ISSN: 2174-7245
Año 2024, Volumen 14, Número 1.doi: 10.7203/Normas.v14i1.29929
Publicado: 2024. Enviado: 2024-11-01. Aceptado: 2024-12-08.

Novedades en los corpus digitales para el estudio del español oral

Recent Developments in Digital Corpora for the Study of Spoken Spanish

Andrea Carcelén Guerrero
Universidad de Helsinki

Abstract

The aim of this paper is to provide an updated overview of Spanish corpora available online for the study of spoken Spanish. Specifically, it highlights recent developments as well as other corpus projects that make their materials publicly accessible to the entire scientific community. In this context, a list of Spanish oral corpora is presented along with a descriptive table, offering a useful resource for researchers interested in working with spoken Spanish. This enables them to become familiar with the available corpora to date and their main characteristics.

Keywords: oral corpora, corpus linguistics, spoken Spanish, synchronic study.

Resumen

El objetivo de este trabajo es presentar un panorama actualizado de corpus de español disponibles en línea para el estudio del español oral. En concreto, se da cuenta de las novedades aparecidas recientemente, así como de otros trabajos de corpus que ponen a disposición pública sus materiales para toda la comunidad científica. En este sentido, se ofrece el listado de corpus orales de español con su correspondiente ficha descriptiva, de manera que sea un recurso útil para que aquellos investigadores e investigadoras que deseen trabajar con español oral puedan conocer los corpus disponibles hasta la fecha, así como sus características fundamentales.

Palabras clave: corpus orales, lingüística de corpus, español oral, estudio sincrónico.

Citar como: Carcelén Guerrero, Andrea (2024). Novedades en los corpus digitales para el estudio del español oral. Normas, 14(1), 241-260, doi: 10.7203/Normas.v14i1.29929.

1. Introducción

Podemos considerar que hoy en día el español es una de las lenguas con mayor desarrollo de corpus orales. Así lo demuestran los trabajos previos que ofrecen con diferente grado de detalle, una amplia visión del panorama de corpus de español[1]. Las recopilaciones más relevantes hasta la actualidad[2], hasta donde hemos podido conocer, son las realizadas por Moreno (2005), Briz y Albelda (2009), Enghels, *et alii* (2015), Rojo (2016), Solís (2018), Parodi y Burdiles (2019), Briz y Carcelén (2019), Llisterri (2021), Briz y Samper (2022), para el español europeo y americano, y Lanza (2023) centrado en el español centroamericano. Estas investigaciones recogen, entre otras variables, cuáles son estos corpus, qué información ofrecen o dónde se pueden consultar, partiendo de la base de que toda recopilación muestra el panorama existente hasta su fecha de publicación y que, además, la elección de unos corpus y no otros puede responder a diferentes objetivos de investigación y adoptar distintos criterios de selección. Así, por ejemplo, el trabajo de Solís (2018) se centra en corpus dialógicos para el análisis de la conversación, Parodi y Burdiles (2019) recopilan aquellos corpus útiles para la enseñanza del español como L2 y Llisterri (2021) reúne trabajos de corpus que pueden contribuir al estudio del componente fónico en español como LE/L2.

Estos recopilatorios muestran la perspectiva del momento en el que se publicaron y, por tanto, no recogen todos los corpus existentes hasta el momento. Así, corpus como el *Corpus del Español del s. XXI* (CORPES XXI), publicado en 2013, no aparece representado en los primeros estudios revisados.

En trabajo tiene como objetivo presentar una recopilación de corpus orales del español que ofrezca una actualización del panorama de corpus a fecha de hoy que complete el listado de corpus orales recogido en los trabajos mencionados anteriormente. Como eje vertebrador de esta selección se encuentra el criterio principal de ser corpus orales disponibles en línea y de acceso abierto para toda la comunidad científica.

A continuación, en el apartado 2, se presentan todos los criterios que han operado en la selección de los corpus orales de español que conforman esta panorámica. En el apartado 3, se da cuenta de las características de cada uno de ellos; se ofrece, en primer lugar, el listado de corpus recogidos junto con el acceso web y, en segundo lugar, se detallan las particularidades de cada corpus, incluyendo, entre otros datos, la tipología, los datos del equipo o la persona responsable de su elaboración o el género discursivo que engloba. Por último, en el apartado 4,

[1]Esta publicación ha sido posible gracias a la ayuda de dos proyectos: el proyecto de I+D+i "Estrategias pragmático-retóricas en la interacción conversacional conflictiva entre íntimos y conocidos: intensificación, atenuación y gestión interaccional (ESPRINT)" (ref. PID2020-114805GB-I00), financiado por MICIU/AEI/10.13039/501100011033/ y el proyecto ECOS-C/N, Estudio de los condicionantes sociales del español actual en el centro y norte de España: nuevas identidades, nuevos retos, nuevas soluciones, (ref. PID2023-148371NB-C42).

[2]No existe hasta ahora ningún repositorio en línea que recopile el panorama de corpus orales del español, plataformas que sí existen para el caso de corpus de aprendices de español (Díaz, 2018, en línea) o para corpus históricos de carácter iberorrománico (Torruella y Kabatek, 2018, en línea). Tampoco es posible la búsqueda combinada en diferentes corpus al mismo tiempo. Más allá de esto, lo que encontramos son pequeños recopilatorios ofrecidos por diferentes grupos de investigación en lingüística en sus respectivas páginas web o recursos como el de la Universidad de Gante que ha publicado en línea un repositorio de corpus con el que pretende ofrecer una visión general de los corpus existentes para las lenguas que se enseñan en el Departamento de Lingüística de dicha universidad, (accesible en https://www.corpusfinder.ugent.be/corpora#edit-language-collapsible--2--content).

las consideraciones finales recogen la reflexión sobre el estado actual del panorama de corpus orales.

