REINGENIERÍA DE CORPUS EN VENEZUELA: UNA PROPUESTA METODOLÓGICA PARA DIVERSIFICAR EL ANÁLISIS DE LOS CORPUS DEL ESPAÑOL HABLADO EN CARACAS

Kristel Guirado
Kristelguirado@gmail.com
Universidad Central de Venezuela
Venezuela

Licenciada en Letras y Magíster Scientiarum en Lingüística de la Universidad Central de Venezuela (UCV). Es profesora agregada de esta casa de estudios y Jefa del Departamento de Dialectología del Instituto de Filología "Andrés Bello".

RESUMEN:

Se propone el término *Reingeniería de Corpus* para designar el proceso de reconfiguración de muestras de habla que permite su reutilización en diversos ámbitos. El proceso se aplicó a cuatro corpus del español hablado en Caracas con el objetivo de construir dos *corpus de propósito especial* para análisis diacrónicos. La metodología incluyó los siguientes pasos: i. hacer una memoria cronológica de la construcción de cada muestra; ii. analizar el patrón distribucional de cada arquitectura; iii. determinar los aspectos quebrantados, los sólidos y los puntos de intervención; iv. crear el rediseño y estimar su representatividad. La aplicación de la metodología produjo dos nuevos corpus para el estudio en tiempo real de los fenómenos lingüísticos en comunidades de habla específicas (universitarios jóvenes y hablantes cultos). Se concluye que la *Reingeniería de Corpus* constituye una práctica provechosa para la Lingüística de Corpus, que proporciona productos útiles e interesantes para la comunidad lingüística.

Palabras clave: Reingeniería de Corpus, corpus de propósito especial, Lingüística de Corpus, español hablado en Caracas, análisis diacrónico.

Recepción: 21/05/2015 **Evaluación:** 27/05/2015 **Recepción de la versión definitiva:** 06/07/2016

CORPUS RE-ENGINEERING IN VENEZUELA: A METHODOLOGICAL PROPOSAL FOR THE DIVERSIFICATION OF ANALYSIS OF CORPORA OF SPANISH SPOKEN IN CARACAS

ABSTRACT

Te term *Corpus re-engineering* is proposed to refer the process of reconfiguration of samples of speech, which allows its reuse in different settings. The process was developed with four corpora of Spanish spoken in Caracas, aiming at the construction of two *special purpose corpora* for diachronic studies. The methodology included the following steps: i. building a chronological memory of each sample construction; ii. analyzing the distributional pattern of each architecture; iii. Determining which aspects are flouted as well as the solid aspects and the points of intervention; iv. creating a re-design and ascertaining its representativeness. Such methodology produced two new corpora for the real time study of linguistic phenomena in specific speech communities (university young people and cultured speakers). It is concluded that *Corpus re-engineering* is a rewarding practice for corpus linguistics that provides useful and interesting products for the linguistic community.

Palabras clave: *Corpus re-engineering*, special purpose corpus, corpus linguistics, Spanish spoken in Caracas, diachronic study.

LA RÉINGÉNERIE DU CORPUS AU VENEZUELA : UNE PROPOSITION MÉTHODOLOGIQUE POUR DIVERSIFIER L'ANALYSE DES CORPUS DE L'ESPAGNOL PARLÉ À CARACAS.

RÉSUMÉ

On propose le terme *Réingénierie de Corpus* pour désigner le processus de reconfiguration d'échantillons du parler permettant sa réutilisation dans des domaines divers. La démarche a été appliquée à quatre corpus de l'espagnol parlé à Caracas avec l'objectif de *construire deux corpus de propos spécial* pour des analyses diachroniques. La méthodologie a compris les étapes suivantes : i. faire une mémoire chronologique de la construction de chaque échantillon ; analyser le patron distributionnel de chaque architecture ; iii. déterminer les aspects transgressés, ceux étant solides et les points d'intervention ; iv. créer le nouveau paramètre et estimer sa représentativité. L'application de cette méthodologie a produit deux nouveaux corpus pour l'étude en temps réel des phénomènes linguistiques dans des communautés spécifiques de parler (de jeunes étudiants et des parlants cultivés). On conclut que la *Réingénierie de Corpus* constitue une pratique profitable à la Linguistique de Corpus fournissant des produits utiles et intéressants pour la communauté linguistique.

Mots clés : Réingénierie de Corpus, corpus de propos spécial, Linguistique de Corpus, espagnol parlé à Caracas, analyse diachronique.

"REINGEGNERIA" DEL CORPUS IN VENEZUELA: UNA PROPOSTA METODOLOGICA PER DIVERSIFICARE L'ANALISI DEL CORPUS DELLO SPAGNOLO PARLATO A CARACAS

RIASSUNTO

Il termine *"reingegneria"* del Corpus intende descrivere il processo di riconfigurazione dei campioni della lingua parlata che ne consente la riutilizzazione in diversi campi. Il processo è stato applicato a quattro corpus dello spagnolo parlato a Caracas, alla fine di costruire due corpus di scopo speciale per l'analisi diacronica. La metodologia ha incluso i seguenti passaggi: a) rendere una memoria cronologica della costruzione di ciascun campione; b) analizzare lo schema distributivo di ogni architettura; c) determinare gli aspetti rotti, gli aspetti solidi e i punti di intervento; d) creare la riprogettazione e stimare la sua rappresentatività. L'applicazione della metodologia ha prodotto due nuovi corpus per lo studio in tempo reale dei fenomeni linguistici nelle comunità col discorso specifico (giovani universitari e parlanti istruiti). Si evince che la *"reingegneria"* del Corpus costituisce una pratica vantaggiosa per la linguistica di corpus, che fornisce prodotti utili e interessanti per la comunità linguistica.

Parole chiavi: *"reingegneria"*. del Corpus. Corpus di scopo speciale. Linguistica di corpus. Spagnolo parlato a Caracas. Analisi diacronica.

REENGENHARIA DE CORPUS NA VENEZUELA: UMA PROPOSTA METODOLÓGICA PARA DIVERSIFICAR A ANÁLISE DOS CORPUS DO ESPANHOL FALADO EM CARACAS

RESUMO

O termo Reengenharia de Corpus designa o processo de reconfiguração de mostras de fala que permitem sua reutilização em diversos âmbitos. O processo foi aplicado a quatro corpus

do espanhol falado em Caracas com o objetivo de construir *dois corpus de propósito especial* para as análises diacrônicas.
A metodologia incluiu os seguintes passos: i. Fazer uma memória cronológica da construção de cada mostra; ii. Analisar o padrão distribucional de cada arquitetura; iii. Determinar os aspectos quebrantados, os sólidos e os pontos de intervenção; iv. Criar o redesenho e estimar sua representatividade. A aplicação da metodologia produziu dois novos corpus para o estudo em tempo real dos fenômenos linguísticos em comunidades de fala específicas (universitários jovens e falantes cultos). Dessa maneira, a reengenharia de Corpus constitui uma pratica proveitosa para a linguística de Corpus, que proporciona produtos úteis e interessantes para a comunidade linguística.

Palavras-chave: reengenharia de Corpus, corpus de propósito especial, Linguística de Corpus, espanhol falado em Caracas, análise diacrônica.

1. INTRODUCCIÓN: LOS CORPUS ELECTRÓNICOS Y SU REINGENIERÍA

Durante las últimas décadas ha habido un aumento en la creación y explotación de corpus lingüísticos para el estudio del español, tanto en España como en América. El auge que ha tomado esta disciplina ha hecho que, actualmente, en la mayoría las instituciones de investigación lingüística, se esté trabajando en la elaboración de algún tipo de corpus. Son muchos los ejemplos que podríamos citar para señalar la tendencia de la lingüística actual al estudio del lenguaje en uso. No obstante, un inventario extenso puede llevar consigo omisiones imperdonables; baste para ello con mencionar la grandes empresas acometidas por la Real Academia desde 1995 (CREA, CORDE, CDH, CORPES XXI).

Se entiende por corpus lingüístico toda colección de textos, diseñada de forma representativa, para el estudio de la lengua. Esta definición básica incluye los parámetros fundamentales de su construcción y el objetivo principal.[39]

En principio, el concepto no parece estar asociado, a priori, con el desarrollo nuevas tecnologías. No obstante, actualmente, la mayoría de los autores incluyen esta característica como criterio fundamental en la descripción,[40] ya que ignorar las propiedades que facilitan el análisis informatizado de los datos lingüísticos sería incurrir en una carencia metodológica. Parodi (2008) presenta una de las definiciones más completa de 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

---

[39] "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 1979: p. 110); "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); "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: p. 449).

[40] Entre otros Sánchez (1995); McEnery y Wilson (1996); (Santalla 2005); Sinclair (2005); McEnery, Xiao y Tono (2006).

> 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 (p. 106).

Esta definición incluye origen, propósito, composición, formato, representatividad y extensión. Adicionalmente, Parodi enfatiza en la suficiencia del corpus para aportar los metadatos de su arquitectura, detalle substancial que muchos teóricos de la Lingüística de Corpus (LC) pasan inadvertido.

De acuerdo con el criterio de representatividad, los corpus pueden ser objeto de diversas tipologías. Atkins, Clear y Ostler (1992) diferencian entre *corpus* y *subcorpus*. Según Sinclair (1996), los subcorpus comparten todas las características del corpus del cual han sido extraídos; no obstante, para Torruella y Llisterri (1999), los subcorpus también pueden ser selecciones estáticas o dinámicas de textos de un corpus en crecimiento. En todo caso, para los tres autores, un subcorpus es cualquier porción seleccionada de un corpus mayor.

