NOTAS

EL NÚMERO DE PALABRAS: UN NUEVO CRITERIO PARA DESCRIBIR TRES CORPUS DEL HABLA DE CARACAS

María José Gallucci
Universidad Central de Venezuela
maria_gallucci@yahoo.es

La revisión bibliográfica y/o electrónica de aquellos materiales con los que cuenta el investigador de la lengua (bien sea un corpus u otro banco de datos)[1] muestra cómo, en la mayoría de los casos, el volumen de dichos materiales se describe no sólo en función del número de textos que se agrupan en ellos sino a partir del número de palabras que contienen. Por ejemplo:

1) *The Brown Corpus*, 1961 (Francis & Kucera, 1967; Kucera, 1992), es un corpus compuesto por quinientos textos de dos mil palabras cada uno; contiene en total un millón de palabras escritas en inglés americano.

2) *The London-Lund Corpus* (LLC), 1960-1970 (Svartvik & Quirk, 1980), contiene quinientas mil palabras del habla culta británica.

3) *The Birmingham Collection of English Text* (BCET), 1980-1985 (Renouf, 1984; Sinclair & Kirby, 1990), está compuesto por veinte millones de palabras de lengua oral y escrita (dieciocho millones quinientas mil y un millón quinientas mil, respectivamente).

1. La distinción entre corpus y banco de datos a la que se hace referencia proviene de Edwards (1993: 282-283). Este autor diferencia ambos materiales de estudio tomando en cuenta dos aspectos que, a su juicio, son fundamentales para tal fin: i) el tamaño y la composición de dichos materiales, y ii) los objetivos del análisis que se pretende realizar. De esta manera, mientras que un corpus es representativo de una población o género específico y se utiliza a gran escala para establecer contrastes sistemáticos (variedades del lenguaje, géneros, modalidades, etc.), un banco de datos tiende a ser un conjunto de textos sin conexión entre sí que puede centrarse en un número restringido de géneros (por lo general, uno solo).

En contraposición a los ejemplos presentados, los corpus lingüísticos venezolanos (Bentivoglio, 1998: 40) no se describen por el número de palabras sino por el número de muestras de habla que los componen. En consonancia con esta desventaja o carencia, el presente trabajo tiene como objetivo principal dar a conocer el número de palabras que tienen tres importantes corpus del habla de Caracas que han sido objeto de numerosos estudios en nuestro país: el *Corpus sociolingüístico* de 1977 (cf. Bentivoglio, 1998), el *Corpus sociolingüístico* de 1987[2] (cf. Bentivoglio y Sedano, 1993) y el *Corpus de habla infantil* de 1996 (cf. Shiro, 1998).[3]

Más allá de ser utilizado como un simple criterio descriptivo, el número de palabras de cada hablante, el promedio de palabras de los hablantes según la edad o el nivel socioeconómico y el total de palabras que contiene cada uno estos corpus, también pueden utilizarse para otros fines; esencialmente para establecer comparaciones con otros corpus.

El procedimiento fundamental para establecer el número de palabras en cada uno de estos corpus consistió en convertir cada una de las muestras que los componen en un texto libre de anotaciones (cf. Caravedo, 1999). En principio, cada texto debía reflejar únicamente las palabras del hablante (el entrevistado).

En el caso de los corpus de 1977 y 1987, convertir las muestras en un texto puro supuso llevar a cabo las siguientes tareas:

1) Eliminar las marcas que identifican el turno (o parlamento) de los entrevistadores (*Enc.1:, Enc.2:*) y del hablante (*Habl.:*). Al eliminar las marcas que identifican a los encuestadores, también debe eliminarse el texto que corresponde a sus intervenciones en la conversación, pues en el corpus no interesa lo que dicen los entrevistadores sino el entrevistado.[4]

2. Este corpus está incorporado a la base de datos de CREA (*Corpus de referencia del español actual*). Para más información al respecto, véase http://www.rae.es

3. Disponible en http://childes.psy.cmu.edu/data/Romance/Spanish/

4. Esta afirmación es válida si se trabaja con los niveles de análisis fonético-fonológico, morfosintáctico y léxico-semántico. Sin embargo, es necesario aclarar que en un estudio prágnatico no se excluyen los turnos de ninguno de los participantes, puesto que este tipo de análisis se centra en la interacción.

