



# La lingüística de corpus en Venezuela: Un nuevo proyecto

*Paola Bentivoglio e Irania Malaver*

*Universidad Central de Venezuela*

*pbentivo@reacciun.ve / imalaver@hotmail.com*

## Resumen

Se presenta el proyecto *Corpus Sociolingüístico de Caracas 2005* (UCV), que forma parte del proyecto sociolingüístico panhispánico *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA). La lingüística venezolana del siglo XX posee una sólida tradición en la recolección de corpus de habla estratificados socialmente en varias comunidades urbanas: Caracas, Cumaná, Maracaibo, Mérida, Puerto Cabello y Valencia. La recolección de un nuevo corpus del habla caraqueña tiene como finalidad, por una parte, continuar estudiando sincrónicamente el habla caraqueña, pero también empezar a realizar análisis diacrónicos, comparando los resultados de las investigaciones anteriores con los que arrojarán las investigaciones sobre la nueva muestra. Se expone el diseño del proyecto, y la distribución de los hablantes según las variables sociodemográficas (sexo, edad y nivel instruccional), así como según la densidad poblacional de los cinco municipios en los que está dividida la ciudad de Caracas.

**Palabras clave:** Español hablado, Venezuela, corpus lingüístico.

## Corpus Linguistics: A New Project

### Abstract

We present the project *Corpus Sociolingüístico de Caracas 2005* (UCV), which is part of the sociolinguistic panhispanic *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA).

Twentieth century Venezuelan linguistics has a solid tradition in the collection of socially stratified speech corpus in several urban communities: Caracas, Cumaná, Maracaibo, Mérida, Puerto Cabello, and Valencia. The collection of a new corpus of Caracas speech is aimed at continuing to study Caracas speech synchronically and also diachronically, comparing new results with those of previous research. We present the design of the project and speakers' distribution according to sociodemographic variables (sex, age and educational level), as well as population density in the five counties that make up the city of Caracas.

**Key words:** Spoken Spanish, Venezuela, linguistic corpus.

## 1. Introducción

En Venezuela, desde hace casi cuarenta años, se han grabado grandes corpus de habla urbana: la primera experiencia –entre 1969 y 1976– se inscribe dentro del macroproyecto conocido como *Estudio coordinado de la norma culta de las principales ciudades de Iberoamérica y de la Península Ibérica* (Lope Blanch 1986). Los corpus siguientes fueron construidos de acuerdo con parámetros sociolingüísticos, con el fin de permitir estudios de variación lingüística, según el modelo del lingüista norteamericano William Labov. Al primer corpus sociolingüístico, el de Caracas (grabado en 1976-77) se añadieron, al poco tiempo, el de Maracaibo (1986), bajo la coordinación de las profesoras Bertha Chela-Flores y Jeannette Gelman, recientemente digitalizado y puesto a disposición de los investigadores por los profesores Carlos y Rosa Sánchez; Caracas 3, grabado en 1987-88 (ver Bentivoglio y Sedano 1993); Mérida, grabado entre 1990 y 1994 (ver Domínguez y Mora 1998); y Cumaná, grabado entre 1993 y 1996, y coordinado por Árida Velásquez. Los cinco corpus están estratificados de acuerdo con los mismos parámetros sociales, lo que permite la comparación de los datos de los análisis lingüísticos que se hagan. Un ejemplo es el estudio sobre la variación entre *para* y *pa'* en el habla de tres ciudades (Guirado 2006).

A los corpus mencionados hay que añadir otros dos, muy importantes porque reflejan el habla de la región central (Valencia y Puerto Cabello), grabados bajo la coordinación del profesor Manuel Navarro, de la Universidad de Carabobo, quien –lamentablemente para nosotros y quizá afortunadamente para él– no está hoy aquí. Estos dos últimos corpus se distinguen de los cinco

mencionados previamente, ya que los parámetros utilizados para la estratificación social no son exactamente los mismos que los empleados para la estratificación de los corpus de Caracas, Cumaná, Maracaibo y Mérida. Esta diferencia no impide, sin embargo, que puedan hacerse comparaciones entre el habla de Valencia y de Puerto Cabello y la de las otras cuatro ciudades, como, en efecto, las ha hecho –y, esperamos, seguirá haciéndolo– el mismo profesor Navarro en muchos de sus estudios.