2. Criterios de selección

La particularidad de los corpus señalados aquí, a diferencia de los trabajos citados en el apartado anterior, radica en que solo se atiende a corpus orales con acceso abierto en línea a toda la comunidad de usuarios y usuarias; no se recogen, por tanto, aquellos que han sido publicados en papel o que cuentan con una versión electrónica no disponible en línea (como, por ejemplo, CD-ROM u otros soportes que requieran de compra o suscripción). Concretamente, para la selección de los trabajos que presentamos se han tenido en cuenta los siguientes criterios:

- son corpus orales que recogen muestras de habla reales en su contexto natural de enunciación;
- sincrónicos;
- de acceso libre y no comercial, aunque se requiera cuenta de usuario para poder acceder a los materiales;
- de cualquier variedad del español;
- con propósitos lingüísticos generales; es decir no se recogen aquellos corpus cuya finalidad sea el desarrollo de tecnologías de habla, corpus de aprendices de español ni otros corpus como los construidos para adquisición y desarrollo del lenguaje o patologías del lenguaje;
- están disponibles en formato digital, aunque sea en grado mínimo, con acceso en línea a los datos, ya sea por medio de plataformas de búsqueda, o bien a través de la descarga directa de los materiales (audio, transcripción y metadatos); se han excluido aquellos que han sido publicados en papel si no cuentan con acceso directo en línea;
- de cualquier género discursivo;
- que se encuentren tanto finalizados como en construcción, pero que tengan la voluntad de ser puestos a disposición pública una vez completados.

Para su descripción, se han seguido consideraciones previas (Briz y Albelda 2009, Briz y Carcelén, 2019, Llisterri, 2021) que permiten entender la terminología empleada, así como el tipo de información aportada para cada material:

- Solo se incluyen corpus orales, aunque en su contexto pertenezcan a otro gran corpus que pueda incluir también material escrito.
- Priorizamos, asimismo, el modo de acceso y consulta de los datos, esto es, que sea en línea, con o sin motor de búsqueda.
- Respecto de la variedad dialectal que se recoge, se distingue entre *corpus panhispánicos*, es decir, que aspiran a recoger muestras de las principales variedades del español; *corpus de diversas variedades* del español, aquellos que incluyen diferentes variedades, sin aspirar a recoger todas ellas; y *corpus de variedades dialectales particulares*. En nuestra

opinión, un corpus sería panhispánico cuando aspire a recoger muestras de habla de las principales ciudades de habla hispánica, tanto en el continente europeo como en el americano. Es el caso de corpus como PRESEEA, CORPES XXI o Ameresco, que, si bien no incluyen aún todos los dialectos, sí que tiene voluntad de recogerlos. Si solo recogen muestras de alguna de estas variedades hablaremos de *corpus multidialectal,* como el caso del corpus COLA, que reúne muestras del español de Madrid, Buenos Aires y Santiago de Chile, o de *una sola variedad*, como el corpus COJEM, que estudia el español de Mallorca.

- Se explica qué materiales ofrece (transcripción, audio, información contextual) cada corpus y si se permite la descarga.
- En cuanto al tamaño, se reconocen por un lado, *macrocorpus* que ofrecen la posibilidad de contrastar resultados entre distintas normas regionales y que pretenden constituirse en grandes bases de datos, y por otro, *microcorpus* cuyo ámbito de acción y dimensiones son más reducidas, así como sus objetivos, que son más concretos (Briz, 2012: 116-117). Más particularmente, en este punto, seguimos la propuesta de Briz (2012), que considera que un material es un macrocorpus si recoge, al menos, tres variedades de habla. Por el contrario, un microcorpus será aquel material que recoja dos variedades o una. Conviene tener en cuenta esta precisión, pues existe la tendencia en la bibliografía sobre el tema a asociar los términos *macrocorpus* y *microcorpus* al número de palabras o formas que componen un corpus. La propuesta de Briz (2012) que aquí seguimos sobre el número de variedades dialectales que recoge el corpus, nos parece más delimitante y objetivable para asignar estas etiquetas.
- Se señala el género discursivo que recogen (entrevista, conversación, miscelánea, etc.).
- Se expone cuál es su objetivo de estudio, la finalidad con la que ha sido concebido inicialmente.

La recopilación de corpus que sigue se ha presentado en orden alfabético, para facilitar su consulta. Se ofrece, en primer lugar, el listado de los corpus en formato de tabla; y, en segundo lugar, la descripción detallada de cada uno de los corpus.

Es necesario hacer una última matización sobre los materiales aquí recogidos. Hoy en día, el desarrollo de las tecnologías y la cantidad de herramientas informáticas de las que disponemos hacen posible la obtención y recolección de grandes cantidades de datos, incluyendo materiales orales que, debidamente clasificados y procesados podrían ser considerados materiales lingüísticos para fines de investigación. Bastaría con descargar, por ejemplo, vídeos y/o audios de redes sociales como Youtube, Tik Tok, o audios enviados a través de aplicaciones de mensajería instantánea como WhatsApp o Telegram, que no tienen un propósito lingüístico por sí mismos, para poder realizar cualquier investigación. Sin embargo, y tal y como ya ha planteado la bibliografía sobre el tema (Torruella y Llisterri, 1999, McEnery *et alii,* 2006, Briz y Albelda, 2009, entre otros), para que un material sea considerado corpus lingüísticos indefectiblemente la selección del material debe haberse realizado atendiendo a unos objetivos de investigación y unos criterios de selección de la muestra. Asimismo, a nuestro juicio, y teniendo en cuenta las obras sobre repertorios de corpus revisadas, dicha muestra debe haber recibido un tratamiento informático que permita el trabajo con el audio/vídeo y la transcripción.

3. Corpus orales del español

A continuación, en la Tabla 1, se presentan por orden alfabético los corpus que conforman la versión actualizada del panorama de corpus orales del español, reunidos aquí bajo los criterios clasificatorios explicados en el apartado anterior.