2) Eliminar toda la información que se da entre corchetes, bien sea: a) la descripción de un sonido que no puede reproducirse mediante la grafía (por ejemplo, *[imita el sonido de un carro que se apaga], [aspiración exclamativa]*); b) la interrupción en el desarrollo de la conversación (por ejemplo, *[ruido], [risas], [se oye la voz de una niña que trata de interrumpir], [Está sonando insistentemente un timbre]*); o c) para indicar que lo que dice el hablante es incomprensible (*[?]*).

3) Eliminar las grafías que aparecen entre paréntesis indicando la forma cómo el hablante pronuncia las palabras que provienen de otro idioma: *shorts* ("chores"), *blue jean* ("bluyín"), *out* ("ao"), etc.

Una vez realizadas todas estas modificaciones, debe verificarse que todos los signos de puntuación (comas, puntos, guiones, interrogación, barras oblicuas, etc.) permanezcan unidos a una palabra y que no haya ningún tipo de espacios entre ellos pues, de ser así, dichos signos también contarían como palabras cuando se ejecute el comando CONTAR PALABRAS de Microsoft Word.

Es importante destacar también algunas precisiones en cuanto a ciertos elementos que, si bien no son palabras en sentido estricto, fueron tomadas como tales en este trabajo, a saber:

1) Las palabras incompletas, por ejemplo: *Bee.. Beethoven, pe.. vocalmente perfecta*.

2) Las vocalizaciones o *palabras marginales* (cf. Du Bois, *et al*. 1993). Según Calsamiglia y Tusón (1999: 54), las vocalizaciones son sonidos o ruidos que no son palabras pero que son significativos en la comunicación pues cumplen distintas funciones; se utilizan para asentir, para mostrar desacuerdo o impaciencia, para pedir la palabra o mantener el turno y para mostrar admiración, desprecio o incomprensión hacia quien habla o hacia lo que dice.[5] Las vocalizaciones más frecuentes son *mjum*, *eh*, *mm*, *ah*.

5. Reconocer la función que cumple una determinada vocalización (como por ejemplo, el mantenimiento del turno) no es posible sólo con el parlamento del hablante, es necesario conocer qué dijo el entrevistador. Esta acotación se hace a propósito de la nota anterior.

En contraposición a lo anterior, establecer el número de palabras del *Corpus del habla infantil* fue más sencillo. Esto en virtud de que las muestras de habla de este corpus se encuentran transcritas en formato CHAT (*Codes for the Human Analysis of Transcriptions*, Sokolov y Snow, 1994). Este formato permite realizar análisis lingüísticos utilizando el programa de computación CLAN (*Computer Language Analysis*, MacWhinney, 1994). Gracias a esto, el cálculo del número de palabras de cada muestra fue prácticamente automático.

Una vez obtenidos los resultados del número de palabras de cada hablante, se calculó el total de palabras según el nivel socioeconómico y el grupo generacional. Como información adicional, a estos datos se les calculó también su promedio.

1. EL CORPUS SOCIOLINGÜÍSTICO DE CARACAS, 1977

Este corpus está conformado por setenta hablantes agrupados según edad, sexo y nivel socioeconómico de la siguiente forma:

i) *Edad*. Treinta y cuatro hablantes[6] en el grupo generacional *A* (de 14 a 29 años) y treinta y seis en el *B* (de 30 a 45 años).
ii) *Sexo*. Treinta y seis hombres y treinta y cuatro mujeres.
iii) *Nivel socioeconómico*. Veinticuatro hablantes en el nivel *alto*, veintidós en el nivel *medio* y veinticuatro en el nivel *bajo*.

Cuadro 1: Distribución de los hablantes del CSC, 1977

GRUPO GENERACIONAL				
NIVEL SOCIOECONÓMICO	**Grupo A 14 - 29 años**		**Grupo B 30 - 45 años**	
	HOMBRES	MUJERES	HOMBRES	MUJERES
ALTO	6	6	6	6
MEDIO	6	4	6	6
BAJO	6	6	6	6
TOTAL	18	16	18	18

6. En principio, este proyecto estuvo ideado para grabar a treinta y seis hablantes por grupo generacional y, en pro de cuotas de afijación uniforme, treinta y seis hombres y treinta y seis mujeres; es decir, seis hablantes por casilla (véase el cuadro 1). Sin embargo, faltaron por grabar dos mujeres del nivel socioeconómico medio y grupo generacional A (ca3fe y ca3ff).

