normas

Recibido: 20 octubre 2023 | Aceptado: 5 noviembre 2023 | Publicado: 15 noviembre 2023
Cita: Lanza, D. F. (2023). Los corpus orales del español centroamericano: compilación y mirada valorativa. *Normas*, 13, 83-111. doi: https://doi.org/10.7203/Normas.v13i1.27658

LOS CORPUS ORALES DEL ESPAÑOL CENTROAMERICANO: COMPILACIÓN Y MIRADA VALORATIVA

THE SPOKEN CORPORA OF CENTRAL AMERICAN SPANISH: COMPILATION AND EVALUATIVE OVERVIEW

Danny F. Lanza

Universitat Jaume I

Resumen

El presente trabajo ofrece un panorama actual sobre los principales corpus orales del español centroamericano, español que es hablado por más de cuarenta y cinco millones de personas en seis países diferentes (Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica y Panamá). El artículo presenta una compilación de los principales corpus orales actuales del español centroamericano. En concreto, su descripción se centra en dar detalles de cinco aspectos: datos generales (coordinador, ciudad representada, año de recolección, etc.), proceso de grabación, muestra recolectada, proceso de transcripción y, por último, acceso y soporte. Finalmente, el artículo reflexiona en torno a cuál es el estado actual de estos corpus y qué déficit y problemas por resolver se hallan en relación con su diseño, construcción, acceso y uso. Este trabajo ha podido recopilar 10 corpus que forman parte de diversos proyectos, 8 recolectan entrevistas semidirigidas y 2 conversaciones coloquiales. La mayoría de las muestras recolectadas pertenecen a ciudades de tres países: Costa Rica, Guatemala y Honduras.

PALABRAS CLAVE: corpus orales, español centroamericano, lingüística de corpus

Abstract

This paper offers a current overview of the main spoken corpora of Central American Spanish, a language spoken by over forty-five million people in six different countries (Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica, and Panama). The article compiles information about the primary existing spoken corpora of Central American Spanish. Specifically, its description focuses on providing details regarding five aspects: general data (coordinator, represented city, collection year, etc.), recording process, collected sample, transcription process, and, finally, access and support. In conclusion, the article reflects upon the current state of these corpora and identifies deficits and unresolved issues related to their design, construction, access, and usage. This study has been able to compile 10 corpora that are part of various projects, 8 of which collect semi-directed interviews, and 2 collect spontaneous conversations. Most of the collected samples are from cities in three countries: Costa Rica, Guatemala, and Honduras.

KEY WORDS: spoken corpora, Central American Spanish, corpus linguistics.

1 INTRODUCCIÓN

En la lingüística actual cada vez es más frecuente encontrar trabajos que parten del análisis de corpus orales, escritos o mixtos. Cuestión que no parece extraña, ya que, según algunos autores como Briz y Samper (2022: 311), «la lingüística científica ha de ser una lingüística de corpus, es decir, que ha de incorporar un conjunto *amplio* y *definido* (*suficiente* y *representativo*) de materiales que le proporcione datos fiables». Este tipo de trabajos, por su parte, han demostrado que, en muchos casos, «las intuiciones del hablante nativo sobre determinados aspectos del uso de su propia lengua no son siempre correctas, o al menos no siempre corresponden a lo que los demás hablantes de la lengua en realidad usan» (Pérez Hernández, 2002). Además, se ha demostrado que el uso de los corpus puede ser útil para la descripción y el análisis de la lengua en cualquier área o ámbito lingüístico (Hincapié y Bernal, 2018; Rojo, 2021; Parodi y otros, 2022, etc.).

Así pues, numerosos investigadores –por ejemplo, del mundo hispánico– han dedicado una enorme cantidad de esfuerzo, tiempo y recursos económicos para construir y publicar una diversidad de corpus, tal como lo han demostrado Briz y Albelda (2009), Briz (2012, 2018), Rojo (2016), Solís García (2018), Briz y Carcelén (2019) y Briz y Samper (2022). Estos trabajos, además, de presentar y recopilar aquellos corpus –sobre todo, los de lengua hablada u oral– del español, ofrecen una mirada crítica sobre el estado, el diseño, la construcción, el acceso y el uso de estos corpus.

Al respecto, por un lado, señalan que, a pesar de que el avance en la construcción de corpus del español ha sido notable, en la mayoría de los corpus más grandes «lo oral ocupa un espacio que no sobrepasa el 10 %» (Briz, 2012:123). Y, por otro, estos estudios destacan que en estos «hay una mínima representación de corpus situacionales […] y dentro de estos las conversaciones (coloquiales) ocupan un lugar nada destacado» Briz y Samper (2022:313).

Además de estos déficits, cabe añadir que no todas las variedades o normas regionales del español se encuentran representadas de igual forma, ni en cantidad ni en calidad. Si revisamos estas recopilaciones, podemos notar cómo hay ciertas regiones, países o ciudades que cuentan con un incipiente desarrollo de corpus orales. Este es el caso, como veremos a continuación, de la región hispanohablante centroamericana.

La población que habla el español como lengua nativa en los países que conforman la denominada América Central asciende, según el Instituto Cervantes (2023), a 45.920.550 personas, tal como se muestra de forma detallada por país en la Tabla 1. En concreto, de los más de 50 millones de centroamericanos, un 94,17 % de la población se encuentra dentro del Dominio Nativo (GDN) (más de 45 millones). Honduras, Nicaragua, El Salvador y Costa Rica tienen porcentajes de población nativa que superan el 97 %, Panamá sobrepasa el 91 % y únicamente Guatemala tiene menos de un 80 %.

País	Población	Hablantes nativos (%)	Grupo de Dominio Nativo (GDN)
Guatemala	17 602 431	78,3 %	13 782 703
Honduras	9 745 149	98,7 %	9 618 462
Nicaragua	7 046 308	97,1 %	6 841 965
El Salvador	6 364 940	99,7 %	6 345 845
Costa Rica	5 262 237	99,3 %	5 225 401
Panamá	4 468 089	91,9 %	4 106 174
Total	50 489 154	94,17 %	45 920 550

Tabla 1. Población de los países hispanohablantes de América Central y sus hablantes nativos. Fuente: elaboración adaptada de la propuesta por el Instituto Cervantes (2023:25-26)

En consecuencia, nos encontramos frente a una región en la que el porcentaje de población hispanohablante es alto en relación con su población total. Lo cual, a su vez, justifica por qué se debe tener en cuenta estas variedades y normas regionales, primero, a la hora de diseñar y construir corpus del español, y, segundo, de cara a la realización de investigaciones que aborden cualquier aspecto de la lengua y que partan de cualquier perspectiva teórica y metodológica.

Para el caso, en el contexto regional, podemos adelantar que la construcción y el acceso a corpus que recojan muestras escritas y, sobre todo, muestras orales que den cuenta de las diversas variedades y normas lingüísticas del español de Guatemala, Honduras, El Salvador, Nicaragua, Costa Rica y Panamá –al que denominaremos de ahora en adelante: español centroamericano– ha sido mínimo y discreto en comparación con otras regiones y países hispanohablantes.

Dado que reconocemos que la elaboración de cualquier corpus –sea escrito u oral– implica la inversión de mucho tiempo y esfuerzo por parte de los investigadores, hemos pensado que sería justo (para ellos) y beneficioso para la lingüística del español centroamericano poder dar cuenta de los principales corpus orales que se han construido. Nos centramos en los corpus orales discursivos –los cuales incluyen material oral habitualmente natural y espontáneo que es transcrito para su posterior manejo y procesamiento (Albelda, 2022:224)–, pues, como hemos comentado, son los que menor atención reciben (Briz y Albelda, 2009) y los que, a su vez, suponen una mayor dificultad en su diseño y elaboración (Briz, 2012: 124).

Así pues, este trabajo, en primer lugar, expone cuáles han sido los criterios de selección de los corpus; en segundo término, recopila los principales corpus orales que se han diseñado y construido sobre el español centroamericano; y, por último, reflexiona sobre cuál es el estado actual de estos corpus y qué déficit y problemas sin resolver se hallan en relación con su diseño, construcción, uso y acceso.

2 CRITERIOS DE SELECCIÓN DE LA COMPILACIÓN

A continuación, se presenta una compilación pormenorizada de los principales corpus[1] orales del español centroamericano. Esta compilación ha tomado como referencia los criterios de selección empleados por Albelda y Briz (2009:166-167) y los ha adaptado:

a) Las muestras de la lengua recolectadas deben ser orales, y deben ser recogidas en su contexto natural y real de producción.

b) Las muestras de la lengua recolectadas deben haber sido producidas por hablantes del español de una variedad o norma regional de alguno de los siguientes países: Guatemala, Honduras, Nicaragua, El Salvador, Costa Rica o Panamá. Además, los hablantes pueden vivir tanto dentro como fuera del país.

c) Los corpus deben haber sido publicados en papel o deben estar disponibles en formato electrónico o digital. En el caso de que no hayan sido publicados en papel y que no se pueda acceder a ellos de forma electrónica o digital, se considerará únicamente a) aquellos que hayan sido recolectados dentro de alguno de los principales macroproyectos de construcción y estudio de corpus orales del español, o b) aquellos que se encuentren en avanzado proceso de construcción y que vayan a ser publicados próximamente. No se consideran las bases textuales (como el CREA) ni los corpus que han sido recopilados por investigadores para estudios individuales, como tesis doctorales, pues, tal como señalan Briz y Albelda (2009:167), puede ser difícil obtener sus datos y, quizá, no se puede ofrecer una información exhaustiva de ellos.

d) Los corpus son sincrónicos del español actual centroamericano, con muestras de mitad del siglo pasado a la actualidad.

Antes de presentar la compilación, cabe resaltar, en primer lugar, que gran parte de estos corpus orales del español centroamericano han sido localizados gracias a los trabajos de Briz y Albelda (2009) y Briz y Carcelén (2019). En estos estudios se da cuenta de forma organizada y sistematizada de los principales macroproyectos hispánicos que se han encargado de construir corpus orales. Sin embargo, no en todos los casos se ofrece información pormenorizada de los microcorpus centroamericanos que forman parte de estos. Este trabajo pretende, en la medida de lo posible, completar esta información y dar algunos datos importantes como la persona y la institución que se encargó de la recolección del microcorpus, la cantidad de material recogido, el período en el que fueron grabados, la cantidad y las características de los hablantes que participan, el tipo de transcripción empleada, su acceso, etc.

[1] Esta compilación entiende un corpus como un conjunto extenso y organizado de textos, que pueden ser orales, escritos o mixtos, que han sido producidos en su contexto natural o real, que aspiran a ser representativos de una lengua o una variedad, que se almacenan –normalmente– en formato digital o electrónico, y los cuales pueden servir para la descripción y el análisis (medio de verificación de hipótesis) de la lengua. Noción que parte de las definiciones que proponen Crystal (1991), Parodi (2010) y Rojo (2021).