Sinclair (1996) hace uso del término *corpus especial* (*special corpus*) para describir aquellos corpus que se diseñan para ser representativos de una variedad lingüística específica o de algún tipo de sublenguaje o lengua especializada. Pearson (1998) señala que Sinclair parece circunscribir los llamados *corpus especializados* (*specialized corpus*) dentro de los corpus especiales, inclusión que no queda del todo clara. No obstante, es un hecho que no constituyen un subcorpus porque "no están diseñados para tener todas las propiedades de un corpus más grande".[41] Además de la tipología descrita en Sinclair (1987) y (1996), Pearson propone un tipo de corpus que no se puede incluir dentro de las clasificaciones previas, el *corpus de propósito especial* (*especial purpose corpus*):

> un corpus cuya composición está determinada por el propósito exacto para el cual se va a utilizar [...] puede derivarse de un corpus de referencia general o de un corpus monitor, pero no constituirá un subcorpus en el sentido definido por Sinclair, porque no tendrá todas las propiedades de un corpus más grande, ya que se pueden imponer restricciones relativas a género, autor, período u otros criterios según la finalidad para la cual se destina el corpus. Tampoco constituirá un corpus especial porque no existe la expectativa a priori de que el lenguaje utilizado se desvíe de la norma [...] como podría ser en el lenguaje de ancianos, afásicos o niños (Pearson, 1998: p. 48).[42]

---

[41] Traducción del texto original, (en adelante, TTO): "it is clearly not possible to classify them under the heading of subcorpus because they are not designed to have all of the properties of a larger corpus" (Pearson 1998: p. 46).

[42] TTO: "This is what we choose to call a special purpose corpus, a corpus whose composition is determined by the precise purpose for which it is to be used. While a special purpose corpus may be derived from a general reference corpus or from a monitor corpus it will not constitute a subcorpus in the sense defined by Sinclair because it will not have all of the properties of a larger corpus. Restrictions relating to genre, author, period or other criteria may be imposed depending on the purpose for which the corpus is intended. Nor will it constitute a

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. Parodi (2008: p. 98) afirma que: "la manera de entender un corpus ha evolucionado y [...] la explotación del mismo enfrenta desafíos y proyecciones jamás antes imaginados". En este sentido, se propone el término *Reingeniería de Corpus* 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.

La gestión de reingeniería se fundamenta en la premisa de que el aprovechamiento de un corpus no depende de su contenido en sí mismo sino de los procesos que dieron lugar a su conformación. Cuando se habla de procesos, se considera que las variables contextuales que intervinieron en la estructuración de una muestra también forman parte de lo estrictamente procedimental, como las características geohistóricas de las poblaciones objeto de estudio, la incidencia de la cultura empírica del momento, e incluso la memoria particular de entornos e intereses que signaron el quehacer científico de una comunidad de investigadores.

Ahora bien, es posible suponer que la configuración inicial de un corpus no resulte del todo atractiva para los usuarios de otra realidad lingüístico-discursiva. Hay que advertir que el ciclo de vida de los "productos" es vulnerable en el marco de la cultura global, hecho del cual no escapan las actividades científicas. No obstante, la eliminación de la frontera comunicacional, que es el soporte fundamental de la cultura actual, representa una ventaja respecto a las posibilidades de expansión de los productos y servicios que ofrece la investigación lingüística.

De esta forma, la labor del investigador debe concentrarse en dos áreas: los usuarios del corpus y/o su manufactura. Para lograr esto, primero, es necesario tener presente los vaivenes teórico-metodológicos y las innovaciones promovidas por el cambio tecnológico en el universo de los estudios lingüísticos, en todas las disciplinas y corrientes, de forma que la naturaleza del corpus responda, en alguna medida, a dichas expectativas. Segundo, no hay que olvidar que son los procesos (sinergia, des/integración, movilización, redistribución, restratificación, eliminación, etc.) y no los corpus en sí los sujetos a reingeniería.

---

special corpus because, in special corpora there is an a priori expectation that the language used will deviate from the norm. This is not the case with the language of special purpose corpora. There may be lexical deviations in the sense that some words are used in a precise and specialized way but this hardly constitutes a contravention of normal rules because these terms are not used incorrectly, as they might be in the language of geriatrics, aphasics or children. As the corpora on which we are working do not fit into any of the categories ascribed by others, we have deliberately chosen to coin this new category, i.e. the special purpose corpus. We plan to use this term whenever the specific purpose for which the corpus is to be used (e.g. retrieval of definition statements, analysis of gender-related issues) is the reason for creating or selecting the corpus (Pearson 1998: p. 48).

Por una parte, para identificar y establecer cuáles procesos van a ser modificados, es necesario verificar las condiciones básicas de viabilidad de un corpus: textos procedentes de entornos naturales, que expliciten los rasgos definitorios y compartidos, disponibles en formato de tipo digital, preferentemente de tamaño extenso, que respondan a principios ecológicos, preferiblemente etiquetados, que respeten principios de proporcionalidad o representatividad (posiblemente estadística), con especificación de la procedencia inicial, de organización textual identificable, con registro de datos cuantitativos que permita la comparación de las muestras entre sí.

Por otra parte, se propone atender los siguientes aspectos antes de seleccionar un proceso de rediseño:

i. *Atributos quebrantados*: depreciación de la sincronía, desequilibrio distribucional, formato, etc.
ii. *Propiedades sólidas*: estratificación social de la muestra, comparabilidad, procesamiento informatizado, datos de procedencia, etc.
iii. *Procesos factibles*: cambio y/o creación de nuevas de unidades de análisis, restratificación, sinergia con otras muestras, disponibilidad, etc.

La ejemplificación anterior no es restrictiva; se puede considerar uno o más aspectos en cada categoría. Asimismo, la planificación heurística de cada reingeniería impondrá la distribución de criterios en cada aspecto, de modo que una propiedad que se presenta sólida en el marco de un rediseño particular (el tamaño, por ejemplo), puede representar un atributo quebrantado en otra reconfiguración.

En el presente artículo, se muestra el proceso para la producción de dos corpus de propósito especial a partir de la reingeniería de una serie de corpus orales para el estudio del español caraqueño que reposan en el Instituto de Filología "Andrés Bello" (IFAB), con el objetivo de disponer de nuevos materiales para el estudio de fenómenos lingüísticos/discursivos en tiempo real y en comunidades de habla específicas. Los materiales que sirven de base para la reingeniería se han recolectado a lo largo de más 50 años de investigación lingüística apoyada en muestras reales del uso de la lengua. Estos corpus fueron estratificados socialmente y diseñados según una rigurosa arquitectura distribucional, haciendo uso de técnicas, herramientas y métodos estadísticamente representativos, de modo que en la etapa de reconstrucción fue posible orientar su empleo hacia mayores niveles de aprovechamiento y reutilización.

El artículo se divide en seis partes. En la presente sección se introduce el tema de investigación y su marco teórico, la propuesta metodológica y los objetivos del proyecto de reingeniería. En el apartado siguiente se presenta una memoria cronológica que permite

recuperar las circunstancias contextuales –colectivas y particulares– que signaron la construcción de cada uno de los corpus escogidos para el rediseño; al mismo tiempo, en este apartado se intenta hacer un reconocimiento a los equipos de trabajo –investigadores, auxiliares y pasantes– que contribuyeron con su constancia y esfuerzo a la culminación de la ardua tarea que implica recoger y sistematizar un corpus, muy especialmente a Paola Bentivoglio, quien coordinó estos equipos de trabajo y dedicó buena parte de su quehacer académico para legar este conjunto de materiales a la comunidad lingüística nacional e internacional. En la tercera se describe la configuración de los cuatro corpus de habla caraqueña que serán objeto de reingeniería. En la cuarta sección se relatan los procesos de reestratificación y rediseño de las muestras para la creación de nuevos corpus. Luego se ofrece un cuadro resumen de las características de todos los corpus reseñados. Y, finalmente, se cierra el artículo con algunas consideraciones sobre el proceso de reingeniería documentado.

2. El IFAB y la Lingüística de Corpus: memoria cronológica

2.1. Ángel Rosenblat y los inicios de la Lingüística de Corpus en Venezuela

En 1947, cuando el profesor Ángel Rosenblat se incorpora como docente e investigador a la recién fundada Facultad de Filosofía y Letras de la Universidad Central de Venezuela, se da a la tarea de crear el IFAB, un instituto dedicado a los estudios de la Filología moderna: "propuse el nombre de Andrés Bello, porque aspiraba a armonizar el movimiento filológico de la época con la más alta expresión gramatical de nuestra lengua" (Rosenblat, 2002: p. 501). El objetivo central del IFAB fue –y sigue siendo– el estudio y la descripción del español de Venezuela.

La línea central de investigación promovida por el Rosenblat fue el estudio sistemático del léxico. Aunque otros temas se llevaron adelante de forma individual (el estudio filológico de obras históricas y literarias, la descripción de algunas lenguas indígenas y del lenguaje coloquial), el proyecto de envergadura que ocupó a los investigadores del instituto durante más de veinte años fue la creación del *Diccionario de venezolanismos*.[43] No obstante, el compromiso de llevar adelante un instituto que floreciera a la par de la investigación lingüística mundial, llevó a Rosenblat, finalmente, al camino de la LC. Es así como, en 1966,

[43] La primera tarea del recién fundado Instituto fue la de recopilar una serie de palabras, identificarlas y clasificarlas para la creación del *Diccionario histórico del español de Venezuela*. Con este propósito inicial comienza la creación de "el más importante y extenso fichero documental sobre la lengua de Venezuela" (Pérez, 1997: web). Aunque el proyecto de diccionario histórico no llegó a concretarse, los documentos compilados alimentaron los artículos publicados por Rosenblat en el *Papel literario* del diario *El Nacional* (de 1954 a 1956), dieron lugar a la creación de la Biblioteca "Ángel Rosenblat" —una unidad de servicios de información especializada en el área de la Lingüística, única en el país, que funciona adscrita al IFAB— y constituyeron la base fundamental para la elaboración del *Diccionario de venezolanismos* (Tejera 1993).

Rosenblat pasa a integrar la primera Sub-Comisión Ejecutiva del *Proyecto para el estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica*, en representación de Caracas.[44]

2.2. Paola Bentivoglio y el estudio sociolingüístico del habla venezolana

Se inicia, entonces, la recopilación de los primeros materiales del habla urbana venezolana: "bajo la mirada de Rosenblat nacería una de las investigaciones más fecundas del Instituto: *El habla culta de Caracas*, que de la mano de Paola Bentivoglio inauguraría los estudios sociolingüísticos modernos [en Venezuela]" (Pérez, 1997: web).[45] Esta primera etapa de recolección de la muestra culmina en 1979 con la publicación impresa de algunas de las entrevistas grabadas (*Vid. infra* §3.1).

