Lexis Vol. XLIII (1) 2019: 195-219

Corpus Oral del Instituto Caro y Cuervo: Reestructuración, diseño y construcción

Ruth Rubio
Julio Bernal
Instituto Caro y Cuervo

Resumen

En este artículo, se describen los avances del proyecto de construcción del Corpus Oral del Instituto Caro y Cuervo, cuyo objetivo es almacenar y sistematizar, en forma electrónica, un corpus oral compuesto por los materiales de audio de tres investigaciones del Instituto: el *Atlas Lingüístico-Etnográfico de Colombia* (ALEC), el *Habla Culta de Bogotá* (HCB) y el *Español Hablado en Bogotá* (EHB). Este texto presenta los fundamentos de la lingüística de corpus y los corpus orales; los parámetros y metodología de las tres investigaciones en las que se recopilaron las muestras y su importancia; el proceso de diseño, reestructuración de los datos y construcción del corpus; las dificultades durante su desarrollo; y las perspectivas futuras.

Palabras clave: corpus, corpus oral, lingüística de corpus, reestructuración de datos

Abstract

The present article describes the advances of the Instituto Caro y Cuervo Spoken Corpus Project. The investigation aims to stored and systematize electronically a spoken corpus composed of the audio files of three

https://doi.org/10.18800/lexis.201901.006

ISSN 0254-9239

investigations of the Institute: The *Atlas Lingüístico-Etnográfico de Colombia* (ALEC), the *Habla Culta de Bogotá* (HCB) and the *Español Hablado en Bogotá* (EHB). This text presents the foundations of corpus linguistics and oral corpora; the parameters and methodology of the three investigations in which the oral samples were collected and their importance; the design process, data restructuring and corpus construction; the difficulties during its development; and the future perspectives.

Keywords: Corpus, spoken corpus, corpus linguistics, data restructuring

1. Introducción

El grupo de investigación de Lingüística de Corpus y Computacional del Instituto Caro y Cuervo (LICC) emprendió, en el año 2013, la tarea de desarrollar un corpus oral que permitiera la sistematización, conservación y divulgación de los archivos de audio de tres investigaciones del Instituto Caro y Cuervo (ICC): el *Atlas Lingüístico-Etnográfico de Colombia* (ALEC), el *Habla Culta de Bogotá* (HCB) y el *Español Hablado en Bogotá* (EHB). Estos archivos contienen más de 1000 horas de grabación de entrevistas realizadas entre 1955 y 1992, con muestras de, aproximadamente, 2400 hablantes de diferentes edades, profesiones, estratos socioeconómicos, etc. Además, cuenta con datos de carácter cultural e histórico, manifiestos en diversos testimonios y relatos que reflejan las costumbres y vida cotidiana de la época. La rigurosidad de las metodologías empleadas para la recolección de las muestras posibilita la realización de investigaciones, tanto en el ámbito de la lingüística como en otras áreas del conocimiento, por lo cual, es fundamental poner estos datos a disposición del público.

La Lingüística de Corpus (LC) conforma un conjunto o colección de principios metodológicos para el estudio de las lenguas y el lenguaje, y se destaca por brindar sustento a la investigación de la lengua en uso a partir de corpus (Parodi 2008). Estos principios metodológicos son base para la recolección, almacenamiento,

organización y explotación de muestras de la lengua en uso. En nuestro caso, la LC constituye una guía para mudar de un formato de investigación a otro, lo que facilitará la reestructuración, la sistematización, la conservación y la divulgación de los materiales de las investigaciones mencionadas. En este artículo se abordan inicialmente la definición y tipología de los corpus, especialmente los orales. Se enumeran algunos corpus orales en español. Posteriormente, se presentan y explican los principios básicos de las investigaciones y su importancia. Luego, se expone el proceso de reestructuración de los datos y las muestras, la sistematización, y los procesos de construcción del corpus. Para continuar, se describe el corpus y las dificultades y retos para su constitución. Finalmente, se explican las perspectivas futuras.

2. Marco referencial

2.1. Lingüística de corpus y corpus orales

El uso de muestras reales de lengua para los estudios lingüísticos, así como las herramientas tecnológicas que automatizan y plantean nuevos procesos para el análisis del lenguaje, han propiciado un fuerte interés por el desarrollo de recursos lingüísticos, entre los cuales sobresalen los corpus. Los corpus son colecciones de datos (orales o escritos) sistematizados en relación con unos criterios y objetivos determinados. Sinclair (1996) los define de la siguiente manera: "A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (171). Por su parte, Gries (2009) señala que los corpus deben presentar las siguientes características: estar estructurados para ser manipulados y procesados por una máquina (*machine-readable*); estar compuestos por datos producidos en una situación real de comunicación; y ser representativos y balanceados, es decir, incluir muestras equilibradas de las variedades objeto de estudio del corpus. En otras palabras, un corpus debe ser una muestra representativa de la lengua, estar formado por textos producidos en situaciones de comunicación, tener criterios lingüísticos para su

organización, ser de naturaleza computacional y permitir un fácil acceso a los datos.

Por otro lado, la tipología de los corpus puede establecerse desde diferentes criterios: el diseño, los métodos utilizados para su constitución, el tamaño, las lenguas de los textos que lo componen y, en el caso del Corpus Oral del ICC, el medio de producción de los textos (orales o escritos). Los corpus orales son aquellos que tienen como objetivo caracterizar desde un punto de vista lingüístico la lengua hablada. Estos corpus se constituyen por transcripciones ortográficas o fonéticas, audios o, en muchos casos, las grabaciones con sus respectivas transcripciones. Llisterri (2003) señala que suele plantearse una disparidad entre los corpus orales o bases de datos orales (*speech corpora, speech databases*) que recogen la señal sonora y los corpus de lengua oral (*spoken language corpora*) que recopilan transcripciones de la lengua escrita. Esta diferenciación se plantea principalmente para determinar los fines de construcción del corpus: los creados para los trabajos en tecnologías del habla (corpus orales) y los usados para estudios lingüísticos (corpus de lengua oral). En nuestro caso, la recopilación inicial de las muestras se realizó con propósitos lingüísticos. Aun así, no descartamos que, debido a su calidad auditiva, algunas grabaciones puedan ser utilizadas con otros fines. Además, el corpus cuenta con audios y transcripciones ortográficas, de ahí que no tengamos en cuenta la distinción planteada por Llisterri y hablemos de corpus orales de forma general para referirnos a los corpus compuestos por transcripciones, audios o ambos, que pueden ser usados para la investigación lingüística, las tecnologías de habla u otras áreas de investigación.