Una vez listados, se ofrece las características técnicas de cada uno de ellos en la que se incluyen sus características más relevantes. No obstante, aparecen referidos tras los cuadros descriptivos otros proyectos que consideramos dignos de mención, pero que o bien no cumplen con todos los criterios establecidos en el apartado 2, o bien poseen unas características diferentes. Por esta razón no siguen la misma línea estructural que los anteriores.

Como hemos comentado anteriormente, los recopilatorios son reflejos del momento en el que se realizaron, y aunque para esta compilación se ha consultado bibliografía específica, así como realizado búsquedas exhaustivas con el objetivo de ofrecer un panorama lo más completo posible, debe entenderse que es imposible recoger todos los corpus existentes, a pesar de contar con la voluntad de hacerlo.

Corpus	**Acceso en línea**
AMERESCO	https://corpusameresco.org/
CE	https://www.corpusdelespanol.org/hist-gen/
CEMC	https://cemc.colmex.mx/
CEMC II	https://cemcii.colmex.mx/
CET	https://corpus.spanishintexas.org/es/sobre-el-corpus
COJEM	PDF Link
COLA	https://blogg.hiof.no/colam-esp/el-corpus-cola/
CoLaGe	https://www.kielipankki.fi/corpora/
COLEH	https://portfolio.umontreal.ca/view/view.php?id=269744
COLEM	https://esp-montreal.jimdo.com/corpus/
CORDIAL	http://lablita.it/app/cordial/corpus.php
CORLEC	http://www.lllf.uam.es/ESP/Corlec.html
CORPES XXI	https://www.rae.es/corpes/
CORPEUU	https://corpeeu.org/
Corpus del habla de Almería	https://www2.ual.es/ilse/corpus/
COSER	http://www.corpusrural.es
CREA	https://www.rae.es/banco-de-datos/crea
El español hablado en Bogotá	https://clicc.caroycuervo.gov.co/corpus/EHB
ESLORA	https://eslora.usc.es/
MESA	http://www.grupoapl.es/materiales-corpus/corpus-mesa
PRESEEA	https://preseea.uah.es/
Val.Es.Co.	https://valesco.es/

Tabla 1. Corpus orales del español disponibles en línea[3][4]

Se describen, a continuación, las principales características de los corpus recogidos en la Tabla 1. De cada uno de ellos se detalla su nombre completo, el acceso a su página electrónica, el nombre de la persona o grupo responsable, la variedad dialectal que recoge, el género o géneros

[3]No se han considerado aquellos corpus orales que tienen acceso en línea, pero que forman parte de otros macrocorpus, como sucede con el *Corpus Sociolingüístico de la ciudad de México* o el *Vernáculo Urbano Malagueño* (VUM) que forman parte de PRESEEA, aunque cuenten con sus propias páginas web donde se puede acceder a los materiales.

[4]Recogemos aquí el acrónimo de cada corpus. El nombre completo del corpus aparece desarrollado en el cuadro descriptivo correspondiente.

discursivos que trabaja, su objeto de estudio, los modos de consulta en línea y los materiales que se ponen a disposición de la comunidad de usuarias y usuarios.

3.1. Corpus Ameresco (América y España Español Coloquial)

El proyecto Ameresco surge de la mano de Antonio Briz (2010) como extensión natural del corpus Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002). Es un corpus aún en construcción y, hasta el momento actual, cuenta con 202 conversaciones coloquiales recogidas en 17 ciudades del ámbito hispánico, tanto del español europeo como del español hispanoamericano, con un número de palabras aproximado de 825 000.

i Corpus Ameresco

- **Enlace**: https://corpusameresco.org/
- **Responsable**:
  - Antonio Briz Gómez (director académico)
  - Marta Albelda Marco y Maria Estellés Arguedas (coordinadoras académicas)
  - Universitat de València
- **Variedad dialectal**: Multidialectal de carácter panhispánico
- **Género discursivo**: Conversación coloquial espontánea
- **Objeto de estudio**: El estudio de la conversación coloquial en español, con especial interés por el estudio de la atenuación pragmática y los fenómenos pragmático-retóricos de la interacción conversacional espontánea.
- **Consulta en línea**: Sí, con motor de búsqueda y con repositorio de archivos.
- **Materiales disponibles**:
  - Permite la descarga completa de los audios anonimizados, las transcripciones tanto alineadas como en formato texto, las fichas técnicas con los metadatos de cada archivo y el Textgrid para su análisis en PRAAT. El motor de búsqueda dispone de consulta básica y consulta avanzada.

3.2. Corpus del Español de Mark Davies (CE)

Este corpus surge en 2001 por iniciativa de Mark Davies, subvencionado por el programa National Endowment for the Humanities de Estados Unidos. Está compuesto por cuatro subcorpus (género/histórico, web/dialectos, NOW 2012-2019 y Google Books n-grams BYU)[5] que, en total contiene 45 500 millones de palabras aproximadamente. En la última actualización de 2022 se han incrementado las funcionalidades del motor de búsqueda, según se indica en la propia página web (CE, en línea).

i Corpus del Español de Mark Davies

- **Enlace:** https://www.corpusdelespanol.org/
- **Responsable:** Mark Davies
- **Variedad dialectal:** Panhispánico

[5]Para el cuadro descriptor únicamente nos referiremos a su parte oral.

- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Obtener las características globales que presenta una lengua en un momento determinado de su historia.
- **Consulta en línea:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Dispone de material procedente del medio oral en la subsección *Genre-Historical*.
  - Recupera la información por medio de concordancias. No permite la descarga del material ni facilita acceso al audio.