En segundo término, como veremos a continuación, la mayoría de los corpus orales del español centroamericano se han construido, precisamente, en el seno de estos macroproyectos. Apenas se encuentran proyectos independientes o no relacionados con estos macroproyectos. Se ha considerado pertinente estructurar la compilación de la siguiente forma: se presentan los corpus por género: entrevistas semidirigidas y conversaciones coloquiales. Segundo, en cada uno de estos grupos se recopila: a) los microcorpus que han nacido en el seno de uno de los macroproyectos y b) los corpus que se han recogido como proyectos particulares o que son de otra naturaleza (por ejemplo, corpus de inmigrantes centroamericanos). Su presentación seguirá un orden cronológico: desde el más antiguo hasta el más reciente (por proyecto).

3 COMPILACIÓN DE LOS CORPUS ORALES DEL ESPAÑOL CENTROAMERICANO

Esta sección se divide en dos: la primera (§ 3.1.) recopila los corpus de entrevistas orales y la segunda (§ 3.2.) reúne los corpus de conversaciones coloquiales.

3.1 Corpus de entrevistas del español centroamericano

A continuación, presentamos los principales corpus de entrevistas del español centroamericano.

3.1.1 El macroproyecto de la norma lingüística culta de las principales ciudades de España y América o Proyecto de estudio de la norma culta hispánica «Juan M. Lope Blanch»

En la actualidad los microcorpus que forman parte de este macrocorpus se aúnan bajo el nombre: «Proyecto de estudio de la norma culta hispánica "Juan M. Lope Blanch"»[2]. No obstante, en un principio el macroproyecto se denominó: «Proyecto del estudio del habla culta de las principales ciudades de Hispanoamérica». Su principal impulsor fue el Dr. Juan M. Lope Blanch. Su objetivo ha sido obtener «un conocimiento riguroso, detallado y completo del habla actual de las grandes urbes modernas de Iberoamérica» (Lope Blanch, 1986: 13). En un principio, de las más de 15 ciudades de habla hispana representadas, se encontraba San José de Costa Rica. En los últimos años se han incorporado otras ciudades, como la Ciudad de Panamá (Panamá).

En 1998, el profesor José A. Samper junto con las profesoras Clara Eugenia Hernández y Magnolia Troya, y gracias a los avances tecnológicos del momento, digitalizaron una parte del macrocorpus (Samper y otros, 1998). El propósito de esta edición consistía «en ofrecer la transliteración de ochenta y cuatro horas de grabación que recogen las voces de 168 hablantes representativos del nivel culto de doce ciudades hispánicas» Samper Padilla (1995:263). Dentro de las 12 capitales que se representaron en este corpus, se incluyó una única ciudad centroamericana: San José de Costa Rica.

- *El microcorpus de la norma culta de San José (Costa Rica)*

[2] A partir de ahora nos referiremos a este como «(Macro)proyecto de la Norma Culta».

La primera ciudad centroamericana que formó parte del macroproyecto fue San José de Costa Rica. La labor de coordinar la recolección y la transcripción del habla culta de la capital costarricense estuvo a cargo del entonces profesor de la Universidad de Granada, Francisco Salvador. Para 1984 el profesor granadino ya había recogido aproximadamente 72 horas de grabaciones y se había dispuesto a transcribirlas.

De este y de todo el material que fue recogido posteriormente por el profesor Salvador, como hemos visto, Samper y otros[3] (1998) seleccionaron y transcribieron catorce entrevistas semidirigidas. El tiempo de grabación de todas estas entrevistas asciende a 420 minutos (7 horas), que se traducen en alrededor de 66 496 palabras.

La muestra representada en estas entrevistas fue de 14 hablantes de San José de Costa Rica, cuyas características sociológicas se pueden ver en la siguiente tabla:

Generación	Sexo	
	Hombre	Mujer
I	3 (60 min)	3 (60 min)
II	2 (90 min)	2 (90 min)
III	2 (60 min)	2 (60 min)
Total	7 (210 min)	7 (210 min)

Tabla 2. Distribución de la muestra del microcorpus de San José de Costa Rica, recogido por Samper, Hernández y Troya (1998). Fuente: adaptada de la propuesta de Samper (1995)

En primer lugar, como se puede apreciar en la tabla anterior, la cantidad de informantes varía según la generación: en la primera generación se ha seleccionado a 3 hombres y 3 mujeres (6 personas en total), en cambio en la segunda y tercera generación se ha elegido a 2 hombres y 2 mujeres (4 personas en total para cada generación). Los hablantes de la muestra pertenecen al nivel de instrucción alto (estudios universitarios), sus edades oscilan entre los 28 a los 68 años y algunas de sus profesiones u oficios son profesores universitarios, licenciados, farmacéuticos, sicoterapeutas, ingenieros, bachilleres, exministros, exdiputados y maestros.

Cabe destacar, por otro lado, que a pesar de que se tiene acceso a la transcripción del microcorpus de San José, no se cuenta con acceso a los audios. Además, es importante señalar que del total del material recogido por el profesor Salvador, lo único que se ha llegado a publicar, tanto de transcripciones como de audios, ha sido lo recopilado por Samper y otros (1998) en un «CD-ROM».

En lo que respecta a la transcripción del corpus, hay que señalar, en primer lugar, que, previo a la publicación de Samper y otros (1998), el equipo del microcorpus de la Norma Culta de San José de Costa Rica había llevado a cabo una transcripción literal y había seguido unas normas de transcripción específicas, sin embargo, Samper y otros (1998), con el objetivo de homogeneizar la transcripción de los diversos microcorpus que recopilaron, aplicaron una serie de lineamientos: mantienen una transcripción ortográfica

[3] A partir de ahora nos referiremos a este corpus como «Corpus de la Norma Culta».

convencional (no fonética), la cual se basa en normas ortográficas generales, y emplean un sistema de convenciones de transliteración: por ejemplo, las vacilaciones se marcan con puntos suspensivos, los corchetes se usan para aportar una aclaración de los encuestadores, los fragmentos ininteligibles se marcan con corchetes que encierran tres puntos suspensivos, el discurso reproducido va entrecomillas dobles (precedidas de dos puntos), etc.[4]. Por último, debido a los pocos avances del momento, como es esperable, en la transcripción no se ha empleado ningún etiquetado para su posible tratamiento digital, como tampoco se ha alineado el texto con el audio.

- *El microcorpus de la norma culta de la Ciudad de Panamá (Panamá)*

Como veremos a continuación, el corpus de la norma culta de la Ciudad de Panamá se construyó en la segunda década de este siglo, sin embargo, explica Spitzová (1991:63) que en 1984 se tenían recogidas veinticinco horas de grabación, pero «debido a dificultades tanto de orden objetivo como subjetivo, se consideró imposible reunir las 400 horas de grabaciones previstas en el cuestionario». Así, más de tres décadas después de la incorporación de la ciudad de San José de Costa Rica a este macroproyecto, en el año 2014, en el marco de la celebración del XVII Congreso Internacional de la Asociación de Lingüística y Filología de América Latina, celebrado en João Pessoa, Brasil, la profesora Fulvia Morales del Castillo, miembro correspondiente de la Academia Panameña de la Lengua y profesora de la Universidad de Panamá, en compañía de un grupo de investigadores, propuso la incorporación de la Ciudad de Panamá a este macroproyecto. Su solicitud fue aceptada por la Comisión Ejecutiva del Proyecto, lo cual dio paso a su recolección y transcripción.

Escobar Samaniego (2017:53-54) explica, respecto al proceso de grabación, que se recolectaron entrevistas tanto de diálogo dirigido como de diálogo libre (siguiendo los lineamientos metodológicos del macroproyecto). Las entrevistas se grabaron con la grabadora a la vista de los informantes. Además, se tuvo cuidado con la selección de los espacios de grabación: se prefirió por los espacios cerrados e interiores con el fin de obtener una mayor calidad en el sonido (sin o con pocos ruidos). Cabe señalar que todas las grabaciones tienen una duración de, mínimo, treinta minutos, aunque hay algunas que duran más del tiempo mínimo establecido. Se desconoce el tiempo total de grabación, así como el número de palabras que aglutina.

Finalmente, después de un proceso de cribado, para el 2017 se tiene constancia de que se recogieron 26 muestras, en las que participaron informantes de las siguientes características[5]:

Generación	Sexo	
	Hombre	Mujer
I	4 (2 libre / 2 dirigido)	4 (2 libre / 2 dirigido)
II	4 (2 libre / 2 dirigido)	4 (2 libre / 2 dirigido)

[4] Para consultar todo el sistema y las normas de transcripción empleadas, véase el trabajo de Samper Padilla (1995).
[5] En la siguiente tabla solamente se da cuenta de las características de 24 hablantes, pues es la información que ofrece Escobar Samaniego (2017:52).

III	4 (2 libre / 2 dirigido)	4 (2 libre / 2 dirigido)
Total	12 (6 libre / 6 dirigido)	12 (6 libre / 6 dirigido)

Tabla 3. Distribución y características de los informantes del microcorpus de la Norma Culta de Ciudad de Panamá: Fuente: adaptada de la propuesta por Escobar Samaniego (2017)

Como podemos ver en la tabla anterior, la distribución ha sido homogénea: por cada generación se recogieron 8 entrevistas (4 dirigidas a informantes mujeres y 4 a informantes hombres). Además, cabe mencionar que de las 4 de cada sexo, 2 eran de diálogo libre y otras 2 de diálogo dirigido (4 en total en cada modalidad). Así, en el cómputo final, el microcorpus de la norma culta de Ciudad de Panamá cuenta con 8 entrevistas en la primera generación, 8 en la segunda y otras 8 en la tercera; 12 entrevistas se han dirigido a hombres y 12 a mujeres; y, por último, 12 han sido en la modalidad de entrevista con diálogo libre y otras 12 en la modalidad de diálogo dirigido.

En lo que respecta a la transcripción, Escobar Samaniego (2017:54) señala que las grabaciones se transcribieron siguiendo los lineamientos del macroproyecto: una transcripción simple en la que se transliteró lo dicho por los participantes de la muestra (tanto entrevistador como informante). Esta transliteración respetó las normas de uso de los signos de puntuación y los ortográficos. Además, no se incluyó ningún etiquetado fónico, pero sí se representó los casos de palabras cortadas (salvo en casos que no permitían comprender el sentido de las intervenciones) y las vacilaciones. Cabe destacar, por último, que las transcripciones pasaron por un proceso de revisión y corrección guiado por criterios homogéneos. No se incluyó etiquetado para su tratamiento informático ni tampoco alineamiento del texto con el audio.