Ahora bien, un corpus puede componerse por subcorpus y componentes: los primeros se refieren a las divisiones que se efectúan dentro del corpus general, y corresponden a un conjunto de datos con características similares; los segundos hacen referencia a colecciones de muestras de la lengua o de textos que comparten un criterio lingüístico, sociolingüístico o cultural específico (Torruela y Llisterri 1999). El Corpus Oral del ICC está compuesto por tres

subcorpus que se relacionan con cada una de las investigaciones: el ALEC, el HCB y el EHB. Con respecto a los componentes, el corpus cuenta con varias colecciones de grabaciones que se pueden organizar de acuerdo con criterios como, por ejemplo, la ubicación geográfica.

Con el objetivo de obtener varias referencias y explorar las estructuras y el manejo de datos en corpus orales de español, realizamos una revisión inicial de corpus en esta lengua. En la tabla 1, presentamos algunos corpus orales públicos, disponibles en línea, que pueden servir de guía sobre los proyectos que se han llevado a cabo en relación con la lengua oral española.

2.2. Contextualización de las investigaciones

Las investigaciones en las que se recopilaron las muestras del Corpus Oral se centraron en el estudio del español de Colombia, y sus materiales se constituyen en una de las principales colecciones de registros de lengua recogidos con propósitos lingüísticos en este país durante el siglo XX. Las características, objetivos principales e importancia de cada uno de estos proyectos son los siguientes:

1. El Atlas Lingüístico-Etnográfico de Colombia (ALEC) es producto de una extensa investigación realizada entre 1955 y 1983 con el objetivo de construir un atlas lingüístico que presentara las variedades léxicas del español de Colombia. El ALEC recogió información de 264 localidades distribuidas a lo largo del territorio nacional, a través de encuestas indirectas basadas en un cuestionario final de 1348 preguntas sobre temas como la vivienda, el vestuario, la ganadería, la vida religiosa, la política, etc. La publicación impresa del ALEC se realizó entre 1981 y 1983, en 6 tomos que contienen 1500 mapas con variaciones léxicas y fonéticas; un manual; un glosario lexicográfico; un índice alfabético; y un suplemento con muestras de habla espontánea, canciones y dos discos de acetato con contenido etnográfico.

Tabla 1
Corpus orales del español

Corpus	Descripción
Corpus de Referencia del Español Actual (CREA)[1]	Está compuesto por más de 160 millones de formas. Contiene datos textuales y orales del español de los años 1975 a 2004. Lo oral está representado por transcripciones de audios obtenidos principalmente de radio y televisión.
Corpus Oral y Sonoro del Español Rural (COSER)[2]	Está compuesto por 1050 horas de grabación y 1739 informantes hasta el 2013. Contiene datos orales del español rural de España. Las grabaciones se han realizado desde 1990 y aumentan anualmente. La mitad de las grabaciones tienen transcripción ortográfica y fonética.
el Proyecto para el Estudio Sociolingüístico del Español de España y de América (PRESEEA)[3]	Está compuesto por entrevistas semidirigidas de unos 45 minutos de duración de una amplia muestra de ciudades de mundo hispano. La selección de los informantes se realizó a partir de criterios sociolingüísticos para conseguir una muestra estratificada. El corpus cuenta con la transcripción de la quinta parte de las grabaciones que corresponden a 100.000 palabras.
Corpus Oral Didáctico Anotado Lingüísticamente (C-Or-DiAL)[4]	Está compuesto por 240 grabaciones de habla espontánea de hablantes de Madrid. También cuenta con 240 transcripciones y sus respectivos metadatos.
Corpus de conversación coloquial-Grupo Val.Es.Co (Valencia español coloquial)[5]	Está compuesto 341 horas de grabación y transcripción ortográfica de grabaciones de conversaciones coloquiales en el entorno familiar del informante.
Corpus para el estudio del español oral (ESLORA)[6]	Contiene 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2014. Los registros sonoros se transcribieron ortográficamente con alineación texto-voz para facilitar un acceso cómodo al audio desde la transcripción, y viceversa.
Corpus Oral de Lenguaje Adolescente (COLA)[7]	Recoge el habla de adolescentes (13 y 19 años) de Madrid, Argentina, Santiago de Chile y Managua.
El corpus del habla en Almería[8]	Está compuesto por 108 entrevistas semidirigidas realizadas a hablantes de diversas edades y niveles socioculturales.

[1] Para una revisión del corpus, véase Real Academia Española: Banco de datos (s/f).
[2] Para una revisión del corpus, véase Fernández (2005-2016).
[3] Para una revisión del corpus, véase PRESEEA (2014-2016).
[4] Para una revisión del corpus, véase Laboratorio de Lingüística Italiana (s/f).
[5] Para una revisión del corpus, véase Grupo de Investigación Val.Es.Co. (1995-2016).
[6] Para una revisión del corpus, véase Grupo de Gramática del Español de la Universidad de Santiago de Compostela (2014-2016).
[7] Para una revisión del corpus, véase Universidad de Bergen (s/f)
[8] Para una revisión del corpus, véase Grupo ILSE (s/f).