Los esfuerzos realizados por los investigadores para documentar el habla de algunas de las ciudades más importantes de Venezuela –lamentablemente, no de todas– constituyen un hito reconocido de la lingüística venezolana, pues pocos países disponen de grandes muestras de habla grabada según los mismos criterios en un arco breve de tiempo (1976-1996) y, además, transcritas y disponibles en formato digital. Del corpus de Caracas 1987-88 también se ha digitalizado el audio. Los cinco corpus sociolingüísticos de Caracas, Cumaná, Maracaibo y Mérida, representan cuatrocientas cincuenta y cuatro (454) horas de grabación, lo que, aproximadamente, corresponde a dos millones (2.000.000) de palabras. Esto sin contar las doscientas (200) horas del habla culta de Caracas y las ciento cuarenta (140) de los dos corpus grabados por el Prof. Navarro. Un tesoro que, hasta ahora, ha sido subexplotado por diferentes razones, una de las cuales es, evidentemente, la escasez de investigadores.

Esta breve descripción de los corpus existentes nos servirá de base para presentarles un nuevo proyecto, el del Corpus Sociolingüístico de Caracas 2006-2007, al que –dentro de nuestra ya establecida tradición– seguirán muy pronto, espero, corpus similares de otras partes de Venezuela. Esta nueva muestra tiene varias finalidades: la principal es obtener nuevas muestras sincrónicas del habla de Caracas, lo que permitirá, por una parte, continuar los estudios (fonético-fonológicos, morfosintácticos, léxico-semánticos y pragmático-discursivos) del habla caraqueña y, por la otra, hacer estudios diacrónicos para determinar los eventuales procesos de cambio lingüístico que se dan en español de Venezuela. El propósito último del proyecto es promover nuevas investigaciones acerca de fenómenos lingüísticos que se manifiestan en el español hablado hoy día, análisis que sería imposible realizar sin datos actualizados.

## 2. El proyecto panispánico PRESEEA

El *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA), coordinado por Francisco Moreno Fernández, de la Universidad de Alcalá de Henares (España), es un megaproyecto, en el que participan los equipos de investigación de distintas universidades de España y de América.<sup>1</sup> La 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). El mismo Moreno Fernández, al presentar el proyecto, reconoce que:

No existe en el mundo hispánico un proyecto de estudio sociolingüístico de intención semejante a la del “Proyecto para el estudio de la norma culta”. Precisamente este proyecto, impulsado y coordinado por Lope Blanch, es el antecedente más claro y cercano del proyecto que aquí presentamos, aunque las diferencias entre uno y otro son palmarias: PRESEEA no se va a limitar a la recogida y estudio de la norma culta sino que aspira a conocer las variedades urbanas del español con toda la profundidad que sea posible. El estudio sociolingüístico completo del español de España y de América deberá suponer un paso adelante en la investigación lingüística.

El proyecto mantiene una página web, cuya dirección es [www.linguas.net/preseea](http://www.linguas.net/preseea), en la que encontrarán toda la información necesaria para participar en él, o bien simplemente informarse acerca de lo que ya se ha hecho, de lo que se está haciendo, etc.

La necesidad de un proyecto de esta magnitud es incuestionable, pues consolida una importante tradición de investigaciones dialectológicas y sociolingüísticas españolas e hispanoameri-

1 De España participan los equipos de las universidades Alcalá de Henares, Barcelona, Cádiz, Granada, Las Palmas, Lérida, Madrid, Málaga, Palma Mallorca, Valencia, Zaragoza. De América, los equipos de Barranquilla, Bogotá y Pereira (Colombia); Caracas (Venezuela), Cipolletti (Argentina); Culiacán, Mérida, México D.F., Monterrey (México); Guatemala; San Juan de Puerto Rico; Quito (Ecuador), Miami (Estados Unidos).

canas, a partir de la conjunción de esfuerzos y experiencias en el campo de la sociolingüística. Las *actividades* propuestas son las siguientes:

1. La creación de un corpus sociolingüístico del español;
2. La creación del Servicio de Documentación Sociolingüística para Hispanoamérica y la Península Ibérica (lenguas española y portuguesa);
3. La creación de un corpus sociolingüístico del portugués (PRESOPO). Los materiales ya producidos por los distintos equipos del proyecto se pueden consultar en la página electrónica <http://www.linguas.net/preseea>.

*Los objetivos* generales atañen a:

- A. La descripción de los procesos lingüísticos más relevantes del español;
- B. La identificación de los fenómenos de variación que se corresponden con cambios lingüísticos en marcha;
- C. Estudios sobre la estructura discursiva.

*Los principios metodológicos* de PRESEEA son los siguientes:

- i. *Principio del dialecto*: “el dialecto se manifiesta en una comunidad”;
- ii. *Principio de la comunidad de habla*: “una comunidad de habla está formada por un conjunto de hablantes que comparten al menos una variedad lingüística, unas reglas de uso, una interpretación de ese uso, unas actitudes y una misma valoración de las formas lingüísticas”;
- iii. *Principio de la variación lingüística*: “los usos lingüísticos variables –fonéticos, gramaticales o léxico-semánticos– pueden covariar con otros elementos lingüísticos o extralingüísticos”;
- iv. *Principio de la cuantificación*: “para establecer relaciones entre variables lingüísticas y variables sociales se requieren métodos cuantitativos”;
- v. *Principio de la representatividad*: “el habla de una muestra debe ser representativa de un grupo social o de una comunidad”.