Por último, es importante mencionar que este microcorpus no ha sido publicado ni en formato físico ni electrónico, por lo cual su acceso se encuentra restringido hasta el momento. A pesar de ello, recientemente, en el XXXI Congreso Científico Nacional-2023, organizado por el Centro de Lectura y Escritura Académica de la Universidad de Panamá, la Dra. Fulvia Morales ha informado que próximamente se estará publicando el corpus, en formato físico y electrónico.

3.1.2 El Proyecto para el estudio sociolingüístico del español de España y de América (PRESEEA)

El Proyecto para el estudio sociolingüístico del español de España y América (PRESEEA) nace con el fin de disponer de material real oral para el estudio sociolingüístico de la lengua española en su contexto geográfico y social. Sus principales impulsores fueron Carmen Silva-Corvalán, Humberto López Morales y Francisco Moreno Fernández. Uno de sus principales objetivos consistía en construir, siguiendo una misma metodología, un macrocorpus de entrevistas provenientes de diferentes ciudades del mundo hispánico y que fuera representativo de los diversos a) niveles socioculturales (alto, medio y bajo), b) sexos y b) grupos generacionales. En la actualidad, este corpus cuenta con muestras completas de más de 20 ciudades hispanas, dentro de las cuales se halla la Ciudad de Guatemala (Guatemala). Además, hay otra serie de ciudades que se encuentran en proceso de incorporación, como es el caso de Tegucigalpa (Honduras).

- *El microcorpus del PRESEEA-Ciudad de Guatemala (Guatemala)*

El macrocorpus del PRESEEA (PRESEEA, 2014) ha contado desde sus inicios hasta la actualidad con una enorme y valiosa representación del mundo hispánico. En el 2003 la profesora de la Universidad Rafael Landívar y académica de número de la Academia Guatemalteca de la Lengua, Lucía Verdugo, en conjunto con las profesoras Ana Acevedo-Halvick y Ana María Palma se proponen recoger, transcribir y etiquetar el corpus PRESEEA de la Ciudad de Guatemala (Guatemala).

En un primer momento se recogieron aproximadamente 108 entrevistas, pero después de su revisión y depuración, el corpus se constituyó de 75 grabaciones. En la página web del macroproyecto (https://preseea.uah.es/) se encuentra disponible una muestra de 18 entrevistas, las cuales fueron recogidas entre el 2003 y el 2005. La duración individual de cada una varía: encontramos entrevistas que duran desde 32 hasta 57 minutos. En total, el tiempo de grabación es de 12 horas con 55 minutos y 21 segundos.

En la siguiente tabla se muestran las características sociológicas de los 18 informantes que están representados:

	Generación 1		Generación 2		Generación 3		Total
	H	M	H	M	H	M	
Nivel educativo 1	1	1	1	1	1	1	6
Nivel educativo 2	1	1	1	1	1	1	6
Nivel educativo 3	1	1	1	1	1	1	6
Total	3	3	3	3	3	3	18

Tabla 4. Distribución y características de los informantes del microcorpus del PRESEEA-Ciudad de Guatemala

Tal como se puede apreciar, la distribución de la muestra que se encuentra accesible en la página web del PRESEEA es homogénea en cuanto al sexo, edad y nivel educativo. Se ha entrevistado a 9 hablantes hombres y a 9 mujeres; de los cuales 6 pertenecen a la primera generación (20-34 años), otros 6 a la segunda generación (35-54 años) y, de igual forma, otros 6 a la tercera generación (55 años en adelante). En lo que concierne a su nivel educativo, 6 de los participantes pertenecen al nivel educativo 1 (analfabetos, sin estudios y enseñanza primaria), otros 6 al nivel educativo 2 (enseñanza secundaria) y 6 más al nivel educativo 3 (enseñanza superior).

Respecto a la transcripción, el corpus del PRESEEA-Guatemala ha seguido los lineamientos metodológicos de transcripción que propone el macroproyecto: una transliteración ortográfica ordinaria que es acompañada de un marcado y etiquetado que sigue las convenciones de la TEI (*Text Encoding Initiave*), aunque cabe destacar que el proyecto seleccionó un número concreto de etiquetas para marcar la lengua hablada, tales

como «<énfasis> </énfasis>», «<cita> </cita>», «<risas = « »/>, entre otras[6]. Los textos no parecen estar alineados con el audio. En la página web del proyecto se puede acceder tanto al audio como a los textos de este corpus, los cuales, a su vez, pueden ser descargados y exportados en diversos formatos.

- *El microcorpus del PRESEEA de Tegucigalpa (Honduras)*

El microcorpus del PRESEEA de Tegucigalpa, capital de Honduras, se encuentra actualmente en proceso de construcción. Nace con el objetivo de ser un punto de contraste con el corpus de conversaciones coloquiales Ameresco-Tegucigalpa, el cual ha sido recogido en años anteriores a este y del cual daremos detalle más adelante. Su impulsor es quien firma este artículo, Danny F. Lanza, estudiante de doctorado del programa en Estudios Hispánicos Avanzados de la Universitat de València. Para su construcción, el corpus ha contado con la estrecha colaboración del Departamento de Letras de la Universidad Pedagógica Nacional Francisco Morazán (Campus Central), de Tegucigalpa, Honduras, especialmente de la profesora Maura Catalina Flores y del profesor Gustavo González Cáceres, así como de estudiantes del Profesorado en la Enseñanza del Español.

En esta primera fase de recolección, llevada a cabo entre el último semestre del 2022 y el primero del 2023, se han grabado 16 entrevistas, de las cuales 14 rellenan un grupo sociológico de la muestra básica exigida por el proyecto. La mayoría de las entrevistas duran como mínimo 45 minutos cada una, aunque hay algunas que superan una hora de grabación. En total, el tiempo de grabación es de 13 horas, con 03 minutos y 07 segundos.

En la siguiente tabla se muestran las características sociológicas de los 16 informantes que están representados:

	Generación 1		Generación 2		Generación 3		Total
	H	M	H	M	H	M	
Nivel educativo 1	0	2	1	1	1	1	6
Nivel educativo 2	1	2	1	0	1	1	6
Nivel educativo 3	1	1	0	1	0	1	4
Total	2	5	2	2	2	3	16

Tabla 5. Distribución y características de los informantes del microcorpus del PRESEEA-Tegucigalpa

A simple vista, a diferencia del corpus del PRESEEA-Ciudad de Guatemala, el corpus del PRESEEA-Tegucigalpa no tiene todavía la muestra básica completa que es exigida por el macroproyecto. Aunque se ha entrevistado a 16 participantes, solo 14 de ellos rellenarían alguno de los grupos sociológicos de la tabla. Hay en total 6 hombres y 10 mujeres, de los cuales 7 pertenecen a la primera generación (20-34 años), 4 a la segunda generación (35-54 años) y 5 a la tercera generación (55 años en adelante). En lo que respecta a su nivel

[6] Para consultar todo el sistema y las normas de transcripción empleadas, véase los documentos de trabajo del PRESEEA, elaborados por Moreno Fernández (2021a, 2021b).

educativo, 6 de los informantes pertenecen al nivel educativo 1 (analfabetos, sin estudios y enseñanza primaria), otros 6 al nivel educativo 2 (enseñanza secundaria) y 4 al nivel educativo 3 (enseñanza superior).

Al igual que el corpus del PRESEEA-Ciudad de Guatemala, el corpus del PRESEEA-Tegucigalpa ha seguido los lineamientos de transcripción propuestos por el proyecto: una transliteración ortográfica ordinaria que es acompañada de un marcado y etiquetado de fenómenos propios de la lengua hablada. Cabe destacar, además, que se han aplicado algunas normas propias referidas a la ortografía y puntuación: por ejemplo, los únicos signos de puntuación que cumplen su función prototípica son los de exclamación «¡!» y los de interrogación «¿?», o las mayúsculas únicamente se añaden en la inicial de nombres propios siglas[7].

Este microcorpus todavía no se encuentra accesible a los usuarios, ni sus audios ni sus transcripciones. Se espera, en primer lugar, acabar de recopilar la muestra básica exigida por el macroproyecto, para, en segundo término, enviarla a la coordinación técnica del PRESEEA y que sea revisada, aceptada (rechazada o mejorada) y, finalmente, incorporada a la página del macroproyecto.

3.1.3 El proyecto EGREHA: Estudio gramatical del español hablado en América

El proyecto EGREHA (Estudio gramatical del español hablado en América) tenía como objetivo llevar a cabo de forma conjunta una serie de estudios gramaticales que dieran cuenta de las variedades lingüísticas del español hablado en América. Su coordinador fue el profesor César Hernández Alonso, de la Universidad de Valladolid. Para cumplir el objetivo principal del proyecto, se construye el corpus EGREHA, el cual, además de incluir los materiales del Macroproyecto de la Norma Lingüística Culta de las principales ciudades de España y América (descrito en el apartado 3.1.1.), también recopiló nuevos archivos. En concreto, recolectó un grupo de entrevistas orales que, en su mayoría, no fueron transcritas. Este corpus no ha sido publicado, ni en formato físico ni electrónico o digital.

De los 12 países hispanoamericanos que formaron parte del proyecto, se encuentran dos centroamericanos: Costa Rica (con 23 entrevistas) y Guatemala (con 9).

- *Los microcorpus EGREHA de Costa Rica y de Guatemala*

Los únicos países centroamericanos de los cuales se grabaron entrevistas para el corpus EGREGA son Costa Rica y Guatemala. En concreto, se recolectaron materiales de la ciudad de San José y de la Ciudad de Guatemala, capitales de ambos países.

Primero, hay que decir que se desconoce con exactitud quiénes fueron los investigadores e instituciones responsables de recogerlos, pero, teniendo en cuenta que el proyecto se basó en gran medida en la metodología del Macroproyecto de la Norma Culta, creemos

[7] Para consultar todo el sistema y las normas de transcripción empleadas, véase los documentos de trabajo del PRESEEA, elaborados por Moreno Fernández (2021*a*, 2021*b*). Además, cabe destacar que estas normas respecto a la puntuación también se aplicaron al corpus del PRESEEA-Ciudad de Guatemala.

que han sido profesores de la región. Segundo, no hemos podido acceder a otros datos concretos como los años de grabación, pero, a partir de algunas de las intervenciones de los participantes podemos intuir que se grabaron entre el 2000 y el 2009.