En 1976, Paola Bentivoglio, Francesco D'Introno y Juan Manuel Sosa –quienes comparten espacios de docencia y perspectivas de análisis– se unen para desarrollar "una investigación sociolingüística del español de Caracas, que Ángel Rosenblat [...] aprobó con beneplácito y Paola puso en marcha" (F. D'Introno, *Semblanzas*, en Sedano, Bolívar y Shiro, 2006: p. 27). En ese momento, el auge de la sociolingüística variacionista en Hispanoamérica convertía esta disciplina en un paradigma teórico atrayente y novedoso.[46] No obstante, su rigurosa metodología establece que los usos lingüísticos deben estudiarse directamente en muestras de habla espontánea, procedentes de hablantes seleccionados de acuerdo con sus características sociales inherentes (sexo, edad o raza) o adquiridas (nivel educativo o clase social); asimismo, el número de hablantes de la muestra debe ser

---

[44] Juan Manuel Lope Blanch presenta el proyecto en 1964, durante el II Simposio del Programa Interamericano de Lingüística y Enseñanza de Idiomas (PILEI), celebrado en Bloomintong/Indiana. El objetivo principal de la propuesta consiste en recolectar y sistematizar muestras de habla con el propósito de estudiar las características y tendencias de las normas del habla culta de las principales ciudades que tienen el español por lengua oficial. Este proyecto constituye la primera empresa colectiva y multinacional que se propone estudiar el español de forma "coordinada, homogénea, común, por encima de preferencias personales o de principios teóricos particulares" (Lope Blanch, 1987: p. 167).

[45] Es importante señalar que, desde finales de los años 60, la mayoría de los investigadores venezolanos se ha preocupado por fundamentar científicamente la investigación del lenguaje al estudiar y documentar la extensión y pertinencia real de los fenómenos en amplias muestras de la lengua en uso. De esta forma, además de los corpus que aquí serán reseñados, 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). Para una descripción de otros corpus de habla venezolanos, Cf. Bentivoglio (1998).

[46] "El año 1976 representa la consolidación de este nuevo enfoque de la realidad lingüística manifestada, sobre todo, en la realización en Río Piedras del I Simposio de Dialectología del Caribe Hispánico que, bajo la experta y hábil dirección de H. López Morales, reunió, por primera vez, a la que podríamos considerar como «plana mayor» de los estudios sociolingüísticos en la América Hispánica" (Granda, 1994: p. 198). A esta reunión tienen el honor de ser invitados, entre otros, Paola Bentivoglio y Francesco D'Introno.

estadísticamente representativo de la comunidad (Labov, 1972).[47] A finales de ese mismo año, los investigadores dan inicio a la recolección de entrevistas.

Este nuevo corpus, enmarcado en el proyecto *Estudio sociolingüístico del habla de Caracas 1977* (*Vid. infra* §3.2), estaba diseñado según las características y parámetros necesarios establecidos por el modelo laboviano; en tal sentido, se incluyeron en la estratificación de la muestra dos características sociales inherentes a los hablantes (sexo y edad) y una adquirida (nivel socioeconómico).

Durante los primeros años de la década de los 80, Paola Bentivoglio y Mercedes Sedano dedican de su actividad profesional al estudio de una serie de fenómenos morfosintácticos documentados en las muestras de habla arriba citadas. De esta forma, lograron consolidar una nueva línea de investigación en el IFAB. En 1987, las investigadoras decidieron coordinar la construcción de un nuevo corpus que les permitiera abordar cabalmente el estudio de algunos fenómenos en el habla caraqueña tanto sincrónica como diacrónicamente.

De esta forma, bajo el nombre de *Estudio sociolingüístico del habla de Caracas 1987* (*Vid. infra* §3.3), las investigadoras construyeron uno de los corpus de habla más emblemáticos con respecto a su estratificación social y representatividad. Este nuevo corpus va a superar las limitaciones y expectativas de los recogidos anteriormente y va a servir de modelo para los proyectos de estudio del español oral en otros dialectos venezolanos y en otras variedades de habla hispana; asimismo, va a motivar los estudios comparados con lenguas como el portugués de Brasil y el francés: "The similarities of our corpuses and our shared belief that linguistic analysis must be firmly anchored in authentic spoken data led naturally to collaborative work comparing spoken Spanish and French" (W. Ashby, *Semblanzas*, en Sedano, Bolívar y Shiro, 2006: p. 25).

En 1996, en el congreso de la ALFAL celebrado en Las Palmas de Gran Canaria, se presentó la metodología para el desarrollo de PRESEEA (Proyecto para el estudio sociolingüístico del español de España y de América).[48] Para el 2003, se había sumado al proyecto un importante número de equipos de diversas ciudades de España, México, Puerto Rico, Colombia y Argentina; sin embargo, no será hasta el 2004 cuando Venezuela y su

---

[47] "Labov prima la recogida de datos procedentes de numerosos individuos seleccionados por medios estadísticos, lo que conlleva un tratamiento también estadístico de esos datos y un abandono conceptual de los individuos concretos que los han proporcionado con el fin de conseguir el componente probabilístico de la competencia sociolingüística" (Herrera Santana, 1994/95: *nota* 61).

[48] La finalidad del Proyecto PRESEEA 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). Para más información, cf. Moreno Fernández (1996, 1997 y 2005a); y la página web: <http://preseea.linguas.net/>.

capital se incorporé formalmente: "la constancia de Irania Malaver y las innegables ventajas del proyecto PRESEEA finalmente convencieron a la primera investigadora de que un equipo de la Universidad Central de Venezuela entrara formar parte del mega proyecto" (Bentivoglio y Malaver, 2006: p. 42). Tras cuarenta años de ejercer la LC en el país, la investigadora inició el proyecto de recolección de un nuevo corpus sociolingüístico de Caracas (*Vid. infra* §3.4): "he tratado siempre de trabajar con alguien, de trabajar en equipo [...] no soy una trabajadora solitaria [...] la individualidad se convierte en una simple desventaja" (Registro Nacional Voz de los Creadores: *Paola Bentivoglio*, p. 164).

3. LOS CORPUS SOCIOLINGÜÍSTICOS DEL HABLA CARAQUEÑA[49]

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

De acuerdo con el proyecto global (*Vid. supra*, *nota* 11), el objetivo lingüístico del estudio de la norma culta caraqueña era disponer de una gran base de datos que permitiera realizar la mayor cantidad de estudios sobre el habla de la ciudad, para reconocer sus características y tendencias normativas y compararla con el resto de las principales ciudades del mundo hispánico. Entre los años 1968 y 1977, el equipo a cargo del proyecto grabó alrededor de 232 conversaciones, con la participación de 310 hablantes aproximadamente, todos caraqueños cultos, residenciados en la ciudad e hijos de padres caraqueños.[50]

Según la metodología preestablecida en el denominado *Proyecto del Habla Culta*,[51] 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.

En el estudio de la norma culta se proyectó un posible análisis tanto de las modalidades de habla informal (habla familiar del coloquio) como formal (conferencias y clases magistrales). Para tales fines se aplicaron cuatro tipos de encuestas: i) *entrevistas*

---

[49] Además de los cuatro corpus que se describen en este apartado, en el IFAB se han recogido otro tipo de muestras de habla como el *Corpus de habla infantil de Caracas 1996* (Shiro, 1998) y el corpus de *Documentos para la Historia del Español de Venezuela de los siglos XVI, XVII y XVIII* (De Stefano y Pérez Arreaza, 2000; Tejera y De Stefano, 2006). No obstante, ellos no son objeto, en esta oportunidad, de ninguna labor de reingeniería, razón por la cual no me extiendo en la descripción de los mismos.

[50] El IFAB 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).

[51] "se han elegido las siguientes variables: 1) la realización oral de la lengua, y no la escrita; 2) el nivel de norma, y no de sistema; 3) la variable estrática culta, y no la inculta; 4) las variables fásicas formal e informal; 5) la variable tópica urbana, y no rural; 6) la variable crónica esencialmente sincrónica, y no diacrónica, y 7) las variables génicas, masculina y femenina" (Rabanales, 1992: p. 254).

con un solo hablante; ii) *diálogos entre dos informantes*; iii) *conferencias*; y, iv) *encuestas secretas*.

El CHCC 68-77 suma un total de 240 horas de grabación, disponibles en formato electrónico. Las grabaciones fueron transliteradas todas ortográficamente y se dispone de la transcripción impresa de las encuestas.[52] Lamentablemente, para este momento, en el IFAB se disponía solo de tecnología para mecanografiar las entrevistas.

La transcripción de una parte de estos materiales, recogidos entre 1973 y 1977, fue publicada en *El habla culta de Caracas. Materiales para su estudio* (Rosenblat y Bentivoglio, 1979). Adicionalmente, las catorce entrevistas incluidas en el citado volumen están contenidas en soporte digital en el *Macro-corpus de la norma lingüística culta de las principales ciudades del mundo hispánico* (Cf. Samper, Hernández Padilla y Troya Déniz, 1998). Estas transcripciones también han sido incorporadas al *Corpus de referencia del español actual* (CREA) de la Real Academia Española. Actualmente, se adelanta un proyecto para transliterar electrónicamente la totalidad de las entrevistas, con la finalidad de ofrecer el material del corpus en un soporte informatizado.

3.2. *Corpus Sociolingüístico de Caracas de 1977* (CSC 77)

El CSC 77 es el primer corpus de habla de la ciudad capital que se diseña siguiendo criterios sociolingüísticos para su estratificación. Bentivoglio, D'Introno y Sosa, coordinadores generales del proyecto, habían concebido un primer corpus que incluía un total de 288 hablantes, los cuales debían estar distribuidos equitativamente en dos zonas geográficas (este y oeste de la ciudad, teniendo por límite la avenida Fuerzas Armadas); cuatro grupos etarios (A: 14 a 29 años; B: 30 a 45; C: 46 a 60; y D: 61 o más años); tres niveles socioeconómicos (alto, medio y bajo) y sexo.[53] El proyecto, sin embargo, completó solo dos de los cuatro grupos etarios que formaban parte la arquitectura original del corpus (los hablantes de los grupos etarios A y B de la zona este).

---

[52] En las grabaciones participaron Paola Bentivoglio, María Caggiano, Delfina Catalá, Carmen Luisa Domínguez, Dexy Galué, Aura Gómez de Ivashevsky, Milagros Guevara, Rosalba Iuliano, Karin Marent, María Jesús Sánchez, Juan Manuel Sosa y Guillermina Suárez. La revisión de los materiales transcriptos estuvo a cargo de Paola Bentivoglio, Liana Gianesin y Rosalba Iuliano. María Teresa Rojas y Jorge Nelson Rojas participaron en la revisión final del volumen. También colaboraron Luciana De Stefano, María Alexandra Álvarez y Fernando Fernández Gómez.