Las comunidades objeto de estudio son centros urbanos iberoamericanos que permitan la obtención de datos de habla real, susceptibles de ser estratificados socialmente.

### **2.1. El corpus sociolingüístico de Caracas 2006-2007**

La decisión de grabar un nuevo corpus (el cuarto, en menos de cuarenta años) ha sido difícil de tomar y de larga incubación para Paola Bentivoglio, quien confesaba abiertamente no tener más fuerzas para emprender una tarea de tal magnitud. Sólo 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 entrar a formar parte del megaproyecto. Para esto era también necesario contar con el apoyo económico del Consejo de Desarrollo Científico y Humanístico de nuestra Universidad, cosa que hemos logrado justamente hace un año.<sup>2</sup> El proyecto está adscrito académicamente al Área de Lingüística de la Comisión de Estudios de Postgrado de la Facultad de Humanidades y Educación y al Instituto de Filología “Andrés Bello” de la UCV.<sup>3</sup>

Para la constitución del corpus de habla, hemos adoptado –como es obvio– los principios metodológicos señalados por PRESEEA, convencidas de que, si bien el diseño responde a los lineamientos sociolingüísticos generales esbozados, esto no impide que se puedan emprender estudios con distintos enfoques y métodos. La nueva muestra, así como ya las anteriores, será un cor-

- 2 El proyecto está inscrito en el CDCH bajo el N° 07-00-5760-2004, de fecha de 16.05.2005. Está previsto para ser realizado en dos etapas, de un año cada una.
- 3 El equipo caraqueño está integrado por las siguientes investigadoras: Paola Bentivoglio, investigadora del Instituto de Filología “Andrés Bello” de la UCV; Coodinadora de la Maestría en Lingüística y, desde 2000, Representante del Área de Lingüística de la Facultad de Humanidades y Educación, es la investigadora responsable. Irania Malaver, profesora de la Escuela de Letras y, actualmente, doctoradanda en Lingüística Hispánica en la Universidad de Alcalá de Henares, y M<sup>a</sup> Alejandra Romero, investigadora del Instituto de Filología “Andrés Bello” son las coinvestigadoras. Las estudiantes de la Maestría en Lingüística, Haronid Blanco, M<sup>a</sup> José Gallucci (también investigadora del IFAB), Greisy Fernández, Zayra Marcano, Vanessa Pérez y Cristina Ruiz participan activamente en el proyecto como asistentes de investigación. Haronid Blanco supervisa las actividades de las asistentes de investigación. En los seminarios y cursos de la Maestría en Lingüística, otros estudiantes también han tenido a su cargo tareas de grabación y transcripción.

pus “abierto”, sin restricciones teórico-metodológicas para su empleo y que, en definitiva, constituirá una base de datos con la finalidad de “ofrecer paquetes preparados de una porción finita y abarcable de la realidad empírica que podrá servir a otros investigadores, con el consiguiente ahorro de energía y tiempo en la recolección de sus propios datos” (Caravedo, 1999: 97).

Creemos así que el nuevo corpus sociolingüístico permitirá continuar la línea teórico-metodológica de las investigaciones sociolingüísticas variacionistas así como el estudio de distintos fenómenos lingüísticos-discursivos para lo cual cada investigador podrá extraer lo que considere relevante para sus fines del material lingüístico del corpus y hacer uso de la información extralingüística inherente a los datos.

## **2.2. Caracas: la comunidad de estudio**

Caracas es la capital del país; esta condición resume la confluencia de factores históricos, políticos, económicos y culturales que la han convertido (principalmente desde mediados del siglo XX) en el centro urbano más complejo e importante de la vida pública nacional. En ella se concentran los entes gubernamentales, los poderes públicos, diversas sedes universitarias; además, en esta ciudad funcionan y se editan los medios de comunicación (periódicos, radios y televisoras) de difusión nacional. Los procesos migratorios han contribuido, desde luego, al alto grado de complejidad y heterogeneidad de su composición demográfica. Cabe agregar que el dialecto caraqueño representa la variedad estándar del español de Venezuela como consecuencia de que el habla de la capital representa la variedad lingüística empleada en los medios de comunicación nacionales y en los medios oficial.