3.3. Corpus del Español Mexicano Contemporáneo-CEMC (I y II)

Ambos corpus, dirigidos por Luis Fernando Lara, nacen con la voluntad de reunir muestras que sean representativas de los usos de la lengua en México en dos periodos de tiempo, de 1921 a 1974 y de 1975 a 2018. Esta recopilación está orientada, principalmente, al estudio del léxico mexicano, tanto del más reciente, como del tradicional, sin perjuicio para los análisis sintácticos que también son posibles con estos materiales.

i Corpus del Español Mexicano Contemporáneo (CEMC)

https://repositorio-cell.colmex.mx/corpus.html

- **Responsable:** Colegio de México, dirigido por Luis Fernando Lara
- **Variedad dialectal:** México
- **Género discursivo:** Miscelánea
- **Objeto de estudio:** Estudio del léxico
- **Consulta en línea:**
  - Sí, con motor de búsqueda.
  - Permite búsquedas combinadas en el CEMC I y II, así como en otros corpus procedentes del medio escrito.
  - Ofrece opciones para filtrar por año y género (tipología de la muestra).
  - Proporciona información estadística sobre la frecuencia.
- **Materiales disponibles:**
  - Solamente se recuperan los resultados por concordancias. No permite la descarga.
  - Incorpora materiales del *Corpus Sociolingüístico de la ciudad de México*.

3.4. Corpus del Español en Texas (CET)

Este corpus, originado en la University of Texas en Austin, cuenta con más de 500 000 palabras procedentes de entrevistas a 97 hablantes bilingües que viven en Texas.

Corpus del español en Texas (CET)

https://corpus.spanishintexas.org/es

- **Responsable:**
  - Barbara E. Bullock y Almeida Jacqueline Toribio, University of Texas at Austin
- **Variedad dialectal:**
  - Español en Texas
- **Género discursivo:**
  - Entrevistas
- **Objeto de estudio:**
  - Desarrollar un corpus de muestras lingüísticas en español o bilingües español-inglés entre hablantes con diversos perfiles personales y provenientes de diferentes regiones en Texas.
- **Consulta:**
  - Sí, previo registro.
  - No cuenta con motor de búsqueda.
- **Materiales disponibles:**
  - Se pueden descargar archivos de vídeo, archivos de audio, las transcripciones completas y las anotaciones de categoría gramatical.

3.4. Corpus Oral Juvenil del Español de Mallorca (COJEM)

La recolección de este corpus surge de los intereses particulares de investigación de Beatriz Méndez Guerrero. Está compuesto por 20 horas de conversaciones coloquiales, mantenidas entre 10 hablantes jóvenes universitarios mallorquines (5 mujeres y 5 hombres). Concretamente, el corpus recoge 7 conversaciones espontáneas de aproximadamente tres horas de duración, recogidas en lugares frecuentados por los informantes (cafeterías, domicilios particulares, vehículos, playas…) con las técnicas de grabación secreta y observación participante (Méndez Guerrero, 2015, en línea).

Corpus Oral Juvenil del Español de Mallorca (COJEM)

Acceso al PDF

- **Responsable:**
  - Beatriz Méndez Guerrero, Universidad Complutense de Madrid
- **Variedad dialectal:**
  - Mallorca (España)
- **Género discursivo:**
  - Conversación coloquial espontánea grabada secretamente.
- **Objeto de estudio:**
  - Pretende mostrar la variedad del español hablado en Mallorca y servir para futuros estudios lingüísticos interesados en cuestiones sociolingüísticas, pragmático-discursivas y dialectales.
- **Consulta:**

  - Sí, sin motor de búsqueda.
- **Materiales disponibles:**
  - Solo las transcripciones en formato digital, presentadas en un archivo, sin motor de búsqueda ni acceso a los audios.

3.5. Corpus Oral de Lenguaje Adolescente (COLA)

El Corpus COLA se centra en el habla juvenil. Recopila conversaciones espontáneas e informales recogidas en Madrid (COLAM), Santiago de Chile (COLAS) y Buenos Aires (COLABA). Contiene un total de 500 000 palabras entre los tres corpus mencionados. La recogida de este corpus se produce entre 2002 y 2004, si bien hasta 2008 no fue posible su acceso a través del motor de búsqueda.

ℹ Corpus Oral de Lenguaje Adolescente (COLA)

https://blogg.hiof.no/colam-esp/el-corpus-cola/

- **Responsable:**
  - Annette Myre Jørgensen, Universidad de Bergen
- **Variedad dialectal:**
  - Multidialectal (Madrid, Santiago de Chile y Buenos Aires)
- **Género discursivo:**
  - Conversación espontánea no secreta
- **Objeto de estudio:**
  - Análisis del lenguaje juvenil
- **Consulta:**
  - Sí, previo registro
  - Cuenta con motor de búsqueda.
- **Materiales disponibles:**
  - Recuperación de la información a través de concordancias con acceso al fragmento de audio correspondiente. Permite la descarga de los resultados de la búsqueda. También se puede acceder a la transcripción completa, con acceso a la grabación de manera segmentada.

3.6. Corpus for the study of Language and Gender in Mexico and Spain (CoLaGe)

El corpus CoLaGe, aún inédito, recopila datos obtenidos en Valencia, España (2021-2022), y Guadalajara, México (2022-2023), dentro del proyecto de investigación *Género, sociedad y uso del lenguaje: evidencias de México y España*, financiado por la Fundación Kone. Está formado por dos subcorpus generales: CoLaGe-V (Valencia), CoLaGe-G (Guadalajara) y un subcorpus exploratorio, CoLaGe-GD (Guadalajara Diversity) que recoge material de informantes pertenecientes a diferentes minorías sexuales y de género. En total, el corpus cuenta con 127 informantes, más de 100 horas de grabación y aproximadamente 1 010 000 palabras.