[53] El nivel socioeconómico de los informantes se determinó a través de un índice creado especialmente para el análisis de fenómenos sociolingüísticos por Max Contasti. Para una explicación más amplia del método, cf. Contasti (1980).

Actualmente, el corpus está conformado por 70 muestras orales,[54] de hablantes oriundos de Caracas, hijos de padres caraqueños y residenciados en el área metropolitana la mayor parte de su vida.[55] En el cuadro 1 puede observarse la distribución de los hablantes del corpus:

Cuadro 1. Distribución de los hablantes del CSC 77

nivel socioeconómico	GRUPO GENERACIONAL				Total
	Grupo A 14 a 29 años		Grupo B 30 a 45 años		
	hombres	mujeres	hombres	mujeres	
ALTO	6	6	6	6	24
MEDIO	6	4	6	6	22
BAJO	6	6	6	6	24
Total	18	16	18	18	70
	34		36		

Las grabaciones, de 30 minutos cada una, suman un total de 35 horas. Cada hablante fue entrevistado individualmente.[56] Se les informó que estaban siendo grabados, pero se les ocultó el objetivo lingüístico de la grabación.[57] El estilo de las entrevistas "puede considerarse como *careful* 'cuidadosa' (Labov, 1972), aunque algunas partes se acercan al habla 'informal' y hasta espontánea" (Bentivoglio, 1987: p. 25). Todas fueron transliteradas ortográficamente. La muestra total suma 285.916 palabras (Gallucci, 2005: p. 112). Todas las encuestas están incorporadas al CREA.

3.3. *Corpus sociolingüístico de Caracas 1987* (CSC 87)

El CSC 87 fue grabado siguiendo los mismos parámetros metodológicos del CSC 77 pero con mayor especificidad en el rango de estratificación (cinco niveles socioeconómicos).[58] Las creadoras y coordinadoras del proyecto fueron Paola Bentivoglio y Mercedes Sedano; posteriormente se unió al equipo coordinador Alexandra Álvarez.[59]

---

[54] Por razones técnicas hubo que prescindir de dos grabaciones, por lo que tampoco fue posible completar una de las casillas de hablantes para una distribución equitativa del corpus (las mujeres del grupo etario A del nivel socioeconómico medio).

[55] El área geográfica considerada como Caracas incluyó, además de los municipios metropolitanos (Libertador, Baruta, Chacao, El Hatillo y Sucre) las zonas aledañas de Los Teques y Valles del Tuy.

[56] Entrevistadores: Juan Manuel Sosa, Irma Chumaceiro, Luis Bruzual, Dexy Galué, Edgar Colmenares del Valle y Zayda Pérez.

[57] En su lugar, se les dijo que el fin era conocer su punto de vista sobre los problemas propios de una metrópolis como la Caracas de entonces (servicios públicos; tráfico; accidentes de tránsito), (Bentivoglio, 1987: p. 25).

[58] El proyecto fue financiado por el CDCH-UCV, bajo los N° H-07.16/86 y H-08.33.1766.88/89.

[59] Se desempeñaron como auxiliares de investigación los estudiantes de la Escuela de Letras: Cleris Malavé, Mercedes Acosta, Antonieta Alario, María Josefina Barajas, Domingo Ledezma, Ricardo Calles, María Fernanda Lander, María Alejandra Calzadilla, Evelyn Castro, Judith Herradas y Ruddy Reyes.

La propuesta se basa en la recolección de 160 muestras de habla a individuos caraqueños, hijos de padres caraqueños, residenciados en la ciudad. Las grabaciones fueron agrupadas proporcionalmente según tres características sociales de los hablantes: cuatro grupos etarios (A: 14 a 29 años; B: 30 a 45; C: 46 a 60; y D: 61 o más años); cinco niveles socioeconómicos (alto, medio alto, medio, medio bajo y bajo) y sexo.[60] Esta distribución se ilustra en el cuadro 2:

Cuadro 2. Distribución de los hablantes del CSC 87

nivel socioeconómico	GRUPO ETARIO								Total
	Grupo A 14 a 29 años		Grupo B 30 a 45 años		Grupo C 46 a 60 años		Grupo D 61 años o más		
	hombres	mujeres	hombres	mujeres	hombres	mujeres	hombres	mujeres	
ALTO	4	4	4	4	4	4	4	4	32
MEDIO ALTO	4	4	4	4	4	4	4	4	32
MEDIO	4	4	4	4	4	4	4	4	32
MEDIO BAJO	4	4	4	4	4	4	4	4	32
BAJO	4	4	4	4	4	4	4	4	32
Total	20	20	20	20	20	20	20	20	160
	40		40		40		40		

Cada entrevista tiene una duración de 30 minutos, lo que suma 80 horas de grabación.[61] En esta oportunidad, se prestó un mayor cuidado a los requisitos que los auxiliares de investigación debían cumplir y seguir para hacer la entrevista.[62] La transcripción de las grabaciones se hizo de forma ortográfica.[63] La muestra tiene 767.868 palabras (Gallucci, 2005:114). Las transcripciones de todas las grabaciones están incorporadas al CREA.

3.4. *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* (PRESEEA), coordinado por Francisco Moreno Fernández.

[60] Se utilizó, esencialmente, el mismo método del CSC 77 para establecer los parámetros socioeconómicos (Contasti, 1980), con una pequeña variante: "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*" (Bentivoglio y Sedano, 1993: p. 8).

[61] Vale la pena señalar que, en la grabación del CSC 87, al igual que en la de los dos corpus anteriores (CHCC 68-77 y CSC 77), se emplearon cintas de un cuarto de pulgadas por 600 pies, con velocidad 9.5 cm/seg y un grabador UHER 4000, un equipo de grabación de auténtica y certificada calidad. Estos equipos eran los mismos que usaban los reporteros de la BBC entonces. El grabador UHER 4000 fue fabricado en Alemania en 1966. Es una grabadora de carrete a carrete portátil, que cuenta con cuatro velocidades (19/9.5/4.75/2.4 cm / seg) y acomoda bobinas de hasta 13 cm (5 pulgadas), lo que hace que sea de dos pistas (algunas veces llamada "vía media") y proporcione grabaciones mono con una rango de frecuencias de hasta 20.000 Hz.

[62] Para consultar la normativa con las pautas detalladas de grabación, cf. Bentivoglio y Sedano 1993: p. 5).

[63] Para consultar la normativa con las pautas detalladas de transcripción, cf. Bentivoglio y Sedano (1993: pp. 16-18).

El desarrollo del proyecto en Caracas está coordinado por Paola Bentivoglio e Irania Malaver. Las co-coordinadoras del proyecto son María José Gallucci y Carla González.[64]

La muestra está formada por 108 entrevistas hechas a informantes nacidos en Caracas, hijos de padres caraqueños. El corpus se dividió según los siguientes factores sociales: 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. La distribución se muestra en el cuadro 3:[65]

Cuadro 3: Distribución de los hablantes de PRESEEA-CSC 04-10

grado de instrucción	GRUPO GENERACIONAL						Total
	Grupo A 20 a 34 años		Grupo B 35 a 54 años		Grupo C 55 años o más		
	hombres	mujeres	hombres	mujeres	hombres	mujeres	
PRIMARIA	6	6	6	6	6	6	36
MEDIA	6	6	6	6	6	6	36
UNIVERSITARIA	6	6	6	6	6	6	36
Total	18	18	18	18	18	18	108
	36		36		36		

Las grabaciones tienen una duración mínima de 45 minutos cada una, para un estimado de 81 horas de grabación. En este corpus se mantiene el uso de la entrevista semidirigida como parte del método para recolectar los datos (Cf. Silva Corvalán, 2001: pp. 52-53).

La pauta de transcripción de los materiales se ajustó a las normas internacionales de la TEI (*Text Encoding Initiative*),[66] lo que constituye una diferencia importante con los corpus precedentes. Las transcripciones son utilizables en dos versiones: con o sin etiquetas; y están estructuradas en dos partes: cabecera y texto transcrito.[67] Una muestra de las

[64] Greisy Fernández, Zayra Marcano, Vanessa Pérez y Cristina Ruiz participaron activamente en el proyecto como asistentes de investigación. Las entrevistas fueron realizadas por estudiantes de la Maestría en Lingüística (muchos de ellos profesores universitarios y de educación media).

[65] La distribución se hizo con cuotas de afijación uniforme de seis hablantes por casilla, este número es suficiente para asegurar la representatividad de la muestra, ya que esta cuota permite alcanzar la proporción de 1/25.000 en las comunidades de dos millones de habitantes aproximadamente (Moreno Fernández, 2005*a*: pp. 127-128). Asimismo, 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 (Cf. Bentivoglio y Malaver, 2006: p. 45; 2012: p. 153).

[66] Haronid Blanco, Katherine Castillo, Mariela Cisneros, Samanta Escalona, Carla González, Andreína Guilarte, Ilse Hernández, Helena Krizmanic, José Alejandro Martínez, Natalia Pericchi, Minerva Ramírez, Simón Ruiz, Zoraida Ruiz, Heidy Sánchez, Alessandra Yerena y María Isabel Yousef han tenido a cargo tareas de grabación y transcripción.

[67] La cabecera permite 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

entrevistas (audio y texto) puede consultarse en la página web del Proyecto PRESEEA, y las transcripciones han sido publicadas en su totalidad en formato digital (Bentivoglio y Malaver, 2014).

4. *Reingeniería de Corpus*: el estudio diacrónico del habla caraqueña

A continuación se reseñan dos corpus diacrónicos construidos a partir de la reingeniería de algunos de los corpus de habla descritos en el apartado anterior, los cuales han sido diseñados con el propósito de estudiar la variación en grupos de entrevistados vinculados entre sí por características comunes; un grupo tiene en común ser hablantes cultos y el otro ser hablantes jóvenes universitarios.

