idealmente, su contraste** (Parodi 2008: 106-107).

1. "una colección de textos que se asume como representativa de una lengua dada, dialecto, u otro subsistema para ser usado en el análisis lingüístico"; nuestra traducción del texto original, (en adelante, NTO): "a corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis" (Francis 1992 [1979, 1982]: 17).
2. "una colección de fragmentos de una lengua que son seleccionados y ordenados de acuerdo a criterios lingüísticos específicos para ser usadas como una muestra de la lengua; NTO: "A *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair 1996: en línea).
3. "El término corpus debe aplicarse propiamente solo a una colección bien organizada de datos, recogidos dentro de los límites de una muestra diseñada para permitir la búsqueda de un determinado rasgo lingüístico (o grupo de rasgos)"; NTO: "The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or set of features)" (McEnery 2003: 449).

Se ha resaltado en negritas un requisito que Parodi señala como necesario en el diseño de un corpus: los metadatos para la recuperación de su arquitectura, propiedad metalingüística que en muchas ocasiones se pasa por alto en la práctica. Gracias a esta propiedad es posible llevar adelante las tareas relacionadas con la *reingeniería de corpus*: "creación de corpus específicos a partir del rediseño radical y la reconcepción de uno o varios corpus" con el propósito de "incrementar y diversificar los ámbitos de análisis de los fenómenos del lenguaje" (Guirado 2013). La autora propone una serie de acciones necesarias para la construcción de un nuevo corpus:

i. Definir el/los corpus objeto de la reingeniería.
ii. Establecer la nueva finalidad: diacrónica, diastrática, genolectal, mixta, etc.
iii. Comprobar la representatividad: definir la población, lo que implica buscar los datos o (lo que suele suceder) reconstruir la estadística; establecer la relación proporcional de hablantes y/u otros criterios de representatividad, como el tamaño de las entrevistas; si es equitativa o no; re-estratificar; diseñar la distribución, llenar las casillas de afijación, criterios de longitud, entre otros.
iv. Registrar el proceso de reconfiguración para garantizar los datos de procedencia.
v. Recodificar las muestras para su empleo ejemplificativo y su ubicación práctica en la distribución.
vi. Diagramar la distribución para hacerla autoexplicativa.
vii. Digitalizar la muestra en diversos formatos para su procesamiento informatizado (*.doc; *.txt; *.pdf; *.htm).
viii. Dar información sobre su disponibilidad.

En el presente artículo, se describe el proceso de producción de un corpus de habla para el estudio de los fenómenos *en tiempo real*,[4] a partir de la reingeniería de dos corpus orales del español caraqueño, diseñados de acuerdo con rigurosos criterios de distribución socioeconómica, haciendo uso, para ello, de técnicas, herramientas y métodos estadísticamente representativos.

4. Cf. Labov (1996), Moreno Fernández (2005b), Silva-Corvalán (2001).

El estudio del español hablado de Venezuela,[5] –especialmente del habla de Caracas– ha dado origen a distintos proyectos para la recolección de corpus con fines diversos:[6] *El Estudio Coordinado de la Norma Culta 1968/1977* (cf. Universidad Central de Caracas 1979); *Estudio sociolingüístico del habla de Caracas 1977* (cf. Bentivoglio 1987: 25); *Estudio sociolingüístico del habla de Caracas 1987* (cf. Bentivoglio y Sedano 1993); *PRESEEA-Caracas –Proyecto para el Estudio Sociolingüístico del Español de España y de América–* (cf. Bentivoglio y Malaver 2006 y 2012); *Corpus de habla infantil de Caracas 1996* (cf. Shiro 1998); *Documentos para la Historia del Español de Venezuela de los siglos XVI, XVII y XVIII* (cf. de Stefano y Pérez Arreaza 2000, Tejera y De Stefano 2006).[7]

Recoger y sistematizar un corpus entraña una ardua labor. Pocas veces queda testimonio material del esfuerzo asociado a las tareas propias de esta disciplina. No es posible publicar una relatoría de las notas de corrección, ni la enumeración de recorridos por callejones y avenidas en busca de un hablante, tampoco el inventario de etiquetas que se han borrado y reescrito o las veces que se ha vuelto sobre un fragmento para descifrar las consonantes aspiradas y las elididas, un balbuceo, un murmullo, una intención. Detrás de cada uno de los corpus que se citan en esta comunicación, hay un equipo de trabajo –investigadores, auxiliares y pasantes– sin cuya constancia y esfuerzo no hubiera sido posible la existencia de esas muestras.

---

5. *El Estudio sociolingüístico del habla de Caracas*, más que un proyecto permanente, constituye una de la líneas de investigación más sólidas del Instituto de Filología "Andrés Bello" de la Universidad Central de Venezuela (IFAB-UCV), la cual ha determinado, incluso, el perfil de la mayor parte de los egresados de la Maestría en Lingüística que ofrece la Comisión de Estudios de Postgrado de esta universidad. De esta forma, bajo el abrigo *de la casa que vence las sombras*, la Lingüística de Corpus y la Sociolingüística (especialmente la variacionista) vienen desarrollándose en el país desde finales de los años sesenta.

6. *El Proyecto del Habla Culta* contó con el apoyo financiero del Consejo de Desarrollo Científico y Humanístico de la Universidad Central de Venezuela (CDCH-UCV), y del, entonces, Consejo Nacional de Investigaciones Científicas (CONICIT). Asimismo, todos los demás proyectos han sido financiados, en su momento, por el CDCH-UCV.

7. Además de los corpus citados, existen otras experiencias útiles para el conocimiento científico del español hablado en Venezuela. Un ejemplo de ello lo constituye el corpus recogido por los investigadores del Instituto Pedagógico de Caracas, el cual ha servido para la realización de varios estudios de tipo dialectal (Cf. Barrera Linares 1978). Asimismo, se pueden nombrar los corpus representativos del habla de otras ciudades: el *Corpus Sociolingüístico de Maracaibo 1986* (cf. Chela-Flores y Gelman 1988); la profesora Alida Velázquez coordinó el proyecto del *Corpus Sociolingüístico de Cumaná* entre *1993* y *1996*; el profesor Manuel Navarro recolectó dos valiosas muestras orales de Valencia y Puerto Cabello, dos ciudades del centro del país (Navarro 1995); y en la Universidad de los Andes, el equipo de *Corpus Sociolingüístico de Mérida 1995* (cf. Domínguez 1996 y Domínguez y Mora 1998), *El habla rural de la cordillera de Mérida*, coordinado por la profesora Jacqueline Clarac de Briceño, (cf. Obediente Sosa 1998); *Documentos para la Historia Lingüística de Mérida* (cf. Obediente Sosa 2002); *Corpus Sociolingüístico de la Ciudad de Mérida 2009-2010* (cf. Briceño, Lee, Fernández, Maldonado, Velazco y Palm 2010).

No obstante, cuando se dice Lingüística de Corpus en Venezuela, no es posible hacerlo sin pensar en Paola Bentivoglio, quien coordinó y/o asesoró a muchos de estos equipos de trabajo y dedicó una parte importante de su vida académica al desarrollo de estos cuerpos de datos, ganada por la necesidad de fundamentar científicamente la investigación del lenguaje.

En el presente artículo, se documenta el proceso de creación y construcción de su más reciente empresa: un corpus para el estudio diacrónico del habla de Caracas. Una aspiración arduamente atesorada que reúne, tal vez, las muestras orales más emblemáticas que se han recogido en el país en virtud de su tamaño, representatividad y estratificación: el *Corpus Sociolingüístico de Caracas 1987* (CSC'87) y el *Corpus Sociolingüístico de Caracas: PRESEEA Caracas 2004/2010* (PRESEEA-Caracas'04/10).

