REVISTA ALED

Corpus de Interacciones de Jóvenes Universitarios: una experiencia para la investigación del lenguaje en contexto

JOSÉ A. MARTÍNEZ LARA
Universidad Central de Venezuela

Resumen. La Lingüística del Corpus (LDC, Caravedo 1999) es un enfoque para estudiar datos provenientes de producciones reales contextualizadas, los cuales son organizados según principios metodológicos precisos y explícitos, con el fin de que el investigador se acerque a la realidad lingüística de una comunidad. Por ende, en el marco de la LDC, es importante plantear una discusión sobre la elaboración de corpus y el análisis y tratamiento de los datos. En tal sentido, el objetivo de este artículo es presentar el *Corpus de Interacciones de Jóvenes Universitarios* (CIJU) como una experiencia en el marco teórico-metodológico de la LDC, además de suscitar una reflexión sobre los alcances de los corpus pequeños en las investigaciones individuales, partiendo de los principios de extensión y representatividad, y siguiendo el criterio de número de palabras.

Palabras clave: *lingüística del corpus, corpus, lenguaje juvenil, número de palabras*

Resumo. A Linguística de Corpus (LC, Caravedo 1999) é uma abordagem para o estudo de dados a partir de discursos autênticos e contextualizados. Seus princípios organizadores são metodologicamente precisos e explícitos, de modo que o pesquisador pode ter uma visão da realidade linguística dos integrantes de uma determinada comunidade. Assim sendo, no âmbito da LC, é importante discutir o processo de construção e análise de dados. Portanto, o objetivo deste artigo é apresentar o *Corpus de Interacciones de Jóvenes Universitários* (CIJU, "O Corpus das Interações de Jovens Universitários") como uma experiência dentro do quadro teórico e metodológico da LC, bem como discutir o alcance de pequenos corpora em pesquisas, destacando os critérios para a extensão do corpus, representatividade e o número de palavras.

Palavras-chave: *linguística de corpus, corpora, o discurso da juventude, número de palavras*

Abstract. Corpus Linguistics (CL, Caravedo 1999) is an approach for studying data from authentic contextualized speech. Its organizing principles are methodologically precise and explicit, so that the researcher can gain insight on a community´s linguistic reality. Therefore, within the framework of LC, it is important to embark on a discussion about corpus building and data analysis. Thus, the purpose of this article is to present the *Corpus de Interacciones de Jóvenes Universitarios* (CIJU, "Corpus of Young University Students´ Interactions") as an experience within the CL theoretical and methodological framework as well as to discuss the scope of small corpora in research, highlighting the criteria for corpus extension, representativity and number of words.

Keywords: *corpus linguistics, corpora, youth speech, number of words*

Recibido: 15 de febrero de 2016 • Aceptado: 12 de mayo de 2016.

Introducción

Desde la segunda mitad del siglo XX ha habido un gran interés por la recolección, organización, transcripción y digitalización de grandes muestras de materiales lingüístico –escritos y orales– que han servido para el análisis de distintos fenómenos de la lengua en cualquiera de los niveles del sistema, de manera que muchos lingüistas contemporáneos, según la propia corriente,[1] han basado sus estudios en datos extraídos de corpus, todo esto gracias a la entrada de una nueva tendencia en esta área: la Lingüística de Corpus (en adelante, LC). Al respecto, Cabré (2007) afirma que:

> Este avance ha abierto a los lingüistas la posibilidad de dar cuenta de forma más adecuada del funcionamiento de las lenguas ya que los análisis han podido basarse por primera vez en muestras representativas y abundantes de producciones lingüísticas, no limitadas ni sesgadas subjetivamente como sucedía anteriormente (Cabré 2007: 89).

En tal sentido, los trabajos cuyos análisis se han basado en grandes muestras auténticas de la lengua se acercan más a la realidad lingüística de una comunidad, puesto que los datos no son manipulados por el investigador, sino que han sido producidos por uno o varios emisores con una intención comunicativa particular. El investigador da cuenta de lo observado. Por lo tanto, puedo acotar que los corpus son una fuente de datos que contienen y reflejan la acción verbal que se supone tiene toda una comunidad lingüística, que deben ser analizados adecuadamente con el fin de determinar cómo funciona el sistema.

En Venezuela, el estudio de la lengua a través de corpus también fue asumido por muchos lingüistas. La mayoría de los corpus recogidos en el país son de tipo sociolingüístico, es decir, las muestras de habla están organizadas según los rasgos extralingüísticos de *sexo*, *edad*, *nivel socioeconómico* y, recientemente, el *grado de instrucción* de los hablantes.[2] Según la bibliografía consultada (Bentivoglio 1998), hay un mayor número de corpus de entrevistas semidirigidas que de interacciones naturales y espontáneas, con los que puedan analizarse fenómenos lingüísticos de conversaciones no mediadas por un guión o esquema previamente establecido.

Así como los corpus sociolingüísticos son útiles para el estudio de múltiples fenómenos, los de conversaciones contextualizadas son necesarios para el estudio de fenómenos específicos de las interacciones naturales tales como el cambio y la recuperación del turno de habla, la resolución de conflictos, la cortesía lingüística, entre otros.

Debo acotar que en la bibliografía sobre LC se hace mayor hincapié sobre los grandes corpus informatizados (cf. Sinclair 1987, McCarthy 1998, Simpson *et al*. 2002, Cheng *et al.* 2003), y se hace poca referencia a los pequeños corpus, que han sido recogidos con criterios metodológicos precisos, según la disciplina en la que se desarrolla una investigación, para sustentar los análisis

de un estudios. Por tal razón, se hace necesario reflexionar sobre los corpus pequeños, su creación y alcances. Por tanto, me pregunto:

- ¿Cómo deben definirse los materiales de análisis cuyos textos no son tan extensos?
- ¿Los corpus pequeños pueden y deben ser tratados con los mismos criterios teórico-metodológicos que los grandes corpus?

Este artículo tiene como objetivos abrir una reflexión sobre los corpus pequeños, partiendo de la presentación del *Corpus de Interacciones de Jóvenes Universitarios* (desde ahora CIJU).

Este texto está organizado en siete partes. En la primera, expongo una breve reseña sobre qué es la LC. En la segunda, presento qué se entiende por corpus. En la tercera, expongo los criterios de extensión y representatividad. En la cuarta, señalo el problema de investigación que me llevó a la creación del CIJU. En la quinta, presento las características generales del CIJU: identificación de las conversaciones, recolección y transcripción. En la sexta, indico el número de palabras por muestra, por grupo y por sexo. Y, por último, ofrezco las conclusiones y recomendaciones derivadas de esta experiencia investigativa.

1. La Lingüística del Corpus

Según algunos investigadores, la LC surge en 1964 con el lanzamiento del *Brown University Standard Corpus of Present-Day American English* de la Universidad de Brown, mejor conocido como *Corpus Brown* (Francis y Kucera 1967).