Se recogieron 23 entrevistas de San José de Costa Rica y 9 de Ciudad de Guatemala, pero se desconoce la duración total e individual de cada una. Un dato que sí conocemos es que, de estas 32 grabaciones, se encomendó al grupo de investigación Val.Es.Co. (Valencia. Español. Coloquial.) y a la empresa Tecnolingüística, empresa especializada en servicios lingüísticos y de comunicación, la transcripción de una parte. Específicamente, se transcribieron 9 entrevistas: 7 de los informantes eran de San José de Costa Rica y apenas 2 de la Ciudad de Guatemala.

A continuación, se muestran algunos datos sociológicos que hemos podido obtener de las transcripciones a las que hemos podido tener acceso[8]:

N.º	Código	Ciudad y país	Nivel educativo	Sexo	Edad
1	CR001	San José (CR)	Alto	M	36
2	CR002	San José (CR)	Alto	M	Desconocida
3	CR023	San José (CR)	Medio	H	84
4	CR006	San José (CR)	Medio	H	36
5	CR008	San José (CR)	Medio	H	43
6	CR011	San José (CR)	Medio	H	23
7	CR014	San José (CR)	Bajo	M	48
8	GU002	Ciudad de Guatemala (GT)	Alto	H	40
9	GU006	Ciudad de Guatemala (GT)	Bajo	H	Desconocida

Tabla 6. Características de los informantes de los microcorpus del EGREHA de San José de Costa Rica y de Ciudad de Guatemala

De la tabla anterior podemos extraer algunos datos relevantes: a) de los de San José de Costa Rica, se hallan 4 hombres y 2 mujeres, de los cuales 2 tienen nivel educativo alto (universitario), otros 4 tiene nivel medio (educación secundaria) y solo 1 tiene nivel bajo (sin estudios o educación primaria). Respecto a la edad de los hablantes, las edades oscilan entre los 23 a los 84 años (así como una edad desconocida); b) de los de Ciudad de Guatemala, encontramos únicamente 2 hombres y ninguna mujer, de los cuales uno pertenece al nivel educativo alto y el otro al bajo. En cuanto a su edad, solo conocemos la de uno (40 años). Un dato que nos ofrecen las transcripciones es la profesión u oficio de los informantes: encontramos dos bibliotecólogas, un operador de radio, un oficial de investigación, un diseñador gráfico, una niñera, un gerente de hotel y un agricultor.

La transcripción de estas 9 entrevistas del EGREHA de San José y de Ciudad de Guatemala se basó en las convenciones y el sistema de transcripción que había desarrollado el grupo de investigación Val.Es.Co. (Valencia. Español. Coloquial.) en ese

[8] En la Tabla 6 no se muestra la distribución por grupos sociológicos (tal como se ha venido mostrando hasta este punto) ya que se desconocen algunos datos de los informantes (como la edad) o cuál es el rango de las diversas generaciones.

momento. En términos generales, el tipo de transcripción consistió en una transliteración ortográfica de los audios que ha sido acompañada de una serie de signos que dan cuenta de algunos fenómenos propios de la lengua hablada, como las pausas, los solapamientos, los susurros, los tonemas, el énfasis, los alargamientos, entre muchos otros[9]. Estas transcripciones no fueron etiquetadas para su posterior tratamiento informático ni tampoco se alineó su texto con los audios.

Por último, habría que señalar que ni las 32 grabaciones de las dos ciudades fueron publicadas ni puestas en ninguna plataforma digital para su acceso, tampoco se publicaron las 9 grabaciones que fueron transcritas por el grupo Val.Es.Co. y Tecnolingüística: ni los audios ni las transcripciones en sí.

3.1.4 El Corpus Oral de la Lengua Española en Montreal (COLEM)

Con el fin de poder obtener un conjunto de muestras amplias de la lengua española hablada en Montreal (*Région métropolitaine de Montréal*) para su descripción y caracterización lingüística, el Dr. Enrique Pato, profesor del *Département de littératures et de langues du monde*, de la Université de Montréal (Canadá); ha diseñado y construido, gracias al financiamiento de diversos proyectos concedidos por entidades públicas canadienses, el Corpus Oral de la Lengua Española en Montreal[10] (COLEM) (Pato, 2023*a*), el cual ha recogido, entre los años 2014-2017 y 2019, una serie de entrevistas semidirigidas a hablantes del español que viven en esa región. Hasta la fecha (septiembre de 2023) se han recolectado 153 entrevistas de aproximadamente una hora de duración individual, que ascienden a más de 170 horas de grabación y que constituyen 1 677 597 palabras. Cada entrevista se ha estructurado en función de un protocolo de encuesta común. Sus informantes son hablantes originarios de diversos países hispanohablantes que viven en Montréal.

En la Tabla 7 podemos ver la cantidad de informantes centroamericanos que forman parte de este corpus. Además, podemos notar, por un lado, su distribución por sexo y, por otro, sus características generales (edad y nivel educativo):

País de origen	Sexo			Edad	Nivel educativo
	H	M	Total		
Costa Rica	3	4	7	29-62	Educación secundaria o universitaria
El Salvador	3	4	7	27-60	
Guatemala	4	5	9	32-61	
Honduras	2	5	7	25-64	
Nicaragua	2	5	7	28-62	
Panamá	3	3	6	21-59	

[9] Para consultar todo el sistema y las normas de transcripción empleadas, véase, por ejemplo, el trabajo de Briz y Grupo Val.Es.Co. (2002).
[10] Toda la información referente al corpus se encuentra disponible en https://esp-montreal.jimdo.com/

Tabla 7. Distribución por sexo y características generales (edad y nivel educativo) de los informantes centroamericanos del corpus COLEM

Como podemos apreciar, el total de hablantes centroamericanos entrevistados son 43: Costa Rica (7), El Salvador (7), Guatemala (9), Honduras (7), Nicaragua (7), Panamá (6). En concreto, de los informantes costarricenses, 3 son hombres y 4 son mujeres, y sus edades oscilan entre 29-62 años; de los informantes salvadoreños, también 3 son hombres y 4 son mujeres, y sus edades oscilan entre 27-60 años; de los hablantes guatemaltecos, 4 son hombres y 5 son mujeres, y sus edades varían de entre 32-61 años; de los participantes hondureños, 2 son hombres y 5 son mujeres, y sus edades varían de 25-64 años; de los informantes nicaragüenses, también 2 son hombres y 5 son mujeres, y tienen entre 28-62 años; por último, de los participantes panameños, 3 son hombres y otros 3 son mujeres, y tienen entre 21-59 años.

Así pues, en términos generales, por un lado, en estos microcorpus centroamericanos del COLEM hay más mujeres (26) que hombres (17), y, por otro, hallamos hablantes de las tres generaciones que ha establecido el proyecto: sus edades se comprenden entre los 21 a los 64 años. Ahora bien, respecto a su nivel educativo, como hemos comentado, no se ofrecen datos concretos, sin embargo, su coordinador nos ha informado, vía medios electrónicos, que los hablantes centroamericanos del COLEM tienen o un nivel educativo secundario o bien un nivel educativo universitario o superior.

En cuanto a la transcripción del corpus COLEM, se ha decidido llevar a cabo una transliteración ortográfica (se siguen las normas ortográficas establecidas por la RAE [2010]) de las entrevistas. La transcripción ha incluido, a su vez, el uso de un sistema propio para dar cuenta, a través de determinados signos y normas, de una diversidad de fenómenos relacionados con la lengua hablada (y otros), por ejemplo, el guion «-» sirve para indicar que una palabra está cortada o no se ha terminado de pronunciar, las comillas inglesas «“”» se usan para representar el discurso reproducido, la barra vertical «|» para representar una autocorrección o reformulación, etc. Además, esta transcripción se ha acompañado de un marcaje y etiquetado para su posterior tratamiento informático: por ejemplo, para las risas se usa [RISAS], para las pausas: [PAUSA], para los ruidos: [RUIDOS], etc. Estos microcorpus estarán disponibles en línea próximamente, según nos ha informado su coordinador. Actualmente, solo se puede acceder, bajo petición vía correo electrónico a la coordinación, a sus transcripciones (no a sus audios). La versión en línea ha sido desarrollada por Anthony Rancourt.

3.1.5 El Corpus Oral de la Lengua Española en Honduras (COLEH)

El último corpus de entrevistas que presentamos en esta compilación de corpus orales del español centroamericano es el Corpus Oral de la Lengua Española en Honduras[11] (COLEH) (Pato, 2023*b*), el cual se encuentra en construcción actualmente. Nace bajo el impulso del doctor Enrique Pato (coordinador del corpus COLEM), profesor de la Université de Montréal

[11] La mayoría de los datos de este corpus han sido extraídos de la página web https://n9.cl/982gi.

(Canadá), debido a la necesidad de llevar a cabo nuevos trabajos descriptivos y análisis gramaticales específicos sobre el español en Honduras.

El corpus COLEH se desarrolla en el marco del proyecto «El español en Honduras en la actualidad». Este corpus recoge entrevistas semidirigidas no de una ciudad en concreto, sino que pretende reunir materiales de las principales ciudades (municipios) de Honduras, así como de algunos enclaves rurales (aldeas y caseríos) en todos los departamentos. El proyecto, además, recolecta entrevistas de los hondureños en la diáspora (muy parecido a lo que se lleva a cabo en el COLEM): encontramos migrantes hondureños que residen en España, Canadá, Estados Unidos, México e Italia.

El COLEH ha contado con la colaboración de profesores y estudiantes, especialmente, de la Universidad de Montreal (UdeM), de la Academia Hondureña de la Lengua (AHL), de la Universidad Pedagógica Nacional Francisco Morazán (UPNFM) y de la Universidad Nacional Autónoma de Honduras (UNAH).

Hasta la actualidad (septiembre de 2023) se han recogido 165 grabaciones, en las que participan 165 hablantes hondureños. Cada grabación dura, al menos, una hora. Se cuenta con 172 horas de grabación en total. Para llevar a cabo la entrevista, se ha seguido un protocolo de encuesta y se ha estructurado a partir de diferentes bloques: pasado, presente, futuro. Así como otros que se han centrado en solicitar la opinión de los informantes y pedirles que expliquen algún tema en particular.

No se cuenta, hasta el momento, con datos específicos de las características sociológicas de los 165 informantes que se han grabado, únicamente se ofrecen datos generales: hay 97 mujeres y 68 hombres, de entre 18 a 85 años; y sin estudios, con estudios de primaria, secundaria y universitarios. Estos informantes proceden de todos los departamentos de Honduras, excepto del departamento de Islas de la Bahía.

El COLEH se encuentra todavía en fase de recolección del material, por lo que no se ha comenzado el proceso de transcripción de las entrevistas, a pesar de ello, se tiene previsto (similar a la metodología de transcripción empleada por el COLEM) llevar a cabo una transliteración ortográfica que se acompañará de un sistema de transcripción propio a partir del cual se representarán ciertos fenómenos propios de la lengua hablada. La transcripción, también, se etiquetará para su respectivo tratamiento informático. Cabe destacar, por último, que el COLEH estará disponible de forma gratuita en una página web: tanto sus audios como sus grabaciones.