El total de palabras que contiene este corpus es de doscientos ochenta y cinco mil novecientos dieciséis (285.916). Si promediamos esta cifra, tenemos que este corpus está compuesto por setenta muestras de habla, a razón de cuatro mil ochenta y cinco (4.085) palabras por hablante. A continuación, se presenta un cuadro que muestra de forma específica el número de palabras que contiene cada una de las grabaciones:

Cuadro 2: Número de palabras por hablante en el CSC 1977[7]

GRUPO GENERACIONAL				
NIVEL SOCIOECONÓMICO	**Grupo A 14 - 29 años**		**Grupo B 30 - 45 años**	
	HOMBRES*	MUJERES*	HOMBRES*	MUJERES*
ALTO	(ca1ma.77) **5.383**	(ca1fa.77) **5.156**	(cb1ma.77) **2.795**	(cb1fa.77) **3.558**
	(ca1mb.77) **4.468**	(ca1fb.77) **3.803**	(cb1mb.77) **3.354**	(cb1fb.77) **3.656**
	(ca1mc.77) **5.611**	(ca1fc.77) **4.442**	(cb1mc.77) **3.804**	(cb1fc.77) **4.339**
	(ca1md.77) **5.242**	(ca1fd.77) **4.538**	(cb1md.77) **3.154**	(cb1fd.77) **4.362**
	(ca1me.77) **2.980**	(ca1fe.77) **3.009**	(cb1me.77) **5.581**	(cb1fe.77) **4.012**
	(ca1mf.77) **3.664**	(ca1ff.77) **3.042**	(cb1mf.77) **4.156**	(cb1ff.77) **2.881**
MEDIO	(ca3ma.77) **4.287**	(ca3fa.77) **2.985**	(cb3ma.77) **4.032**	(cb3fa.77) **4.762**
	(ca3mb.77) **3.860**	(ca3fb.77) **3.542**	(cb3mb.77) **4.427**	(cb3fb.77) **2.871**
	(ca3mc.77) **4.703**	(ca3fc.77) **4.018**	(cb3mc.77) **4.310**	(cb3fc.77) **4.433**
	(ca3md.77) **3.590**	(ca3fd.77) **4.268**	(cb3md.77) **4.475**	(cb3fd.77) **4.733**
	(ca3me.77) **4.592**	(ca3fe.77) **--------**	(cb3me.77) **5.146**	(cb3fe.77) **3.118**
	(ca3mf.77) **4.901**	(ca3ff.77) **--------**	(cb3mf.77) **4.141**	(cb3ff.77) **3.540**
BAJO	(ca5ma.77) **4.088**	(ca5fa.77) **3.284**	(cb5ma.77) **3.957**	(cb5fa.77) **2.544**
	(ca5mb.77) **4.348**	(ca5fb.77) **5.063**	(cb5mb.77) **6.047**	(cb5fb.77) **3.343**
	(ca5mc.77) **4.024**	(ca5fc.77) **4.689**	(cb5mc.77) **4.427**	(cb5fc.77) **4.832**
	(ca5md.77) **4.736**	(ca5fd.77) **4.854**	(cb5md.77) **2.142**	(cb5fd.77) **5.455**
	(ca5me.77) **4.023**	(ca5fe.77) **3.317**	(cb5me.77) **4.205**	(cb5fe.77) **3.833**
	(ca5mf.77) **4.572**	(ca5ff.77) **3.609**	(cb5mf.77) **4.133**	(cb5ff.77) **2.667**

* Lo que aparece entre paréntesis corresponde a la identificación del hablante; por su parte, las cifras en negritas indican el número de palabras por hablante.

7. En este corpus, al igual que en el de 1987, las letras que identifican a cada una de las muestras indican lo siguiente: c (Caracas); a o b (grupo generacional); 1, 3 ,5 (nivel socioeconómico alto, medio o bajo); m o f (masculino o femenino); a, b, c, d, e, f (identificación del hablante en la casilla correspondiente); y 77 (año en el que se grabó el corpus).