El artículo se divide en cuatro (4) partes. En la primera sección se describe la configuración de los dos corpus de habla caraqueña que han sido objeto de reingeniería. En el segundo apartado se relatan los procesos de restratificación y rediseño de los corpus «fuente» y se describe la arquitectura del corpus «meta» con fines diacrónicos. Luego se ofrece un cuadro que resume las características de los tres corpus reseñados. Finalmente, el artículo se concluye con algunas consideraciones sobre el proceso de reingeniería documentado.

1. El estudio sociolingüístico del habla de Caracas: descripción de los corpus «fuente»

Guirado (2013) establece que el primer paso en el proceso de reingienería consiste en confirmar que los corpus «fuente» cumplan con las características que garantizan su factibilidad: i. recolección de textos en entornos naturales; ii. explicitud de los rasgos definitorios y compartidos por los textos constitutivos; iii. formato final de tipo digital plano (*txt.) para cada texto o documento; iv. tamaño, preferentemente, extenso; v. respeto a principios ecológicos; vi. etiquetaje computacional semiautomático de naturaleza morfosintáctica u otra para cada texto; vii. disponibilidad a través de medios computacionales; viii. acceso a visualización completa de los textos que lo componen en formato plano; ix. búsqueda de principios de proporcionalidad o representatividad (posiblemente estadística); x. sustento o procedencia inicial especificada; xi. identificación de una organización en torno a temas, tipos de textos, registros, géneros, etc.; xii. registro de

datos cuantitativos que permita la comparación y posible normalización de cifras (Parodi 2008).

A continuación, se describen los detalles estructurales de los dos corpus sociolingüísticos del habla de Caracas que han sido objeto de reingienería, de modo que los usuarios puedan confirmar que ambos cumplen con la mayor parte de los requisitos fundamentales que acaban de enumerarse.

1.1. *Corpus Sociolingüístico de Caracas 1987* (CSC'87)[8]

El corpus consta de ciento sesenta (160) muestras de habla de individuos caraqueños, hijos de padres caraqueños, residenciados en la ciudad. Se grabó entre enero de 1987 y abril de 1988. Las grabaciones fueron agrupadas equitativamente según tres características sociales de los hablantes: cuatro (4) grupos etarios (A: 14 a 29 años; B: 30 a 45; C: 46 a 60; y D: 61 o más años); cinco (5) niveles socioeconómicos (alto, medio alto, medio, medio bajo y bajo) y sexo (masculino y femenino). Esta distribución se ilustra en el cuadro 1:

Cuadro 1. Distribución de los hablantes: matriz de datos del CSC'87

	GRUPO ETARIO								
	Grupo A		Grupo B		Grupo C		Grupo D		
nivel	14 a 29 años		30 a 45 años		46 a 60 años		61 años o más		Total
socioeconómico	masculino (M)	femenino (F)	masculino (M)	femenino (F)	masculino (M)	femenino (F)	masculino (M)	femenino (F)	
ALTO (1)	4	4	4	4	4	4	4	4	32
MEDIO ALTO (2)	4	4	4	4	4	4	4	4	32
MEDIO (3)	4	4	4	4	4	4	4	4	32
MEDIO BAJO (4)	4	4	4	4	4	4	4	4	32
BAJO (5)	4	4	4	4	4	4	4	4	32
Total	20	20	20	20	20	20	20	20	160
	40		40		40		40		

El nivel socioeconómico de los informantes se determinó a través de un índice creado especialmente para el análisis de fenómenos sociolingüísticos (Contasti 1980), utilizado anteriormente en el diseño del *Corpus Sociolingüístico de Caracas 1977* (CSC'77).

8. El proyecto fue financiado por el CDCH-UCV, bajo los números H-07.16/86 y H-08.33.1766.88/89. Las creadoras y coordinadoras del proyecto fueron Paola Bentivoglio y Mercedes Sedano; posteriormente se unió al equipo coordinador Alexandra Álvarez.

El estilo en que se llevó a cabo las encuestas puede considerarse como *careful* 'cuidadoso' (Labov 1972). Los tópicos de las entrevistas giran alrededor de la infancia de los hablantes (educación, juegos tradicionales, disciplina familiar); festividades caraqueñas; celebraciones religiosas y la historia política de la ciudad.

La transcripción de las grabaciones se hizo de forma ortográfica, con apego a todos los signos de puntuación convencionales que permiten reproducir la organización discursiva del texto, el sentido y la intencionalidad.[9]

La selección de encuestas se codificaron conforme a dígitos alfanuméricos asociados a la ubicación de las grabaciones en la matriz de distribución (cuadro 1); de modo que, además de permitir identificar al emisor, sirven de metadatos para recuperar la arquitectura de los materiales: la inicial de la ciudad (C), la letra que distingue al hablante en el grupo etario (A/B/C), el número del nivel socioeconómico (1/2/3/4/5), el sexo (M/F), y una letra para identificar al hablante dentro de las cuatros grabaciones incluidas en la casilla común (A/B/C/D); asimismo, cada código culmina con los dos últimos dígitos del año de inscripción del proyecto, para el cual se toma como referencia del período de grabación (87), de modo que, en los diversos textos de divulgación, los casos que ilustran los fenómenos estudiados terminan con dicho código, como puede observarse en el siguiente ejemplo –tomado de un estudio de la alternancia *tú~uno* como sujetos impersonales (Guirado 2011:39)– y que corresponde a una hablante caraqueña, entre 46 y 60 años, de nivel socioeconómico alto:

(1) Caracas es ciudad, pero lo demás es pueblo; donde TÚ *te metas* no hay comodidad (CC1FD.87)

La duración de cada entrevista es de media (½) hora, lo que suma ochenta (80) horas de grabación. La muestra en su totalidad tiene setecientas sesenta y siete mil ochocientas sesenta y ocho (767.868) palabras (Gallucci 2005:114). Las transcripciones de todas las grabaciones están incorporadas

9. Para consultar la normativa con las pautas detalladas de transcripción, véase Bentivoglio y Sedano (1993:16-18).

10. El Corpus de Referencia del Español Actual (CREA) de la Real Academia Española es un conjunto de textos de diversas fuentes (almacenados en soporte informático). Se compone de una amplia variedad de textos escritos y orales, producidos en todos los países de habla hispana desde 1975 hasta 2004. Para una información más amplia, véase en línea: http://www.rae.es/recursos/banco-de-datos/crea.

al Corpus de Referencia del Español Actual (CREA)[10] y disponibles en audio y texto digital en el IFAB–Dialectología.

1.2. *Corpus Sociolingüístico de Caracas: PRESEEA Caracas 2004/2010 (PRESEEA-Caracas'04/10)*

El PRESEEA-Caracas'04/10 forma parte del *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA), coordinado por el catedrático Francisco Moreno Fernández.[11] El desarrollo del proyecto en Caracas está coordinado por Paola Bentivoglio e Irania Malaver.[12]

El PRESEEA-Caracas '04/10 está conformado por ciento ocho (108) entrevistas hechas a informantes nacidos en Caracas, hijos de padres caraqueños. Este corpus se grabó entre junio de 2004 y diciembre de 2010. Los materiales se clasificaron según los siguientes factores sociales: tres (3) grupos generacionales (1: 20 a 34 años, 2: 35 a 54, y 3: 55 o más años); tres (3) grados de instrucción (1: primaria, 2: secundaria, 3: superior); y sexo (hombres y mujeres). La distribución se muestra en el cuadro 2:

Cuadro 2. Distribución de los hablantes: matriz del PRESEEA-Caracas '04/10.

grado de instrucción	GRUPO GENERACIONAL						Total
	Grupo 1 20 a 34 años		Grupo 2 35 a 54 años		Grupo 3 55 años o más		
	Hombres (H)	Mujeres (M)	Hombres (H)	Mujeres (M)	Hombres (H)	Mujeres (M)	
PRIMARIA (1)	6	6	6	6	6	6	36
MEDIA (2)	6	6	6	6	6	6	36
UNIVERSITARIA (3)	6	6	6	6	6	6	36
Total	18	18	18	18	18	18	108
	36		36		36		

El diseño del corpus sigue los principios sociológicos y sociolingüísticos del *Proyecto para el Estudio Sociolingüístico del Español*

11. Un megaproyecto coordinado por el catedrático Francisco Moreno Fernández (auspiciado por la ALFAL), cuya finalidad es "coordinar las investigaciones sociolingüísticas de Iberoamérica y de la Península Ibérica para facilitar la comparabilidad de los estudios y el intercambio de información básica" (Moreno Fernández 1997:137-138). En el proyecto participan equipos de investigación de varias universidades de España y de América, representativos de distintas comunidades dialectales. Para una información más amplia, consultar en línea: http://preseea.linguas.net/.