Corpus for the study of Language and Gender in Mexico and Spain (CoLaGe)

https://clarino.uib.no/comedi/editor/lb-2024030603
https://www.kielipankki.fi/corpora/

- **Responsable:**
  Gloria Uclés Ramada, Peka Posio, Sven Kachel, Grecia González-Guzmán, Andrea Carcelén-Guerrero, University of Helsinki.
- **Variedad dialectal:**
  - Español de Valencia (España) y Guadalajara (México)
- **Género discursivo:**
  - Entrevista sociolingüística, juego de rol y tarea de descripción de imágenes con fines fonéticos.
- **Objeto de estudio:**
  - Estudiar las interconexiones entre el género del hablante, los roles y las expectativas de género en la sociedad y la variación en el lenguaje hablado, combinando metodologías sociolingüísticas y de psicología social.
- **Consulta:**
  - En preparación.
  - Se podrán descargar los materiales completos desde el repositorio The Language Bank of Finland, sin motor de búsqueda.
- **Materiales disponibles:**
  - En preparación.
  - Estarán disponibles para descarga los audios anonimizados, la transcripción alineada en ELAN, la transcripción en formato CSV y el archivo Textgrid de la tarea fonética para su análisis en PRAAT.

3.7. Corpus Oral de la Lengua Hablada en Honduras (COLEH)

El corpus COLEH, aún en construcción, está dirigido por Enrique Pato, en coordinación con otros investigadores de distintas universidades. Según la información actualizada en su página a fecha de septiembre de 2024, por el momento cuenta con 166 informantes (98 mujeres y 68 hombres, de edades comprendidas entre los 18 y los 85 años, de diferentes grados de instrucción (sin estudios, primaria, secundaria y universitaria). En total hay recogidas por el momento 174 horas de grabación.

Corpus Oral de la Lengua Hablada en Honduras (COLEH)

- https://portfolio.umontreal.ca/view/view.php?id=269744
- **Responsable:** Enrique Pato, Université de Montréal
- **Variedad dialectal:** Español de Honduras
- **Género discursivo:** Entrevista
- **Objeto de estudio:**
  - Obtener datos lingüísticos de las principales ciudades (municipios) del país, así como de algunos enclaves rurales (aldeas y caseríos) en todos los departamentos. El proyecto también se interesa por el español de los hondureños

en la diáspora. En concreto por los migrantes en España, Canadá, Estados Unidos, México e Italia.
- **Consulta:** En preparación.
- **Materiales disponibles:** En preparación.

3.8. Corpus oral de la lengua española en Montreal (COLEM)

Dirigido también por Enrique Pato, este corpus, según la información proporcionada en línea, cuenta con hablantes de habla español residentes en Montreal procedentes de 20 países distintos. En total, cuenta con la participación de 65 hombres y 88 mujeres, de rangos de edad comprendidos entre los 19-34 años, 35-54 y 55-81, residentes en esta ciudad con un mínimo de 4 años. Los hablantes se encuentran, así mismo, estratificados según nivel de estudios y el tipo de migración (política, económica o sociocultural).

ℹ Corpus Oral de la Lengua Española en Montreal (COLEM)

- https://esp-montreal.jimdo.com/corpus/
- **Responsable:** Enrique Pato, Université de Montréal
- **Variedad dialectal:** Español en Montreal
- **Género discursivo:** Entrevista
- **Objeto de estudio:**
  - Ilustrar numerosos rasgos gramaticales y léxicos de las diferentes normas del español actual. Así mismo, permiten documentar fenómenos de contacto lingüístico, fruto de la convivencia del español con el francés y el inglés en la Región metropolitana de Montreal (RMM).
- **Consulta:** Se prevé que esté disponible próximamente. Ofrece un acceso en pruebas al motor de búsqueda.
- **riales disponibles:** Las transcripciones completas pueden solicitarse en formato PDF.

3.9. Corpus Oral Didáctico Anotado Lingüísticamente (C-Or-DiAL)

C-Or-DiAL es un corpus de lengua oral espontánea que contiene 118 756 palabras procedentes de la transcripción de unas diez horas de grabaciones. En ellas se recogen muestras de español en situaciones cotidianas. Se concibe como recurso lingüístico utilizable en la investigación para el análisis general de la lengua oral, y específicamente en el ámbito de la didáctica de la lengua. (Nicolás, 2012).

ℹ Corpus Oral Didáctico Anotado Lingüísticamente (C-Or-DiAL)

- http://lablita.it/app/cordial/corpus.php
- **Responsable:** Carlota Nicolás Martínez, Università degli Studi di Firenze
- **Variedad dialectal:** Peninsular (Madrid, España)
- **Género discursivo:**

  - Conversación espontánea
  - Conversación semidirigida
  - Formal
- **Objeto de estudio:**
  - Investigación para el análisis general de la lengua oral, y específicamente en el ámbito de la Didáctica de la lengua.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Audio y transcripción consultable en línea y descargable. Además, cuenta con un motor de búsqueda para localizar funciones comunicativas concretas. No permite filtrar por criterios sociolingüísticos.

3.10. Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)

Este corpus conforma una base de datos textual (corpus de lengua hablada) que incluye la transliteración de textos grabados en cintas de audio del registro oral, con un total de un millón de palabras transliteradas aproximadamente en soporte informático, según la información obtenida de su portal electrónico.

ℹ Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)

- http://www.lllf.uam.es/ESP/Corlec.html
- **Responsable:** Francisco Marcos Marín, Universidad Autónoma de Madrid
- **Variedad dialectal:** Peninsular
- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Proporcionar material para el estudio de la lengua en general en sus diversas variantes, géneros y canales.
- **Consulta:** Sí, sin motor de búsqueda.
- **Materiales disponibles:** Descarga de las transcripciones sin acceso al audio desde su página electrónica.

3.11. Corpus del Español del siglo XXI (CORPES XXI)

Por iniciativa académica, continuando los primeros trabajos de corpus iniciados con CREA, surge CORPES XXI, esta vez coordinado por la Real Academia y la Asociación de Academias de la Lengua Española. Del total de los materiales recogidos, el género oral representa un 10 %, frente al 90 % que supone el escrito.

ℹ Corpus del Español del siglo XXI (CORPES XXI)

- https://www.rae.es/corpes/
- **Responsable:** RAE y ASALE

- **Variedad dialectal:** Panhispánico
- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Obtener las características globales que presenta una lengua en un momento determinado de su historia.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Motor de búsqueda. Recuperación de la información por medio de concordancias, con acceso al fragmento de audio correspondiente (no en todos los documentos). Permite descargar los resultados de la búsqueda, pero no el material completo.