De esta manera las grabaciones del ALEC se consolidan en importantes muestras para la realización de estudios sobre el español de Colombia y su diversidad lingüística. También, los relatos, historias, cantos, descripciones, entre otros, pueden servir para el estudio de la historia, la cultura, las costumbres y la vida rural de Colombia entre 1955 y 1978.

2. El Habla de la Ciudad de Bogotá o Habla Culta de Bogotá (HCB) es una investigación que se realizó en el marco del proyecto vinculado al "Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica", cuyo principal promotor fue M. Lope de Blanch. El objetivo central de la investigación fue estudiar "el habla culta media (habitual), con referencias a las actitudes formal (habla esmerada) e informal (habla familiar)" (Spitzová 1991: 62). La recopilación de las encuestas se realizó entre 1972 y 1984 a partir del "Cuestionario para el estudio coordinado de la norma lingüística culta", organizado en tres cuadernillos de acuerdo con los niveles de lengua: fonético y fonológico, morfosintáctico y léxico. Algunas de las publicaciones producto de esta investigación son las siguientes: Los libros de González y Otálora, *Muestras del habla culta de Bogotá* de 1985 y *El habla de la ciudad de Bogotá* de 1986, y el libro de Otálora *Léxico del habla culta de Santafé de Bogotá* de 1997. Las grabaciones del HCB suponen un material de apoyo para la descripción y el análisis del habla culta de la ciudad Bogotá de este periodo, y un archivo de testimonios sobre las costumbres bogotanas, la estructura de la ciudad, la educación y la descripción de varios hechos políticos e históricos importantes como el 9 de octubre o el Frente Nacional.

3. El Español Hablado en Bogotá (EHB) se inició en 1987 como una iniciativa del Departamento de Dialectología del ICC para realizar un trabajo enfocado en "dialectología urbana con métodos sociolingüísticos" (Montes et al. 1998: 18). La recolección de las muestras se realizó entre 1990 y 1992 en 60 barrios de Bogotá, distribuidos de forma equitativa entre los puntos cardinales de la capital. Se diferenciaron tres niveles educativos: "nivel alto (universitarios y especializados), nivel medio (bachillerato hasta carreras

intermedias), nivel bajo (analfabetos o primaria)" (Montes et al. 1998: 735). Algunos de los trabajos publicados sobre este tema son los libros de Montes y colegas, *El español hablado en Bogotá. Relatos semilibres de informantes pertenecientes a tres estratos sociales* de 1997 y *El español hablado en Bogotá. Análisis previo de su estratificación social* de 1998.

Los audios del EHB son base para el estudio sociolingüístico del español de Bogotá. Sus muestras y los datos que las acompañan pueden servir para estudios lingüísticos en distintos niveles de lengua. Como lo mencionan Montes et al.

> Estamos seguros de que este conjunto de materiales será recibido con satisfacción por investigadores de varias ramas, en primer lugar, por supuesto, por los estudiosos del español, de su estructura y sus variantes dialectales, pero también por lingüistas de más amplios, intereses (textolingüística, gramática del texto, oralidad, etc.), e incluso por sociólogos y políticos para quienes las diversas historias narradas en estas cintas, serán de interés por constituir un reflejo bastante fiel de la profunda crisis social que Bogotá y desde luego el país, han soportado en estos últimos años (1997).

Adicionalmente, destacamos que como varias de las muestras fueron recogidas a través de la encuesta indirecta pueden ser valiosas para el estudio del español hablado, la entonación, el uso de interjecciones, la toma de turnos, los solapamientos, las interrupciones, entre muchos otros aspectos relaciones con el español oral.

3. El corpus

3.1. Diseño y reestructuración de los datos

En primer lugar, es importante resaltar que si bien los datos y muestras recogidas durante las investigaciones cumplen con la mayoría de criterios para su constitución en corpus, es decir, textos producidos en situaciones de comunicación con criterios lingüísticos bien definidos para su recolección y que componen una muestra de un tipo de lengua específico. En la actualidad, una de las características principales de los corpus es la posibilidad de tratamiento a través de

herramientas tecnológicas que permitan fácil acceso y explotación de los datos. De esta manera los archivos de audio e información de las investigaciones han pasado por varios procesos de actualización y reestructuración para garantizar su constitución en un corpus oral informatizado, sin olvidar la naturaleza inicial y criterios de cada una de las investigaciones en que fueron recogidos. Estos pasos se podrían comparar con el de un diccionario que pasa de un formato impreso a un formato digital interactivo.

Una primera fase del proceso consistió en la digitalización de los archivos de audio que se encontraban almacenados en cintas de carrete abierto y casetes. Este proceso fue realizado por la Fundación Patrimonio Fílmico Colombiano, que entregó los archivos de audio digitalizados en formato *wav* y almacenados en discos duros y CD con la misma organización que tenían en su primera versión de almacenamiento. Posteriormente, la Biblioteca del ICC realizó un primer proceso de sistematización en una tabla de Excel que distribuía los datos en relación con tres categorías: datos completados y corregidos por la Fundación Patrimonio Fílmico, datos verificados y corregidos por los investigadores del ICC y datos completados por la Biblioteca del ICC. Al respecto, también es importante mencionar que la Biblioteca es la que se encarga actualmente de almacenar los audios en cintas de carrete abierto y casetes, como también, los discos duros y CD. El almacenamiento y organización de estos archivos siguen los criterios archivísticos de documentación planteados por la biblioteca para garantizar su conservación y consulta.

La siguiente fase se fundamentó en la creación de copias de seguridad y de trabajo para garantizar la conservación y facilitar la manipulación de los archivos digitalizados. Una vez realizado este proceso, los investigadores del grupo LICC, un lingüista con estudios en ingeniería y profesionales en humanidades, iniciaron la exploración del material y la búsqueda de información de cada subcorpus. Esto sirvió para plantear varias propuestas de metadatos, a partir de las cuales se definió el conjunto general de metadatos del Corpus Oral del ICC, los nombres de archivo y la organización de

las tablas de Excel dinámicas para su posterior migración a una base de datos.