Berber (2000) señala que los objetivos y las características de la LC son:

> A Lingüística de Corpus ocupa-se da coleta e exploração de corpora, ou conjuntos de dados lingüísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem através de evidências empíricas, extraídas por meio de computador (Berber 2000: 325).

Por su parte, Parodi (2008: 96) afirma que la LC es "una metodología para la investigación de las lenguas y del lenguaje, la cual permite llevar a cabo investigaciones empíricas en contextos auténticos y que se constituye en torno a ciertos principios reguladores poderosos".

Según estos autores, la LC es una metodología cuyo objetivo es estudiar las lenguas naturales a través de datos provenientes tanto del lenguaje oral como del escrito, los cuales son recogidos, agrupados y organizados con unos criterios específicos, según los objetivos propuestos por el/los investigador/es.

Caravedo (1999), sin embargo, hace una distinción entre la *Lingüística de Corpus* (LC) y la *Lingüística del Corpus* (LDC). La primera (LC, Aarts y Meijs 1984) es una tendencia cuyo objetivo principal es la creación de grandes

muestras de materiales de la lengua oral o escrita, almacenados en computadoras para el análisis de los datos a través de programas informáticos. En cambio, la segunda, LDC, se refiere a:

> […] toda orientación que, en la formulación y en el desarrollo de su programa de investigación (comprendidos la teoría y el sistema de corroboraciones o refutaciones desprendido de la actividad analítica), *dependen* de la observación de un conjunto de datos extraídos de la *producción real* de los individuos, y ordenados según criterios metodológicos diferentes pero *explícitos* de investigación (Caravedo 1999:19)

La autora señala que esta no excluye a la otra, puesto que con este término (LC) incluye a todo enfoque que estudie la lengua en cualquiera de sus niveles y/o dimensiones de análisis, tomando en consideración datos reales proveniente tanto de la lengua oral como de la escrita. Caravedo (1999) resalta que los rasgos más característicos de la LDC son:

- La *dependencia empírica*, que implica que todo análisis lingüístico se desprende de la observación sistemática de datos reales, escogidos y organizados previamente por el investigador.
- La *producción*, se refiere a los actos de habla contextualizados que pueden ser identificables y analizables.
- El *realismo*, que consiste en la realización efectiva de un texto en una situación comunicativa concreta sin que el investigador manipule los datos ni los cree a través de su conocimiento del sistema de la lengua.
- La *explicitud*, se refiere a la comunicabilidad del proceso de análisis de los datos y no solo de los resultados, además de la naturaleza de los datos y de cómo estos han sido procesados para el análisis.

En tal sentido, la LDC no es una mera metodología de trabajo cuyo fin último es simplemente organizar esquemáticamente grandes cantidades de materiales lingüísticos, sino una tendencia que conlleva la discusión y formulación de teorías sobre los corpus, su clasificación y tipología; la manera de abordar el análisis cuantitativo y cualitativo de los fenómenos de la lenguas naturales basados en muestras reales, dar cuenta de la naturaleza de los datos y la metodología usada para su obtención y estudio.

En este trabajo, me suscribo a la propuesta teóricas de Caravedo (1999); es decir, considero que hay una corriente en la lingüística cuyo objetivo es la creación de grandes corpus informatizados, esta es la LC, y otra (LDC) –en la que se incluye la primera– cuyo objetivo es elaborar todo un abordaje teórico y metodológico sobre qué es un corpus, cómo debe ser organizado, por qué debe tener una forma específica, cuál es su tipología y cómo deben ser tratados los datos para los distintos estudios.

En virtud de lo expuesto anteriormente, puedo decir que la LDC es una tendencia lingüística que permite a los investigadores estudiar las lenguas

naturales de forma empírica a través de datos, provenientes tanto del lenguaje oral como del escrito, que son recogidos, agrupados y organizados con unos criterios específicos, según los objetivos propuestos por el investigador, enmarcado en una tradición discursiva y científica particular, y estos datos son almacenados de manera electrónica con el fin de ser analizados a través de programas informáticos –dependiendo del tipo de fenómeno y del tamaño del corpus– y, según la cual, cada investigador debe exponer concisamente la forma en que han sido tratado dichos datos.

2. ¿Qué es un corpus?

Actualmente, entre los lingüistas, no hay unanimidad a la hora de definir corpus. Son distintos los criterios tomados en consideración al momento de definirlo. Algunos se basan en el de cantidad, otros en la representatividad o bien en la forma de almacenamiento y otros, la disciplina o postura teórica. En las siguientes líneas presento algunas definiciones de corpus:

- "una colección de textos de ocurrencias de lenguaje natural, escogidos para caracterizar un estado o una variedad de lengua" (Sinclair 1991: 171).
- "Se conoce como *corpus* una muestra representativa de lenguaje reunido con propósitos de análisis lingüístico" (Crystal 1994: 410).
- "una colección de partes de una lengua que son seleccionados y ordenados de acuerdo a explícitos criterios lingüísticos, con el fin de ser empleados como ejemplos de esa lengua" (Monachini y Calzolari 1996).
- "un conjunto homogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como modelo de un estado o nivel de lengua predeterminado" (Torruella y Llisterri 1999: 8).
- "un corpus est une collection de donnés langagières qui sont seleccionées et organisées selon de critères linguistique explicites pour servir d'échantillon du langage" (Charaudeau y Mainguenau 2002: 148).
- "Un corpus, por su parte, reúne un conjunto de unidades textuales y no es una única instancia comunicativa, tampoco cuenta con cierre de ningún tipo. En este sentido, un corpus busca entregar datos acerca de la lengua en una proyección mayor que la que busca un texto como instancia de habla" (Parodi 2008: 106).
- "el corpus de la investigación se define como el conjunto de materiales lingüísticos o no, que conforman el objeto de estudio en una investigación o en una parte de ella" (Bolívar 2013: 3).

Como puede observarse, a pesar de la postura de cada autor, todas estas definiciones tienen un eje en común: el objetivo del corpus. Los textos orga-

nizados bajo unos mismos parámetros deben servir para estudiar y dar cuenta de un estado de lengua específica y natural.

En consonancia con estas afirmaciones, y para fines de este trabajo, considero que un corpus es el conjunto de textos –orales o escritos– recogidos de manera sistemática, en su estado natural y sin ningún tipo de manipulación, bajo criterios y objetivos específicos, tomando en consideración el contexto de uso, almacenados de forma electrónica, que sirven para el estudio de una variedad lingüística determinada.