Si agrupamos las muestras de habla según el nivel socioeconómico, observamos que el mayor número de palabras se registra en el nivel bajo, seguidas muy de cerca por el nivel alto:

Cuadro 3: Número de palabras CSC 1977 según el nivel socioeconómico

Nivel Socioeconómico	N° palabras
Nivel Alto	96.990
Promedio	4.041
Nivel Medio	90.734
Promedio	4.124
Nivel Bajo	98.192
Promedio	4.091
TOTAL	285.916

En cambio, si agrupamos el número de palabras de este corpus según el grupo generacional, no encontramos diferencias considerables; tan sólo quinientas treinta y cuatro (534) palabras de más en el grupo *B*:

Cuadro 4: Número de palabras CSC 1977 según el grupo generacional

Grupo generacional	N° palabras
Grupo A (14 - 29 años)	142.691
Promedio	4.197
Grupo B (30 - 45 años)	143.225
Promedio	3.978
TOTAL	285.916

2. EL CORPUS SOCIOLINGÜÍSTICO DE CARACAS, 1987

Este corpus está conformado por ciento sesenta hablantes. A diferencia del anterior, tiene más especificidad en cuanto a su estratificación. En este caso, se distinguen cuatro grupos generacionales y cinco niveles socioeconómicos. El corpus está agrupado como sigue (Bentivoglio y Sedano, 1993: 4):

i) *Edad*. Cuarenta hablantes en cada grupo generacional, a saber: *A* (de 14 a 29 años), *B* (de 30 a 45 años), *C* (de 46 a 60 años) y *D* (de 61 años en adelante).
ii) *Sexo*. Ochenta hombres y ochenta mujeres.
iii) *Nivel socioeconómico*. Treinta y dos hablantes en cada nivel socioeconómico: alto, medio alto, medio, medio bajo y bajo.

Cuadro 5: Distribución de los hablantes del CSC, 1987

GRUPO GENERACIONAL								
NIVEL SOCIOECONÓMICO	**Grupo A 14 - 29 años**		**Grupo B 30 - 45 años**		**Grupo C 46 - 60 años**		**Grupo D 61 años y +**	
	H	M	H	M	H	M	H	M
ALTO	4	4	4	4	4	4	4	4
MEDIO ALTO	4	4	4	4	4	4	4	4
MEDIO	4	4	4	4	4	4	4	4
MEDIO BAJO	4	4	4	4	4	4	4	4
BAJO	4	4	4	4	4	4	4	4
TOTAL	20	20	20	20	20	20	20	20

Luego de eliminar aproximadamente ciento quince mil palabras, pudo observarse que el total de palabras en este corpus es de setecientos sesenta y siete mil ochocientos sesenta y ocho (767.868). El promedio de palabras por hablante es de cuatro mil setecientos noventa y nueve (4.799). A continuación, se presenta un cuadro que muestra detalladamente el número de palabras de cada hablante:

Cuadro 6: Número de palabras por hablante en el CSC 1987

GRUPO GENERACIONAL								
NIVEL SOCIO-ECONÓ-MICO	**Grupo A 14 - 29 años**		**Grupo B 30 - 45 años**		**Grupo C 46 - 60 años**		**Grupo D 61 años y +**	
	HOMBRES*	MUJERES*	HOMBRES*	MUJERES*	HOMBRES*	MUJERES*	HOMBRES*	MUJERES*
ALTO	(ca1ma.87) **4.555**	(ca1fa.87) **4.063**	(cb1ma.87) **5.424**	(cb1fa.87) **5.653**	(cc1ma.87) **4.131**	(cc1fa.87) **4.971**	(cd1ma.87) **4.914**	(cd1fa.87) **5.915**
	(ca1mb.87) **4.665**	(ca1fb.87) **5.510**	(cb1mb.87) **6.009**	(cb1fb.87) **3.954**	(cc1mb.87) **5.049**	(cc1fb.87) **4.269**	(cd1mb.87) **3.494**	(cd1fb.87) **3.576**
	(ca1mc.87) **6.073**	(ca1fc.87) **5.997**	(cb1mc.87) **3.557**	(cb1fc.87) **4.793**	(cc1mc.87) **4.989**	(cc1fc.87) **4.968**	(cd1mc.87) **4.038**	(cd1fc.87) **4.984**
	(ca1md.87) **4.731**	(ca1fd.87) **6.165**	(cb1md.87) **5.061**	(cb1fd.87) **3.999**	(cc1md.87) **6.331**	(cc1fd.87) **5.660**	(cd1md.87) **3.857**	(cd1fd.87) **5.167**
MEDIO ALTO	(ca2ma.87) **5.685**	(ca2fa.87) **4.048**	(cb2ma.87) **4.465**	(cb2fa.87) **4.584**	(cc2ma.87) **3.943**	(cc2fa.87) **4.483**	(cd2ma.87) **3.248**	(cd2fa.87) **4.032**
	(ca2mb.87) **5.081**	(ca2fb.87) **4.175**	(cb2mb.87) **4.437**	(cb2fb.87) **5.610**	(cc2mb.87) **4.164**	(cc2fb.87) **5.554**	(cd2mb.87) **4.379**	(cd2fb.87) **5.860**
	(ca2mc.87) **5.259**	(ca2fc.87) **4.590**	(cb2mc.87) **5.006**	(cb2fc.87) **4.535**	(cc2mc.87) **5.202**	(cc2fc.87) **5.402**	(cd2mc.87) **4.650**	(cd2fc.87) **5.660**
	(ca2md.87) **5.450**	(ca2fd.87) **5.244**	(cb2md.87) **5.368**	(cb2fd.87) **5.000**	(cc2md.87) **5.354**	(cc2fd.87) **4.980**	(cd2md.87) **4.409**	(cd2fd.87) **4.351**
MEDIO	(ca3ma.87) **5.383**	(ca3fa.87) **4.863**	(cb3ma.87) **4.398**	(cb3fa.87) **5.643**	(cc3ma.87) **4.002**	(cc3fa.87) **5.772**	(cd3ma.87) **5.009**	(cd3fa.87) **4.508**
	(ca3mb.87) **5.185**	(ca3fb.87) **5.685**	(cb3mb.87) **5.293**	(cb3fb.87) **5.489**	(cc3mb.87) **4.658**	(cc3fb.87) **4.877**	(cd3mb.87) **4.403**	(cd3fb.87) **5.113**
	(ca3mc.87) **4.798**	(ca3fc.87) **4.635**	(cb3mc.87) **6.488**	(cb3fc.87) **4.105**	(cc3mc.87) **4.905**	(cc3fc.87) **5.046**	(cd3mc.87) **4.538**	(cd3fc.87) **5.343**
	(ca3md.87) **7.054**	(ca3fd.87) **5.637**	(cb3md.87) **5.079**	(cb3fd.87) **5.386**	(cc3md.87) **4.058**	(cc3fd.87) **4.679**	(cd3md.87) **5.447**	(cd3fd.87) **5.343**
MEDIO BAJO	(ca4ma.87) **5.121**	(ca4fa.87) **5.397**	(cb4ma.87) **4.566**	(cb4fa.87) **5.230**	(cc4ma.87) **5.004**	(cc4fa.87) **5.832**	(cd4ma.87) **4.158**	(cd4fa.87) **4.302**
	(ca4mb.87) **3.765**	(ca4fb.87) **4.143**	(cb4mb.87) **4.774**	(cb4fb.87) **4.195**	(cc4mb.87) **4.344**	(cc4fb.87) **4.882**	(cd4mb.87) **3.065**	(cd4fb.87) **4.258**
	(ca4mc.87) **5.393**	(ca4fc.87) **5.850**	(cb4mc.87) **4.774**	(cb4fc.87) **5.061**	(cc4mc.87) **2.963**	(cc4fc.87) **4.464**	(cd4mc.87) **4.726**	(cd4fc.87) **5.966**
	(ca4md.87) **4.362**	(ca4fd.87) **4.190**	(cb4md.87) **4.387**	(cb4fd.87) **4.702**	(cc4md.87) **3.533**	(cc4fd.87) **4.559**	(cd4md.87) **4.740**	(cd4fd.87) **4.921**
BAJO	(ca5ma.87) **4.754**	(ca5fa.87) **5.088**	(cb5ma.87) **4.705**	(cb5fa.87) **4.933**	(cc5ma.87) **4.896**	(cc5fa.87) **4.670**	(cd5ma.87) **4.571**	(cd5fa.87) **4.927**
	(ca5mb.87) **4.050**	(ca5fb.87) **4.561**	(cb5mb.87) **4.337**	(cb5fb.87) **5.158**	(cc5mb.87) **3.134**	(cc5fb.87) **5.008**	(cd5mb.87) **4.174**	(cd5fb.87) **4.958**
	(ca5mc.87) **5.626**	(ca5fc.87) **3.883**	(cb5mc.87) **4.178**	(cb5fc.87) **3.536**	(cc5mc.87) **4.032**	(cc5fc.87) **3.422**	(cd5mc.87) **5.677**	(cd5fc.87) **3.565**
	(ca5md.87) **6.443**	(ca5fd.87) **4.250**	(cb5md.87) **6.433**	(cb5fd.87) **4.815**	(cc5md.87) **4.707**	(cc5fd.87) **4.249**	(cd5md.87) **4.154**	(cd5fd.87) **3.799**

* Lo que aparece entre paréntesis corresponde a la identificación del hablante; por su parte, las cifras en negritas indican el número de palabras por hablante.