12. El proyecto de investigación fue financiado por el CDCH-UCV bajo el número PG-07-32-5760-04. La investigadora responsable del proyecto es Paola Bentivoglio, y las co-investigadoras, Irania Malaver y María Alejandra Romero (en la 1ª etapa), y María José Gallucci y Carla González (en la 2ª etapa).

de España y de América (PRESEEA). Como puede observarse en el cuadro 2, la distribución se hizo por cuotas de afijación uniforme de seis hablantes por casilla. De acuerdo con las pautas metodológicas establecidas en el proyecto, este número es suficiente para asegurar la representatividad de la muestra, ya que esta cuota permite alcanzar la proporción entre dos y tres millones de habitantes aproximadamente (Moreno Fernández 2005a: 127-128).[13] Se mantiene el uso de la entrevista semidirigida para recolectar los datos (Silva-Corvalán 2001). Los tópicos de conversación son comunes a los del CSC'87: la infancia, los cambios de Caracas, las tradiciones y costumbres, la familia, el trabajo y/o los estudios.

La pauta de transcripción de los materiales, en cambio, difiere de manera importante con los corpus precedentes en el país, ya que se buscó reflejar las características (condiciones e incidencias) del discurso en cada conversación. Para ello, se etiquetaron las entrevistas con una serie de marcas o etiquetas (< >) establecidas en las normas internacionales de la TEI (*Text Encoding Initiative*).[14]

El código alfanumérico con el cual se identifican los hablantes del PRESEEA-Caracas '04/10 también está asociado al diseño de la distribución en la matriz de datos (cuadro 2): la ciudad (CARA), el sexo (H/M), el grupo etario (1/2/3), el grado de instrucción (1/2/3); finalmente, cada código culmina con la identificación del hablante a través de tres dígitos correspondientes al número que el equipo le asignó a la entrevista. El ejemplo (2), –escogido de una investigación sobre estilo directo e indirecto (Galluci 2013: 95)– procede de una hablante caraqueña, con más de 55 años y nivel de instrucción universitaria:

(2) Ella sacaba las morocotas, CONTABA MI ABUELA, y se las daba a los esclavos (CARA_M33_105)

Las grabaciones tienen una duración mínima de 45 minutos (¾ de hora) cada una, para un estimado de ochenta y un (81) horas de

13. No solo se aseguró la representatividad poblacional sino inclusive la geográfico-cultural que configura a la ciudad. Para ello, también se tomó en cuenta la densidad poblacional por municipio con el propósito de garantizar el equilibrio en la distribución espacial de los hablantes (Cf. Bentivoglio y Malaver 2012).
14. Las transcripciones son utilizables en dos versiones: con o sin etiqueta; y están estructuradas en dos partes: cabecera y texto transcrito. Gracias a la cabecera es posible recuperar los datos propios del archivo; los de la grabación; los de la transcripción y revisión; y, los datos sobre los participantes de la entrevista; cf. Bentivoglio y Malaver (2012), para una amplia explicación acerca de los métodos de recolección y procesamiento de los datos, que incluye detalladas descripciones y ejemplos ilustrativos.

grabación. Se dispone del audio de las entrevistas y de su transcripción en formato digital.

2. El estudio sociolingüístico diacrónico del habla de Caracas: construcción del corpus «meta»

Guirado (2013) advierte que "son los procesos (sinergia, des/integración, movilización, redistribución, restratificación, eliminación, etc.) y no los corpus en sí los sujetos a reingeniería", de modo que, antes de seleccionar el rediseño de un proceso para la creación del corpus «meta», es preciso determinar los siguientes aspectos: i. los *atributos quebrantados* de cada corpus fuente; ii. las *propiedades sólidas*; y, iii. los *procesos factibles de crear/redimensionar* en el corpus «meta». Asimismo, se debe tener presente que "la planificación heurística de cada reingeniería impondrá la distribución de criterios en cada aspecto […] una propiedad que se presenta sólida en el marco de un rediseño particular […] puede representar un atributo quebrantado en otra reconfiguración".

Uno de los objetivos principales que justifica la recolección y elaboración de un tercer corpus sociolingüístico de Caracas es "llevar a cabo estudios diacrónicos/diastráticos con el fin de determinar los procesos de cambio lingüístico que se dan en el español de Venezuela" (Bentivoglio y Malaver 2012: 151). No obstante, la comparabilidad del CSC'87 con el PRESEEA-Caracas'04/10 no es posible en el marco de las configuraciones originales de cada uno, ya que ambos difieren en la cantidad de hablantes, la forma de categorizar la información socioeconómica y los modos de clasificación: el número de entrevistas y su distribución en cuotas de afijación uniforme no siguen los mismos criterios; la estratificación de los hablantes por edad y la relevancia del nivel educativo también es distinta. Debido a estas razones, ha sido necesario elaborar un modelo estratificatorio compartido que –a partir de elementos comunes a ambos corpus– permita su comparación íntegra, en el cual se tomen en cuenta los cambios del modo de vida en la población, es decir, los cambios sociales y económicos que modifican la relevancia de la información necesitada.

2.1. *Post-estratificación socioeconómica del PRESEEA-Caracas'04/10*

A través de los años, ha habido diversas críticas a los modelos estratificatorios, entre las que destaca la falta de universalidad (López

Morales 1989). No obstante, la búsqueda de un modelo global para todos los corpus se aleja de los objetivos propios de la sociolingüística: el estudio de las lenguas en su contexto social. En este sentido, cada comunidad de habla tiene una cantidad de particularidades sociales que la diferencian de otra, de modo que sería utópico pensar en un modelo estratificatorio que permita ponderar por igual la realidad heterogénea de todas las comunidades de habla. En todo caso, de existir, sería necesario cuestionar su capacidad descriptiva.

En una comunidad específica, el valor de los factores que se utilizan para describir los estratos sociales puede variar de una comunidad de habla a otra. Por ejemplo, la importancia de ciertas ocupaciones no siempre es la misma en todos los países: en Venezuela, generalmente, las enfermeras son consideradas personal auxiliar que asiste al médico; mientras que, en España, califican como profesionales con un rol más independiente de la figura del galeno.

Como se dijo en §1.1, el nivel socioeconómico de los informantes del CSC'87 se determinó a través de un índice creado especialmente por Max Constanti para el análisis de fenómenos sociolingüísticos. El método permite ubicar al hablante en un nivel de los cinco (5) propuestos (alto, medio alto, medio, medio bajo o bajo). Para su cálculo, se tomaron en cuenta siete factores ajustados a la "realidad venezolana" de entonces: i. ocupación del hablante; ii. ocupación de la madre; iii. ocupación del padre; iv. grado de instrucción del hablante; v. condiciones de alojamiento; vi. ingreso total familiar; vii. ingreso promedio familiar. El cálculo consistió en asignar un peso a cada uno de los siete (7) factores y ponderarlos según la técnica de análisis factorial. Se propuso originalmente una escala de nueve valores, puntuados entre el uno (1) y el nueve (9), ambos inclusive.