Un corpus no es la suma de un grupo de textos recogidos al azar y sin ningún tipo de criterios que los asemeje y los oriente hacia un fin específico, y mucho menos un conjunto de ejemplos creados por el investigador para afirmar sus hipótesis. Por el contrario, la definición y la construcción de un corpus de estudio responden al problema y a los objetivos planteados por el investigador, quien, además, debe tomar en cuenta la postura teórica y la disciplina en la que se enmarca el estudio (Bolívar 2013). Puesto que, si bien es cierto que en la actualidad muchas disciplinas lingüísticas usan corpus en sus investigaciones, los materiales usados en los trabajos no siempre coinciden con los mismos criterios. De esto se desprende, como afirma Bolívar (2013), que hay una diferencia entre "los corpus" y "el corpus de estudio": el primer término se refiere a los materiales en general que usan las distintas disciplinas; y el segundo, al particular usado en un estudio. Por ejemplo, para un sociolingüista, un corpus de artículos de investigación en el área de las humanidades o de ingeniería o de cualquier otra ciencia no le sirve para hacer un estudio prototípico en su disciplina. Este tipo de corpus será más adecuado para un estudioso de los géneros discursivos. Sin embargo, no por esto deja de ser un corpus.

En este mismo orden de ideas, un corpus no es la colección de textos sin axiomas. Una cosa es coleccionar un conjunto de textos periodísticos, por ejemplo, escogidos al azar con el fin de observar algún aspecto de la lengua escrita y otra cosa es recoger sistemáticamente, según unos criterios metodológicos precisos, este mismo tipo de textos para analizar un fenómeno lingüístico específico. La colección de textos muy difícilmente le serviría a otro investigador, mientras que un corpus puede ser usado en otra investigación por otro analista. Entonces, puedo decir, considerando los criterios de Bolívar (2013), que el primer paso que debe tener en cuenta el investigador a la hora de recoger y armar su corpus es tener los objetivos claros que lo lleven a resolver el problema de la investigación.

En este punto es importante señalar dos metodologías utilizadas en la investigación con corpus. La primera es *corpus-based*, descrita por Tognini-Bonelli (2001:65) de la siguiente manera "to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study". En tal sentido, este enfoque es de tipo deductivo, es decir, el investigador parte de una presuposición o un patrón lingüístico que luego es verificada a

través de los datos emanados de un corpus. La segunda, *corpus-driven*, por el contrario, es de tipo inductivo; es decir, el lingüista construye sus categorías y describe los patrones de la lengua a partir de los casos que surgen del análisis del corpus. Al respecto, Biber (2009: 276) señala que "In its most basic form, corpus-driven analysis assumes only the existence of words; co-occurrence patterns among words, discovered from the corpus analysis, are the basis for subsequent linguistic descriptions". Los investigadores pueden asumir una de estas dos orientaciones metodológicas en sus trabajos. No obstante, ambas pueden ser utilizadas en un mismo estudio (Bolívar 2013). En el caso del estudio de los insultos (Martínez Lara 2006) se utilizaron ambos enfoque.

3. La extensión y la representatividad

En la bibliografía consultada, la extensión y la representatividad son dos criterios recurrentes entres las características que deben tener los corpus. Para muchos autores mientras más grande es el corpus más representativo es de la comunidad lingüística de la que se ha obtenido; y, mientras más grande sea, más se acerca a la realidad o al estado de lengua. Sin embargo, no hay unanimidad sobre qué tan extenso debe ser un corpus para ser representativos. Para esto, muchos autores recurren a los datos demográficos, matemáticos y probabilísticos para determinar estos criterios.

Al tratar la representatividad, Berber (2000) indica que este es el requisito más complicado en la creación de un corpus. Para este lingüista, de este problema surgen dos cuestionamientos: el primero, el corpus debe ser representativo ¿de qué? Y el segundo, debe ser representativo ¿para quién? Es decir, este problema debe tratarse cuantitativa y cualitativamente.

En cuanto al primer cuestionamiento, Berber indica que es difícil determinar el tamaño de muestras o extensión que debe tener un corpus, puesto que es muy difícil saber el número de hablantes de una misma lengua y/o dialecto; o, en términos de las frecuencias de uso de algunas formas con relación a otras en el sistema, es muy difícil saber que tan extenso debe ser un corpus en el que se encuentren formas muy frecuentes al igual que formas poco frecuentes. Pese a esto, lo que sí es probable es que a mayor número de palabras, habrá mayor posibilidad de que aparezcan formas lingüísticas menos frecuentes. Por esto, Berber (2000) señala que la extensión de un corpus tiene tres dimensiones: i) el número de palabras, ii) el número de textos y iii) el número de géneros o tipos de textos. En este artículo, haré hincapié en la extensión de un corpus según la dimensión del número de palabras, esto sin ignorar las otras dos dimensiones.

En el caso de los pequeños corpus, ciertamente conllevan, al menos, los tres siguientes problemas generados de la extensión: i) que no se hallen las formas poco frecuentes de la lengua, ii) que no se hallen las formas que creemos que están en aumento y iii) solo podrán ser usados para estudiar algunos pocos fenómenos.

Con relación a la segunda pregunta, ¿para quién debe ser representativo un corpus? A este respecto, vuelvo a señalar que es importante el problema y el objetivo de la investigación, puesto que estos permiten que el investigador decida si la muestra que está analizando es representativa o no. Pero esta decisión no es azarosa, ya que quien estudia un corpus debe tomar en consideración el origen, la naturaleza y el tratamiento de los textos; los criterios de recolección y los objetivos de su creación; además, de verificar la compatibilidad con la perspectiva teórica asumida en el estudio.

4. El CIJU como respuesta a un problema de investigación

El CIJU es el resultado de una investigación particular, titulada: *Estudio sociopragmático del uso de los insultos en la comunidad juvenil universitaria*. El objetivo del estudio era "identificar las palabras insultantes usadas por jóvenes universitarios venezolanos en sus interacciones cotidianas, y describir las funciones lingüísticas y discursivas que ellos les otorgan en la conversación" (Martínez Lara 2006: 10). Por tanto, se hacía necesario contar con muestras de habla naturales y en contextos específicos.

En 2006, había muestras de habla de jóvenes caraqueños dentro de los corpus de las universidades venezolanas. Sin embargo, estas presentaban los siguientes problemas:

- Habían sido recopiladas según criterios sociolingüísticos, mientras que la investigación sobre los insultos estaba enmarcada en los estudios pragmáticos de cortesía lingüística, por lo que debía afrontar los problemas de ¿qué relación tienen el entrevistado y el entrevistador? ¿Dónde y cómo se desarrolló la entrevista? Entre otras cuestiones.
- Las transcripciones de las entrevistas tenían un etiquetado mínimo necesario para los estudios sociolingüísticos, por lo que no podría saber –salvo al escuchar la entrevista– si había solapamientos, perdidas y recuperación del turno de la conversación, énfasis en la entonación, etc.
- Eran entrevistas semidirigidas, es decir, eran conversaciones mediadas indirectamente por un cuestionario preestablecido.
- Las entrevistas se enfocaba en recoger la producción de habla de un solo informante, quien desarrollaba un tema sin esperar una retroalimentación o respuesta del entrevistador.