3.2 Corpus de conversaciones coloquiales del español centroamericano. El macroproyecto AMERESCO (América y España. Español coloquial.)

El macroproyecto AMERESCO (América y España. Español coloquial.) nace con el fin de profundizar en el estudio de la variedad coloquial del español en geolectos europeos y americanos. Su director es el profesor Antonio Briz y su coordinación ha estado a cargo de las profesoras Marta Albelda y Maria Estellés, todos de la Universitat de València. En el seno del proyecto se impulsa la creación del corpus de conversaciones coloquiales

AMERESCO[12] (Albelda y Estellés, 2023), cuyo fin es recopilar muestras reales de conversaciones coloquiales de la mayor cantidad de ciudades y países hispanohablantes. En la actualidad, de las ciudades que forman parte del corpus, hay dos centroamericanas: Ciudad de Panamá (Panamá) y Tegucigalpa (Honduras).

- *El microcorpus AMERESCO de Ciudad de Panamá (Panamá)*

La primera ciudad centroamericana en formar parte del macrocorpus AMERESCO es la Ciudad de Panamá. Su impulsora y coordinadora es la profesora Fulvia Morales del Castillo, miembro correspondiente de la Academia Panameña de la Lengua y profesora de la Universidad de Panamá, quien, a su vez, como hemos expuesto en apartados anteriores, también es la coordinadora del Corpus de la Norma Culta de Ciudad de Panamá.

La colaboración nace en el marco del «Seminario Internacional "Enseñanza de la Lengua Española: Gramática, Escritura y Oralidad. Los Diccionarios», organizado por la Academia Cubana de la Lengua, la Facultad de Artes y Letras de la Universidad de La Habana, y la Consejería Cultural de la Embajada de España en La Habana. AECID-Cuba, del 09 al 12 de enero de 2017. En este seminario la profesora Marta Albelda, una de las coordinadoras del corpus AMERESCO, propuso a la profesora Morales su incorporación, propuesta que fue aceptada.

El corpus AMERESCO-Ciudad de Panamá (Morales del Castillo, 2023) empezó a recogerse ese mismo año. Actualmente, en la página web del proyecto se encuentra a disposición 5 conversaciones coloquiales. Cabe destacar que cada grabación dura, como mínimo, 20 minutos. La duración total de las cinco entrevistas disponibles es de 2 horas (19 590 palabras). En estas participan activamente 13 hablantes originarios de la Ciudad de Panamá, y cuyas características sociológicas se encuentran detalladas en la siguiente tabla:

Código	N.º de hablantes activos	Sexo		Edad			Nivel educativo		
		H	M	≤25	26-55	>55	Alto	Medio	Bajo
PTY_001_04_17	4	1	3	4	0	0	0	4	0
PTY_003_02_17	2	0	2	2	0	0	0	2	0
PTY_004_03_17	2	0	2	0	1	1	0	1	1
PTY_007_02_17	2	1	1	2	0	0	0	2	0
PTY_009_04_17	3	1	2	1	2	0	0	2	1
Total	13	3	10	9	3	1	0	11	2

Tabla 8. Distribución y características de los informantes del microcorpus AMERESCO-Ciudad de Panamá

Del total de los 13 hablantes que participan activamente en el corpus AMERESCO-Ciudad de Panamá, 10 son mujeres y apenas 3 son hombres, de los cuales 9 tienen entre 18 a 25

[12] Toda la información del corpus y su acceso están disponibles en https://esvaratenuacion.es/.

años, 3 tienen entre 26-55 y solamente 1 tiene más de 55 años. En cuanto a sus niveles de estudio, la mayoría (11) poseen estudios medios (educación secundaria) y únicamente 2 cuentan con estudios bajos (sin estudios o estudios primarios).

Todos los equipos que forman parte del macrocorpus de conversaciones coloquiales AMERESCO han desarrollado una transcripción que sigue unos mismos lineamientos homogéneos. Los equipos de las diversas ciudades, como es el caso del equipo de Ciudad de Panamá, llevan a cabo una primera transcripción ancha en Word, la cual consiste en una transliteración ortográfica acompañada de otros signos del sistema de transcripción del grupo Val.Es.Co. Este sistema contempla, por ejemplo, que para los alargamientos vocálicos o consonánticos se debe duplicar la vocal o consonante que se alarga, también establece que el inicio y el final del habla simultánea debe marcarse con dos corchetes «[ ]» o que los susurros se deben marcar con el siguiente símbolo «º()º», entre otros[13]. Posterior a ello, la coordinación técnica del corpus AMERESCO lleva a cabo una segunda transcripción que tiene como objetivo, en primer lugar, el alineamiento del texto con el audio, labor que se realiza en el programa ELAN©; y, en segundo término, el etiquetado y marcado pragmático discursivo. Algunas de las etiquetas empleadas y que sustituyen algunos símbolos de la transcripción ancha son, por ejemplo: para los fragmentos ininteligibles se usa la etiqueta simple: «<ininteligible/>», para los fragmentos proferidos entre risas se emplea la etiqueta doble: «<entre_risas> </entre_risas>», para indicar que se está reproduciendo el estilo directo se emplea la etiqueta doble: «<cita></cita>», o para marcar la pronunciación marcada se usa la etiqueta doble: «<énfasis t=“ ”></énfasis>», etc. Además, en este proceso se anonimizan los nombres propios de personas, lugares y mascotas que permitan identificar a los hablantes, tanto de los audios como de los textos, mediante la etiqueta: «<anónimo></anónimo>».

El corpus de conversaciones coloquiales AMERESCO-Ciudad de Panamá tiene cinco conversaciones accesibles en la página web del proyecto (https://esvaratenuacion.es/), tanto sus audios, transcripciones, como sus fichas técnicas.

- *El microcorpus AMERESCO de Tegucigalpa (Honduras)*

La segunda ciudad centroamericana en tener representación en el corpus AMERESCO es la ciudad de Tegucigalpa, Municipio del Distrito Central, de Honduras. El corpus AMERESCO-Tegucigalpa (Murillo Lanza, 2023) nació gracias al impulso de Danny Lanza, doctorando del programa de Doctorado en Estudios Hispánicos Avanzados de la Universitat de València. Su incorporación, en 2018, tanto al Departamento de Filología Española –como profesor e investigador en formación– y al grupo de investigación Val.Es.Co. (Valencia. Español. Coloquial.) le permitieron poder comenzar el proyecto.

La primera recogida de materiales, llevada a cabo en el año 2019, fue posible gracias a la colaboración del Departamento de Letras del Campus Central de la Universidad Pedagógica Nacional Francisco Morazán, de Tegucigalpa. Se recogieron en el marco de la asignatura de «Seminario del Español de Honduras» de la carrera del Profesorado en la

[13] Para consultar todo el sistema, las normas de transcripción y el etiquetado empleado, véase a Carcelén y Uclés (2019).

Enseñanza del Español, asignatura que era dirigida por la profesora Sandra Liz Irías. Se recolectaron 17 conversaciones, de las que se aceptaron 10.

En una segunda fase, ejecutada en el año 2022, en el marco de esa misma asignatura, pero, en este caso, con la colaboración del profesor Gustavo González Cáceres, se recolectaron 19 conversaciones, de la que se aceptaron 13. De manera que el corpus total recogido hasta el momento consta de 23 conversaciones. A pesar de ello, a la página web del proyecto solamente se tiene previsto dar acceso[14] a 19 de ellas, pues en cada una de estas se encuentra, al menos, un hablante que rellena algún grupo sociológico de la muestra básica exigida por el proyecto. Estas 19 conversaciones constituyen 6 horas con 44 minutos y 14 segundos.

Como podemos ver en la Tabla 9, el número total de hablantes que participan en estas 19 conversaciones es de 56. De ese número total, 37 son mujeres y 19 son hombres, de los cuales 22 tienen entre 18-25 años, otros 25 tienen entre 26-55 años y apenas 9 tienen más de 56 años. Respecto a sus estudios, 13 poseen estudios universitarios, 33 cuentan con estudios medios (secundaria, bachillerato, etc.) y únicamente 10 cuentan con estudios bajos (sin estudios o primaria):

Edad	Sexo	Nivel sociocultural			Total
		alto	medio	bajo	
18-25	Varón (4-4-4)	1	6	1	8
	Mujer (4-4-4)	1	13	0	14
26-55	Varón (4-4-4)	4	4	0	8
	Mujer (4-4-4)	4	8	5	17
≥ 56	Varón (4-4-4)	1	1	1	2
	Mujer (4-4-4)	2	1	3	6
Total		13	33	10	56

Tabla 9. Distribución y características de los informantes del microcorpus AMERESCO-Tegucigalpa

En lo que respecta a la transcripción, el corpus AMERESCO-Tegucigalpa, como hemos comentado anteriormente, sigue los lineamientos que ha establecido el macroproyecto, los cuales hemos explicado con más detalle en la descripción del corpus AMERESCO-Ciudad de Panamá. A pesar de ello, sí que cabría puntualizar algunas otras cuestiones importantes de la transcripción: por ejemplo, no se usan las mayúsculas por razones de puntuación, solamente en los nombres propios y siglas; los únicos signos de puntuación que se emplean según su función canónica son los signos de exclamación «¡!» y los de interrogación «¿?», o, por ejemplo, los números y símbolos se transcriben con letras.

Actualmente, el corpus AMERESCO-Tegucigalpa tiene disponible en la página web del proyecto 10 conversaciones (audios, transcripciones y fichas técnicas). Un aspecto que es

[14] Cabe acotar que actualmente se encuentran disponibles 10 conversaciones (audio, transcripción y fichas técnicas) y pronto se subirán las restantes 9.

imperativo apuntar es que, a diferencia de otros corpus, el corpus AMERESCO permite a los investigadores poder descargar la transcripción y el audio de las conversaciones completas y no solo de los fragmentos que se ofrecen al realizar una búsqueda.

A continuación, procedemos a reflexionar y determinar cuáles son algunos de los principales problemas sin resolver que manifiestan estos corpus.

4 UNA BREVE MIRADA VALORATIVA EN TORNO A LOS CORPUS ORALES DEL ESPAÑOL CENTROAMERICANO

Como hemos podido ver hasta aquí con la exposición de la compilación anterior, los países hispanohablantes del istmo centroamericano cuentan con diversos corpus orales, tanto de entrevistas como de conversaciones, que han sido recogidos desde finales del siglo pasado hasta la actualidad. Es justo reconocer la labor tanto de sus coordinadores e impulsores, sus colaboradores, sus participantes, etc. al ofrecer este material lingüísticos. Asimismo, es fundamental destacar y valorar el auspicio de diversas instituciones, la mayoría de ellas universidades, así como el apoyo financiero de los diversos proyectos, pues gracias a todos estos actores ha sido posible diseñar, grabar, transcribir y publicar (en la mayoría de los casos) estos corpus. Por todo ello, hemos querido recoger, en formato de compilación toda esta inversión de tiempo, esfuerzo y de recursos económicos.