El empleo de este instrumento de medición ofrece tres ventajas: i. está diseñado para investigaciones sociolingüísticas, especialmente de tipo variacionista; ii. refleja la realidad socioeconómica venezolana del momento histórico de obtención de la muestra; y, iii. con pequeños ajustes en lo relativo a los niveles de ingreso es aplicable en diferentes épocas. Gracias a este último atributo, a lo largo de los años, ha sido posible ajustar el modelo a los cambios contextuales propios de cada período de grabación:

> En el proyecto de 1987, la escala de nueve valores (1, 2, 3, 4, 5, 6, 7, 8, 9) fue reducida a cinco (1, 3, 5, 7, 9) con el fin de ajustarla al modelo de cinco valores utilizado por el *Proyecto Venezuela*, auspiciado por Fundacredesa. La numeración del 1 al 9 se mantuvo para conservar el paralelismo con el proyecto de 1977. No se tomaron en cuenta los valores intermedios (2, 4, 6, 8), salvo en la variable 4, que es la correspondiente al grado de instrucción del hablante, donde sí se conservaron los nueve valores originales porque a través de ellos quedaba mejor reflejado el nivel educativo de la persona seleccionada (Bentivoglio y Sedano 1993: 8).

Igualmente, para la postestratificación socioeconómica de las entrevistas del PRESEEA-Caracas'04/10, fue necesario ajustar el modelo a las fluctuaciones del sistema económico venezolano de principios del siglo XXI. Para ello, Max Contasti propuso tomar solo los cinco primeros factores del modelo original y eliminar el *ingreso total* y el *ingreso promedio familiar*. A su vez, el investigador introdujo la variable *tipo* de ingreso; este nuevo factor tiene dos ventajas: agrupa los factores descartados en una sola categoría y, principalmente, la hace comparable en el tiempo, lo que era imposible cuando se tomaba como referencia el monto de los salarios. Así, las variantes de la nueva categoría permiten ponderar la frecuencia y la forma de obtener los ingresos: i. sin ingresos; ii. diario o a destajo; iii. mensual o quincenal; iv. honorarios profesionales; y, v. fortuna heredada o adquirida.

A continuación, se muestra un cuadro con los seis (6) factores socioeconómicos y los pesos de ponderación de cada una de sus categorías. El índice socioeconómico de cada hablante se calculó multiplicando los pesos de ponderación de cada uno de los factores estratificatorios por los valores del 1 al 9 descritos en el cuadro. Los valores más cercanos a 1 pertenecen a los estratos sociales altos o de mayor instrucción, mientras que los valores más cercanos a 9 hacen referencia a hablantes con menor escolaridad y que forman parte de los estratos sociales de tipo bajo:

Cuadro 3. Factores de estratificación socioeconómica aplicados al PRESEEA-Caracas'04/10

Factor	Pesos	Valor	Descripción
I. Ocupación del hablante	0,12 x	1	Altos funcionarios del gobierno; altos oficiales del ejército; empresarios privados; hacendados; altos ejecutivos (sectores público y privado); autoridades universitarias.
		3	Profesionales universitarios de libre ejercicio; gerentes medios del sector público y privado; oficiales de rango medio; industriales y productores medios; docentes universitarios; artistas reconocidos.
II. Ocupación de su padre	0,12 x	5	Profesionales universitarios no liberales; profesores de educación media y básica; oficiales de rango bajo; pequeños empresarios y productores; técnicos superiores; secretarias ejecutivas; supervisores; enfermeras graduadas; miembros de la farándula.
III. Ocupación de su madre	0,14 x	7	Pequeños comerciantes; secretarias y oficinistas; obreros especializados; artesanos; mecánicos; vendedores; cobradores; ayudantes técnicos; policías y agentes de tránsito; deportistas profesionales; regulares de las fuerzas armadas.
		9	Buhoneros y vendedores ambulantes; obreros no especializados; servicio doméstico; mesoneros, bedeles y vigilantes.
IV. Grado de instrucción del hablante	0,10 x	1	Doctorado
		2	Maestría
		3	Pregrado universitario completo
		4	Pregrado incompleto / T.S.U.
		5	Bachillerato completo / Técnico medio
		6	Bachillerato incompleto / Cursos de capacitación
		7	Primaria completa
		8	Primaria incompleta
		9	Analfabeta
V. Condiciones de alojamiento	0,25 x	1	Casa o apartamento lujoso o espacioso
		3	Casa o apartamento menos lujoso o espacioso
		5	Casa o apartamento sin lujo
		7	Casa o apartamento modesto
		9	Vivienda sin comodidades sanitarias y de difícil acceso
VI. Tipo de ingreso	0,27 x	1	Fortuna heredada o adquirida
		3	Honorarios profesionales
		5	Mensual o quincenal
		7	Diario o a destajo
		9	Sin ingresos

El producto de la multiplicación de los pesos de ponderación por las categorías estratificatorias dio como resultado un índice por hablante entre 1,00 y 9,00. Para calcular los índices del PRESEEA-Caracas'04/10 fue necesario contar con los datos de los hablantes, los cuales estaban disponibles al comienzo de cada una de las transcripciones (cabeceras). De allí se tomó la información necesaria para su clasificación y postestratificación, según el modelo descrito. Cuando alguno de estos datos no estaba especificado en la cabecera, se procedió a leer toda la entrevista para buscar la información que faltaba. De no encontrarla de esta forma, los auxiliares de investigación que

entrevistaron a los hablantes ayudaron a contactar a los entrevistados para recuperar la información faltante.

En virtud de los cambios experimentados en el sistema económico nacional a lo largo de los últimos 30 años (devaluaciones de la moneda, disminución de la producción interna, desarrollo del mercado de servicios, crecimiento de la economía informal, entre otros), se tomó la decisión de reducir la clasificación de cinco estratos (alto, medio alto, medio, medio bajo y bajo) a tres (alto, medio bajo), ya que estos reflejaban con más precisión la realidad socioeconómica actual. Adicionalmente, esta redistribución garantizó una mayor disponibilidad de hablantes comparables por grupo socioeconómico, ya que el número de entrevistados del CSC PRESEEA-Caracas'04/10 (108) es inferior a la muestra de CSC'87 (160). Los hablantes cuyo índice oscila entre 1,00 a 3,66 se clasificaron dentro del nivel alto; entre 3,67 a 6,33, dentro del nivel medio; y entre 6,34 y 9,00, dentro del nivel bajo. Por ejemplo, el nivel socioeconómico del hablante CARA_H33_097, –oftalmólogo de 74 años, de padre comerciante, madre ama de casa, con un apartamento con comodidades–, se calculó de la siguiente forma (ver cuadro 4):

Cuadro 4. Composición del índice socioeconómico del hablante CARA_H33_0

Categoría	Datos del hablante	Ponderación
OCUPACIÓN DEL HABLANTE	Oftalmólogo	3 x 0,12 = 0,36
OCUPACIÓN DEL PADRE	Comerciante	7 x 0,12 = 0,84
OCUPACIÓN DE LA MADRE	Ama de casa	7 x 0,14 = 0,98
ESTUDIOS	Postgrado en oftalmología	2 x 0,10 = 0,20
VIVIENDA	Apartamento cómodo o espacioso	3 x 0,25 = 1,25
TIPO DE INGRESO	Honorarios profesionales	3 x 0,27 = 0,81
	Total	3,94

El índice 3,94 ubica a este hablante dentro del nivel medio (índices de 3,67 a 6,33) y por su edad pertenece al grupo etario 3 (55 o más años). El proceso que se acaba de ilustrar se realizó con todos los hablantes del PRESEEA-Caracas'04/10.[15]

2.2. *Comparación del CSC'87 con el PRESEEA-Caracas'04/10*

Al finalizar la postestratificación de los 108 hablantes del corpus más reciente, el siguiente paso consistió en seleccionar las entrevistas comparables de cada uno de los corpus.