Una primera solución fue aplicar cuestionarios o *tests* de hábitos sociales. Aunque estos son usados en algunas investigaciones sobre cortesía lingüística (Hernández Flores 2002, Boretti 2002), no permitirían observar detalladamente el problema en cuestión: ¿qué función tienen las palabras insultantes en las interacciones juveniles? Por esta razón, decidí recoger un corpus que tuviera las siguientes características:

1. Interacciones naturales y espontáneas.
2. Las grabaciones debían hacerse en contextos familiares para los interactuantes, con el fin de obtener producción lingüística real en contexto.
3. Sin un guión preestablecido de temas y/o tópicos de conversación.
4. Sin la intervención del investigador en la escogencia ni en el desarrollo del/los tema/s.
5. Sin informarles a los interactuantes que se analizarían los insultos, puesto que de lo contrario podrían moldear sus intervenciones y podrían preestablecer su vocabulario.
6. Hablantes jóvenes universitarios, de esta manera limitaba la comunidad estudiada.
7. Tres grupos: uno de hombres, uno de mujeres y uno de ambos sexos, de esta manera podría tener tres escenarios distintos de conversaciones.
8. El corpus debía tener el mismo número de conversaciones por grupo, y cada grupo debía tener un mínimo de dos conversaciones, ya que así se obtenía un número homogéneo de muestras, lo que mantendría la equidad y el equilibrio en el análisis.
9. Las transcripciones debían tener etiquetas que especificaran los rasgos de la oralidad, con el fin de reflejar de la manera más precisa posible lo que ocurría en la conversación; y con marcas que indicaran el contexto situacional.

Como puede observarse, la elaboración de un corpus para fines específicos requiere un proceso de reflexión teórica y metodológica; uno de planificación, otro, propiamente, de recolección, y otro de transcripción y etiquetado, antes de llegar al objetivo de la investigación: analizar el fenómeno en cuestión, con un claro sentido del tratamiento de los datos, y la explicitación concisa y precisa de todo el proceso.

5. Características generales del CIJU

El CIJU está compuesto por seis conversaciones naturales y espontáneas de estudiantes regulares de la UCV, sede Caracas. Las conversaciones sostenidas por estos fueron grabadas y transcritas por el investigador.

La duración de cada conversación no es igual, puesto que cada grupo tenía dinámicas de interacción distintas: estudios, juegos, conversaciones ocasionales, discusiones. La conversación más corta dura quince minutos, mientras que las más largas duran treinta minutos. Igualmente, el número de interactuantes por conversación es distinto; el menor número fue tres y el máximo fue siete participantes.

Las seis conversaciones están organizadas en tres grupos: dos conversaciones por cada grupo:

- Grupo mixto: en este se incluyen las dos grabaciones en las que interactuaron hombres y mujeres de manera conjunta.
- Grupo femenino: las dos conversaciones entre mujeres.
- Grupo masculino: este comprende las dos conversaciones en la que solo intervinieron hombres.

A continuación presento la distribución de cada una de las conversaciones:

Corpus de Interacciones de Jóvenes Universitarios					
Grupo mixto		Grupo femenino		Grupo masculino	
G1MA	**G2MB**	**G3FA**	**G4FB**	**G5HA**	**G6HB**
4 participantes: 2 hombres y 2 mujeres	7 participantes: 3 hombres y 4 mujeres	3 participantes	4 participantes	6 participantes	6 participantes
30’	30’	20’02”	15’	15’	17’20”

Cuadro 1. Distribución de los hablantes del CIJU

5.1. Identificación de las conversaciones

Cada una de las conversaciones del corpus las identifiqué con un código alfanumérico. El código está compuesto por cuatro caracteres, entre letras y números. Los dos primeros caracteres son la letra **G** seguida de un número (**1, 2, 3**…), lo que significa el número de la grabación, es decir: **G1**=grabación uno, **G2**= grabación dos, y así hasta la sexta.

El tercer carácter del código son las letras **M**, **F** o **H**, las cuales indican el sexo de los interactuantes: **M**, de conversación mixta, es decir, en la que intervinieron tanto hombres como mujeres; **F**, de femenino y **H**, de hombre.

Por último, le agregué al código la letra **A** o **B** según el lugar que ocupaba la conversación en la casilla del corpus: **A** para la primera casilla y **B** para la segunda.

De tal manera, el código **G3FA** significa que es la tercera grabación del corpus cuyos participantes son mujeres y es la grabación de la primera casilla del grupo femenino; mientras que el código **G6HB** especifica que se trata de la sexta grabación cuyos participantes son hombres y ocupa la segunda celda del grupo de hombres.

Cada una de las muestras que componen un corpus debe estar debidamente identificada con un código que indique las características y/o categorías con las que se ha recogido cada texto, de manera que el investigador pueda seleccionar las muestras que le interese de forma sencilla e inequívoca. Además, la identificación de cada muestra permite visualizar la organización interna del corpus.

5.2. Recolección del corpus

Con el fin de obtener muestras de habla naturales, decidí grabar a los estudiantes en sus entornos cotidianos. Es decir, en los espacios que habitualmente usan en la universidad, tanto para sus actividades académicas como lúdicas, por ende, grabé las seis conversaciones en: pasillos de Escuelas y bibliotecas, jardines y cubículos de estudios.

La primera conversación, G1MA, la grabé en el pasillo de la Escuela de Letras, momentos antes de que los interactuantes entraran en sus respectivas clases. Todos estaban sentados en el piso del pasillo. Los temas que trataron giraron en torno a programas de televisión, situaciones del día a día en la calle y relaciones familiares y amistades.

El escenario de la segunda grabación, G2MA, fue en uno de los cubículos de estudio de la Facultad de Ciencias. Estos cubículos son espacios semiabiertos, cada uno tiene un mesón, bancos y un pizarrón, y están ubicados en uno de los jardines de la facultad. Los estudiantes suelen ir a estos cubículos a estudiar o a charlar. El grupo entrevistado estaba estudiando y charlando al mismo tiempo. Los temas de la conversación se refieren a cuestiones académicas y experiencias del día a día vividas por ellos o por otros compañeros.

La tercera conversación, G3FA, la grabé en uno de los jardines de la Facultad de Medicina, específicamente entre el Instituto de Medicina Experimental y el de Medicina Tropical. Las tres mujeres grabadas hablaron sobre algunos compañeros de clases, algunas experiencias lúdicas vividas en la universidad, algunas experiencias académicas, el deseo de estudiar otras carreras y hacer actividades extracurriculares.