Si calculamos el número de palabras según el nivel socioeconómico, es posible observar que el mayor número de palabras se ubica en el estrato medio (162.828 palabras); a éste le siguen los niveles alto, medio alto, medio bajo y bajo con 156.525, 154.195, 147.627 y 146.693 palabras, respectivamente. Como se desprende de esta información, el número de palabras del nivel *medio* está muy por encima de los otros niveles; en estos últimos llama la atención la cercanía numérica que existe entre el número de palabras de los estratos alto, medio alto, bajo y medio bajo:

Cuadro 7: Número de palabras CSC 1987 según el nivel socioeconómico

Nivel Socioeconómico	N° palabras
Nivel Alto	156.525
Promedio	4.891
Nivel Medio Alto	154.195
Promedio	4.819
Nivel Bajo	162.828
Promedio	5.088
Nivel Medio Bajo	147.627
Promedio	4.613
Nivel Bajo	146.693
Promedio	4.584
TOTAL	767.868

En cuanto al grupo generacional, vemos que el que tiene mayor número de palabras es el *A* (201.400 palabras). A partir de este grupo, el número de palabras comienza a descender: *B*= 195.123, *C*= 187.146 y *D*= 184.199:

Cuadro 8: Número de palabras CSC 1987 según el grupo generacional

Grupo generacional	N° palabras
Grupo A (14 - 29 años)	201.400
Promedio	5.035
Grupo B (30 - 45 años)	195.123
Promedio	4.878
Grupo C (46 - 60 años)	187.146
Promedio	4.679
Grupo D (61 años y +)	184.199
Promedio	4.605
TOTAL	767.868

3. EL CORPUS DE HABLA INFANTIL, 1996

Este corpus está conformado por ciento trece hablantes agrupados según edad y nivel socioeconómico del siguiente modo:

i) *Edad*. Cincuenta y siete niños en el grupo generacional *A* (de 6 a 8 años) y cincuenta y seis en el *B* (de 9 a 11 años).

ii) *Sexo*. Cincuenta y nueve niños y cincuenta y cuatro niñas.

iii) *Nivel socioeconómico*. Cincuenta y cuatro niños provenientes de un nivel socioeconómico bajo y cincuenta y nueve niños de un nivel alto. Las muestras del habla infantil fueron recopiladas en seis escuelas de Caracas: tres públicas y tres privadas; indicadoras de los niveles bajo y alto, respectivamente. En este caso, las escuelas tomadas para obtener los datos son, como apunta Shiro (1998:17), representativas de polos extremos de la pirámide social. En consonancia con la elección de este criterio de estratificación, es importante precisar que "las escuelas públicas y privadas sirven como contexto para conformar la muestra, según una variable que refleja el nivel socioeconómico al que pertenece la familia del niño" (Shiro 1998:20).

Cuadro 9: Distribución de los hablantes del *Corpus infantil*, 1996[8]

Nivel socioeconómico / **Edad**	**Nivel bajo**		**Nivel alto**	
	niños	niñas	niños	niñas
Grupo A (6 a 8 años de edad)	14	13	17	13
Grupo B (9 a 11 años de edad)	12	15	16	13
TOTAL	26	28	33	26

En cuanto a las características de este corpus, habría que señalar también que la selección de los niños se llevó a cabo según otro criterio: el grado de estudio. De esta manera, la muestra quedó restringida a niños del ciclo básico que cursaran primero o cuarto grado.