15. Los índices del grado de instrucción 1 fueron calculados por Muñoz (2010) y los índices de los grados de instrucción 2 y 3 por Lárez (2012).

Los hablantes del CSC'87 (clasificados originalmente en cinco niveles socioeconómicos), se reordenaron según su índice en los tres nuevos niveles establecidos para la comparación (alto, medio, bajo). Una vez reclasificados, se revisaron los datos de origen para descartar a los hablantes no comparables. Primero se excluyeron los nacidos y/o residentes fuera del Área Metropolitana de Caracas,[16] ya que en el PRESEEA-Caracas'04/10 no hay hablantes con estas características. Luego, se eliminaron los jóvenes menores de 18 años con el propósito de equilibrar, en lo posible, el rango inferior de edad del grupo etario más joven en el nuevo corpus.

En el cuadro 5 se muestran los hablantes excluidos del CSC'87, que suman catorce (14) en total:

Cuadro 5. Hablantes excluidos del CSC'87 por edad y/o municipio de residencia[17]

Código hablante	índice	edad	Estudios/Ocupación	Grado de Instrucción		Municipio
CA2FD.87	3,20	17	Estudiante	B	I	Baruta
CA3FD.87	5,38	22	Estudiante	U	I	Guaicaipuro
CA4MA.87	6,04	27	Vendedor	U	I	Carrizal
CA4FD.87	6,90	17	Estudiante	B	I	Libertador
CA5FA.87	7,64	17	Del hogar	B	I	Libertador
CA5FB.87	7,80	17	Del hogar	B	I	Libertador
CA5FD.87	7,94	17	Doméstica	B	I	Sucre
CA5MA.87	7,64	15	Estudiante	B	I	Libertador
CB2MA.87	3,98	43	Lic. Economía/Gerente	U	C	Los Salias
CB4MD.87	6,16	40	Dibujante	T	I	Guaicaipuro
CB5FD.87	7,62	39	Del hogar	P	I	Guaicapuro
CC2MA.87	3,56	49	Medico-profesor	M	C	Carrizal
CC5MD.87	8,90	59	Agricultor	P	I	Guaicapuro
CD4MB.87	6,16	66	Carpintero	B	I	Guaicapuro

La muestra debería estar formada por treinta y seis (36) cuotas o casillas con afijación uniforme, creadas a partir del producto de las variantes de las variables sexo, grupo etario, nivel socioeconómico y período de grabación (2 x 3 x 3 x 2 = 36). Para asegurar la representatividad de este nuevo corpus diacrónico, sería necesario seleccionar una cuota de afijación de cuatro (4) hablantes por casilla: "normalmente cada cuota es representada

16. El CSC'87 incluía hablantes que vivían fuera de los cinco municipios del Área Metropolitana, en las ciudades *aledañas, pilotos* o *dormitorio* que conforman, junto a la zona Metropolitana, la llamada *Gran Caracas*: Los Valles del Tuy, Altos Mirandinos (Guaicaipuro, Carrizal, Los Salias), Guarenas, Guatire, La Guaira.

17. En el recuadro de "grado de instrucción" la primera columna indica el nivel máximo obtenido: "P" para primaria; "B", bachillerato; "T", técnico superior; "U", universitario; "M", postgrado. La segunda columna indica si los estudios son "C" completos o "I" incompletos.

por entre tres y cinco informantes" (Moreno Fernández 2005b: 312), para un total de setenta y dos (72) hablantes por corpus. Sin embargo, vale la pena tener presente que, en el caso de 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 y Moreno Ortiz 2009: 76). En el cuadro 6, se puede observar la delimitación del universo de habitantes para cada población objeto de estudio y los porcentajes respecto a la totalidad de la población. Los datos aportados por los censos de 1981, 2001 y 2011 se usan de referencia para las muestras:

Cuadro 6. Estimación de la proporción de la población objeto de estudio por entrevistado[18]

Corpus y año	Población (Dtto. Metropolitano)	Pobl. del Dtto. Capital oriunda, mayor de 20 años	Total de hablantes (36 casillas = *2s* x 3*ge* x 3*ne* x 2*a*)	Proporción por entrevistado (*f*=*población/hablantes*)
CSC'87	2.583.396 (Censo 1981)	---	4 x casilla = 72	1/12.500 → 900.000
PRESEEA-Caracas'04/10	2.762.759 (Censo 2001)	649.337 (Censo 2001)	4 x casilla = 72	1/20.500→ 1.500.000
	2.904.376 (Censo 2011)	931.045 (Censo 2011)		

En el cuadro 6 se presentan varios datos por columna: i. la relación corpus-año de grabación; ii. la población de la región por año según censo de referencia; iii. datos específicos de los empadronados del municipio Libertador, nacidos en esta entidad, mayores de veinte (20) años;[19] y, finalmente, vi. el cálculo de la proporción de habitantes por hablante (*f*=población/72 hablantes). De esta forma, la proporción de habitantes por hablante se calculó considerando la población de la tercera columna más un 30% estimado correspondiente a los municipios mirandinos.

La distribución ideal de la muestra comparativa se observa en el cuadro 7, en el que el "Período 1" remite al corpus grabado en 1987, y el "Período 2", al corpus más reciente:

18. El área metropolitana de la ciudad de Caracas es una unidad político-territorial que integra cinco municipios: Libertador del Distrito Capital y los municipios Baruta, Chacao, El Hatillo y Sucre del estado Miranda. Fuente: Delgado Linero (2005: 200) e INE: http://www.redatam.ine.gob.ve.

19. Los usuarios de la web del INE cuentan con la aplicación Redatam+SP, un programa para procesar y mapear datos de censos y encuestas para análisis local y regional que permite solicitar información estadística sobre el Censo 2001 y el Censo 2011; no obstante, la información sobre censos de años anteriores aún no puede ser procesada del todo con esta refinada herramienta. Este recurso ha permitido circunscribir los datos del Dtto. Capital a la población mayor de 20 años, nacida en Caracas. A pesar del alcance de esta herramienta, no es posible solicitar la misma información de los empadronados de los municipios del estado Miranda; sin embargo, en virtud de que los habitantes del municipio Libertador representan, aproximadamente, entre el 70% y el 75% de la población del Dtto. Metropolitano, se ha considerado importante ofrecer las estadísticas reducidas de acuerdo con la población objeto de estudio del corpus diacrónico, ya que estas nos permiten sustentar con más fuerza la representatividad del diseño de la muestra.