La G4FB la hice en el pasillo principal de la Escuela de Psicología. La conversación de las jóvenes se basó en una crítica a la postura que tuvo una de sus profesoras sobre el deber ser de un profesional, y de las connotaciones políticas que acarreaban esa opinión, por lo que pueden observarse algunas secuencias textuales argumentativas.

La quinta grabación, G5HA, es la primera del grupo masculino. Esta interacción tuvo lugar en los espacios abiertos de la biblioteca de la Facultad de Arquitectura. Los jóvenes estaban estudiando. Los temas de la conversación se centraron en sus estudios y en experiencias personales y grupales, las cuales cayeron en lo lúdico.

La última conversación, G6HB, la grabé en uno de los pasillos de la Facultad de Ciencias. Los estudiantes estaban jugando cartas, por lo que la conversación estuvo centrada en el juego y en comentarios jocosos sobre los participantes.

Debo acotar que escogí los grupos al azar. Después de haber escogido el grupo, les entregué a los jóvenes una carta de presentación en la que se avalaba y explicaba el fin de las grabaciones. A los jóvenes no les dije cuál era el tema de la investigación, puesto que esto hubiera hecho que ellos adecuaran su

vocabulario y/o discurso. En vez de esto, les dije que la investigación trataba sobre el léxico de los universitarios caraqueños. Después de esto, y con la autorización de los jóvenes, procedí a hacer la grabación.

En virtud de que la grabación era solo de voz, no podía limitarme solo a grabar, sino que también tomé apuntes y anotaciones sobre cómo se iba desarrollando la interacción, ya que en este tipo de grabaciones se pierden detalles tales como: i) cambio de turno de conversación (cuando no se conoce a los participantes, se puede confundir la voz de alguno de ellos), ii) cambio de lugar, iii) la llegada de alguna otra persona, iv) intercambio de objetos, etc. Por tanto, para cada grabación, tenía apuntes sobre comportamientos paralingüísticos y también sobre actos de habla específicos que me permitieran, a la hora de transcribir, recordar bien el momento, identificar al hablante y conocer la secuencia de la conversación.

Para llevar a cabo las grabaciones, utilicé una grabadora de voz marca AIWA modelo TP-VA300.

5.3. Transcripción de las grabaciones

Las transcripciones de las grabaciones las hice siguiendo el modelo ortográfico del español, tomando en consideración los signos de puntuación necesarios como: *coma*, *punto y coma*, *punto y seguido*, *punto y aparte*, *punto final*, *comillas* y *signos de interrogación* y de *exclamación*. Debo mencionar que en la transliteración de las grabaciones hice algunas anotaciones de índole fonéticas-prosódicas, por considerarlas pertinentes para el análisis y el tratamiento de los casos, tales como: i) alargamientos vocálicos (*noo*) y consonánticos (*verrsiale*), ii) supresión de sílabas en la preposición *para* (*pa'*)y iii) las distintas formas de pronunciar la palabra *huevón* (*güevón* y *gwón*).

El proceso de transcripción lo hice en dos grandes pasos. El primero consistió en transcribir las grabaciones en formato digital en *Microsoft Word*; una vez hecha esta primera transcripción, escuché nuevamente la grabación con el fin de verificar que lo transcrito concordara con la grabación de voz, de manera que la transcripción fuera lo más fiel posible a lo que habían dicho los participantes. El segundo paso consistió en agregar los signos que indican algunos de los rasgos lingüísticos y extralingüísticos. Los signos que utilicé fueron adaptados de la propuesta de Briz (2001).

El sistema de transcripción propuesto por Briz (2001:14) intenta codificar, en gran medida, algunas realizaciones de la lengua hablada, aunque es imposible reproducirla tal cual, puesto que en la transliteración de lo oral a lo escrito se pierden algunas informaciones como los gestos y señales, los movimientos y cambios de lugares, etc. Sin embargo, con los signos propuestos por el autor pueden reflejarse algunos hechos concretos, pertinentes y observables de la oralidad; en tal sentido, este sistema de transcripción facilita el análisis de los textos orales. Los signos de transcripción atienden a:

> "[…] fenómenos relacionados con la alternancia del turnos, la sucesión inmediata de emisiones, solapamientos, reinicios y autointerrupciones, escisiones conversacionales, pausas y silencios, entonación […] fenómenos de énfasis, problemas relacionados con emisiones dudosas o indescifrables, de fonosintaxis, alargamientos fonéticos, preguntas retóricas, expresiones irónicas, estilo directo, referencias contextuales, etcétera" (Briz 2001: 14).

A pesar de que el sistema de transcripción de Briz y su equipo es bastante completo y atiende a muchos rasgos del habla, no los usé todo. Los signos utilizados para las seis transcripciones del corpus son los siguientes:[3]

- **Nombre**, le puse un seudónimo a cada participante de las conversaciones, con el fin de preservar sus identidades.
- **:** señala cambio de turno de conversación
- **MAYÚSCULA** significa que la pronunciación es marcada, enfática o el hablante levantó la voz.
- ↑ entonación ascendente
- ↓ entonación descendente
- → entonación mantenida
- **§** sucesión inmediata de los turnos de habla.
- - vacilación y/o mantenimiento del turno de habla.
- [lugar donde comienza un solapamiento o superposición de voces
-] final de habla simultánea o solapamiento
- = mantenimiento y/o recuperación del turno de habla de un participante durante un solapamiento.
- **/** pausa cortas, inferiores a medio segundo
- // pausa entre medio segundo y un segundo
- **///** pausa de más de un segundo.
- **(( ))** fragmento indescifrable
- **((audio dudoso))** transcripción dudosa
- **(en)tonces** reconstrucción de una unidad léxica que se ha pronunciado incompleta.
- º( )º pronunciación muy baja, próxima al susurro.
- **aa**, **ee**, **ii**, **oo** y **uu**, alargamiento vocálico.
- **rr, ll, nn,** o cualquier otra consonante, alargamiento consonántico.
- **[[ ]]** este símbolo doble lo he agregado con el fin de hacer las anotaciones y comentarios etnográficos pertinentes. Briz (2001) utiliza la nota al pie de página para hacer este tipo de anotaciones. Yo, sin embargo, preferí que este tipo de anotación estuviera en el texto, tal como se presenta en (1) y (2):

  (1) **Cecilia:** ¿Por qué tiene que tronar? **[[Tronó en ese momento]]** G1MA

  (2) **Érica:** O sea, ella puede durar **[[Se dirige a María señalando a Amalia quien es gorda]] [[Risas]]** ella no necesita [(( ))]G4FB

6. *Número de palabras del corpus*

Como he mencionado en párrafos anteriores, la extensión es uno de los criterios con los que se definen los corpus, siendo el número de palabra una de las dimensiones métricas para atender a este criterio, y, en la actualidad, es una de las formas para determinar el tipo de corpus, además de la cantidad de muestras o textos que contenga. Al respecto, Gallucci (2005) indica:

> Más allá de ser utilizado como un simple criterio descriptivo, el número de palabras de cada hablante, el promedio de palabras que contiene cada uno de estos corpus, también pueden utilizarse para otros fines; esencialmente para establecer comparaciones con otros corpus (Gallucci 2005: 109).