El total de palabras de este corpus es de ciento veintiséis mil setecientos noventa y ocho (126.798). El promedio de palabras por niño es de mil ciento veintidós (1.122). A continuación, se presenta un cuadro que muestra el número de palabras de cada niño:

8. Este cuadro es una adaptación del presentado por Shiro (2000: 319).

Cuadro 10: Número de palabras por hablante en el *Corpus de habla infantil* 1996[9]

EDAD				
NIVEL SOCIO-ECONÓMICO	**Grupo A 6 a 8 años de edad**		**Grupo B 9 a 11 años de edad**	
	NIÑOS*	NIÑAS*	NIÑOS*	NIÑAS*
NIVEL ALTO	(036.A.84.M) **885** (040.A.90.M) **1.249** (043.A.85.M) **600** (046.A.86.M) **822** (048.A.88.M) **707** (056.A.89.M) **481** (075.A.93.M) **520** (076.A.93.M) **675** (112.A.93.M) **765** (080.A.94.M) **893** (088.A.92.M) **1.393** (090.A.90.M) **1.727** (100.A.84.M) **1.108** (102.A.78.M) **972** (108.A.87.M) **837** (109.A.117.M) **513** (111.A.87.M) **780**	(039.A.89.F) **767** (041.A.87.F) **1.733** (044.A.86.F) **397** (051.A.87.F) **1.116** (055.A.89.F) **993** (078.A.92.F) **1.285** (087.A.94.F) **1.362** (092.A.95.F) **1.590** (094.A.90.F) **1.480** (099.A.85.F) **724** (103.A.80.F) **782** (104.A.86.F) **508** (106.A.78.F) **1.254**	(042.A.124.M) **3.163** (045.A.127.M) **2.175** (047.A.126.M) **2.304** (050.A.124.M) **922** (052.A.125.M) **1.951** (077.A.128.M) **1.021** (081.A.129.M) **1.228** (082.A.128.M) **1.259** (083.A.128.M) **1.450** (084.A.128.M) **1.254** (086.A.129.M) **1.744** (089.A.126.M) **2.305** (093.A.128.M) **1.173** (095.A.121.M) **1.151** (098.A.118.M) **1.538** (101.A.117.M) **1.605**	(037.A.123.F) **1.446** (038.A.120.F) **1.661** (049.A.125.F) **549** (053.A.114.F) **2.042** (054.A.124.F) **2.047** (079.A.126.F) **2.120** (085.A.129.F) **1.878** (091.A.127.F) **984** (096.A.125.F) **727** (097.A.117.F) **1.357** (105.A.126.F) **1.097** (107.A.118.F) **1.773** (110.A.124.F) **1.574**
NIVEL BAJO	(004.B.86.M) **1.212** (006.B.79.M) **907** (010.B.85.M) **1.431** (016.B.85.M) **798** (018.B.88.M) **1.498** (021.B.90.M) **1.463** (023.B.85.M) **1.391** (024.B.82.M) **175** (033.B.82.M) **576** (061.B.78.M) **632** (062.B.78.M) **1.278** (065.B.77.M) **1.077** (074.B.77.M) **931** (113.B.83.M) **275**	(001.B.90.F) **1.147** (011.B.85.F) **684** (012.B.81.F) **727** (017.B.89.F) **1.807** (022.B.82.F) **1.141** (025.B.78.F) **848** (028.B.87.F) **1.606** (035.B.80.F) **516** (057.B.81.F) **1.169** (059.B.77.F) **411** (064.B.77.F) **1.743** (067.B.84.F) **1.081** (070.B.84.F) **808**	(002.B.119.M) **427** (003.B.126.M) **1.271** (007.B.120.M) **962** (020.B.117.M) **626** (029.B.120.M) **477** (032.B.125.M) **805** (034.B.129.M) **928** (060.B.114.M) **1.364** (063.B.122.M) **615** (066.B.121.M) **733** (071.B.125.M) **2.363** (072.B.120.M) **1.033**	(005.B.117.F) **1.043** (008.B.116.F) **1.804** (009.B.124.F) **1.391** (013.B.116.F) **795** (014.B.122.F) **619** (015.B.114.F) **806** (019.B.117.F) **749** (026.B.121.F) **767** (027.B.121.F) **483** (030.B.126.F) **663** (031.B.118.F) **620** (058.B.117.F) **1.570** (068.B.122.F) **505** (069.B.119.F) **775** (073.B.109.F) **856**

* Lo que aparece entre paréntesis corresponde a la identificación del hablante; por su parte, las cifras en negritas indican el número de palabras por hablante.

9. En este corpus, la primera cifra indica el número de grabación; la letra A o B se utiliza para indicar el nivel socioeconómico; el número que le sigue corresponde a la edad del niño en meses y la letra final indica masculino o femenino.