Cuadro 7. Distribución ideal de hablantes en la comparación

	GRUPO GENERACIONAL												
	18-34 años (A)				35-54 años (B)				55 o más años (C)				
	Período 1		Período 2		Período 1		Período 2		Período 1		Período 2		Total
nivel socioeconómico (índices)	H	M	H	M	H	M	H	M	H	M	H	M	
Alto (1-3,66)	4	4	4	4	4	4	4	4	4	4	4	4	48
Medio (3,67-6,33)	4	4	4	4	4	4	4	4	4	4	4	4	48
Bajo (6,34-9)	4	4	4	4	4	4	4	4	4	4	4	4	48
Total	12	12	12	12	12	12	12	12	12	12	12	12	**144**
	24		24		24		24		24		24		

El cuadro 7 muestra justamente una "distribución ideal de la muestra", que no ha sido posible alcanzar, ya que faltan diecisiete (17) hablantes de las muestras comparables para llenar completamente los nichos requeridos. Este problema es la causa de que la distribución de los hablantes quedara, en cierta forma, irregular:

Cuadro 8. Casillas de afijación por completar en la comparación total

	GRUPO GENERACIONAL												
	18-34 años (A)				35-54 años (B)				55 o más años (C)				
	Período 1		Período 2		Período 1		Período 2		Período 1		Período 2		Total faltantes
nivel socioeconómico (índices)	H	M	H	M	H	M	H	M	H	M	H	M	
Alto (1-3,66)	4	4	1	2	4	4	4	4	4	4	1	3	9
Medio (3,67-6,33)	4	4	4	4	4	4	0	1	4	4	4	3	8
Bajo (6,34-9)	4	4	4	4	4	4	4	4	4	4	4	4	0
Total faltantes	0	0	3	2	0	0	4	3	0	0	3	2	17
	0		5		0		7		0		5		

Como se observa en el cuadro 8, para completar la muestra comparativa, faltan diez (10) hablantes de nivel alto y cinco (5) de nivel medio; en el nivel alto faltan tres (3) mujeres y tres (3) hombres de 18 a 34 años y tres (3) hombres de 55 o más años; mientras que en el nivel medio faltan tres (3) mujeres de 35 a 54 años y otra de 55 o más años. En el caso de los hombres de nivel medio, solo hace falta grabar un (1) hablante de 35 a 54 años.

Ya se ha dado inicio a las grabaciones de estos hablantes. No obstante, el desarrollo de esta última tarea se ha planificado en dos etapas. La primera consiste en completar la mitad de hablantes por casilla, de forma que sea posible iniciar los estudios comparativos con el corpus. La segunda radica, obviamente, en completar el resto de las entrevistas. De esta forma, para completar la primera meta, solo hacía falta grabar los cinco (5) hablantes que se observan en el cuadro 9:

Cuadro 9. Casillas de afijación por completar para la comparación parcial

	GRUPO GENERACIONAL												
	18-34 años (A)				35-54 años (B)				55 o más años (C)				Total faltantes
	Período 1		Período 2		Período 1		Período 2		Período 1		Período 2		
nivel socioeconómico (índices)	H	M	H	M	H	M	H	M	H	M	H	M	
Alto (1-3,66)	2	2	1	2	2	2	2	2	2	2	1	2	2
Medio (3,67-6,33)	2	2	2	2	2	2	0	1	2	2	2	2	3
Bajo (6,34-9)	2	2	2	2	2	2	2	2	2	2	2	2	0
Total faltantes	6	6	5	5	6	6	5	5	6	6	5	6	5

Durante la etapa de construcción del PRESEEA-Caracas'04/10, el equipo dejó fuera de la configuración final algunas de las grabaciones hechas, en virtud de que, al ser clasificadas, la ubicación de las mismas excedía la cuota de afijación de seis (6) hablantes en una determinada casilla (cf. §1.2). Por suerte, tres de estas grabaciones podían ser utilizadas para completar algunas de las casillas del cuadro 8, puesto que coincidían el sexo, la edad y el cálculo del índice socioeconómico. De esta forma, se completaron las casillas correspondientes a los hombres del grupo etario A de nivel alto (CARA_H13_110) y del grupo etario B de nivel medio (CARA_H22_109 y CARA_H22_111). Una vez solucionadas estas ausencias, se procedió a grabar los dos hablantes que faltaban: una (1) mujer del grupo generacional B de nivel socioeconómico medio (CARA_M23_113) y un (1) hombre de más de 55 años del nivel alto (CARA_H33_112).

A continuación, se reseña el resultado de esta primera etapa del proceso de reingeniería que permitió crear un corpus diacrónico de setenta y dos (72) encuestas, el cual ha sido diseñado con el propósito de estudiar

los fenómenos de variación en el habla de Caracas en dos períodos de tiempo entre los cuales median 30 años aproximadamente.

2.3. *Corpus Diacrónico del Habla de Caracas 1987/2013 (CDHC'87/13)*[20]

Actualmente, el CDHC'87/13 está conformado por setenta y dos (72) entrevistas hechas a informantes nacidos en Caracas, hijos de padres caraqueños. Las grabaciones que constituyen el corpus están distribuidas equitativamente según tres (3) características sociales de los hablantes, dos inherentes y una adquirida: tres (3) rangos de edad (A: 18 a 34 años; B: 35 a 54; C: 55 o más años); sexo (hombre y mujer) y tres (3) niveles socioeconómicos (alto, medio y bajo). Asimismo, las entrevistas fueron seleccionadas proporcionalmente de dos corpus sociolingüísticos grabados en dos (2) períodos distintos (1987 y 2013). La distribución socioeconómica de las entrevistas, identificadas con nuevos códigos, se muestra en el cuadro 10:

Cuadro 10. Distribución socioeconómica y codificación de las entrevistas del CDHC'87/13

Edad	A:20-34 años				B:35-54 años				C:55 años o +			
Sexo	hombres		mujeres		hombres		mujeres		hombres		mujeres	
NSE/Año	1987	2013	1987	2013	1987	2013	1987	2013	1987	2013	1987	2013
Alto (1)	CA1HA.87 CA1HC.87	CA1HG.09 CA1HD.05	CA1MA.87 CA1MD.87	CA1ME.04 CA1MF.05	CB1HB.87 CB1HC.87	CB1HE.08 CB1HF.08	CB1MB.87 CB1MC.87	CB1MC.05 CB1MF.07	CC1HC.87 CC1HD.87	CC1HF.10 CC1HG.13	CC1MC.87 CC1MD.87	CC1ME.09 CC1MF.09
Medio (2)	CA2HB.87 CA2HD.87	CA2HE.05 CA2HA.04	CA2MA.87 CA2MD.87	CA2MB.04 CA2MC.04	CB2HB.87 CB2HD.87	CB2HG.08 CB2HI.09	CB2MA.87 CB2MB.87	CB2MD.06 CB2MG.13	CC2HC.87 CC2HD.87	CC2HE.08 CC2HA.04	CC2MA.87 CC2MD.87	CC2MA.04 CC2MB.04
Bajo (3)	CA3HA.87 CA3HB.87	CA3HB.08 CA3HC.08	CA3MA.87 CA3MB.87	CA3MA.07 CA3MB.07	CB3HA.87 CB3HD.87	CB3HD.08 CB3HE.09	CB3MC.87 CB3MD.87	CB3MF.09 CB3MA.06	CC3HC.87 CC3HD.87	CC3HD.09 CC3HC.09	CC3MB.87 CC3MD.87	CC3MD.08 CC3MA.04
Minutos	180"	270"	180"	270"	180"	270"	180"	270"	180"	270"	180"	270"

Las entrevistas se postcodificaron con una nomenclatura que describe la reconfiguración de los materiales. El nuevo código resume los datos en ciudad (C), grupo etario (A/B/C), nivel socioeconómico (1/2/3),

20. En Guirado (2013) también se describe el proceso de creación de dos corpus diacrónicos: *Habla culta de Caracas 1974-2004*. *Corpus diacrónico* (HCC-CD'74-09) y *Habla de jóvenes universitarios caraqueños 1977-1987-2005*. *Corpus diacrónico* (HJUC-CD'77-87-05). No obstante, existe una diferencia fundamental entre estos corpus y el que aquí se describe. El ámbito de los dos primeros está restringido al estudio de comunidades de habla específicas dentro de la comunidad de habla caraqueña; mientras que el diseño del corpus diacrónico del presente artículo tiene la ambición de representar la generalidad de hablantes de la ciudad a través de sus características socioeconómicas.

sexo (H/M), distinción del hablante (A-I), y año de grabación. Como se señaló anteriormente en §1.1, para las grabaciones del primer período, se toma como referencia el año 1987 (87); en cambio, en las grabaciones más recientes, se ha optado por conservar la referencia a los dos últimos dígitos del año de grabación (04, 05, 06, …13), en virtud de que el período que abarcan las entrevistas es más amplio que el de las grabaciones de hace treinta años.