En este apartado, presento el número de palabras que contiene el CIJU. Para hacer esto, convertí las transcripciones en "textos libres de anotaciones" (Caravedo 1999, Gallucci 2005); es decir, eliminé todo los signos del sistema de transcripción, de tal forma que quedaran solo las palabras de los interactuantes.

Grupo mixto: el número de palabras total de la primera grabación, G1MA, es de 1.765. Los hombres de esta conversación produjeron 1.068 (60%), mientras que las mujeres, 697 (40%). Por su parte, la G2MB tiene un total de 4.753 palabras, de las cuales los hombres dijeron 3.123 (66%), y las mujeres 1.630 (34%). El Grupo mixto tiene un total de 6.518 palabras, de las cuales los hombres tiene mayor porcentaje (4.191/ 64%); mientras que las mujeres tienen 2.327 (36%). En el gráfico 1, se ilustra la distribución del número de palabras en el grupo mixto según el sexo de los hablantes.

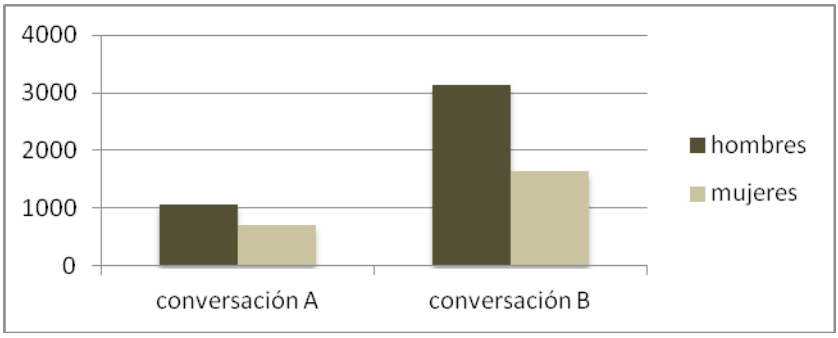

Gráfico 1: Distribución del número de palabras según el sexo de los hablantes en el grupo mixto

En cuanto al número de palabras del Grupo femenino, la tercera grabación, G3FA, tiene 3.359 (57%) palabras, mientras que la G4FB tiene 2.496 (43%) palabras, lo que da un total de 5.855 palabras. Por su parte, las conversaciones del Grupo de los hombres tiene un total de 3.802 palabras, distribuidas de la siguiente manera: la G5HA tiene 1.773 palabras (47%), y la G6HB tiene

2.029 palabras (53%).El número total de palabras del CIJU es 16.175: el *Grupo mixto* tiene el mayor porcentaje de palabras (40%), seguido del *Grupo femenino* (36%) y, por último, el *Grupo de los hombres* (24%). En el gráfico 2, puede apreciarse la distribución de palabras por cada grupo del corpus.

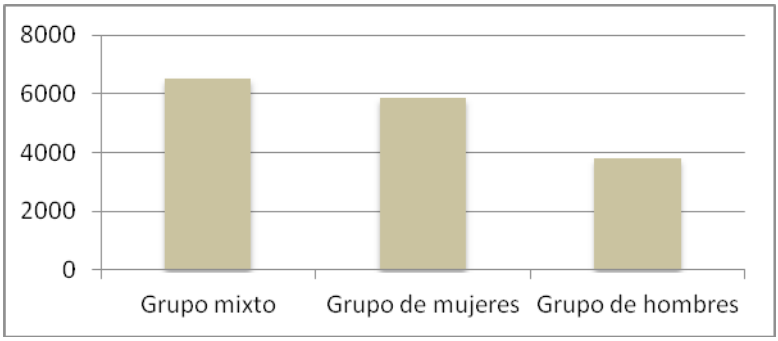

Gráfico 2: Distribución del número de palabras por grupo

Con respecto al número de palabras por sexo de los hablantes, los hombres tienen 7.993 palabras (49%) y las mujeres, 8.182 palabras (51%). En el gráfico 3, se ilustra esta descripción.

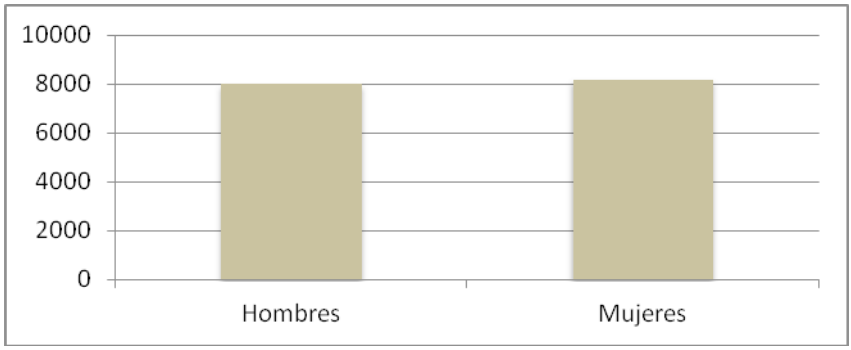

Gráfico 3: Distribución del número de palabras según el sexo de los hablantes

Berber (2000) afirma que hay tres enfoques para definir un corpus según la extensión: i) impresionista, que depende de las posturas que tengan las autoridades y especialistas en el tema; ii) histórico, basados en la observación de corpus usados en los últimos años por los investigadores, iii) estadístico, referidos a los criterios estadísticos y matemáticos. Según el primer enfoque, el CIJU estaría fuera por tener menos palabras que las consideradas como mínimas por especialistas como Leech (1991) y Aston (1997); por el contrario, el CIJU sería considerado como un corpus pequeño por Berber, por tener menos de 80.000 palabras; y estadísticamente, podría decir que el CIJU es pequeño por el número de muestra por grupo.

7. Los insultos en el CIJU

En este apartado describo a grandes rasgos los hallazgos sobre los insultos en el CIJU (Martínez Lara 2006, 2009a y 2009b). Del corpus se extrajeron 511 insultos, groserías y tabúes lingüísticos. Estos enunciados tenían distintos rasgos formales (categorías gramaticales y tipo de cláusulas) y discursivos (funciones) y, según el contexto de enunciación, distintos grados de amenazas. Del total de enunciados con insultos, 74% fueron emitidos por los hombres y 26%, por las mujeres.