Por otro lado, para la organización y delimitación de las muestras, los archivos de audio se fraccionaron por sesiones con el criterio de acto comunicativo, es decir, por temáticas específicas de conversación entre el encuestador y el informante en un periodo continuo de grabación. Generalmente, cada sesión está compuesta por una encuesta o entrevista de una duración variable de acuerdo con cada subcorpus y el tipo de entrevista. Los metadatos se organizaron en un documento de Excel con cuatro hojas de cálculo, como se observa en la tabla 2.

Como se puede ver en la tabla 2, hay varios metadatos que son específicos de un solo subcorpus: por ejemplo, los datos de estrato social y barrio se utilizan únicamente en EHB. También hay metadatos que varían de acuerdo con cada subcorpus. En el caso de "lugar", la localidad en el subcorpus ALEC hace referencia a los lugares de Colombia en los que se realizaron las entrevistas; y en EHB se relaciona con la localidad de Bogotá donde se encontraba ubicado el barrio en el que se realizó la encuesta.

Los datos técnicos de la hoja de documentos fueron extraídos de manera automática y masiva por medio de una de las aplicaciones desarrolladas por el grupo LICC: Herramienta para la extracción masiva de datos en archivos *wav* (HEMADAW). Esta aplicación utiliza la librería *PyQT5*[9] para crear la plataforma gráfica y la librería *wavio*[10] para manipular archivos *wav* de 24 bits, lo que permite seleccionar un directorio en la máquina del usuario, buscar todos los archivos *wav* que hay dentro del directorio y los subdirectorios, y generar un archivo separado por comas que contiene los datos técnicos de cada audio.

[9] Se puede acceder a ella a partir del siguiente enlace: https://riverbankcomputing.com/software/pyqt/download5

[10] Se puede acceder a ello a través del siguiente enlace: https://github.com/WarrenWeckesser/wavio

Tabla 2
Descripción del documento general de metadatos

Hoja	Descripción de los datos	Metadatos
Documentos	Datos técnicos y de archivo de la cinta	Identificación del archivo Ubicación en el sistema Tamaño en bytes Duración en segundos Número de canales Profundidad de muestreo Frecuencia de muestreo Comprensibilidad (Dato subjetivo, aportado por la persona que escucha el audio) Comentarios
Informantes	Datos del informante o informantes encuestados en la grabación	Identificación del archivo Identificación del informante Nombres y apellidos Descripción Edad, Sexo, Estrato (en EHB) Lugar de nacimiento, Fecha (en EHB) Nivel educativo, profesión u oficio (Varía de acuerdo a cada subcorpus) Procedencia del padre y de la madre Otras lenguas y viajes (en HCB) Estudios no sistematizados (en HCB) Lecturas habituales (en HCB) Estudios y ocupación del padre, la madre o cónyuge (en HCB)
Sesiones		Identificador del archivo Identificador del informante Identificador del encuestador Título Descripción Fecha Lugar de la grabación: departamento, localidad, barrio, etc. (varía de acuerdo a cada subcorpus Tipo de encuesta (varía de acuerdo a cada subcorpus) Temas
Fragmentos		Nombre del archivo Tiempo de inicio y finalización de la cabecera[11] (en EHB y HCB) Información sobre las cintas que se deben pegar o cortar

[11] La cabecera o cabezote hace referencia a la primera parte del audio en la que se graba la información sobre la muestra y el informante, por ejemplo, la fecha de la grabación, el nombre del informante, etc.

La hoja de fragmentos registra principalmente los tiempos de inicio y finalización de las cabeceras, lo que permitirá que posteriormente una herramienta automática los elimine al presentarlos al usuario, protegiendo el anonimato de los hablantes. Asimismo, en caso de que un archivo de audio contenga varias sesiones o que dos archivos contengan una sola sesión, se ingresa el tiempo de inicio y finalización de estos fragmentos, y la tarea específica necesaria (cortar, pegar). De esta manera, se podrá realizar de forma automática la tarea requerida de acuerdo con el archivo de audio y la sesión. Además de las tablas principales de metadatos, se crearon tablas anexas para documentar información adicional de cada subcorpus, como, por ejemplo, las grabaciones repetidas. En el subcorpus HCB se creó un documento que contiene las mismas cuatro hojas de cálculo con información de la segunda parte de las grabaciones, y se asignó un código que las asocia con la primera parte.

Una vez definidos los metadatos, se realizó un proceso de ingreso y sistematización de los audios en el que se escuchaban las grabaciones una a una y se ingresaba la información a las tablas de Excel. Para facilitar este proceso y mantener la seguridad y el cuidado de los audios, se utilizó *Ampache*[12], una aplicación web basada en el servidor *web Apache*, los lenguajes de programación *PHP* y *Javascript* y la base de datos *MySQL*. *Ampache* permite la creación y consulta de catálogos de audio y video y posee una interfaz muy parecida a la de otros servicios para escuchar música en línea, como *Spotify*. Aunque los metadatos que maneja esta aplicación están pensados para la organización de catálogos musicales, se pueden utilizar para generar estructuras básicas de un corpus lingüístico, al menos para uso interno de los investigadores. Otras de las herramientas usadas para el trabajo con los audios fueron las aplicaciones *Audacity* y *PRAAT*. Al finalizar el ingreso de metadatos se realizó un proceso de revisión y normalización de las tablas con el fin de corregir posibles errores

[12] Se puede acceder a dicha aplicación a través del siguiente enlace: http://ampache.org

durante el ingreso y garantizar que los datos fueran homogéneos. Asimismo, constantemente se examinó y comparó la información con los datos archivísticos básicos que registraron la biblioteca del ICC y la *Fundación Patrimonio Fílmico Colombiano* durante la etapa de digitalización.