El CDHC'87/13 suma un total de cuarenta y cinco (45) horas de grabación. Si bien es cierto que la diferencia en la duración de las grabaciones de cada corpus «fuente» influye, obviamente, en la extensión de las entrevistas, se ha tomado la decisión de ofrecer las transcripciones en toda su extensión ya que, dependiendo del tipo de investigación, los métodos de selección de datos pueden variar de un estudio a otro. Por ejemplo, en los estudios de variación léxica, en los que es determinante el número de palabras, los autores tendrán la libertad de unificar las entrevistas en el marco de este criterio.[21] Otros estudios, en cambio, pueden optar por diversos criterios: unificar las transcripciones a los primeros 30 minutos de grabación; reasignar porcentajes de acuerdo con la longitud discursiva según la edad de los hablantes; o bien usar el corpus en sus proporciones reales, ya que la opción de una distribución ascendente se justifica bajo el argumento de que la mayor densidad discursiva de CDHC'87/13 se corresponde con el aumento de la población en el tiempo.

El audio y el texto de las entrevistas están disponibles en formato digital en el IFAB-Dialectología.

3. Digitalización y almacenamiento del material

A continuación, se presenta un cuadro que resume las principales características de los corpus citados en este artículo:

21. El equipo dispone de una versión ajustada a 252.147 palabras. En esta versión, el número total de palabras por entrevista oscila entre 3.500 y 3.505, de modo que es posible analizar entre 7.000 y 7.010 palabras por cada casilla de clasificación socioeconómica.

Cuadro 11. Resumen de las características de los corpus descritos

CORPUS	Nº DE GRABACIONES	ESTRATIFICACIÓN	Nº DE PALABRAS	FORMATO TRANSCRIPCIONES	RESPONSABLES
CSC'87	160	· Nivel socioeconómico · Edad · Sexo	767.868	Digitalizadas *.doc; *.txt; *.pdf; *.html	Paola Bentivoglio Mercedes Sedano Kristel Guirado
CSC-PRESEEA' 04/11	108	· Grado de instrucción · Edad · Sexo	Por calcular	Digitalizadas *.doc; *.txt	Paola Bentivoglio Irania Malaver
CDHC'87/13	72	· Nivel socioeconómico · Edad · Sexo · Período de grabación	Por calcular	Digitalizadas *.doc; *.txt; *.pdf; *.html	Paola Bentivoglio Kristel Guirado

4. Consideraciones finales

Al principio se hizo referencia al término *reingeniería de corpus* para designar las tareas relacionadas con la reconfiguración de materiales de habla (orales y escritos) bajo el supuesto de que un corpus lingüístico que ha sido elaborado con criterios de representatividad contiene, en su propia estructura, la posibilidad de ser comparado con otros corpus de similar diseño. La utilidad de este proceso radica en que "el contraste hace emerger características distintivas y prototípicas que, de otro modo, sería imposible llegar a descubrir" (Parodi 2008: 108).

En esta oportunidad, la comparación demandó la intervención en la arquitectura original de los dos corpus sometidos a reingeniería. La postestratificación del PRESEEA-Caracas'04/10 permitió clasificar a sus hablantes según índices socioeconómicos con criterios similares a los de CSC'87, pero adaptados a la época. Asimismo, el análisis de los contextos sociales e históricos permitió repensar los rangos de las variables inherentes a los hablantes (edad) y constreñir el ámbito descriptivo de las adquiridas (nivel socioeconómico), para orientar su empleo como variables diacrónicas con capacidad descriptiva de los fenómenos en el tiempo.

Si bien no siempre ha sido posible resolver las dificultades que implica transferir la representatividad de un corpus mayor a una muestra prototipo, el subcorpus que acaba de presentarse constituye una base de datos útil para el estudio exploratorio de tendencias y se puede afirmar, con

seguridad, que al completar la totalidad de hablantes del proyecto CDHC'87/13, la comunidad lingüística contará con una imponderable fuente de datos para el análisis exhaustivo, en *tiempo real*, de aquellos fenómenos del lenguaje asociados a factores socioculturales específicos.

Referencias bibliográficas

Barrera Linares, Luis. 1978. Las áreas dialectales de Venezuela. *Letras* 34-35. 18-31.

Bentivoglio, Paola. 1987. *Los sujetos pronominales de primera persona en el habla de Caracas*. Caracas: Universidad Central de Venezuela.

Bentivoglio, Paola y Mercedes Sedano 1993. Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. *Boletín de Lingüística* 8. 3-35.

Bentivoglio, Paola e Irania Malaver. 2006. La lingüística de corpus en Venezuela: un nuevo proyecto. *Lingua Americana* 19. 37-46.

Bentivoglio, Paola e Irania Malaver. 2012. Corpus sociolingüístico de Caracas: PRESEEA Caracas 2004-2010. Hablantes de instrucción superior. *Boletín de Lingüística* XXIV, 37-38. 144-180.

Briceño, Diana Lee; María Fernanda Fernández; Jhon Maldonado; Juan Velazco; Pamela Palm. 2010. [En línea]. Un nuevo corpus sociolingüístico del habla de Mérida: PRESEEA-MÉRIDA-VE. *Lengua y Habla* 14. Disponible en http://erevistas.saber.ula.ve/index.php/lenguayhabla/article/view/1080/1041[Consulta: 30 abril 2014].

Chela-Flores, Bertha y Jeannette Gelman. 1988. *El habla de Maracaibo: Materiales para su estudio*. Maracaibo: Universidad del Zulia.

Contasti, Max. 1980. Metodología para la medición del nivel socio-económico para la población venezolana. *Boletín de la AVEPSO* 3, 2. 13-17.

de Stefano, Luciana y Laura Pérez Arreaza. 2000. Estudio histórico del español de Venezuela: recolección del corpus y rasgos lingüísticos más resaltantes de los documentos. *Lingua Americana* 7. 5-22.

Delgado Linero, Manuel. 2005. Crecimiento de la población y proceso de urbanización en el Distrito Metropolitano de Caracas: efectos ambientales. En Anitza Freitez, María Di Brienza, Genny Zúñiga, Rhayza Carvallo, Mauricio Phélan y Thaís García (ed.), *Cambio demográfico y desigualdad social en Venezuela al inicio del tercer*

milenio. II Encuentro Nacional de Demógrafos y Estudiosos de la Población, 197-212. Caracas: AVEPO.

Domínguez, Carmen Luisa. 1996. El habla de Mérida: un corpus de estudio. *Lengua y Habla* 1. 2, 46-55.

Domínguez, Carmen Luisa y Elsa Mora. 1998. *El habla de Mérida*. Mérida: Universidad de Los Andes.

El habla culta de Caracas. Materiales para su estudio. 1979. Dirección y presentación de Ángel Rosenblat. Selección de muestras de Paola Bentivoglio. Caracas: Facultad de Humanidades y Educación, Universidad Central de Venezuela.

Francis, W. Nelson. 1979. Problems of assembling and computerizing large corpora. En Henning Bergenholtz y Burkhard Schaeder (eds.), *Empirische Sprachwissenschaft. Aufbau und Auswertung von Text-Corpora*, 110-123. Konigstein/Ts.: Scriptor.