4.1. *Habla culta de Caracas 1973-2011. Corpus diacrónico* (HCC/CD 73-11)

La idea de crear un corpus para el estudio diacrónico del habla culta de Caracas tiene su antecedente en una actividad de publicación integral, concebida para conmemorar los 50 años de la creación de la primera gran empresa de recolección de materiales orales de habla hispana: *Estudio coordinado de la norma lingüística culta hispánica "Juan M. Lope Blanch"*. Durante el XVI Congreso de la ALFAL 2011 celebrado en Alcalá de Henares (España), los delegados de la Comisión Ejecutiva del proyecto acordaron confeccionar un corpus diacrónico con características comunes para un estudio inicial de los marcadores discursivos en cada una de las ciudades:

> Con ese propósito, se realiza un estudio comparativo-diacrónico de las muestras de habla hispánica recogidas hace cincuenta años (las más antiguas), contrastándolas con muestras que obedecen a los mismos parámetros anteriores, pero grabadas con una distancia temporal que va de los 30 a los 45 años, aproximadamente (Valencia Espinoza, 2014: p. 6).

La muestra de cada ciudad estuvo conformada por 12 entrevistas, con una longitud de 5.800 palabras cada una, distribuidas equitativamente entre los dos períodos de grabación de las submuestras y las variables edad y sexo de los hablantes, estipuladas en el estudio inicial. En el caso de Caracas, se seleccionaron seis entrevistas del CHCC 68-77 y seis del PRESEEA-Caracas 04-10.[68] A pesar de los substanciales y valiosos resultados obtenidos con el estudio coordinado de los marcadores discursivos en este corpus, en el XVII Congreso de la ALFAL 2014 celebrado en Paraíba (Brasil), los equipos de investigación de cada una de las ciudades coincidieron en la necesidad de construir una muestra de mayor representatividad para el desarrollo de los futuros temas de investigación conjunta.

---

explicación acerca de los métodos de recolección y procesamiento de los datos, que incluye detalladas descripciones e ilustrativos ejemplos.

68 Para más detalles, cf. Bentivoglio, Guirado y Malaver (2014) y Guirado (2015)

La decisión de ampliar el número de hablantes por casilla no era una tarea fácil, ya que la mayoría de los equipos –como el caso de Caracas– acudió a las grabaciones recogidas por el proyecto PRESEEA de cada ciudad para obtener las entrevistas actuales, lo que representa una economía de recursos y tiempo, pero supone algunos inconvenientes metodológicos. Como se explica en los apartados §2.1 y §2.4 (*vid. supra*), los criterios de selección de los hablantes difiere 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 un hablante de esta modalidad: "convenimos en considerar [...] como "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: p. 258).

A lo anterior, se suma la limitación de no disponer de datos relevantes que permitieran obtener un cálculo del número de entrevistados proporcional al universo relativo, de modo que fue difícil establecer criterios precisos de representatividad de la muestra. No obstante, en el caso de los corpus de propósito especial, es importante considerar que "el criterio de representatividad debe restringirse a la del dominio de estudio específico para el que son creados" (Pérez y Moreno, 2009: p. 76). En este sentido, se optó por determinar una cuota de afijación de tres hablantes por casilla: "normalmente cada cuota es representada por entre tres y cinco informantes" (Moreno Fernández 2005*b*: p. 312). Con este número se logró, 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.[69]

Además del número de hablantes por casilla, también se reconsideró el criterio de longitud de las entrevistas para la distribución del nuevo corpus. En lugar del número de palabras, se tomó en cuenta la duración de las mismas. Para ello, se reasignaron los

---

[69] Como se mencionó antes, no se dispuso de una cifra exacta de la totalidad de hablantes "cultos" caraqueños; sin embargo, se contó con algunos indicadores que indirectamente permiten hacer algunas estimaciones. Por ejemplo, según los datos del censo de 2006, una población de 314.291 encuestados laboraba en ámbitos profesionales, técnicos, gerenciales, administrativos y otros de categoría directiva en el Distrito Capital. En este mismo censo, se registra que el promedio de escolaridad de la población mayor de 25 años era aproximadamente de 9 de años de estudio, lo que permite suponer que solo una minoría de la población de la zona había completado los estudios universitarios (Fuente: Instituto Nacional de Estadística (INE): <http://www.ine.gov.ve/documentos/see/sintesisestadistica2007/estados/distritocapital/index.htm>).
Estos datos resultan interesantes porque, como podrá verse de inmediato, las encuestas de la muestra más actual que se usan en la estructuración del corpus fueron grabadas, en su mayoría, durante los años 2004-2005. Otro dato importante se registra en el censo 2011; de un universo de 1.026.706 habitantes caraqueños que dieron respuesta a los datos educativos, apenas 84.091 tienen estudios universitarios, y solo 67.549 completaron estos estudios (5 a 6 años); (Fuente INE: <http://www.redatam.ine.gob.ve/Censo2011/index.html>).

porcentajes establecidos para los grupos etarios en el proyecto original a los minutos de grabación (30%: 25 a 35 años; 45%: 36 a 55 años; y, 25%: 56 años en adelante); de modo que, en lugar de más hablantes, se adjudicó mayor tiempo o, lo que es igual, mayor longitud discursiva para el grupo 2:

> El lapso de 36 a 55 años para la segunda generación se basa en que estas personas –en la plenitud ahora de sus facultades– tienen en mayor numero y grado injerencia decisiva en la marcha del país, tanto en lo político como en lo económico y en lo cultural. De ella dependen, en su mayor parte, el prestigio y expansión de la norma culta que ella misma contribuye a formar (Rabanales, 1992: pp. 264-5).

Vale destacar que la reasignación de los porcentajes de acuerdo con la longitud discursiva tiene su origen en la distribución que Rosenblat hiciera para la publicación de las entrevistas de un solo informante en los materiales de *El habla culta de Caracas* (Rosenblat y Bentivoglio, 1979: p. 10). En una muestra limitada como la del libro, ciertamente, era imposible establecer una diferencia entre 25% y 30% de hablantes en los grupos 1 y 3; en este caso, Rosenblat tomó la decisión de incluir la misma cantidad de informantes en estos dos grupos y marcó la diferencia porcentual con el tiempo transcrito. De esta forma, Rosenblat previó que la preeminencia sociocomunicativa de este grupo etario culto no depende exactamente del número de hablantes en este rango de edad sino, sustancialmente, de una mayor cantidad de discurso, ya que este grupo suele estar expuesto con más frecuencia a contextos comunicativos y de interacción más amplios.

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. cuadro 5, *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 el cuadro 4, se puede observar la distribución del nuevo corpus:

Cuadro 4. Distribución en hablantes y minutos del HCC/CD 73-11

Período de grabación		GRUPO ETARIO Grupo 1 20 a 35 años hombres	mujeres	Grupo 2 36 a 55 años hombres	mujeres	Grupo 3 56 años o más hombres	mujeres	Total ∑	*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 30% = 360		12 45% = 540		12 25% = 300		**36**	1200

En la submuestra correspondiente al período grabado entre 1973 y 1975, se incluyeron 14 entrevistas publicadas en Rosenblat y Bentivoglio (1979: pp. 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 ampliar las seis entrevistas que conformaban los materiales del segundo período se seleccionaron 12 transcripciones de hablantes pertenecientes al grupo con grado 3 de instrucción del corpus PRESEEA-CSC 04-10.

Una vez seleccionadas las entrevistas, 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 cuadro 5:

Cuadro 5. Postcodificación y distribución de encuestas del HCC/CD 73-11

Período de grabación	GRUPO ETARIO Grupo 1 20 a 35 años hombres	mujeres	Grupo 2 36 a 55 años hombres	mujeres	Grupo 3 56 años o más 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.05	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.

4.2. *Habla de jóvenes universitarios caraqueños 1977-1987-2006. Corpus diacrónico* (HJUC/CD 77-87-06)

Entre los años 2004 y 2006, el equipo de investigación de PRESEEA-Caracas culminó las grabaciones de los 12 hablantes correspondientes al grado 3 de instrucción universitaria del grupo etario 1 (20 a 34 años). Inmediatamente, el equipo del Departamento de Dialectología del IFAB se propuso diseñar una muestra que permitiera hacer estudios comparativos con los otros dos corpus sociolingüísticos del habla caraqueña (CSC 77 y CSC 87). De esta forma, la selección de hablantes estuvo signada por dos características sociales, una inherente (edad) y otra adquirida (nivel educativo). No se dispone del espacio para reseñarlos todos, pero un simple arqueo en un buscador informático común permite comprobar la existencia de abundantes estudios sobre el habla juvenil y de universitarios que se han llevado a cabo desde la hipótesis de que el habla de este segmento etario puede presentar rasgos peculiares y distintivos. Asimismo, la experiencia previa de corpus similares en otras variedades del español validó esta iniciativa.[70]

Particularmente, ya se he expuesto en una publicación anterior las razones que motivan el uso de un corpus de habla de jóvenes universitarios: "representan, sin duda, un grupo de gran interés para el estudio de la variación genolectal; sobre todo, porque su habla se ve afectada por la aspiración a la identidad y solidaridad grupal" (Guirado, 2011: p. 61). Este deseo se ha fortalecido en los últimos cincuenta años por las circunstancias históricas,[71] hecho que justifica que toda disciplina intente, especulativa o empíricamente, dar una mirada retrospectiva sobre los hechos que caracterizan al joven profesional como una categoría de análisis.[72] En este caso, se habla de un corpus que comprende muestras orales recogidas en el lapso de treinta años y que permite su análisis diacrónico en tres períodos específicos.

La muestra que sirvió de base para la selección fue la del PRESEEA-CSC 04-10. Lo primero que se hizo fue seleccionar 12 entrevistas del CSC 77 y 12 del CSC 87, de

---

[70] El habla juvenil de Valdivia (Cepeda y Barrientos, 1989); Corpus Oral de la Variedad Juvenil Universitaria del Español de Alicante (COVJA), subcorpus del ALCORE (Azorín 2002 y Azorín y Jiménez Ruiz, 1997); El habla culta de la generación joven de San Juan, La Habana y Santo Domingo (Reyes Benítez, 2001); Corpus de habla de los universitarios salmantinos (CHUS), dirigido por Julio Borrego Nieto y Carmen Fernández Juncal, en fase de elaboración por el Departamento de Lengua Española de la Universidad de Salamanca; Corpus Oral del Lenguaje Adolescente COLA (jóvenes de Madrid, Santiago de Chile, Buenos Aires, Guatemala, La Habana), coordinado por Annette Myre Jorgensen (Universidad de Bergen) (http://www.colam.org/). Para más detalles de cada uno de los corpus, Cf. Briz Gómez y Abelda Marco (2009).