En cuanto a los rasgos formales, pueden mencionarse los siguientes:

- Fonético-fonológicos: modificación fonética, ejemplo: *gwón* en vez de *huevón*.
- Morfológicos: modificación por diminutivos, ejemplo: *mamá>mam**ita***. *Marica>mari**quita***.
- Léxicos: los insultos correspondían mayormente a sustantivos (*carajo*, *culito*, *mierda*, *Pantaletica*), adjetivos (*loco*, *maldito*, *menso*, *peludito*) y verbos (*escoñetar*, *joder*).
- Sintácticos: sintagmas (*una vaina*, *de bolas*) y cláusulas (*eres una vaina seria*)

Con respecto a las funciones discursivas de los insultos pueden mencionarse las siguientes:

1. Tipo de enunciados:
   - Enunciados exclamativos, ejemplo: "¡COÑO pero tengo sed!" (G2MB)
   - Enunciados asertivos, ejemplo: "No mi amor/ mi culito huele limpio" (G1MA)
   - Enunciados imperativos, ejemplo: "Púdrete lentamente entre mis ojos" (G3FA)
   - Enunciados interrogativos, ejemplo: "entonces, ¿pa' qué coño (d)e la madre nosotras estamos estudiando universitarios si tenemos que conformarnos con el resto?" (G4FB)

2. Cortesía lingüística:
   - Acto amenazador de la imagen: en el corpus se encontraron actos de habla cuyo objetivo es atacar la imagen positiva del destinatario, como se aprecia a continuación:

     (3)(G5HA)

     Enso: bueno, güevón [[Dirigiéndose a Raúl. No se escucho lo que Raúl le había dicho]]
     Oscar: juega/ juega pues/ juega [[Se dirige al Gordo]]

Gordo: ((pero sin chicote))
Raúl: pero/ ¿por qué? [[Los demás le discuten al Gordo el hecho de que él quiera jugar dudo sin chicote]]
Gordo: porque SÍ
Raúl: **si eres gay [º(güevón)º]**

- Insulto ritual: también se encontraron insultos rituales, es decir, aquellos que no buscaban atacar la imagen positiva del interlocutor, sino mostrar cercanía y camaradería. Muchos con función de vocativo.

(4)(G6HB)
Omar: ¿qué dice Rey?
Rey: ya descubrí el tabúl [[Risas]] dame un segundo para ver
Manuel: ¡ah! Descu- descubrió los papiros
Rey: ¿tú ves?
Omar: Rey sacó los papiros, **güevón**
Saúl: hacía la derecha
Rey: son las columnas
Jesses: (( )) de la derecha, **marico**
Rey: o sea, o sea

3. Grado de amenaza:

Para la investigación, propuse una escala de amenaza compuesta por cuatro (4) grados, a saber: i) *nada amenazante*; ii) *poco amenazante*; iii) *amenazante* y iv) *muy amenazante*. La mayoría de los enunciados tenían un valor de *poco amenazante*, es decir, no atacaban directamente la imagen positiva del destinatario; por el contrario, muchos enunciados con insultos fueron utilizados como formas de tratamiento y camaradería. En segundo lugar, se encuentran los enunciados con insultos cuyo valor era *nada amenazante*; es decir, aquellos que eran mencionados repetitivamente en la interacción sin ser apreciadas como un peligro por los participantes y que tenían una función de muletillas o de marcadores conversacionales. En el tercer lugar están los actos de habla *amenazantes*, aquellos que afectaban la interacción y la sana convivencia, pero no rompían completamente la conversación. Son aquellos que tocaban temas argüidos y producían discusiones. Y, por último, los *muy amenazantes*. Es decir, aquellos que atacaban la imagen positiva del otro y, por tanto, rompían la interacción.

8. A manera de conclusión

El CIJU es un corpus muy pequeño en comparación con otros existentes. Por tal motivo, en este artículo he querido abrir la reflexión sobre cómo denominar a este tipo de materiales, partiendo de los criterios de extensión y

representatividad, sin olvidar otros criterios tales como el tratamiento de los datos y los fines para los que fueron seleccionados.

La representatividad y la extensión de un corpus son muy importantes. Sin embargo, debo señalar que en muchos casos los investigadores no toman en consideración la extensión que debe tener el corpus, sino el objetivo del estudio. En tal sentido, en un primer momento, es más importante para el investigador que las muestras del corpus tengan las características necesarias, mencionadas en este trabajo: i) recolección de textos en entornos naturales, ii) explicitud de los rasgos definitorios, iii) formato digital, iv) etiquetado, v) sustento o procedencia inicial especificada, vi) indicar la disciplina lingüística en la que se inscribe el corpus, vii) textos completos y viii) detallar el tipo de texto: oral, escrito o mixto y el contexto de la situación en que se produjo.

En virtud de esto, considero que al hablarse de la representatividad y la extensión de un corpus para una investigación individual, debe tomarse en cuenta también el problema y el objetivo del estudio. Siendo así, un conjunto de textos con características similares, recogidos y organizados con criterios específicos, previamente establecidos, usados en un estudio y cuyo número de palabra es igual o menor a 80.000, pude denominarse corpus pequeño, ya que permite que el investigador se acerque a una parte de la realidad lingüística de una comunidad.

A pesar de que el corpus contenga menos de 100.000 palabras y sea utilizado en un solo trabajo, el investigador no puede dejar a un lado todos los criterios pertinentes para la elaboración de este tipo de material; más bien debe amoldarse a la metodología exigida, de manera que el corpus cumpla satisfactoriamente el objetivo para el cual fue creado.

El CIJU es un corpus pequeño, en comparación con otros como, por ejemplo, el PRESEEA-Caracas 2004-2011 (Bentivoglio y Malaver 2012), pero se puede considerar representativo de la comunidad juvenil universitaria de la UCV, ya que: i) ha permitido responder algunas preguntas sobre el uso de los insultos, y ii) ha reflejado, a grandes rasgos, el comportamiento léxico de los jóvenes de la UCV. Por lo que puedo afirmar que un corpus, a pesar de ser pequeño, es una herramienta útil que permite observar algunos usos lingüísticos, sobre todo aquellos que no han sido estudiados.

En virtud de todo lo antes planteado, enumero las siguientes inquietudes y recomendaciones:

- Es necesario crear un proyecto institucional para la elaboración de un corpus de interacciones más extenso.
- Además de organizar los grupos por el sexo de los hablantes, también es recomendable hacerlo según la edad.
- Escoger los lugares de grabación, de manera que se evite el exceso de ruido, ya que esto complica el proceso de transcripción.
- Contar con muestras de habla de dos o más universidades de la misma ciudad.

- Tomar en consideración el número de palabras, de muestras y de tipo de textos.
- Poner el corpus a la disposición de otros investigadores.