Hasta ahora, todos los procesos que hemos mencionado están relacionados con el diseño previo y reestructuración de los datos. Sin embargo, para que estos se fundamenten en un corpus electrónico deben pasar a una fase de desarrollo tecnológico que garantice el fácil acceso y explotación de los datos. Esta fase aún está en proceso de desarrollo.[13] Básicamente, consiste en diseñar y desarrollar una base de datos flexible con base en los metadatos planteados, y una plataforma conformada por la interfaz de visualización del usuario y un motor de búsqueda del corpus. Mientras se finaliza este proceso, las búsquedas básicas de las grabaciones se realizan con las tablas de metadatos y el sistema de búsqueda de *Ampache.*

3.2. Descripción del corpus y subcorpus

El corpus oral del ICC está compuesto por tres subcorpus[14] que contienen 2010 grabaciones recopiladas entre 1955 y 1992. En la tabla 3 se encuentra el número de grabaciones, los años de recolección y el número de hablantes de cada subcorpus.

[13] De la fecha de construcción del artículo al presente año (2019) se diseñó y desarrolló una plataforma de consulta para los corpus del Instituto denominada Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) que se puede consultar en el siguiente enlace: clicc.caroycuervo.gov.co

[14] Es importante resaltar que en la plataforma actual (CLICC) estos subcorpus no hacen parte de un solo corpus oral, sino que son independientes.

Tabla 3
Composición de cada uno de los subcorpus

Subcorpus	Número de archivos	Años de recopilación	Hablantes
Atlas Lingüístico-etnográfico de Colombia (ALEC)	763 grabaciones	1955-1978	1176
Habla culta de Bogotá (HCB)	600 grabaciones principales 124 grabaciones que complementan las grabaciones principales	1972-1984	760
Español hablado en Bogotá (EHB)	522 grabaciones	1990-1992	522

3.2.1. Subcorpus del Atlas lingüístico-etnográfico de Colombia – ALEC

El subcorpus ALEC está formado por 763 grabaciones que pueden tener una duración de entre 2 y 60 minutos. La recopilación de las muestras fue realizada en 264 localidades por 23 encuestadores que se desplazaban a dichas zonas para realizar las grabaciones. La colección contiene encuestas fonéticas, relatos, entrevistas o muestras folclóricas relacionadas con los siguientes campos semánticos:

- Alimentación
- Animales Domésticos
- Animales Silvestres
- El Campo - cultivo y otros vegetales
- Cuerpo humano
- Embarcaciones y Pesca
- Familia y ciclo de vida
- Festividades y distracciones
- FONÉTICA
- Ganadería
- GRAMÁTICA
- Instituciones y vida religiosa
- Industrias relacionadas con la agricultura
- Oficios y empleo

- Transporte
- Tiempo y espacio
- Vivienda
- Vestido
- ONOMÁSTICA

El ALEC cuenta con 1176 informantes de edades entre los 5 y 100 años (Ver tabla 4). Durante la investigación se buscaba que los hablantes fueran nativos de su localidad y que tuvieran un mismo nivel sociocultural, es decir: trabajadores del campo o relacionados, con niveles similares de formación académica, que estuvieran entre los 40 y 60 años. En las grabaciones, sin embargo, los hablantes no siempre cumplen con estas características: se pueden encontrar personas de diversas edades con estudios superiores a la primaria, e incluso con cargos no relacionados con la vida rural. Esto también se debe a que, en varias grabaciones, cuando se estaba entrevistando a una persona, llegaban otras e intervenían en la conversación.

Tabla 4
Distribución de los informantes del subcorpus ALEC

Edad		Nivel educativo		Sexo	
5-15 años	87	Analfabeto(a)/sin escolarización	56	Femenino	425
16-35 años	66	Primaria o algunos años de escuela primaria	125	Masculino	689
36-55 años	180	Secundaria o algunos años de secundaria	18	No identificado	62
56-75 años	126	Superior o universitaria	12		
75 en adelante	33	Estudiante	131		
No identificado	684	No identificado	814		
		Lee y escribe	19		
Total	1176	Total	1176	Total	1176

3.2.2. Subcorpus del Habla de Culta de Bogotá-HCB

El subcorpus HCB está compuesto por 1360 archivos de audio que contienen fragmentos y grabaciones de 600 encuestas recopiladas por 21 encuestadores principales entre 1972 y 1984 en la ciudad de Bogotá. Las encuestas están organizadas de acuerdo con 4 tipos de entrevista: diálogo entre informante y encuestador, diálogo entre dos informantes, grabación secreta en un diálogo espontáneo y elocuciones en actitudes formales como clases, conferencias, discursos, etc. De acuerdo con la investigación, las encuestas debían ser de 30 minutos, por lo cual, aunque la grabación tuviera una duración mayor, se cortaba para cumplir con los requerimientos del proyecto. En consecuencia, la mayoría de los archivos tienen una duración de entre 28 y 32 minutos, y algunos de estos tienen una segunda parte de entre 2 y 30 minutos que hace referencia al trozo que fue eliminado.

El subcorpus HCB cuenta con 759 hablantes (ver tabla 5) de tres generaciones: la primera de 25 a 35 años, la segunda de 36 a 55 años y la tercera de 56 años en adelante. Según González y Otalora (1986), las características que se tuvieron en cuenta para la elección de los informantes son las siguientes:

- Factores socioculturales: ambiente familiar, tanto paterno como conyugal; instrucción recibida a través de estudios regulares o asistemáticos; profesión u ocupación; y viajes y otras experiencias culturales.
- Requisitos: ser nacido o residente en la ciudad de objeto de estudio desde los cinco años; haber vivido en ella al menos durante las tres cuartas partes de su vida; ser hijo de hispanohablantes, preferentemente nacidos en la misma ciudad; y haber recibido sus instrucciones primaria y superior en la propia ciudad.