[71] "Una época particularmente significativa en la historia del siglo XX fue la década de los sesenta [...] Aunque los cambios sociológicos producidos afectaron a la población en su totalidad, fueron sin duda los jóvenes los principales beneficiarios así como sus máximos protagonistas, un protagonismo sin precedentes en la historia del movimiento juvenil, que hizo que se hablara de los jóvenes de los sesenta [...] Con estos y otros ingredientes la juventud pasó a considerarse casi como una nueva clase o estamento social, que tomó conciencia de sí misma y de su poder rompiendo con la atonía de épocas pretéritas. Tal vez el mayor exponente fue la fuerza que en ese momento cobró la contestación estudiantil, especialmente universitaria" (Rodríguez González, 2002: p. 29).

[72] Camacho (2011) habla de *edad social*: "que se entiende, desde la sociología, como el papel que cumplen en la sociedad las personas según condiciones típicas asociadas a la edad" (Camacho, 2011: p. 5, *nota* 4).

hablantes con características similares (con estudios universitarios –concluidos o no– dentro del mismo rango de edad y residenciados exclusivamente en el Distrito Metropolitano).[73] En virtud de que estos dos corpus están estratificados con criterios diferentes al corpus PRESEEA, fue necesario buscar los datos relativos al nivel de instrucción en la ficha descriptiva del hablante almacenada en la base de datos original.[74]

En esta oportunidad, fue posible contar con datos y cifras más específicas sobre el nivel educativo de la población caraqueña, con los cuales se intentó reconstruir la estadística necesaria para definir el universo relativo en términos cuantitativos, a saber: la población entre 20 y 34 años, con estudios universitarios, oriundos de la ciudad y residentes la mayor parte de su vida en ella, en tres periodos diferentes (1977, 1987 y 2006). Se adoptó el *Censo Nacional de Población y Vivienda* inmediatamente anterior a la fecha de construcción de cada corpus como fuente para la estimación de la población a la cual hace referencia cada uno. Esta opción no resultó efectiva en todos los casos debido a la especificidad de la información requerida. Por ello, cuando fue útil y esclarecedor, se acudió a datos extraídos de trabajos de investigación de expertos en la materia.

A continuación se presenta un cuadro con la información sobre el universo de estudio correspondiente a los años 2001 y 2011 con la cual fue posible establecer un porcentaje referencial de esta misma población en años anteriores:[75]

Cuadro 6. Relación porcentual de la población universitaria joven de Caracas en dos censos

Año del censo	Población de Caracas (Dtto. Metropolitano)[50]	Población con estudios universitarios (Entre 20 y 34 años)[51]
2001	2.762.759	144.958 (5%)
2011	2.904.376	198.887 (7%)[52]

En el cuadro 6 se puede observar la delimitación del universo de habitantes y los porcentajes de universitarios respecto a la totalidad de la población. Los datos aportados por el censo de 2001 se usan como referencia para la muestra procedente de PRESEEA-CSC

[73] Se excluyeron las entrevistas de hablantes residenciados en las zonas de influencia —conocidas como *ciudades piloto* o *dormitorio*— de la Gran Caracas: Guarenas-Guatire, Altos Mirandinos y Valles del Tuy.

[74] Vale la pena aclarar que dicha base de datos consiste en un registro manual de la información en una planilla mimeografiada diseñada para tales fines. Esto ha permitido caer en cuenta de la necesidad de digitalizar esta información en un futuro, de modo que sea posible su procesamiento informatizado. Asimismo, se ha proyectado trasladar parte de esta información a las transcripciones en un formato similar a la cabecera de las transcripciones del proyecto PRESEEA.

[75] 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.

04-10. Adicionalmente, se tomó el 4% del total de la población del área metropolitana según los censos 1971 y 1981 –base para la muestra de los otros dos corpus (CSC 77 y CSC 87)– como referencia para la estimación de la población joven universitaria de ambos períodos:[76]

Cuadro 7. Estimación de la proporción de la población objeto de estudio por hablante entrevistado

Corpus y año	Población de Caracas (Dtto. Metropolitano)	Población con estudios universitarios (Entre 20 y 34 años)	Proporción por entrevistado (*f=población/hablantes*)
PRESEEA 04-06	2.762.759 (Censo 2001)	144.958 (Censo 2001)	1/12.500 → 150.000
CSC'87	2.685.901 (Censo 1981)	107.436 (estimado)	1/10.000 → 120.000
CSC'77	2.583.396 (Censo1971)	103.336 (estimado)	1/10.000 → 120.000

En el cuadro 7, se presentan varios datos por columna: i. la relación corpus-año de grabación de las muestras; ii. la población de la región y la fuente; iii. el número de universitarios entre 20 y 34 años según censo 2001 y la estimación para los años anteriores; y, finalmente, vi. el cálculo de la proporción de habitantes jóvenes caraqueños por hablante (*f*=población/12 hablantes de cada muestra).[77]

De acuerdo con la metodología del Proyecto PRESEEA, la muestra está formada por 6 cuotas o casillas con afijación uniforme, creadas a partir del cruce de las variables sexo y períodos de grabación (2 x 3 = 6). De esta forma, los 12 hablantes de cada corpus se

[76] Aunque este porcentaje se deriva de un dato no disponible, su uso se apoya con otras cifras sobre el incremento discreto de esta población en el ámbito nacional durante los períodos anteriores a la muestra más reciente. Centeno de Figueroa (1990: p. 87) destaca que: "de una matrícula de de 85.675 estudiantes en nivel superior para 1970-71, se pasó a 298.884 en 1979-80, lo cual supone un incremento de hasta 3,5 veces en esos diez años". Adicionalmente, la autora también advierte que esta demanda social no se concentra en las instituciones del área metropolitana sino en Los Andes-ULA (50%) en 1980/81, en la región Centro-norte costera-Universidad de Carabobo (61,05%) en 1985/86 y en la región Zuliana que absorbe casi el 100% de su población (93). Asimismo, la investigadora reporta que 583.494 personas en todo el ámbito nacional declararon tener nivel educativo superior en el Censo de 1981 (100). Morillo Moreno (2007: web) también ofrece un dato importante sobre el incremento moderado del nivel educativo de la población activa durante el período 1994-2004: "lo mismo sucedió pero con menor rapidez con el nivel Universitario, al pasar durante el mismo período del 9,9% al 11,4%".

[77] Adicionalmente, en el Censo 2011, a través de la opción de "cruce" —que concede la oportunidad de introducir una tercera variable control— fue posible delimitar la búsqueda a los habitantes caraqueños nacidos en la entidad (*grupo de edad-nivel educativo-entidad de nacimiento*). Asimismo, se pudo aplicar el filtro *jóvenes*, preestablecido en el programa de cálculo. Según estos datos específicos, en el 2011, el total de la población oriunda de la ciudad con estudios universitarios se reduce de 198.887 a 88.270 jóvenes (∇56%). Esto implica que, en la actualidad, las entrevistas de 19 hablantes sería una proporción representativa de esta población, incluso en términos labovianos (0,025%). Desafortunadamente, la opción de "cruce" que permite la introducción de la variable control *entidad de nacimiento* no está disponible en el Censo 2001.Tampoco es posible filtrar la información con ningún indicador. No obstante, esta información permite presuponer que una reducción similar debe ocurrir en las poblaciones de universitarios de años anteriores, por lo cual es muy probable que 12 hablantes por período constituya una muestra representativa del universo objeto de estudio en esos años.

distribuyen equitativamente en dos cuotas fijas de 6 por casilla para un corpus total de 36 entrevistas. En el cuadro 8 se visualiza la distribución del nuevo corpus que se ha denominado *Habla de jóvenes universitarios caraqueños 1977-1987-2006. Corpus diacrónico* (HJUC/CD 77-87-06), para hacer referencia a su propósito especial y al período de grabación, en este caso, expresado por año (en el caso de las encuestas procedentes de PRESEEA-CSC 04-10, se ha usado 2006 como año de grabación de las entrevistas utilizadas más próximo a la investigación):

Cuadro 8. Distribución de los hablantes del HJUC/CD 77-87-06

Período de grabación	SEXO		Total
	hombres	mujeres	
HJUC-1977	6	6	12
HJUC-1987	6	6	12
HJUC-2006	6	6	12
Total	12	12	36

Posteriormente, la información adicional de los hablantes, obtenida de la ficha de identificación original, permitió postestratificar el corpus de acuerdo con dos rasgos específicos: edad y zona de procedencia.

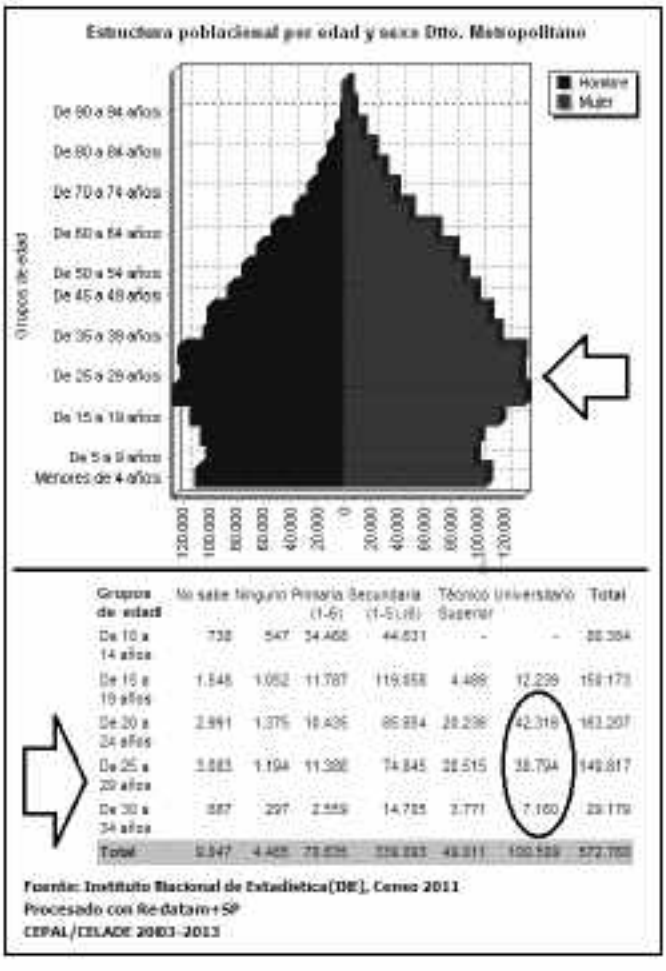
Por una parte, el dato sobre el domicilio del hablante permitió establecer una división socioterritorial como unidad de análisis geohistórica. El área metropolitana de la ciudad de Caracas es una unidad político-territorial que integra cinco municipios: Libertador del Distrito Capital y los municipios Baruta, Chacao, El Hatillo y Sucre del estado Miranda. En la figura 1 se puede apreciar la distribución de los municipios en el mapa del área metropolitana:

Figura 1. Mapa de la zona metropolitana de Caracas

Históricamente la ciudad se ha dividido en dos zonas, oeste y este. El municipio Libertador incluye todas las zonas del oeste, mientras que los otros cuatro municipios están ubicados al este. Esta división de la ciudad ha estado relacionada, como en muchas otras ciudades, con una diferenciación social. En el oeste están las parroquias fundadoras, alrededor de las cuales se han desarrollado grandes zonas de miseria y pobreza (Catia, Caricuao, La Candelaria, El Valle, San Agustín, etc.); los poderes públicos y el comercio formal e informal también se encuentran en esta parte de la ciudad. En el este se desarrollaron los complejos residenciales de la clase media en la década de 1980 y fue el asiento de las grandes urbanizaciones de la clase alta en la ciudad; asimismo, la actividad empresarial y de negocios, los grandes centros comerciales y las telecomunicaciones están altamente asociada a la zona. De esta forma, es posible creer que estas diferencias, arraigadas profundamente en la memoria colectiva del caraqueño, se traduzcan en conductas lingüísticas diferenciadas.

Por otra parte, con el detalle de la edad en años cumplidos, declarada por el hablante en el momento de la grabación, fue posible reclasificar las muestras dentro de nuevos rangos. Para ello, se tomó como referencia, nuevamente, los datos del Censo 2011. Véase la figura 2:

Figura 2. Distribución de las variables por grupos de edades en el Censo 2011



Grupos de edad	No sabe ninguno	Primaria (1-6)	Secundaria (1-5 Ud)	Técnico Superior	Universitario	Total
De 10 a 14 años	738	547	34.468	44.631	-	-	80.384
De 15 a 19 años	1.548	1.052	11.787	119.658	4.489	12.239	150.173
De 20 a 24 años	2.991	1.375	10.426	86.884	20.236	42.316	163.207
De 25 a 29 años	3.883	1.194	11.382	74.845	22.515	35.794	149.817
De 30 a 34 años	887	297	2.559	14.755	3.771	7.160	29.179
Total	9.947	4.465	70.626	340.693	49.011	100.509	572.760

Fuente: Instituto Nacional de Estadística(INE), Censo 2011
Procesado con Redatam+SP
CEPAL/CELADE 2003-2013

En la pirámide de edad de la población de la zona metropolitana de Caracas (parte superior de la figura 2) se verifica que la mayor densidad poblacional se concentra en el rango de edades comprendidas entre los 20 y los 34 años. Asimismo, en la parte inferior de la figura, se observa la distribución de la población joven universitaria oriunda de los cincos municipios caraqueños. Nótese que el programa, al introducirse el filtro *jóvenes*, limita el conteo de personas un par de años antes de los 34 años, de modo que las estadísticas que sustentan la muestra están construidas dentro de estos parámetros.

En la realidad, la edad constituye un continuum que el investigador segmenta con la intención de verificar de qué forma la lengua es reflejo de los procesos del cambio psíquico y cultural que experimenta un sujeto social. En tal sentido, se proponen dos nuevos rangos: jóvenes entre 20 y 25 años y jóvenes entre 26 y 33 años (20 y 33 son límites reales en la muestra), a partir del supuesto de que hay una diferencia respecto a la conciencia que se tiene sobre la construcción de la identidad, de la imagen social, entre un joven veinteañero y uno que ya superó los 30 años: "Pregunta social: no solo por las características de una edad, que importa básicamente a quienes la atraviesan. Es la sociedad que trata de saber

cómo comienza su futuro. Cuántos torneros o médicos va a haber, cuántos con educación universitaria" (García Canclini, 2004: p. 167).

La variable geohistórica está distribuida equitativamente en las tres submuestras, no así la variable edad, ya que la muestra de hablantes más reciente tiene 4 mujeres/2 hombres en el rango de 20 a 25 años, mientras que en el de 26 a 33 años tiene 2 mujeres/4 hombres. De esta forma, los universitarios cuyas grabaciones constituyen la muestra analizada, están distribuidos equitativamente según sexo (18 mujeres y 18 hombres) y municipio de procedencia (habitante del municipio Libertador y habitante de otros municipio metropolitanos), y clasificados de acuerdo a dos rangos de edad (20 a 25 años y 26 a 33 años). En el cuadro 9 se observa la distribución de las entrevistas según la postestratificación:

Cuadro 9. Distribución de la post-estratificación de los hablantes del HJUC/CD 77-87-06

Periodo de grabación	Libertador				Otros municipios				Total
	hombres		mujeres		hombres		mujeres		
	20-25	26-33	20-25	26-33	20-25	26-33	20-25	26-33	
HJUC-1977	2	1	1	2	1	2	2	1	12
HJUC-1987	1	2	2	1	2	1	1	2	12
HJUC-2006	1	2	2	1	1	2	2	1	12
Total	9		9		9		9		36
	18				18				

La selección de encuestas se postcodificaron conforme a dígitos alfanuméricos con los que se abstrae la nueva arquitectura de los materiales. La identificación de las encuestas del HJUC/CD 77-87-06 se puede observar en el cuadro 10:

Cuadro 10. Postcodificación y distribución de encuestas del HJUC/CD 77-87-06

Periodo de grabación	Zona	Sexo	
		hombres	*mujeres*
HJUC-1977	*oeste*	UCHL1-A.77	UCML1-A.77
		UCHL1-B.77	UCML2-B.77
		UCHL2-C.77	UCML2-C.77
	este	UCHO1-D.77	UCMO1-D.77
		UCHO2-E.77	UCMO1-E.77
		UCHO2-F.77	UCMO2-F.77
HJUC-1987	*oeste*	UCHL1-A.87	UCML1-A.87
		UCHL2-B.87	UCML1-B.87
		UCHL2-C.87	UCML2-C.87
	este	UCHO1-D.87	UCMO1-D.87
		UCHO1-E.87	UCMO2-E.87
		UCHO2-F.87	UCMO2-F.87
HJUC-2006	*oeste*	UCHL1-A.05	UCML1-A.04
		UCHL2-B.04	UCML1-B.04
		UCHL2-C.05	UCML2-C.04
	este	UCHO1-D.05	UCMO1-D.05
		UCHO2-E.05	UCMO1-E.05
		UCHO2-F.06	UCMO2-F.04

Con los códigos nuevos se identifica la variedad de habla del corpus (U) y la ciudad (C), sexo (H/M), el municipio de residencia del hablante (L/O) y grupo de edad (1/2), la letra que lo distingue en el grupo (A/B/C) y el año de grabación (77/87/04-06).

Asimismo, debido a la diferencia establecida en cada corpus para el tiempo de grabación de las entrevista, la extensión de las mismas se unificó a los primeros treinta minutos de grabación. El total de palabras es de 172.264, distribuido en una proporción ascendente de acuerdo con el aumento de la población para cada periodo (1977 = 30%; 1987 = 34%; 2005 = 36%); en tal sentido, el promedio por hablante dependerá del año de grabación (1977 = 4305; 1987 = 4915; 2005 = 5136).

5. DIGITALIZACIÓN Y ALMACENAMIENTO DE LOS MATERIALES

El audio y las transcripciones de las grabaciones de los corpus descritos están disponibles en el Departamento de Dialectología del IFAB. A continuación, se presenta un cuadro resumen de las principales características de los corpus:

Cuadro 11. Resumen de las características de los corpus descritos

Corpus	Nº de grabaciones	Estratificación	Nº de palabras	Formato transcripciones	Contactos responsables
CHCC 68-77	232	- Registro - Edad - Sexo	Por calcular	232 Mecanografiadas 26 digitalizadas *.doc; *.txt; *.pdf	Paola Bentivoglio Kristel Guirado
CSC 77	70	- Nivel socioeconómico - Edad - Sexo	285.916	Digitalizadas *.doc; *.txt	Paola Bentivoglio Kristel Guirado
CSC 87	160	- Nivel socioeconómico - Edad - Sexo	767.868	Digitalizadas *.doc; *.txt; *.pdf; *.html	Paola Bentivoglio Mercedes Sedano Kristel Guirado
PRESEEA CSC 04-11	108	- Grado de instrucción - Edad - Sexo	Por calcular	Digitalizadas *.doc; *.txt	Paola Bentivoglio Irania Malaver
HCC CD 73-09	36	- Periodo de grabación - Edad - Sexo	Por calcular	Digitalizadas *.doc; *.txt; *.pdf; *.html	Kristel Guirado
HJUC CD 77-87-06	36	- Periodo de grabación - Sexo - Edad - Zona de procedencia	172.264	Digitalizadas *.doc; *.txt; *.pdf; *.html	Kristel Guirado

6. CONSIDERACIONES FINALES

Al principio de este artículo se propuso acuñar el término *Reingeniería de Corpus* para designar las tareas relacionadas con la reconfiguración de materiales de habla (orales y escritos) bajo el supuesto de que el diseño de un corpus lingüístico entraña el estudio de los fenómenos del lenguaje a partir de la comparación y el contraste de sus características distintivas con otras muestras de naturaleza diversa.

Desde esta perspectiva, la actividad de reingeniería supone conocer los procesos objetivos y subjetivos que confluyeron en la construcción del corpus. La intervención en la arquitectura original requiere de una revisión fundamental para generar cambios que permitan ofrecer un producto útil, rentable y disponible a la comunidad de usuarios. Por ejemplo, un corpus de extensión media, que garantice el análisis exploratorio de textos reales y cuya representatividad no arriesgue la proyección de algunas tendencias de uso de una variedad de lengua en el tiempo, representa un producto de interés y utilidad, bien para los estudiantes de pregrado –quienes se inician en el aprendizaje de la investigación– o para el profesor que necesita ejemplificar el estudio de un fenómeno en el corto lapso de un curso de postgrado.