Ahora bien, con el único objetivo de poder avanzar, mejorar y poder realmente aprovechar estos corpus, ofrecemos una sucinta reflexión respecto a qué problemas sin resolver hallamos en relación con el diseño, la construcción, el uso y el acceso de estos corpus. El fin último es poder ofrecer –a los investigadores que están construyendo estos corpus o que pretendan llevar a cabo este cometido– algunas líneas de trabajo que permitan, de alguna forma, desarrollar esta labor de mejor manera.

Para llevar a cabo esta reflexión, hemos decidido partir del trabajo del profesor Briz (2012:115), quien argumentaba que la lingüística con corpus orales del español manifiesta una imagen imperfecta todavía, a pesar de que, como bien señala, ha habido un enorme avance en el área. Es así como detecta 8 principales problemas, déficit o sesgos: a) la falta de muestras orales, b) la falta de materiales de conversaciones, c) la necesidad de crear corpus de géneros discursivos con registros más coloquiales y más formales, d) el acceso, e) la heterogeneidad de métodos, características y fines, f) la muestras, su diseño y elaboración; g) la transcripción y codificación, y, h) el soporte.

En consecuencia, a continuación, intentaremos corroborar si en los corpus orales del español centroamericano que hemos recopilado se manifiestan estos –u otros– déficit o problemas no resueltos todavía.

- *La falta de materiales de conversación*

Briz (2012:125) señala que «las conversaciones (coloquiales o formales) están representadas minoritariamente en los corpus actuales de lengua hablada en España e Hispanoamérica». En la compilación que hemos presentado este déficit se mantiene, pues, como habremos podido notar, en Centroamérica únicamente se han construido dos

corpus de conversaciones: el corpus AMERESCO-Ciudad de Panamá (5) y el corpus AMERESCO-Tegucigalpa (19).

Por el contrario, hallamos cinco corpus que han recolectado entrevistas: el corpus de la Norma Culta: San José (14) y Ciudad de Panamá (26), el corpus del PRESEEA: Ciudad de Guatemala (18) y Tegucigalpa (16), el corpus del EGREHA: San José (7) y Ciudad de Guatemala (2), el corpus COLEM: Costa Rica (7), El Salvador (7), Guatemala (9), Honduras (7), Nicaragua (7) y Panamá (6); y el corpus COLEH (165). Así pues, tal como se puede ver en la Figura 1., del total de muestras individuales que recogen todos estos corpus (315), un 94 % de estas corresponden a entrevistas semidirigidas o libres (291) frente a un 6 % que corresponden a conversaciones (24).

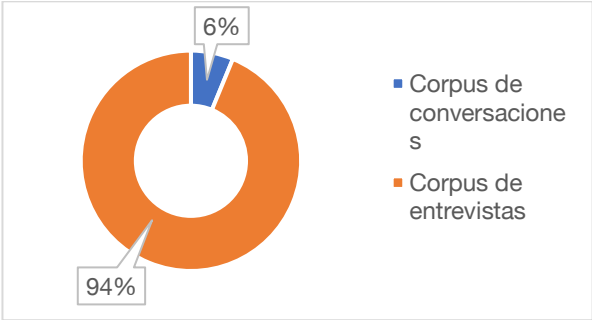

Figura 1. Gráfico del porcentaje de muestras recolectas en los corpus orales del español centroamericano según el género discursivo

- *La necesidad de crear corpus orales que reflejen diversos registros*

El autor considera que es necesario construir corpus de géneros discursivos que den cuenta tanto registros más informales o coloquiales como de registros más formales, pues hay muy pocos. Esta carencia podría hacer que pensemos que ciertos fenómenos lingüísticos no existan, pero, como él señala: «lo que no está… a veces sí existe». En particular, propone que se incorporen más muestras de géneros orales más coloquiales (más conversaciones coloquiales, por ejemplo) y más material de géneros orales más formales, pues sostiene que están poco representados.

En nuestra compilación sobre los corpus orales del español centroamericano se confirma, nuevamente, este sesgo, pues los únicos dos géneros recolectados son la entrevista (semidirigida y libre) y, de forma casi testimonial todavía, la conversación (coloquial). No hallamos, por ejemplo, corpus de noticieros, de tertulias televisivas y radiales, ni de debates en los respectivos congresos legislativos, tampoco encontramos representadas conversaciones grabadas en situaciones de mayor formalidad, por ejemplo, durante la consulta de médicos-pacientes o las que pueden llegar a producirse en una sala de profesores, etc.

- *Dificultad de acceso*

El profesor Briz (2012:126) explica que «si además de ser pocos los corpus orales, el acceso a estos es difícil, el problema aumenta considerablemente», cuestión que afirma que ocurre con los corpus del español construidos hasta ese momento. El caso de los corpus orales del español centroamericano no está exento, todavía, de este problema, pues, como se ha notado, muchos de los corpus recopilados en este trabajo:

a) no son accesible por completos (ni audios ni grabaciones)

b) solo se puede acceder a ellos parcialmente (bien a sus transcripciones, pero no a sus grabaciones o bien a una parte de su muestra total recogida)

c) aunque pronto se podrá acceder a ellos, todavía no es posible

En la siguiente tabla, mostramos, en concreto, el acceso que ofrecen los corpus orales del español centroamericano:

Corpus	Ciudad/país	Acceso				
		Total	Parcial		Sin acceso	En breve
			Audio	Transcripción		
Norma Culta	San José			X		
	Ciudad de Panamá					X
PRESEEA	Ciudad de Guatemala	X				
	Tegucigalpa					X
EGREHA	San José y Ciudad de Guatemala				X	
COLEM	CRI, SLV, GTM, HND, NIC y PAN			X		X
AMERESCO	Ciudad de Panamá	X				
	Tegucigalpa	X				X
COLEH	Honduras					X

Tabla 10. Acceso a los corpus orales del español centroamericano

Como se desprende de la tabla anterior, del total de corpus recopilados, únicamente tres microcorpus dan acceso total (audio y transcripciones): el corpus del PRESEEA-Ciudad de Guatemala, el corpus AMERESCO-Ciudad de Panamá y el AMERESCO-Tegucigalpa –si bien este último todavía no da acceso al total de muestras recogidas–. Un corpus no ofrece ningún acceso: el corpus EGREHA de San José y de Ciudad de Guatemala.

Por su parte, hay dos corpus que dan un acceso parcial (únicamente a las transcripciones): el corpus de la Norma Culta de San José y el corpus COLEM –aunque este último pondrá accesible en breve tanto sus audios como transcripciones–. Y, por último, hallamos tres corpus que todavía no han dado acceso ni a sus audios ni a sus transcripciones, pero que

se espera que lo hagan próximamente: el corpus COLEH, el corpus del PRESEEA-Tegucigalpa y el corpus de la Norma Culta de Ciudad de Panamá.

Ante tal panorama, podemos afirmar que los corpus orales del español centroamericano que aquí hemos recopilado manifiestan todavía una dificultad en lo que respecta a su acceso, pues, hasta el momento, únicamente podemos acceder totalmente (audio y transcripciones) a tres microcorpus, los cuales representan apenas un 10.47 % (33) del total de muestras individuales recogidas (315). Al ser material oral, deberían estar disponibles no solo las transcripciones, sino también los archivos de audio.

- *La heterogeneidad de métodos, características y fines*

La variedad en los métodos, las características y las finalidades de los corpus puede llegar a ser enriquecedor, pero, en palabras de Briz (2012:126) y otros autores que él cita, esta cuestión «dificulta el uso de estos desde otra perspectiva que no sea aquella para la que fueron diseñadas». Así pues, el autor sugiere que se debe tender «a la homogeneidad de los corpus», pues considera que, hasta ese momento, «ningún corpus oral del español es autosuficiente».

En nuestra compilación sobre corpus orales del español centroamericano, como hemos podido ver, la heterogeneidad de los corpus, en cuanto a sus métodos, por ejemplo, de selección de la muestra, de grabación, de transcripción, de acceso, de revisión y validación, etc. es palpable. Asimismo, aunque muchos de ellos comparten ciertas características, como el tipo de género discursivo recolectado, también manifiestan diferencias.

Esta última cuestión es razonable si partimos de la idea de que cada uno ha sido diseñado y construido para finalidades distintas. Para el caso, los corpus de la Norma Culta, aunque recogen entrevistas, tienen interés en el estudio lingüístico de un estrato sociocultural: el nivel alto; los corpus del PRESEEA, por su parte, pretenden no solo centrarse en un nivel sociocultural, sino que aspiran a llevar a cabo estudios de corte sociolingüísticos, para lo cual parten del análisis de la influencia de variables como el sexo, la edad o el nivel educativo de los hablante; los microcorpus del EGREHA aspiraban a recoger material oral que sirviera principalmente para su estudio gramatical; el corpus COLEM, por ejemplo, no se ha centrado en representar una ciudad en concreto, sino que recoge muestras de hablantes de todos los países hispanohablantes de Centroamérica que viven en Montreal (Canadá), para poder conocer, de forma general, cómo es el español de los inmigrantes en esta región; los corpus AMERESCO pretenden, sobre todo, llevar a cabo estudios pragmáticos y discursivos a partir de conversaciones coloquiales; mientras que el corpus COLEH, de entrevistas semidirigidas, pretende ofrecer una visión amplia del español actual de Honduras a partir de estudios sobre todo de tipo gramatical y léxico.