Tabla 5
Distribución de los informantes del subcorpus HCB

Edad		Nivel educativo		Sexo	
25 a 35 años	317	Primarios	0	Femenino	361
36 a 55 años	281	Secundarios	102	Masculino	387
56 años en adelante	136	Superiores	619	No identificado	11
No identificado	25	No identificado	38		
Total	759	Total	759	Total	759

3.2.3. Subcorpus del Español Hablado en Bogotá - EHB

El subcorpus EHB cuenta con 522 grabaciones recogidas entre 1990 y 1992, de las cuales 241 contienen relatos semilibres de entre 45 y 60 minutos, y 281 son cuestionarios fonéticos de entre 10 y 30 minutos. Los relatos semilibres son conversaciones entre un encuestador y un informante en las que el primero busca obtener una narración espontánea e intervenir lo menos posible. Los cuestionarios fonéticos siguieron un sistema similar al de las encuestas léxicas del ALEC, es decir, elicitación a partir de definiciones y sinónimos para la recogida de datos (Montes et al. 1998). En la tabla 6 se pueden observar algunos ejemplos del tipo de preguntas que se hacía a los hablantes para las encuestas fonéticas.

Las grabaciones fueron realizadas en 60 barrios organizados por niveles socioeconómicos: alto, medio y bajo. Para la selección del nivel socioeconómico se tuvo en cuenta la formación académica de los hablantes y la elección de los barrios se realizó con base en el *Plano estratificado de los barrios de Bogotá* del DANE de 1981 y la investigación de las sociólogas Medina y Rincón titulada *Estratificación social de la ciudad de Bogotá, D.C.* de 1985 (Montes et al. 1998).

El subcorpus EHB cuenta con 477 informantes, hombres y mujeres entre los 15 y los 60 años de edad, nacidos en Bogotá o con más de 15 años viviendo en ella. En la mayoría de los casos los informantes son los mismos para los dos tipos de grabación, y las entrevistas fueron realizadas el mismo día o en días cercanos.

Tabla 6
Ejemplo de variables y preguntas de una encuesta fonética del EHB

Variable (expresada en ortografía)	Palabra a elicitar	Pregunta inicial	Sugerencia de último recurso
S	Asno	¿Qué otro nombre se le da al burro?	Comienza por A y termina en no
RR	Rana	¿Cómo se llama el animal que se parece al sapo?	
LL/Y	Gallina	¿Cómo se llama el ave que pone los huevos que nos comemos?	
LL/Y	Yema	¿Cómo se llama la parte amarilla del huevo?	
R	Carne	¿En las famas uno compra la ____?	
F	Café	¿Cuál es el producto que más exporta Colombia?	
P	Pastas	Los fideos, los espaguetis, los ravioles, son ¿qué?	
R	Trigo	¿Cómo se llama la planta que no es el maíz de la cual se saca la harina para hace el pan?	
¿Acentuación?	Papa	¿Qué nombre recibe el jefe máximo de la iglesia católica que vive en roma?	
F/(¿S?)	Frutas	La manzana, la pera...son ¿qué?	
Extranjerismo	Sándwich	A un pedazo de queso o de jamón metido entre dos rebanas de pan ¿cómo se le llama?	
Grupo PT	Egipto	¿Cómo se llama el barrio Bogotano donde se hace cada año la fiesta de los reyes magos a lo vivo?	
Grupo TR/R	Transversales	¿Por dónde circulan los carros en la ciudad que no son calles ni carreras?	
S	Seis	¿Cuánto es cuatro más dos?	
Grupo XTR	Extraordinario	En navidad hay un sorteo especial de la lotería, es el sorteo…	
BJ	Objeto	¿De qué otra forma se puede decir cosa?	

Tabla 7
Distribución de los informantes del Subcorpus EHB

Edad		Sexo		Procedencia		Nivel educativo	
15-34	289	Masc.	230	Nativos	263	Analfabetos-primaria	184
35-59	150	Femen.	247	Inmigrantes	214	Bachillerato-Carrera intermedia	242
60-	38	Total	477	Total	477	Universitarios-	51
Total	477					Total	477

Nota. Información adaptada de Montes et al. (1998: 19)

4. Dificultades para el desarrollo del corpus y necesidades actuales

La elaboración del Corpus oral del ICC ha tenido una serie de dificultades de diferente naturaleza que se pueden distribuir de la siguiente manera: dificultades metodológicas, dificultades tecnológicas y de almacenamiento, y dificultades de carácter legal.

En el primer grupo cabe mencionar las dificultades de compatibilidad de los tres subcorpus, pues como se ha evidenciado cada uno contiene muestras recopiladas con distintos objetivos y metodologías. Por tanto, fue necesario revisar los puntos en común y disparidades, especialmente para facilitar consultas ágiles y amables cuando se desarrolle la plataforma de búsqueda. De esta manera, al definir la tabla de metadatos fue necesario juntar varios datos en una sola categoría, aunque no coincidieran de forma exacta en cada subcorpus. Asimismo, fue ineludible dejar algunos metadatos particulares que solo aparecen en un solo subcorpus por su importancia, por ejemplo, el estrato socioeconómico en el EHB.

Con respecto al segundo grupo, para iniciar, es importante indicar que la institución no estaba preparada para almacenar cantidades de información como las del corpus, dado que los archivos de audio son pesados y deben estar en distintos formatos (wav y mp3) para garantizar que los usuarios puedan acceder a estos de acuerdo con sus necesidades, sean estas de consulta general, académicas,

de formación o para labores investigativas. Por lo cual, fue fundamental solicitar servidores de alta capacidad y seguridad y contar con discos externos y aplicaciones como *Ampache* para el fácil acceso a los audios por parte de los investigadores.