Con respecto a la experiencia específica de Reingeniería de Corpus descrita en este artículo, cabe puntualizar algunas de las acciones ejecutadas:

i. Definir el o 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 implicó buscar los datos o –lo que suele suceder– reconstruir la estadística; definir el método para establecer la relación proporcional de hablantes u otros criterios de representatividad, como el tamaño de las entrevistas; la re-estratificación; el diseño la distribución, la afijación por casilla, el criterios de longitud, etc.
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.

Los criterios para la construcción de los dos corpus productos de la labor de reingeniería no estuvo exenta de los mismos dilemas metodológicos propios de la LC; aunque se tuvo la ventaja de que las muestras proceden de textos reales y todas se encuentran disponibles en formato digital –o son susceptible de ser informatizadas– todavía hubo que enfrentarse a la complejidad que implica decidir los criterios de selección, representatividad y tamaño. Tal y como afirma Parodi (2008):

> En este sentido, en lingüística, el universo de estudio (en el giro técnico) no es en muchas investigaciones fácilmente determinable ni calculable, por ende tampoco lo es la población o muestra estadísticamente representativa que de él se desprende. Por ejemplo, esto se aplica al trabajo con los corpus orales correspondientes, digamos, a una ciudad, cuyo universo no resulta del todo fácil de estimar. Es muy cierto que se podría determinar el tipo y cantidad de hablantes por estratos específicos, pero otra cosa es decidir el tamaño de cada entrevista, de cada grabación o de cada muestra textual. En otras palabras: ¿cuántas horas de entrevistas son necesarias para alcanzar la representatividad estadística del discurso oral en un registro específico de los hablantes de una ciudad cualquiera? Ciertamente es un asunto de complejidades. Algunos podrían decir que no existe límite. Otros pueden sostener que se deben hacer opciones y definir claramente los parámetros, variedades y estratos a abordar [...] Otra opción es que, más bien, se busque una proporcionalidad adecuada del corpus y que ello conduzca a solo ciertas proyecciones. Por supuesto que no será posible realizar generalizaciones, como desde otros modelos estadísticos inferenciales (pp. 105-6).

Si bien no siempre fue posible zanjar todos los escollos, los corpus que presentamos constituyen muestras útiles 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

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

Azorín, D. (coord.). (2002). *Alicante corpus oral del español*. Alicante: Universidad de Alicante. CD-rom.

Azorín, D. y J. L. Jiménez Ruiz. (1997). *Corpus oral de la variedad juvenil universitaria del español hablado en Alicante*. Alicante: Instituto de Cultura Juan Gil-Albert.

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

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

Bentivoglio, P. (1998). La variación sociofonológica. *Español actual*, (69), 29-42.

Bentivoglio, P. e I. Malaver. (2006). La lingüística de corpus en Venezuela: un nuevo proyecto. *Lingua Americana*, (19), 37-46.

Bentivoglio, P. e I. Malaver. (2012). Corpus sociolingüístico de Caracas: *Preseea* Caracas 2004-2010. Hablantes de instrucción superior. *Boletín de Lingüística*, (24), 37-38.

Bentivoglio, P. e I. Malaver. (2014). Corpus sociolingüístico *Preseea* Caracas 2004-2010. Caracas: UCV-FHE-IFAB. CD-rom.

Bentivoglio, P. y M. Sedano. (1993). "Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana". *Boletín de Lingüística* (8), 3-35.

Bentivoglio, P., K. Guirado e I. Malaver. (2014). Marcadores del discurso de Caracas. En A. Valencia Espinoza (coord.). *Marcadores discursivos en la norma culta hispánica: 1964-2014. Cuadernos del ALFAL*, (5), 43-68. Disponible en http://www.dialogoseducativos.cl/revista/papeldigital/.

Briz Gómez, A. y M. Albelda Marco. (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D. *El español en el mundo. Anuario del Instituto Cervantes*, 165-226. <http://cvc.cervantes.es/lengua/anuario/anuario_09/briz_albeida/p03.htm>.

Camacho, M. (2011). *Análisis pragmático de los apelativos empleados por jóvenes universitarios en el español de Costa Rica*. Tesis de Maestría. Universidad de Costa Rica, Costa Rica.

Centeno de Figueroa, A. B. (1990). Perfil socioeconómico de los recursos humanos de nivel educativo superior en Venezuela: Década de los 80. *Informe de Investigaciones Educativas* (IV, 1-2), 83-104.

Cepeda, G. y A. Barrientos. (1989). *El habla juvenil de Valdivia. Entrevistas informales*. Valdivia: Central de Publicaciones de la Universidad Austral de Chile.

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

De Stefano, L. y L. 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, M. (2005). Crecimiento de la población y proceso de urbanización en el Distrito Metropolitano de Caracas: efectos ambientales. En A. Freitez, M. Di Brienza, G. Zúñiga, R. Carvallo, M. Phélan y T. García (eds.). *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.

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

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

García, N. (2004). *Diferentes, desiguales y desconectados. Mapas de la interculturalidad*. Barcelona: Gedisa.

Guirado, K. (2011). *Allá yo vivía pa' estudiar*: Un estudio de variación en el habla de jóvenes universitarios caraqueños. *Boletín de Lingüística* (35-36), 57-80. En línea: http://saber.ucv.ve/ojs/index.php/rev_bl/article/view/2682/2572.

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

Granda, G. de (1994). Observaciones metodologicas sobre la investigacion sociolingüistica en hispanoamerica. *Lexis* (XVIII, 2), 197-210.

Herrera, J. (1994-95). *Estudio sociolingüístico de los relativos en el español de Santa Cruz de Tenerife*. Tesis doctoral. Universidad de la Laguna, España. <ftp://tesis.bbtk.ull.es/ccssyhum/cs18.pdf>.

Instituto Nacional de Estadística. *INE. Instituto Nacional de Estadística*. <http://www.ine.gov.ve/>.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lope, J. (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, 163-67. Santo Domingo: Asociación de Lingüística y Filología de América Latina.

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

McEnery, T. y A. Wilson (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

McEnery, T., R. Xiao y Y. Tono (2006). *Corpus-Based Language Studies. An advanced resource book*. London y New York: Routledge.

Moreno, F. (1996). Metodología del 'Proyecto para el estudio sociolingüístico del Español de España y de América' (PRESEEA). *Lingüística* (8), 257-287.

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

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

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

Morill, M. (2007). El sistema educativo y el trabajo en Venezuela. *Saber* (19, 2). <http://ojs.udo.edu.ve/index.php/saber/issue/view/15>.

Osuna, Z. (2007). *Boletín de indicadores de Educación Superior 1. 2000-2005*. Caracas: MPPES. CNU. OPSU.

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *Revista de lingüística teórica y aplicada* (46, 1), 93-119.

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

Pérez, C. y A. M. Ortiz. (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*, 67-96. Gijón: TREA.

Pérez, F. (1997). Instituto de Filología Andrés Bello de la UCV. 50 Aniversario en homenaje a Ángel Rosenblat. *El ucabista* (22). <http://w2.ucab.edu.ve/tl_files/sala_de_prensa/recursos/ucabista/nov97/10.htm>.

Ponce, M. (2010). La diversidad de la pobreza en Venezuela: desarrollo urbano, educación y trabajo. 2003-2005. *Revista Venezolana de Análisis de Coyuntura* (XVI, 1), 77-109.

Proyecto para el estudio sociolingüístico del Español de España y de América (PRESEEA). *PRESEEA. Proyecto para el estudio sociolingüístico del Español de España y de América*. <http://preseea.linguas.net/>.

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.

Real Academia Española. *Real Academia Española. Banco de datos. CREA*. <http://www.rae.es/recursos/banco-de-datos/crea>.

Registro Nacional Voz de los Creadores. *Catálogo I. Antropología y Lingüística. Paola Bentivoglio*. Fundación Casa del Artista: Caracas.

Reyes, I. (ed.) (2001). *El habla culta de la generación joven de La Habana, Cuba. Materiales para su estudio*. Tomos I-III. San Juan: Universidad de Puerto Rico.

Rodríguez, F. (2002). Lenguaje y contracultura juvenil: anatomía de una generación. En F. Rodríguez G. (coord.). *El lenguaje de los jóvenes*, 29-56. Barcelona: Ariel.

Rosenblat, Á. (2002). *El español de América*. Caracas: Fundación Ayacucho.

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

Samper, J., C. E. Hernández Padilla y M. Troya Déniz (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Las Palmas de Gran Canaria: Universidad de Las Palmas de Gran Canaria. Disco compacto.

Sánchez, A. (1995). Definición e historia de los corpus. En A. Sánchez, R. Sarmiento, P. Cantos y J. S. (eds.). *CUMBRE-Corpus Lingüístico de Español Contemporáneo: fundamentos, metodología y aplicaciones*, 7-24. Madrid: SGEL.

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

Sedano, M., A. Bolívar y M. Shiro (comp.) (2006). *Haciendo lingüística. Homenaje a Paola Bentivoglio*. Caracas: Comisión de Estudios de Postgrado. Facultad de Humanidades y Educación. Universidad Central de Venezuela.

Shiro, M. (1998). *Los pequeños cuentacuentos. El desarrollo de las habilidades narrativas de niños en edad escolar*. Trabajo de ascenso. Universidad Central de Venezuela, Venezuela.

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

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

Sinclair, J. (2005). Corpus and text - basic principles". En M. Wynne (ed.). *Developing linguistic corpora: A guide to good practice*, 1-16. Oxford: Oxbow Books. <http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>.

Tejera, M. (dir.) (1993). *Diccionario de Venezolanismos* I, II y III. Caracas: Universidad Central de Venezuela, Academia Venezolana de la Lengua, Fundación Edmundo y Hilde Scnoegass.

Tejera, M. y L. de Stefano (2006). *Documentos para la historia del español de Venezuela - Siglos XVI-XVIII*. Caracas: Fondo Editorial de Humanidades y Educación, Universidad Central de Venezuela. Disco compacto.

Torruella, J. y J. Llisterri (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*, 45-77. Barcelona: Milenio. <http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf>.

Valencia, A. (coord.) (2014). *Marcadores discursivos en la norma culta hispánica: 1964-2014. Cuadernos del ALFAL* 5. <http://www.dialogoseducativos.cl/revista/papeldigital/>.