Notas

1 La lingüística generativa no basa sus estudios en corpus, por el contrario, lo rechaza, puesto que en esta corriente de pensamiento se considera que su objetivo es la lengua en sí misma, en la gramática que posee el hablante-oyente ideal, y que la forma de acceder a esa gramática es a través de la introspección, por lo que no es necesario analizar la lengua mediante corpus, ya que con estos, el lingüista solo da cuenta de las construcciones identificables en estos materiales (cf. Rojo 2002)

2 El *grado de instrucción* es un criterio metodológico asumido por los investigadores del macro Proyecto para el Estudio Sociolingüístico del Español de España y América, cuya sigla es PRESEEA (ver Moreno Fernández 2006) En tal sentido, los corpus más recientes se basan en este criterio y no en el nivel socioeconómico de los hablantes, como en los corpus anteriores. Sin embargo, Guirado (2014) señala que pueden calcularse los índices socioeconómicos de los hablantes de PRESEEA con el fin de estratificarlos según ese criterio y poder así equipararlos con los de corpus anteriores y, de esta manera, hacer estudios comparativos y contrastivos con hablantes de distintos períodos, pero de iguales características socioeconómicas.

3 Los ejemplos que presento a continuación son extraídos del CIJU.

Referencias Bibliográficas

Aarts, J. y Meijs W. (eds.) 1984. *Corpus Linguistics. Recent developments in the use of computer corpora in English language reseach*. Amsterdam: Rodopi.

Aston, G. 1997. Small and large corpora in language learning. Paper presented at the *PALC Conference*, University of Lodz, Poland.

Bentivoglio, P. 1998. La variación sociofonológica. *Español Actual* 69, 1: 29-42.

Bentivoglio, P. y Malaver, I. 2012. Corpus sociolingüístico de Caracas: PRESEEA. Caracas 2004-2010. Hablantes de instrucción superior. *Boletín de Lingüística* 24, 37-38:144-180.

Berber, T. 2000. Lingüística de corpus: histórico e problemática. *D.E.L.T.A.* 16, 2: 323-367.

Biber, D. 2009. A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics* 14, 3: 275-311.

Bolívar, A. 2013. La definición del corpus en los estudios del discurso. *Revista Latinoamericana de Estudios del Discurso* 13, 1: 3-7.

Boretti, S. 2002. Tests de hábitos sociales y la investigación de la cortesía. En D. Bravo (ed.). *Actas del primer coloquio del programa EDICE: La perspectiva no etnocentrista de la cortesía: identidad sociocultural de las comunidades hispanohablantes*, pp. 198-202. Estocolmo: Universidad de Estocolmo.

Briz, A. 2001. *El español coloquial en la conversación*. Barcelona: Ariel.

Cabré, M. T. 2007. Constituir un corpus de textos de especialidad: condiciones y

posibilidades. En M. Ballard y C. Pineira-Tresmontant (eds.). *Les corpus en linguistique et en traductologie*, pp. 89-106. Arras: Artois Presses Université.

Caravedo, R. 1999. *Gramática española: enseñanza e investigación. Apuntes metodológicos: Lingüística del corpus*. Salamanca: Ediciones Universidad de Salamanca.

Charaudeau, P. y Mainguenau, D. 2002. *Dictionnaire d'analyse du discours*. Paris: Éditions du Seuil.

Cheng, W., Warren, M., Xun-feng, X. 2003. The language learner as language researcher: putting corpus linguistics on the timetable. *System* 31: 173-186.

Crystal, D. 1994. *Enciclopedia del lenguaje*.Madrid: Taurus.

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

Gallucci, M. J. 2005. El número de palabras: un nuevo criterio para describir tres corpus del habla de Caracas. *Boletín de Lingüística* 17, 24: 108-121.

Guirado, K. 2014. Corpus Diacrónico del Habla de Caracas. *Boletín de Lingüística* 26, 41-42: 17-42.

Hernández Flores, N. 2002. Los tests de hábitos sociales y su uso en el estudio de la cortesía: una introducción. En D. Bravo (ed.) *Actas del primer coloquio del programa EDICE: La perspectiva no etnocentrista de la cortesía: identidad sociocultural de las comunidades hispanohablantes*, pp. 186-197. Estocolmo: Universidad de Estocolmo.

Leech, G. 1991. The state of the art in corpus linguistics. En K. Aijmer y B. Altenberg (org.). *English Corpus Linguistics. Studies in honour of Jan Svartvik*. London: Longman.

McCarthy, M. 1998. *Spoken language and Applied Linguistics*. Cambridge: Cambridge University Press.

Martínez Lara, J. A. 2006. *Estudio sociopragmático del uso de los insultos*. Trabajo especial de licenciatura de la Escuela de Letras. Caracas: Universidad Central de Venezuela.

Martínez Lara, J. A. 2009a. Los insultos y palabras tabúes en las interacciones juveniles. Un estudio sociopragmático funcional. *Boletín de Lingüística* 21, 31: 59-85.

Martínez Lara, J. A. 2009b. El uso del vocativo como estrategia de cortesía entre jóvenes universitarios de Caracas. Una primera indagación. *Lingua Americana* 13, 25: 100-120.

Monachini, M. y Calzolari, N. 1996. EAGLES *Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages*. Paris: Centre National de la Recherche Scientifique.

Parodi, G. 2008. Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada* 46, 1: 93-119.

Rojo, G. 2002. [Disponible en línea en http://www.bachelor-issen.de/download/bachelor wissen/spanische_sprachwissenschaft/rojo.pdf]. *Lingüística de corpus y lingüística del español* [Consulta: 19 de septiembre de 2015].

Simpson, R., Briggs, S., Ovens, J. y Swales, J. 2002. *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: University of Michigan Press.

Sinclair, J. (Ed.), 1987. *Looking up: an account of the COBUILD Project in lexical computing*. London and Glasgow: Collins ELT.

Sinclair, J. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at work*. Amsterdam/Philadelphia: John Benjamins.

Torruella, J. y Llisterri J. 1999. Diseño de corpus textuales y orales. En J. M. Blecua, C. Sánchez y J. Torruella (eds.). *Filología e informática. Nuevas tecnología en los estudios filológicos*, pp. 45-77. Barcelona: Universidad Autónoma de Barcelona.

José Alejandro Martínez Lara es *Magister Scientiarum* en Lingüística y Licenciado en Letra de la Universidad Central de Venezuela. Es profesor-investigador del Instituto de Filología "Andrés Bello". Sus investigaciones han estado centradas en fenómenos morfosintácticos y de (des)cortesía lingüística. Dicta asignaturas obligatorias de pregrado en la UCV. Ha sido ponente en eventos nacionales e internacionales, sus ponencias han versado sobre (des) cortesía lingüística, morfosintaxis del español y lexicografía, especialmente del habla de Caracas. Ha publicado en revistas científicas arbitradas. Es socio de la ALFAL y de la ALED; y es investigador PEII, nivel A2, del ONCTI.

Correo electrónico: jose.m.lara@ucv.ve