Adicionalmente, a pesar de que la *Fundación Patrimonio Fílmico Colombiano* es una institución especializada en el área y, por tanto, implementó los protocolos requeridos para la digitalización del material patrimonial de audio, muchas de las grabaciones originales se realizaron con tecnología de mitad del siglo XX y en condiciones ambientales y de ruido que no favorecen su optima comprensibilidad. Sumado a esto, medio siglo de manipulaciones de los carretes y casetes incidieron en cierto nivel de deterioro, y varias cintas quedaron digitalizadas con una velocidad diferente a la de su formato original. Todo esto lleva a afectar la calidad de los audios, muchos de ellos tendrán que pasar por una etapa de reparación de tal forma que sirvan para estudios en los diferentes niveles de análisis de lengua.

Asimismo, debido a los diferentes procesos de sistematización, cambios de formato y catalogación de los archivos del corpus, que responden a lógicas de épocas y circunstancias disímiles, varias grabaciones no corresponden con la marcación que tenían en catálogos anteriores. Por este motivo, se evidencian problemas como la repetición de audios, la desaparición de archivos y la falta de datos para clasificar las grabaciones. Por otro lado, hay archivos de audio que contienen varias entrevistas en una sola cinta o cintas que contienen un acto comunicativo que continúa en otros audios. Este hecho requiere un trabajo para seccionar y pegar los contenidos, con el fin de que el usuario final del Corpus Oral del ICC pueda escuchar las sesiones completas.

Por otro lado, una gran falencia en la elaboración del Corpus ha sido la falta de especialistas en el área de ingeniería, más específicamente en lingüística computacional, para el desarrollo de la interfaz de consulta. Esto nos ha llevado a una dinámica de ensayo-error, de implementación de programas y aplicaciones de *software* libre que no están interrelacionadas de manera sistemática y coherente, y que

generan la necesidad de programar o crear nuevas aplicaciones que interconecten las que hemos usado. Tal vacío ha impedido que tengamos un diseño final que permita a los usuarios acceder a los audios a través de una interfaz que permita búsquedas estándares y complejas.

Finalmente, en el tercer grupo se encuentran las dificultades relacionadas con los aspectos legales de la información que se archiva en el Corpus oral del ICC relacionadas con el anonimato de los informantes y la carencia de permisos documentados para el acceso público a los audios. La época en las que se recogieron las muestras no había políticas sobre la recolección de permisos de uso, por ende, se requiere de asesoría profesional, ya que se conjugan leyes de patrimonio lingüístico y cultural y épocas con políticas inexistentes o poco claras respecto al tema. Por ahora, es importante mencionar que se realizarán todos los procesos necesarios para garantizar el anonimato de los informantes en los subcorpus EHB y HCB, pues en el subcorpus ALEC todos los datos siempre han sido públicos. Asimismo, cabe resaltar que la información y muestras solo se usan con fines académicos y se solicita a los investigadores y usuarios de los audios firmar compromisos de manejo de privacidad y uso solo con los fines permitidos.

5. Perspectivas futuras

En cuanto a los retos del futuro, uno de los objetivos principales para el desarrollo del Corpus es la consolidación de la base de datos, la plataforma de búsqueda y la interfaz gráfica para usuarios finales. Esto permitirá que los interesados puedan acceder al corpus fácilmente, con el uso de determinados criterios de búsqueda que estén relacionados con sus intereses. Para lograr este objetivo es fundamental el trabajo conjunto entre ingenieros, desarrolladores y lingüistas en todos los procesos.

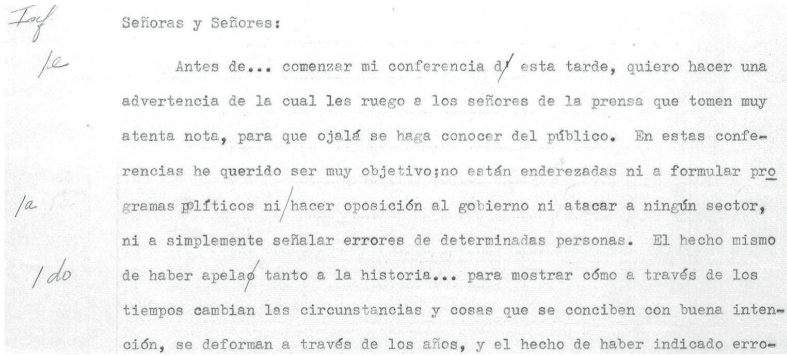
Por otro lado, Torruella y Llisterri (1999) plantean que una vez se han definido los aspectos de diseño del corpus (finalidad, objetivos, licencias, población, etc.) y se han recogido los materiales, se continúa con una fase de procesamiento que permita su poste-

rior uso. En el caso de los corpus orales, el primer paso suele ser la transcripción ortográfica,[15] seguida de la alineación y del etiquetado de los datos. De ahí que las siguientes fases estén relacionadas con procesos que alimenten la información, y que faciliten el manejo y explotación de los datos: la transcripción ortográfica, la alineación y el etiquetado. En relación con estas últimas fases, se espera que los usuarios que realicen trabajos relacionados con las actividades que acabamos de citar compartan y permitan la adición de esta información al Corpus. Asimismo, uno de los primeros pasos consistirá en la creación de un protocolo de transcripción ortográfica que garantice que todas las muestras sean transcritas con los mismos criterios.

Agregado a lo anterior, es importante resaltar que el sucorpus HCB cuenta con las transcripciones ortográficas de 600 entrevistas, realizadas a máquina y con algunas anotaciones a lápiz. En la figura 1 se muestra un fragmento de una transcripción ortográfica de una entrevista tipo 4. En la actualidad, se cuenta con 20 transcripciones digitalizadas por OCR (*Optical Character Recognitión*) y se está llevando a cabo un proceso para gestionar la digitalización total de todas las transcripciones y de los libros que cuentan con información de este subcorpus. Los siguientes procesos consistirán en la revisión de las transcripciones digitalizadas y su almacenamiento en los formatos necesarios para su explotación. Asimismo, será necesario definir si las transcripciones se mantendrán con los criterios anteriores o se actualizarán de acuerdo al protocolo de transcripción.