3.12. Corpus del Español en los Estados Unidos (CORPEEU)

Bajo la dirección de Francisco Moreno Fernández, se inician los trabajos para la construcción del CORPEEU, apoyados por el Observatorio de la lengua española y las culturas hispánicas del Instituto Cervantes en la Universidad de Harvard y con la colaboración de la Academia Norteamericana de la Lengua Española (ANLE). Pretende registrar la lengua española hablada y escrita que está documentado en Estados Unidos desde 1960. Estas muestras se clasifican según el origen geográfico y social de los hablantes, la fecha de producción de las muestras, así como según los estilos, géneros y contextos de la comunidad hispanohablante en Estados Unidos, tal y como se referencia en su web.

Corpus del Español en los Estados Unidos (CORPEEU)

- https://corpeeu.org/
- **Responsable:** Francisco Moreno Fernández, Instituto Cervantes y Universidad de Harvard
- **Variedad dialectal:** Multidialectal (Español de los Estados Unidos)
- **Género discursivo:** Miscelánea (incluye entrevistas de PRESEEA Nueva York)
- **Objeto de estudio:**
  - Documentar el español en los Estados Unidos.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Recuperación de la información por medio de concordancias. Permite descargar los resultados de la búsqueda, pero no el material completo. No hay acceso al audio.

3.13. Corpus del Habla de Almería

Este corpus lo desarrolla el grupo de investigación en Análisis del Discurso Oral en Español ILSE, de la Universidad de Almería, dirigido por Luis Cortés. Está compuesto por 108 entrevistas a hablantes seleccionados por criterios de representatividad sociolingüística (grupos etarios de 18 a 35 años, 36 a 55 años y más de 55; nivel sociocultural alto, medio y bajo y sexo).

Corpus del Habla de Almería

- https://www2.ual.es/ilse/corpus/
- **Responsable:** Grupo ILSE, Universidad de Almería
- **Variedad dialectal:** Peninsular (Almería, España)
- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Análisis del discurso oral.
- **Consulta:** En línea (en actualización).
- **Materiales disponibles:** Actualmente se observa que la página web está siendo actualizada, si bien, en los inicios de esta investigación el material sí estaba disponible.

3.14. Corpus Oral y Sonoro del Español Rural (COSER)

Dirigido por Inés Fernández Ordóñez, este corpus está formado por grabaciones de la lengua hablada en zonas rurales de la península ibérica que se obtuvieron con el propósito de ofrecer una muestra representativa de la variedad dialectal, pero también permiten conocer los modos de vida en el campo en la época previa a la mecanización agraria y a la despoblación rural (COSER, en línea). La duración media de las grabaciones es de una hora y cuatro minutos por enclave y, como se indica en su metodología de trabajo, aunque se han registrado unos 2910 informantes, la inmensa mayoría de las veces solo ha sido encuestado uno informante por enclave.

Corpus Oral y Sonoro del Español Rural (COSER)

- http://www.corpusrural.es/
- **Responsable:** Inés Fernández Ordóñez, Universidad Autónoma de Madrid
- **Variedad dialectal:** Multidialectal (Español rural de España)
- **Género discursivo:** Entrevistas
- **Objeto de estudio:**
  - Dialectológico.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Cuenta con motor de búsqueda a través del cual puede accederse no solo a la forma buscada en su contexto inmediato, sino también al audio y a la transcripción completa. Puede descargarse la transcripción, no así el audio.

3.15. Corpus de Referencia del Español Actual (CREA)

Este corpus de referencia fue el primero que surge por iniciativa de la Real Academia de la Lengua Española que contiene material oral, aunque solamente un 10 % del total se corresponde a este medio, frente al 90 % de materiales escritos. Además, el porcentaje de representación de las distintas variedades del español era de 50 % para España y 50 % para Hispanoamérica, cifras que se encuentran lejos de ser equilibradas con respecto al número de

hablantes que se engloban en cada una de estas zonas.

i Corpus de Referencia del Español Actual (CREA)

- https://www.rae.es/banco-de-datos/crea
- **Responsable:** Real Academia Española de la Lengua
- **Variedad dialectal:** Panhispánico
- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Estudio global de la lengua en su historia reciente.
- **Consulta:** Sí (corpus anotado y no anotado).
- **Materiales disponibles:**
  - Cuenta con motor de búsqueda que recupera la información a través de concordancias, pero no se puede acceder al audio. La versión anotada, por el momento, no incluye los documentos orales.

3.16. El español hablado en Bogotá

Dirigido por José Joaquín Montes Giraldo y coordinado por Jennie Figueroa Lorza, del Instituto Caro y Cuervo, este corpus se ha recogido en tres fases desde 2013 a 2019 en las que se han realizados las tareas de digitalización de las grabaciones, su sistematización y almacenamiento, así como su revisión. Esta institución ha desarrollado diversos trabajos sobre el español y las lenguas de Colombia a través de proyectos como el *Atlas Lingüístico Etnográfico de Colombia* (ALEC), el *Habla Culta de Bogotá* (HCB) y el *Español Hablado en Bogotá* (EHB), entre otros (Bejarano *et al.*, 2018, p. 5).

i El español hablado en Bogotá

- https://clicc.caroycuervo.gov.co/corpus/EHB
- **Responsable:** Instituto Caro y Cuervo
- **Variedad dialectal:** Bogotá (Colombia)
- **Género discursivo:** Encuestas semilibres
- **Objeto de estudio:**
  - Dialectología y sociolingüística.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Cuenta con motor de búsqueda que da acceso a los resultados a través de concordancias, si bien puede consultarse la transcripción y el audio completos, aunque no permite la descarga.