Sin lugar a duda, el hecho de que cada corpus se plantee objetivos concretos y distintos entre sí genera notables diferencias en lo que concierne, por ejemplo:

a) sus métodos de selección de muestra: mientras los corpus del PRESEEA o de AMERESCO establecen una muestra básica que cubra un determinado número de estratos sociolingüísticos (18 y 72 hablantes respectivamente), por su parte, hay otros que, por sus finalidades, buscan obtener una mayor cantidad de muestras, que contemplan los estratos sociolingüísticos, pero sin establecer una muestra básica: como puede ser el caso del COLEH (165 hablantes).

b) Sus métodos de grabación: en los corpus que hemos recopilado, de forma general, las grabaciones varían totalmente, por ejemplo, en relación con su duración individual: mientras que las entrevistas del PRESEEA deben durar, como mínimo, 45 minutos, las del COLEM y COLEH deben durar una hora, las del corpus de la Norma Culta de Ciudad de Panamá duran, en su mayoría, treinta minutos; y las grabaciones de conversaciones coloquiales de los corpus AMERESCO duran, como mínimo, veinte minutos. Otro ejemplo de esta heterogeneidad en el proceso de grabación tiene que ver con el tipo de encuestas o guías que se han usado, sobre todo en los corpus de entrevistas, pues en cada corpus se ha empleado una distinta.

c) Sus métodos de transcripción: como veremos más adelante, el tipo de transcripción desarrollada en cada corpus es distinta entre sí, en el único aspecto en el que coinciden es en que las transcripciones son transliteraciones ortográficas –y no transcripciones fonéticas–, pero en lo que respecta a las normas y signos de transcripción empleados, por ejemplo, hay notables diferencias entre sí.

d) Sus métodos de acceso y explotación: como hemos podido notar en el anterior subapartado, el acceso y la explotación que ofrecen los corpus recopilados varía: unos ofrecen un acceso completo, otros un acceso parcial, otros ningún acceso; unos corpus han desarrollado buscadores o están en proceso de desarrollarlos (AMERESCO, PRESEEA o COLEH), mientras que otros simplemente ofrecen sus materiales sin posibilidad de llevar a cabo búsquedas automáticas (por ejemplo, el corpus de la Norma Culta de San José).

En definitiva, si bien reconocemos que construir un corpus oral representa un enorme avance y aporte, como así ha sido con estos corpus orales del español centroamericano recopilados aquí, creemos, sin embargo, que se deben aunar fuerzas, tiempo y recursos para aprovecharlos, intentado homogeneizar, en la medida de lo posible, cuestiones que tienen que ver, principalmente, con la representatividad, las muestras, la grabación, la transcripción y el acceso.

Esto, en última instancia, posibilitará llevar a cabo estudios más completos, pues permitiría a los investigadores poder usar todas las muestras y, con ello, realizar contrastes más fiables.

- *Las muestras. Diseño y elaboración*

Señala el profesor Briz (2012:127-128) que la «amplitud (*de un corpus*[15]) se mide, sobre todo, con la representatividad o suficiencia». Así pues, la amplitud de un corpus dependerá de si, por ejemplo, las muestras quieren ser representativas de una ciudad, en particular, o si contemplan serlo de un país o de varias normas regionales. Ahora bien, en cualquiera de los casos, «deberían ser igualmente exhaustivos y representativos».

Como se puede ver en la Tabla 11, las muestras de los corpus orales del español centroamericano varían entre sí: mientras el corpus del PRESEEA de Ciudad de Guatemala recoge 18 entrevistas, del corpus EGREHA solo se puede acceder a 2 entrevistas; otro ejemplo puede ser el de los corpus de conversaciones coloquiales AMERESCO: mientras el de Ciudad de Panamá tiene disponible 5 conversaciones (13 hablantes), el de Tegucigalpa tiene disponible el doble (10) y pretende incorporar 9 más, que sumarían en total 19 (56 hablantes). Esto no representa un problema en sí, porque muchos de ellos se encuentran en proceso de construcción y pueden llegar a homogeneizar sus muestras, salvo algunos de ellos, como los corpus centroamericanos del EGREHA o el corpus de la Norma Culta de San José, que son corpus acabados.

Corpus	Ciudad/país	N.º de muestras del corpus	N.º de informantes
Norma Culta	San José	14	14
	Ciudad de Panamá	26	26
PRESEEA	Ciudad de Guatemala	18	18
	Tegucigalpa	16	16
EGREHA	San José	7	7
	Ciudad de Guatemala	2	2
COLEM	Costa Rica	7	7
	El Salvador	7	7
	Guatemala	9	9
	Honduras	7	7
	Nicaragua	7	7
	Panamá	6	6
AMERESCO	Ciudad de Panamá	5	13
	Tegucigalpa	19	56
COLEH	Honduras	165	165

Tabla 11. Número de muestras e informantes de los corpus orales del español centroamericano

Sin embargo, como bien argumenta Briz (2012:127), tanto la cantidad de las muestras debe ser un aspecto que se debe tener en cuenta, pero, quizá, es más importante que estas muestras sean representativas, por ejemplo, de los diversos estratos sociales (sexo, edad y nivel educativo), de las diversas variedades diatópicas del país que se pretenda representar o, bien, de las distintas situaciones comunicativas (géneros discursivos y registros) que puedan hallarse.

15 El comentario añadido es nuestro.

En cuanto a esto, de todos los corpus recopilados aquí, solamente el corpus del PRESEEA-Ciudad de Guatemala nos ofrece una muestra básica que integra, al menos, un informante de cada uno de los estratos sociales. Otros corpus, como los de AMERESCO o el COLEH, también han tomado seriamente en consideración estos parámetros para diseñar y construir sus muestras.

En otro orden, hemos detectado, por nuestra parte y relacionado con la representatividad de la muestra, que estos corpus orales del español centroamericano se centran en representar determinadas zonas geográficas del país. En particular, la mayoría se centran en los hablantes de las capitales: Tegucigalpa, San José, Ciudad de Guatemala y Ciudad de Panamá. Así que estos (o nuevos) corpus deberían tener en cuenta otras ciudades que aglutinan importantes núcleos poblacionales centroamericanos: por ejemplo, Alajuela o Cartago (Costa Rica); Mixco o Villa Nueva (Guatemala); San Pedro Sula o Choloma (Honduras); y San Miguelito o Arraiján (Panamá).

Por otro lado, sobre la representatividad diatópica de los corpus del español centroamericano, destaca la falta de muestras de hablantes que residan en dos países: El Salvador y Nicaragua. Por consiguiente, habría que impulsar la construcción de corpus orales del español de ciudades tan importantes como Managua o León, en Nicaragua; o de San Salvador, Santa Ana y Soyapango, en El Salvador. Para lo cual es necesario, tal como arguye el profesor Briz (2012:128), primero, el trabajo en equipo, y, segundo, la formación de expertos en la lingüística de corpus.

- *La transcripción y codificación*

Comenta el profesor Briz (2012:128) que «cualquier sistema de transcripción es adecuado siempre que se ajuste al objeto de estudio y a la finalidad para la que se emplee y, por supuesto, cumpla los principios de exhaustividad y pertinencia de los signos». Esto nos permitiría afirmar que el hecho de que los corpus recopilados hayan empleado diferentes sistemas de transcripción. Ahora bien, de lo que no se da cuenta en muchos de estos corpus orales del español centroamericano es del proceso de control, revisión o validación de las transcripciones: por ejemplo, los filtros de selección puestos en marcha o el tipo de personas que transcriben las muestras recolectadas (¿estudiantes, profesores?).

Un segundo aspecto relacionado con la transcripción es el método que se adopta para llevar a cabo tal cometido, en todos los corpus orales del español centroamericano, como hemos explicado, el tipo método de transcripción empleado ha sido el que se basa en una transliteración ortográfica (no transcripción fonética) que es enriquecida con signos que marcan fenómenos propios de la lengua hablada o fenómenos de interés para el proyecto que construye el corpus.

Por último, una tercera cuestión en torno a la cual si hallamos un problema es la que tiene que ver con el marcaje o el etiquetado para su tratamiento informático en los buscadores respectivos. Según el profesor Briz (2012:130), «sería ideal que un corpus dispusiera de una transliteración con signos y convenciones específicas, junto con la codificación a través de marcas automáticas». En nuestra recopilación, hasta el momento solo encontramos

disponibles dos corpus que sigan esta recomendación: el del PRESEEA y el de AMERESCO, tal como lo hemos ejemplificado en los apartados correspondientes. A estos corpus, se añadirá próximamente tanto el corpus COLEM como el COLEH, pues ambos estarán disponibles, afortunadamente, en plataformas digitales. Por el contrario, el corpus de la Norma Culta y el corpus EGREHA son los únicos dos que no disponen de un etiquetado o marcaje automático.

- *El soporte*

El último de los problemas que detecta el profesor Briz (2012) es el que tiene que ver con el soporte en el que se almacenan los corpus. El autor reconoce, en primera instancia, que es obvio que en la actualidad los datos deben almacenarse en soporte digital o en formato electrónico (por ejemplo, en DVD o en discos de CD-ROM). Al respecto, en nuestra opinión, el soporte digital (páginas web con buscadores automáticos) es la mejor opción hoy en día. Pero, por otro lado, Briz insiste en que, además, de almacenarse de esta forma, los corpus deben publicarse en corpus de papel o en formato de libros digitales –aunque sin marcas o etiquetas–.

Sobre estas dos cuestiones, los corpus orales del español centroamericano compilados también manifiestan ciertos problemas. Como hemos comentado, del total de microcorpus, únicamente tres se encuentran almacenados y disponibles actualmente en formato digital (páginas web y buscadores que incluyen filtros avanzados y de gran ayuda): el del PRESEEA-Ciudad de Guatemala, el de AMERESCO-Ciudad de Panamá y la primera parte del corpus AMERESCO-Tegucigalpa. El resto o bien están almacenados en formato electrónico (CD-ROM): el corpus de la Norma Culta de San José, o bien están en proceso de estar almacenados y accesibles en formato digital: COLEM, COLEH, el PRESEEA-Tegucigalpa y el corpus de la Norma Culta de la Ciudad de Panamá, o, en su caso extremo, no están almacenados ni disponibles en ningún formato: ni digital, ni electrónico, ni en papel: los corpus del EGREHA de San José y Ciudad de Guatemala.

Y, por último, respecto a su publicación en papel o en formato de libros digitales, cabría señalar que, actualmente, ninguno está publicado de esta forma. No obstante, sí es justo mencionar que hay algunos que ofrecen las transcripciones en formato Word, PDF o TXT, como es el caso del corpus de la Norma Culta de San José (Word), los microcorpus centroamericanos del COLEM (PDF) y los de AMERESCO, tanto el de Ciudad de Panamá como el de Tegucigalpa, los cuales se pueden descargar en formato TXT. Habría que mencionar que los corpus del PRESEEA, en su página web anterior, también ofrecían esta posibilidad, sin embargo, parece que en su nueva plataforma esta opción no está disponible.

- *Difusión y uso*

Finalmente, con el objetivo de cerrar estas reflexiones valorativas respecto de los corpus orales del español centroamericano que hemos incluido y expuesto en este trabajo, queremos añadir un problema que hemos notado. Este no tiene tanto que ver con la construcción, el diseño, etc., sino más bien con el uso y la explotación que se les ha dado.

En términos generales, aunque podemos hallar algunos trabajos publicados que han partido de los materiales que han sido recolectados en estos corpus, creemos que la cantidad es totalmente inferior si lo comparamos con el tiempo, el esfuerzo y la financiación que se pudo haber invertido en todos ellos. Así pues, creemos que es fundamental poder usar, explotar, difundir y llevar a cabo más estudios a partir de ellos.