De igual manera, el subcorpus EHB cuenta con un libro digitalizado por OCR que contiene 30 transcripciones organizadas por estrato socioeconómico que se usarán para complementar la información de este subcorpus. En lo que respecta al subcorpus ALEC, se espera que una vez definido el protocolo se inicie el proceso de transcripción y alineación a través de *ELAN* o *PRAAT*.

[15] Muchas veces y dependiendo de la tipología y objetivos del corpus la transcripción ortográfica va acompañada de la transcripción fonética o fonológica (Torruella y Llisterri 1999).


Señoras y Señores:

Antes de... comenzar mi conferencia d/ esta tarde, quiero hacer una advertencia de la cual les ruego a los señores de la prensa que tomen muy atenta nota, para que ojalá se haga conocer del público. En estas conferencias he querido ser muy objetivo;no están enderezadas ni a formular programas plíticos ni/hacer oposición al gobierno ni atacar a ningún sector, ni a simplemente señalar errores de determinadas personas. El hecho mismo de haber apelad/ tanto a la historia... para mostrar cómo a través de los tiempos cambian las circunstancias y cosas que se conciben con buena intención, se deforman a través de los años, y el hecho de haber indicado erro-

Figura 1. Fragmento de una transcripción ortográfica de una entrevista tipo 4

En último lugar, vale decir que en la actualidad la capacidad de almacenamiento y tratamiento de datos puede hacerse automáticamente, y con mayor rapidez y precisión. La información que antes podría tardar años en ser analizada manualmente puede manipularse con programas y aplicaciones diseñadas para facilitar el trabajo de los investigadores. En este sentido, esperamos que la elaboración de este corpus permita el desarrollo de varias investigaciones y proyectos sobre el español de Bogotá y Colombia; la aplicación pedagógica del corpus en la formación de la diversidad lingüística y la enseñanza de español como primera, segunda y lengua extranjera; y la historia y cultura de Colombia.

Referencias bibliográficas

Buesa Oliver, Tomás y Luis Flórez
1954 El Atlas Lingüístico-Etnográfico de Colombia (ALEC). Cuestionario preliminar. *Thesaurus: boletín del Instituto Caro y Cuervo* 10.1-3 (1954), 147-315.

Fernández, Inés (dir.)
2005-2016 *Corpus Oral y Sonoro del Español Rural*. Consultado: febrero de 2016. <www.uam.es/coser>.

González, Alonso y Hilda Otalora
1986 *El habla de la ciudad de Bogotá: materiales para su estudio*. Bogotá: Instituto Caro y Cuervo.

Gries, Stefan
2009 *Quantitative Corpus Linguistics with R. A Practical Introduction.* Nueva York y Londres: Routledge. https://doi.org/10.4324/9780203880920

Grupo de Gramática del Español de la Universidad de Santiago de Compostela
2014-2016 *El corpus para el estudio del español oral, ESLORA*. Consultado: febrero de 2016. <http://eslora.usc.es/>.

Grupo de Investigación Val.Es.Co. (Valencia, Español Coloquial) de la Universidad de Valencia
1995-2016 *Corpus Val. Es. Co.: Corpus anotado del español coloquial*. Consultado: febrero de 2016. <https://www.uv.es/corpusvalesco/index.html>.

Grupo ILSE
S/f El Corpus del Habla en Almería. Almería: Universidad de Almería. Consultado: febrero de 2016. <http://nevada.ual.es/otri/ilse/corpus.asp>.

Laboratorio de Lingüística Italiana
S/f *C-OR-DIAL, Corpus Oral Didáctico Anotado Lingüísticamente*. Firenze: Universidad de Firenze. Consultado: febrero de 2016. <http://lablita.dit.unifi.it/app/cordial/>.

Montes, José, Jennie Figueroa, Siervo Mora, Mariano Lozano y Ricardo Aparicio
1997 *El español hablado en Bogotá. Relatos semilibres de informantes pertenecientes a tres estratos sociales*. Bogotá: Instituto Caro y Cuervo.

Montes, José, Jennie Figueroa, Siervo Mora, Mariano Lozano, Ricardo Aparicio, María Espejo y Gloria Duarte
1998 *El español hablado en Bogotá. Análisis previo de su estratificación social*. Bogotá: Instituto Caro y Cuervo.

Parodi, Giovanni
2008 "Lingüística de corpus: una introducción al ámbito". *RLA. Revista de lingüística teórica y aplicada*. 46, 1, 93-119. https://doi.org/10.4067/S0718-48832008000100006

Proyecto para el Estudio Sociolingüístico del Español de España y de América (PRESEEA)

2014-2016 *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. Consultado: febrero de 2016. <http://preseea.linguas.net>.

Real academia española: Banco de datos

s/f *Corpus de referencia del español actual* (CREA). Consultado: febrero de 2016. <http://www.rae.es>

Sinclair, John

1996 "Preliminary recommendations on corpus typology". En *EAGLES Document EAG-TCWG-CTYP/P*. Consultado: marzo de 2016. <http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html>.

Spitzová, Eva

1991 *Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la península ibérica: proyecto y realización.* República Checa: s/d.

Torruela, Joan y Llisterri, Joaquim

1999 "Diseño de corpus textuales y orales". En *Filología e informática: Nuevas tecnologías en los estudios filológicos*. Barcelona: Universidad Autónoma de Barcelona - Editorial Milenio, 45- 77 Consultado: enero de 2016. <http://liceu.uab.cat/~joaquim/publicacions/Torruella_Llisterri_99.pdf>.

Universidad de Bergen

S/f *Corpus COLA: Corpus Oral de Lenguaje Adolescente*. Consultado: febrero de 2016. <http://www.colam.org/om_prosj-espannol.html>.

Recepción: 23/06/2017
Aceptación: 13/04/2018