## **2.3. Preestratificación de la muestra**

La muestra se preestratifica según tres variables sociodemográficas: sexo (hombres y mujeres en igual número); tres grupos generacionales: **1.** 20-34; **2.** 35-54; **3.** 55 o más años; y nivel de instrucción: **1:** primaria; **2:** secundaria; **3:** superior. La distribución se hace con cuotas de afijación uniforme de seis hablantes por casilla, número suficiente para satisfacer las condiciones

necesarias de representatividad de la muestra.<sup>4</sup> La distribución se muestra en el Cuadro 1.

A cada hablante se le asigna un código que lo identifica de acuerdo con las tres variables, por ejemplo: HABLANTE 1H2: hombre, de edad comprendida entre 20 a 34 años, y de grado de instrucción media o HABLANTE 3M2: mujer de 55 años o más de edad y con un grado de instrucción superior.

#### **2.4. Distribución geográfica de los hablantes**

Para asegurar una distribución espacial equilibrada y contar con hablantes de todos los micromundos geográficos que configuran la ciudad capital, los hablantes se seleccionan de acuerdo con los porcentajes señalados para los cinco municipios caraqueños: Libertador (61,2%); Sucre (24,8%); Baruta (9,9%); Chacao (2,2%); y El Hatillo (1,9%). En el Cuadro 2 se muestra la distribución de los hablantes preestratificados por sexo, edad y nivel de instrucción de acuerdo con la densidad poblacional de los cinco municipios:

Chacao, el más pequeño de los municipios (y prácticamente el más rico en ingresos), donde habita solamente el 2,2% de la población, contrasta con el Municipio Sucre, que –con el 24,8% de la población, su problemática topografía, sus quinientos y más barrios pobres, y treinta y nueve (39) urbanizaciones– tiene graves dificultades educativo-sanitarias y es el escenario de continuas expresiones de violencia social.

### **Conclusión**

Hemos presentado de manera muy esquemática la creación y el desarrollo de los corpus de habla venezolanos, con el fin de dar a conocer la extraordinaria riqueza de materiales de que disponen los lingüistas interesados en el estudio del español hablado. Asimismo, hemos querido informar acerca del nuevo corpus que se está actualmente grabando según los lineamientos del Proyecto panhispánico PRESEEA, con la esperanza de que otros equipos de investigación, de todas partes de Venezuela, se unan

4 Cf. Moreno Fernández 2005. PRESEEA propuso esta cuota fija por casilla para alcanzar la proporción de 1/25 000 en las poblaciones de alrededor de dos millones de habitantes.

CUADRO 1  
*Distribución por cuota de los hablantes*

Nivel de instrucción	Grupo gener. 1 20–34 años		Grupo gener. 2 35–54 años		Grupo gener. 3 55 o más		Total
	H	M	H	M	H	M	
1: Primaria	6	6	6	6	6	6	36
2: Media	6	6	6	6	6	6	36
3: Superior	6	6	6	6	6	6	36
Total	18	18	18	18	18	18	108
	36		36		36		108

CUADRO 2  
*Distribución de los hablantes por municipios*

Municipios/ Nivel de instrucción	Libertador		Sucre		Baruta		Chacao		El Hatillo		Total
	H	M	H	M	H	M	H	M	H	M	
1	10	10	4	4	2	2	1	1	1	1	36
2	10	10	4	4	2	2	1	1	1	1	36
3	10	10	4	4	2	2	1	1	1	1	36
Total	30	30	12	12	6	6	3	3	3	3	108
	60		24		12		6		6		108
	56%		21%		11%		6%		6%		100

al nuestro, para así obtener una muestra similar a las de muchísimas otras ciudades, grandes y pequeñas, de Hispanoamérica y de España.

### Referencias bibliográficas

- Bentivoglio, Paola y Mercedes Sedano (1993). Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. *Boletín de Lingüística* 8.3-35.
- Caravedo, Rocío (1999). *Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español*. Salamanca: Ediciones Universidad.

- Chela-Flores, Bertha y Jeannette Gelman (1987). *Hacia un análisis fonético del habla de Maracaibo*. Trabajo inédito, Universidad del Zulia, Maracaibo.
- Chela-Flores, Berta y Godsuno Chela-Flores (1994). *Hacia un estudio fonetológico del español hablado en Venezuela*. Caracas: Editorial Tropykos.
- Domínguez, Carmen Luisa y Elsa Mora (coordinadoras) (1998). *El habla de Mérida*. Mérida (Venezuela): Universidad de Los Andes.
- Lope Blanch, Juan M. (1986). *El estudio del español hablado culto. Historia de un proyecto*. México: Universidad Nacional Autónoma de México.
- Moreno Fernández, Francisco (2005). Corpus para el estudio del español en su variación geográfica y social. El corpus «PRESEEA». *Oralia* 8, 123-139.
- Navarro C., Manuel (1995). *El español hablado en Puerto Cabello*. Valencia (Venezuela): Universidad de Carabobo.