El número de palabras según el nivel socioeconómico revela que los hablantes del nivel alto superan por una diferencia considerable (22.034 palabras) a los del nivel bajo:

Cuadro 11: Número de palabras del *Corpus infantil* según el nivel socioeconómico

Nivel socioeconómico	N° palabras
Nivel Alto	74.416
Promedio	1.261
Nivel Bajo	52.382
Promedio	970
TOTAL	126.798

Por su parte, el número de palabras según la edad también resulta significativo; en este caso son los niños del grupo *B* (entre 9 y 11 años de edad) quienes producen más palabras (70.548 frente a 56.250):

4. CONSIDERACIONES FINALES

Cuadro 12: Número de palabras del *Corpus infantil* según la edad

Grupo etario	N° palabras
Grupo A (6 - 8 años)	56.250
Promedio	987
Grupo B (9 - 11 años)	70.548
Promedio	1.260
TOTAL	126.798

Lo expuesto hasta ahora, además de constituir un nuevo criterio para la descripción de tres importantes corpus sociolingüísticos del habla de Caracas, se perfila como una herramienta útil para establecer comparaciones con otros corpus; estos pueden estar definidos por los mismos criterios que se han descrito aquí o bien sólo por el número de palabras que contienen. Asímismo, los datos estadísticos presentados se traducirán en una mayor precisión metodológica a la hora de comparar dos o más corpus.

REFERENCIAS BIBLIOGRÁFICAS

Bentivoglio, Paola y Mercedes Sedano. 1993. Investigación sociolingüística: sus métodos aplicados a una experiencia venezolana. *Boletín de Lingüística,* 8.3-35.

Bentivoglio, Paola. 1998. La variación sociofonológica. *Español Actual,* 69. 29-42.

Calsamiglia, Helena y Amparo Tusón. 1999. *Las cosas del decir. Manual de análisis del discurso*. Barcelona: Ariel.

Caravedo, Rocío. 1999. Lingüística del corpus. Cuestiones teórico-metodológicas aplicadas al español. [Gramatica española, enseñanza e investigación, Josse Dekock, director, Vol. 6]. Salamanca: Universidad de Salamanca.

Du Bois, John; Stephan Schuetze-Coburn; Danae Paolino y Susanna Cumming. 1993. Outline of discourse transcription. En Jane Edwards y Martin Lampert (eds.), *Talking data. Transcription and coding in discourse research*, 45-89. Hillsdale: Lawrence Erlbaum Associates.

Edwards, Jane. 1993. Survey of electronic corpora and related resources for language researchers. En Jane Edwards y Martin Lampert (eds.), *Talking data. Transcription and coding in discourse research*, 263-309. Hillsdale: Lawrence Erlbaum Associates.

Kucera, Henry. 1992. Brown corpus. En Stuart Shapiro (ed.). *Encyclopedia of artificial intelligence*, Vol 1, 128-130. New York: John Wiley & Sons.

Kucera, Henry y Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence: Brown University Press.

MacWhinney, Brian. 1994. *The CHILDES Project: Tools for analyzing talk*. Hillsdale: Lawrence Erlbaum Associates.

Real Academia Española: Banco de datos (CREA) [En línea]. *Corpus de referencia del español actual*. Disponible en: <http://www.rae.es> [Consulta: 10 de noviembre de 2005].

Renouf, Antoinette. 1984. Corpus development in Birmingham University. En Jan Aarts y Willem Meijs (eds.), *Corpus linguistics: Recent developments in the use of computers corpora in English language research*, 3-39. Amsterdam: Rodopi.

Shiro, Martha. 1996. *CHILDES database. Romance Languages Corpora, Spanish-Shiro* [En línea]. Disponible en: http://childes.psy.cmu.edu/data/Romance/Spanish/ [Consulta: 10 de noviembre de 2005].

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

Shiro, Martha. 2000. Los pequeños cuentacuentos. *Cuadernos Lengua y Habla*, 2. 319-337.

Sinclair, John y David Kirby. 1990. Progress in English computational lexicography. *World Englishes*, 9. 21-36.

Sokolov, Jeffrey y Catherine Snow. 1994. *Handbook of research in language development using CHILDES*. Hillsdale: Lawrence Erlbaum Associates.

Svartvik, Jan y Randolph Quirk (eds.). 1980. *A corpus of spoken English*. Lund: Lund University Press.