Francis, W. Nelson. 1982. Problems of assembling and computerizing large corpora. En Stig Johansson (ed.), *Computer corpora in English language research*, 7-24. Bergen: Norwegian Computing Centre for the Humanities.

Francis, W. Nelson. 1992. Language corpora B.C. En Jan Svartvik (ed.), *Directions in Linguistics: Proceedings of Nobel Symposium* 82, 17-32. Berlin y New York: Mouton de Gruyter.

Gallucci, María José. 2005. El número de palabras: un nuevo criterio para describir tres corpus del habla de Caracas. *Boletín de Lingüística* 24. 108-121.

Gallucci, María José. 2013. Más sobre el estilo directo e indirecto en el español de Caracas. *Lengua y Habla* 17. 89-117.

Guirado, Krístel. 2011. [En línea]. La alternancia tú~uno impersonal en el habla de Caracas. *Lingüística* 26. 26-54. Disponible en http://www.mundoalfal.org/sites/default/files/revista/26_linguistica_026_054.pdf [Consulta: 14 junio 2012].

Guirado, Krístel. 2013. *Reingeniería de corpus en Venezuela: una propuesta metodológica para diversificar el análisis de los corpus del español hablado en Caracas*. Caracas: Instituto de Filología "Andrés Bello", Universidad Central de Venezuela. Documento inédito.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1996. *Principio del cambio lingüístico*. Madrid: Gredos.

Lárez Barrios, Moisés. 2012. *Postestratificación del Corpus Sociolingüístico de Caracas*, 2004-2010 y comparación con el Corpus Sociolingüístico de Caracas, 1987. Caracas: Universidad Central de Venezuela. Trabajo especial de Licenciatura. Informe de Pasantía.

López Morales, Humberto. 1989. *Sociolingüística*. Madrid: Gredos.

McEnery, Tony 2003. Corpus Linguistics. En Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, 448-463. Oxford: Oxford University Press.

McEnery, Tony y Andrew Wilson. 1996. *Corpus linguistics*. Edinburgh: Edinburgh University Press

McEnery, Tony, Richard Xiao y Yukio Tono. 2006. *Corpus-based language studies. An advanced resource book*. London y New York: Routledge.

Moreno Fernández, Francisco. 1997. Metodología del «Proyecto para el Estudio Sociolingüístico del Español de España y América». En Francisco Moreno Fernández (coord.), *Trabajos de sociolingüística hispánica*, 137-166. Alcalá de Henares: Universidad de Alcalá.

Moreno Fernández, Francisco. 2005a. Corpus para el estudio del español en su variación geográfica y social. El corpus «PRESEEA». *Oralia* 8. 123-139.

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

Muñoz, Jhon. 2010. *Transcripción y postestratificación de muestras de habla, como aporte al Corpus sociolingüístico de Caracas 2004-2010, adscrito al PRESEEA*. Caracas: Universidad Central de Venezuela. Trabajo especial de Licenciatura. Informe de Pasantía.

Navarro, Manuel. 1995. *El español hablado en Puerto Cabello*. Valencia: Universidad de Carabobo.

Obediente Sosa, Enrique (comp.) 1998. *El habla rural de Mérida*. Mérida, Venezuela): Universidad de los Andes.

Obediente Sosa, Enrique (comp.) 2002. [En línea]. *Documentos para la Historia Lingüística de Mérida (Venezuela). SIGLOS XVI-XVII*. Mérida (Venezuela): Universidad de los Andes. Disponible en http://www.serbi.ula.ve/serbiula/libros-electronicos/Libros/doc_hist_ling_merida/pdf/librocompleto.pdf [Consulta: 5 agosto 2013].

Parodi, Giovanni. 2008. Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada* 46, 1.93-119.

Pérez Hernández, Chantal y Antonio Moreno Ortiz. 2009. Lingüística computacional y Lingüística de corpus. Potencialidades para la investigación textual. En Nuria Rodríguez Ortega (dir.), *Teoría y literatura artística en la sociedad digital: construcción y aplicabilidad de colecciones textuales informatizadas*, 67-96. Gijón: TREA.

Sánchez, Aquilino. 1995. Definición e historia de los corpus. En Aquilino Sánchez, Ramón Sarmiento, Pascual Cantos y José Simón (eds.), *CUMBRE-Corpus lingüístico de español contemporáneo: fundamentos, metodología y aplicaciones*, 7-24. Madrid: SGEL.

Santalla del Río, María Paula. 2005. La elaboración de corpus lingüísticos. En Mario Cal, Paloma Núñez e Ignacio M. Palacios (eds.), *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, 45-63. Santiago de Compostela: Universidad de Santiago de Compostela.

Shiro, Martha. 1998. *Los pequeños cuentacuentos. El desarrollo de las habilidades narrativas de niños en edad escolar*. Caracas: Universidad Central de Venezuela. Trabajo de ascenso.

Silva-Corvalán, Carmen. 2001. *Sociolingüística y pragmática del español*. Washington, D.C.: Georgetown University Press.

Sinclair, John McHardy. 1996. [En línea]. Preliminary recommendations on Corpus Typology. *EAGLES* EAG-TCWG-CTYP/P. Pisa: ILC-CNR. Disponible en http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html. [Consulta: 13 noviembre 2014].

Sinclair, John McHardy. 2005. [En línea]. Corpus and text - basic principles. En Martin Wynne (ed.), *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. Disponible en http: //icar.univ-lyon2.fr/ecole_thematique/contaci/documents/Baude/ wynne.pdf [Consulta: 5 enero 2013].

Tejera, María Josefina y Luciana de Stefano. 2006. [En línea] Documentos para la historia del español de Venezuela - Siglos XVI-XVIII. [Disco compacto]. Caracas: Fondo Editorial de Humanidades y Educación, Universidad Central de Venezuela. Disponible en http://saber.ucv.ve/jspui/handle/123456789/2051 [Consulta: 4 mayo 2010].

KRÍSTEL GUIRADO

Licenciada en Letras de la Universidad Central de Venezuela y Magíster Scientiarum en Lingüística de la misma universidad. Es profesora agregada de esta casa de estudios y jefa del Departamento de Dialectología del Instituto de Filología "Andrés Bello". Ejerce labores de docencia en las Escuelas de Letras y de Idiomas Modernos y, en el Postgrado en Lingüística de la Facultad de Humanidades y Educación. En el área de lenguaje, la ocupa el estudio de la Gramática y sus usos en el español de Venezuela. Sus investigaciones personales denotan su interés por la sociolingüística, especialmente el estudio de las intenciones comunicativas que puedan explicar ciertos fenómenos gramaticales. Hasta la fecha de su vigencia, contó con la acreditación como Nivel I del Programa de Promoción al Investigador (PPI). Actualmente, está acreditada como Investigador "A-2" del Programa de Estímulo a la Innovación e Investigación (PEII). Su Tesis de Postgrado fue reconocida con el Premio a la Investigación 2005-2006 de la Facultad de Humanidades y Educación de la UCV y ha recibido Mención de Honor por sus trabajos de Ascenso a los escalafones de Asistente y Agregado. Paralelamente, desarrolla una labor literaria y su trabajo creador ha sido merecedor de premios y menciones en bienales y concursos literarios nacionales. Además de los artículos en revistas de investigación arbitradas nacionales e internacionales, cuenta con la publicación impresa de varios de sus trabajos en ambas áreas. La Comisión de Estudios de Postgrado de la Facultad de Humanidades y Educación publicó en el 2009 su libro *(De)queísmo: uso deíctico y distribución social en el habla de Caracas*.