5 A MODO DE CONCLUSIÓN

Este trabajo ha ofrecido un panorama actual de los principales corpus del español de los países del istmo centroamericano: Costa Rica, El Salvador, Guatemala, Honduras, Nicaragua y Panamá. Como hemos podido ver, son 10 los corpus que se han construido (o están en proceso): dentro del proyecto de la Norma Culta, el de San José de Costa de Rica y el Ciudad de Panamá; dentro del PRESEEA, el de Ciudad de Guatemala y el de Tegucigalpa; dentro del EGREHA, el de San José de Costa Rica y el de Ciudad de Guatemala; el del COLEM (con muestras de los cinco países centroamericanos); dentro del proyecto AMERESCO, el de Ciudad de Panamá y el de Tegucigalpa; y, por último, el corpus COLEH. Los corpus han sido construidos desde finales del siglo pasado hasta la actualidad. Sus coordinadores han sido investigadores que han formado parte o pertenecen a universidades de la región o extranjeras. El número total de muestras recolectadas varía según el corpus, pero, en total, hemos podido dar cuenta de aproximadamente 315 (24 conversaciones coloquiales y 291 entrevistas), en las cuales participan aproximadamente 360 hispanohablantes centroamericanos (317 residen en los diversos países y 43 son centroamericanos en la diáspora).

En lo que respecta a la transcripción de este material, habría que señalar que ha sido transcrito a partir de una transliteración ortográfica que ha sido acompañada de unas normas y signos de transcripción propias, en su mayoría diferentes entre sí. Además, únicamente cuatro de los corpus contemplaron o han contemplado la incorporación de un etiquetado de fenómenos propios de la lengua oral –aunque también, en algunos casos, pragmáticos, discursivos, léxicos, prosódicos, etc.–: PRESEEA, AMERESCO, COLEM y el COLEH.

Entre tanto, en la actualidad únicamente se puede acceder por completo (grabaciones y transcripciones) a un 10.4 % de las muestras totales. Aunque se espera que pronto pueda estar disponible completamente otro 82,2 % del material recopilado –el del PRESEEA-Tegucigalpa, el del COLEM, el del COLEH, una segunda parte del corpus AMERESCO-Tegucigalpa y el corpus de la Norma Culta de Ciudad de Panamá–. El otro porcentaje restante, solo podrá estar disponible de forma parcial (por ejemplo, las transcripciones del corpus de la Norma Culta de San José de Costa Rica) o bien se podrá acceder a ellos únicamente a través del contacto personal con su coordinador (los corpus del EGREHA de San José de Costa Rica y de Ciudad de Guatemala).

Finalmente, como hemos podido notar, los corpus orales del español centroamericano manifiestan la mayoría de los problemas que ya detectó el profesor Briz (2012) en su trabajo sobre sobre los déficits de los corpus orales del español y de algunos análisis. Así pues, en términos generales, urge a) que estos corpus puedan ofrecer un acceso completo tanto

a sus grabaciones como transcripciones, en la medida de lo posible, b) que las muestras sean ampliadas (que se recojan más) y que sean representativas (tanto de los estratos sociales, de las variedades diatópicas más importantes [Managua y San Salvador, por ejemplo], como de las situaciones y géneros discursivos [más coloquiales, como las conversaciones; y más formales, como los debates en los congresos de los diputados o los noticieros, etc.]; c) que las transcripciones incorporen un sistema combinado de símbolos y etiquetas (enfocándose, además, en un marcaje pragmático o prosódico, por ejemplo); d) que se almacenen y estén disponibles en medios digitales (páginas web con buscadores automáticos que incluyan filtros de búsqueda), pero, que a la vez, ofrezcan la posibilidad de acceder a los textos completos de forma simplificada (sin etiquetas). Cuestiones que pueden ser superadas con mayor facilidad si son abordadas de forma exhaustiva, metódica y rigurosa por equipos de trabajo (profesores, estudiantes e instituciones) y no de forma particular.

Y, sobre todo, urge que los investigadores interesados en el estudio del español centroamericano lleven a cabo trabajos a partir de estos materiales, con el objetivo de ser publicados respectivamente o difundidos en congresos o jornadas científicas, ya que, desde nuestra perspectiva, no sirve en absoluto la inversión de tanto tiempo, esfuerzo y dinero, si los materiales no son aprovechados para el objetivo final de todo corpus: su estudio y explotación.

6 REFERENCIAS BIBLIOGRÁFICAS

Albelda Marco, Marta (2022): «Los corpus del español hablado y los estudios pragmáticos», en Parodi, Giovanni y otros, eds., *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, Londres, Routledge, 222-238, https://doi.org/10.4324/9780429329296-18

Albelda, Marta y Antonio Briz (2009): «Estado actual de los corpus de lengua española hablada y escrita: I+D», en Instituto Cervantes, ed., *El español en el mundo. Anuario del Instituto Cervantes 2009*, Madrid, Santilla, 165-225.

Albelda, Marta y Maria Estellés (2023): *Corpus Ameresco*, Valencia, Universitat de València [en línea]: www.corpusameresco.com

Briz, Antonio y Marta Samper Hernández (2022): «Estudio de variación situacional en corpus orales del español», en Parodi, Giovanni y otros, eds., *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, Londres, Routledge, 309-324, https://doi.org/10.4324/9780429329296-24

Briz Gómez, Antonio (2012): «Los déficits de los corpus orales del español (y de algunos análisis)», en Jiménez Juliá, Tomás Eduardo y otros, coords., *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*, Santiago de Compostela, Servizo de Publicacións da Universidade de Santiago de Compostela, 115-137.

Briz Gómez, Antonio (2018): «Los corpus de conversaciones coloquiales. La elaboración del corpus AMERESCO (Español coloquial de América y España)», Ponencia de clausura en el I Col·loqui Internacional de Lingüística de Corpus (LingCor 2018), Valencia, del 13 al 14 de diciembre de 2018.

Briz, Antonio y Andrea Carcelén Guerrero (2019): «El futuro iberoamericano del español: la investigación del español oral y en español», en Instituto Cervantes, ed., *El español en el mundo. Anuario del Instituto Cervantes 2019*, Madrid, Bala Perdida e Instituto Cervantes, 189-217.

Briz, Antonio y Grupo Val.Es.Co. (2002): «La transcripción de la lengua hablada: el sistema del grupo Val.Es.Co.», *Español Actual*, 77, 1-30.

Carcelén Guerrero, Andrea y Gloria Uclés Ramada (2019): «Diseño y construcción de un corpus oral multidialectal. El corpus Ameresco», *Normas. Revista de Estudios Lingüísticos Hispánicos*, 9, 1, 17-36, https://doi.org/10.7203/Normas.v9i1.16007

Crystal, David (1991): *A Dictionary of Linguistics and Phonetics*, London, Blackwell, 3.ª ed.

Escobar Samaniego, Linier Enrique (2017): *Estudio de los marcadores del discurso en muestras de habla culta de Panamá*, Ciudad de Panamá, Repositorio Institucional Digital de la Universidad de Panamá, http://up-rid.up.ac.pa/1544/

Hincapié Moreno, Diana Alejandra y Julio Alexánder Bernal Chávez (2018): *Lingüística de corpus*, Bogotá, Instituto Caro y Cuervo.

Instituto Cervantes (2023): *El español en el mundo 2023. Anuario del Instituto Cervantes*, Madrid, Instituto Cervantes.

Lope Blanch, Juan (1986): *El estudio del español hablado culto. Historia de un proyecto*, Ciudad de México, Instituto de Investigaciones Filológicas de la UNAM.

Morales del Castillo, Fulvia (2023): «Corpus de conversaciones Ameresco-Ciudad de Panamá», en Albelda, Marco y Maria Estellés, coords., *Corpus Ameresco*, Valencia, Universitat de València [en línea]: www.corpusameresco.com

Moreno Fernández, Francisco (2021*a*): *Metodología del "Proyecto para el estudio sociolingüístico del español de España y de América" PRESEEA*, Alcalá de Henares, Editorial Universidad de Alcalá, https://doi.org/10.37536/preseea.2021.doc1

Moreno Fernández, Francisco (2021*b*): *Marcas y etiquetas mínimas obligatorias para materiales de PRESEEA*, Alcalá de Henares, Editorial Universidad de Alcalá, https://doi.org/10.37536/preseea.2021.doc2

Murillo Lanza, Danny Fernando (2023): «Corpus de conversaciones Ameresco-Tegucigalpa», en Albelda, Marco y Maria Estellés, coords., *Corpus Ameresco*, Valencia, Universitat de València [en línea]: www.corpusameresco.com

Parodi, Giovanni (2010): *Lingüística de Corpus*: *de la teoría a la empiria*, Madrid, Iberoamericana.

Parodi, Giovanni y otros, eds. (2022): *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics*, Londres, Routledge, https://doi.org/10.4324/9780429329296

Pato, Enrique (2023*a*): *COLEM (Corpus oral de la lengua española en Montreal)*, Montréal, Université de Montréal [en línea]: https://esp-montreal.jimdo.com/

Pato, Enrique (2023*b*): *Corpus oral de la lengua española en Honduras (COLEH)*, Montréal, Université de Montréal [en línea]: https://n9.cl/982gj

Pérez Hernández, Chantal (2002): «Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento», *Estudios de Lingüística del Español*, 18.

PRESEEA (2014): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*, Alcalá de Henares, Universidad de Alcalá [en línea]: http://preseea.uah.es

Real Academia Española (2010): *Ortografía de la lengua española*, Madrid, Espasa.

Rojo Sánchez, Guillermo (2016): «Los corpus textuales del español», en Gutiérrez-Rexach, Javier, ed., *Enciclopedia lingüística hispánica*, London, Routledge, 285-296, https://doi.org/10.4324/9781315713441-99

Rojo, Guillermo (2021): *Introducción a la lingüística de corpus en español*¸ Oxon-New York, Routledge, https://doi.org/10.4324/9781003119760

Samper Padilla, José Antonio (1995): «Criterios metodológicos del "Macro-corpus" de la Norma lingüística culta de las principales ciudades del mundo hispánico», *Lingüística*, 7, 263-293.

Samper, José Antonio y otros (1998): *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, Las Palmas de Gran Canaria, Universidad de Las Palmas de Gran Canaria.

Solís García, Inmaculada (2018): «Corpus españoles dialógicos para el análisis de la conversación», *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, 5, 1, 117-129, https://doi.org/10.15366/chimera2018.5.1.010

Spitzová, Eva (1991): «Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica : proyecto y realización», *Études romanes de Brno*, 21, 1, 61-66.