3.17. ESLORA

Este corpus ha sido elaborado por el Grupo de Gramática del Español de la Universidad de Santiago de Compostela a través de diferentes proyectos. En su versión actual contiene 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes de Galicia grabadas

entre los años 2007 y 2015. Los registros sonoros se encuentran transcritos ortográficamente con alineación texto-voz para facilitar el acceso inmediato al audio desde la transcripción. En el proceso de anotación del corpus se han desarrollado recursos para la lematización y el etiquetado morfosintáctico de los textos que permiten realizar diversos tipos de búsquedas (ESLORA, en línea).

ESLORA

- https://eslora.usc.es/
- **Responsable:** Grupo de Gramática del Español, Universidad de Santiago de Compostela
- **Variedad dialectal:** Peninsular (Galicia)
- **Género discursivo:**
  - Entrevistas (PRESEEA)
  - Conversaciones
- **Objeto de estudio:**
  - Estudio del español en una región bilingüe (gallego-español).
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Cuenta con motor de búsqueda que da acceso a las concordancias. Se puede descargar el corpus en formato textual, pero para acceder al corpus etiquetado, a los audios y a la información sociolingüística de los hablantes se ha de pedir autorización al equipo responsable.

3.18. Macrosintaxis del Español Actual (MESA)

El grupo de investigación Argumentación y Persuasión en la Lingüística, dirigido por Catalina Fuentes, es el responsable de la recolección de este corpus a través de diversos proyectos desde 2013. Está centrado en el análisis y el estudio de la unidad básica de la sintaxis del discurso, el enunciado, por medio de material público disponible en Internet (Corpus MEsA 2.0, 2021, en línea).

Macrosintaxis del Español Actual (MESA)

- http://www.grupoapl.es/materiales-corpus/corpus-mesa
- **Responsable:** Catalina Fuentes Rodríguez, Universidad de Sevilla
- **Variedad dialectal:** Multidialectal (España)
- **Género discursivo:** Miscelánea
- **Objeto de estudio:**
  - Sintaxis del discurso.
- **Consulta:** Sí (sin motor de búsqueda).
- **Materiales disponibles:** Pueden descargarse las transcripciones en formato PDF, sin acceso al audio.

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

La decisión de iniciar este proyecto se toma en abril de 1993, durante la celebración del X Congreso Internacional de la Asociación de Lingüística y Filología de la América Latina (ALFAL) y en 1996, durante el XI Congreso de la ALFAL celebrado en Las Palmas de Gran Canaria, se presentó el primer borrador de metodología para el desarrollo del proyecto (Moreno Fernández, 2021a, p. 5). Actualmente, agrupa a más de 40 equipos de investigación sociolingüística que trabajan con una metodología común para reunir un banco de materiales coherente que posibilite su aplicación con fines educativos y tecnológicos (PRESEEA, en línea).

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

- https://preseea.uah.es/corpus-preseea
- **Responsable:** Francisco Moreno Fernández, Universidad de Alcalá de Henares
- **Variedad dialectal:** Panhispánico
- **Género discursivo:** Entrevistas semidirigidas
- **Objeto de estudio:**
  - Sociolingüística comparada.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Resultados obtenidos por concordancias en pantalla, permite la descarga de la transcripción completa del archivo en formato TXT y el audio en formato MP3. Es posible exportar los resultados de la búsqueda.

3.20. Valencia Español Coloquial (Val.Es.Co.) versión 3.0

El grupo de investigación Val.Es.Co. (Valencia, Español Coloquial) surge en el seno del Departamento de Filología Española de la Universidad de Valencia en 1990, con el objetivo principal de estudiar del español coloquial. Esta tarea se ha sustentado en dos pilares básicos: la creación y desarrollo de un corpus de conversaciones coloquiales y la descripción y explicación de los principios rectores de una conversación desde un acercamiento pragmático y de corte funcional. El corpus Val.Es.Co 3.0. (en línea) presenta una muestra de español coloquial para la cual se han transcrito sesenta y seis conversaciones, además, un subcorpus de quince de estas conversaciones ha sido segmentado en diferentes unidades de análisis: discursos, diálogos, turnos, intervenciones, actos y subactos siguiendo la metodología establecida por Briz y Grupo Val.Es.Co., (2002)[6].

ℹ Valencia Español Coloquial (Val.Es.Co.) versión 3.0

- https://www.valesco.es/
- **Responsable:** Salvador Pons Bordería, Universitat de València

[6]Las transcripciones completas procedentes de versiones anteriores (2002) solamente se pueden consultar en papel.

- **Variedad dialectal:** Peninsular (Valencia, España)
- **Género discursivo:** Conversación coloquial espontánea
- **Objeto de estudio:**
  - Análisis pragmático del discurso coloquial.
- **Consulta:** Sí, con motor de búsqueda.
- **Materiales disponibles:**
  - Búsqueda por concordancias sin acceso al audio. Permite la descarga de los resultados, pero no de los materiales.

3.21. Otros

Cabe señalar, en último lugar, otros trabajos dignos de mención que, si bien trabajan con material oral, no cumplen con todos los requisitos clasificatorios establecidos en el apartado 2.

En este sentido, se han quedado fuera aquellas plataformas que recogen atlas lingüísticos, entre ellos, el *Atlas interactivo de la entonación española* (Prieto y Roseano, 2009-2013) que, aunque recoge material oral, su construcción responde a otros criterios ajenos a la construcción de corpus. Sucede también con el *Archivo de textos hispánicos de la Universidad de Santiago de Compostela* (ARTHUS), aunque nace como un trabajo de corpus, sus materiales se han incorporado a la base de datos para el estudio sintáctico y verbal ADESSE, dirigido por García Miguel. El *Corpus Oral del Español de México* (COEM) (Martín Butragueño, Mendoza y Orozco), solo permite recuperar enunciados aseverativos e interrogativos, sin poder recuperar otras informaciones.

El caso de C-ORAL-ROM*, Corpus Oral de Referencia del Español en Contacto* solo permite el acceso a una muestra de los materiales, el material completo debe comprarse. El proyecto *Voices of Hispanics World* (Terrell A. Morgan) desarrollado desde la Ohio State University, se constituye como un recurso audiovisual de muestras dialectales del español que conforma más bien un catálogo y no un corpus, aunque las muestras dialectales que recoge están acompañadas de su correspondiente transcripción.

El grupo de investigación Vernáculo Urbano Malagueño (Universidad de Málaga) recoge proyectos reseñables como el *Corpus Oral Telemático* (COLEMA) que contiene material oral obtenido por medio de audios de WhatsApp, pero solo hay disponible al público una muestra; igual sucede con el *Corpus Oral de Inmigrantes de Buenos Aires residentes en Málaga* (CORINBAS), recogido por María Clara von Essen como parte de su tesis doctoral.

4. Consideraciones finales

A lo largo de este trabajo se ha dado cuenta de más de una veintena de corpus orales de español disponibles en línea y de acceso abierto para cualquier persona interesada en el estudio lingüístico de la oralidad. Si incluyéramos también aquellos que han sido publicados en papel o no están disponibles al público, obtendríamos una cifra nada desconsiderada teniendo en cuenta que, en el contexto hispánico, la creación de corpus suele ser una tarea altruista, con ciertas dificultades para acceder a fondos subvencionados. Trabajos como el de Briz (2012)

ponen de manifiesto los déficits existentes en el panorama de corpus orales, como son la falta de muestras orales, en particular de materiales conversacionales, las dificultades de acceso a materiales existentes que no se encuentran disponibles públicamente o la heterogeneidad en los procesos de transcripción y codificación de las muestras. Recalde y Vázquez (2009) y Carcelén (2024), entre otros, han señalado, además, los problemas metodológicos que conlleva realizar corpus escritos, motivo por el cual existen desequilibrios en la representación de géneros discursivos a favor de aquellos procedentes del medio escrito.

No obstante, es loable el esfuerzo por facilitar a la comunidad científica el acceso al material oral recopilado en estos corpus, favoreciendo la reusabilidad y el intercambio de los datos para el análisis lingüístico desde cualquier disciplina. Dada la inversión económica y en tiempo de trabajo que implica la construcción de corpus orales, compartir con la comunidad científica el trabajo realizado debería ser una práctica habitual.

5. Referencias bibliográficas

Briz Gómez, Antonio (2012): «Los déficits de los corpus orales del español (y de algunos análisis)», en Tomás E. Jiménez *et alii,* coord., *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*, Universidade de Santiago de Compostela, 115-137.

Briz Gómez, Antonio y Marta Albelda Marco (2009): «Estado actual de los corpus de lengua española hablada y escrita: I+D», en *Anuario del Instituto Cervantes, El español en el mundo*, Instituto Cervantes, 165-226.

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

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

Carcelén Guerrero, Andrea (2024): *Bases teórico-metodológicas para la construcción de un corpus multidialectal de conversación coloquial: el corpus Ameresco.* Tesis doctoral. [En línea, https://hdl.handle.net/10550/92265].

Díaz Sánchez, Alicia (2018): Indexador de corpus de aprendices de español. [En línea, http://repositorios.fdi.ucm.es/corpus_aprendices_espa%c3%b1ol/view/paginas/view_paginas.php?id=1)].

Enghels, Renata *et alii* (2015): «Panorama de los corpus y textos del español peninsular contemporáneo», en Maria Iliescu y Eugeen Roegiest, ed., *Manuel des anthologies, corpus et textes romans*, Berlin, München, Boston, De Gruyter, 147-170. https://doi.org/10.1515/9783110333138-012

Llisterri, Joaquim (2021): «Corpus para investigar sobre el componente fónico en español como LE/L2». En Mar Cruz y J. Muñoz, eds., *e-Research y español LE/L2.: investigar en la era digital*, Routledge, 164-196.

McEnery, Tony *et alii* (2006): *Corpus-Based Language Studies*, London, Routledge.

Moreno Fernández, Francisco (2005): «Corpora of Spoken Spanish Language. The Representativeness Issue», en Yuji Kawaguchi *et alii*, eds., *Linguistic Informatics, State of the Art and the Future*, Amsterdam/Philadelphia, John Benjamins, 120-144.

Lanza, Danny Fernando (2023): «Los corpus orales del español centroamericano: compilación y mirada valorativa», *Normas*, 13, 83-111. https://doi.org/10.7203/Normas.v13i1.27658

Parodi, Giovanni y Gina Burdiles (2019): «Corpus y bases de datos (Corpora and databases)», en Javier Muñoz *et alii*, coord*., The Routledge Handbook of Spanish Language Teaching: metodologías, contextos y recursos para la enseñanza del español*, Londres/Nueva York, Routledge, 596-612.

Recalde, Montserrat. y, Victoria Vázquez (2009): «Problemas metodológicos en la formación de corpus orales», en Pascual Cantos y Aquilino Sánchez, eds., *A survey of corpus-based research*, Murcia, Asociación Española de Lingüística del Corpus, 51-64.

Rojo, Guillermo (2016): «Corpus textuales del español», en Javier Gutiérrez Rexach, coord., *Enciclopedia de Lingüística Hispánica,* vol. 2, Londres, Routledge, 285-296.

Solís García, Inmaculada (2018): «Corpus españoles dialógicos para el análisis de la conversación», *CHIMERA: Romance Corpora and Linguistic Studies*, 5 (1). 117-129. https://doi.org/10.15366/chimera2018.5.1.010

Torruella, Joan y Johannes Kabatek (2018): *Portal de Corpus Históricos Iberorrománicos* (CORHIBER). [En línea http://www.corhiber.org/].

Torruella, Joan y Joaquim Llisterri (1999): «Diseño de corpus textuales y orales», en José Manuel Blecua *et alii,* eds., *Filología e Informática. Nuevas tecnologías en los estudios filológicos*, Universidad Autónoma de Barcelona, Milenio, 45-77.