normas

Recibido: 15 octubre 2019 | Aceptado: 20 noviembre 2019 | Publicado: 1 diciembre 2019
Cita: Carcelén, A. y Uclés, G. (2019). Diseño y construcción de un corpus oral multidialectal. El corpus Ameresco. *Normas*, 9, 17-36. doi: https://doi.org/10.7203/Normas.v9i1.16007

DISEÑO Y CONSTRUCCIÓN DE UN CORPUS ORAL MULTIDIALECTAL. EL CORPUS AMERESCO

DESIGNING AND DEVELOPING A MULTIDIALECTAL ORAL CORPUS. THE AMERESCO CORPUS

Andrea Carcelén Guerrero

Gloria Uclés Ramada

Universitat de València

Resumen

En este artículo se describe el protocolo que se ha seguido para la construcción del corpus Ameresco (América Español Coloquial). La recopilación de un corpus multidialectal presenta una serie de retos. Por una parte, la gestión de un gran número de equipos externos requiere un proyecto que metodológicamente sea sólido. Por otra parte, la metodología debe ser coherente con los objetivos del proyecto y con los parámetros esenciales en el diseño de corpus como es establecer las características de las grabaciones, el sistema de transcripción y etiquetado y aspectos relacionados con la anonimización de datos sensibles. Todas estas cuestiones deben provenir de una decisión razonada que garantice que el corpus cumpla con unos estándares de calidad aceptables por la comunidad científica.

PALABRAS CLAVE: corpus de español, oralidad, corpus lingüísticos, conversación coloquial, sistema de transcripción

Abstract

This paper describes the protocol used to build the Ameresco corpus (America Colloquial Spanish). Collecting a corpus containing more than one dialect poses a series of challenges. On the one hand, managing a large number of external teams requires that the methodology used is sound. On the other hand, the methodology should be in line with the goals that the project aims to reach and with essential corpus design features such as issues when recording, the transcription and labelling system and the anonymisation of sensitive data. All these aspects should be thoughtfully chosen so that the quality standards set by the scientific community are reached.

KEY WORDS: Spanish corpora, orality, linguistic corpora, colloquial conversation, transcription system

1 INTRODUCCIÓN[1]

Este artículo describe el protocolo de trabajo para la construcción del corpus Ameresco (América Español Coloquial). El proyecto Ameresco (Briz, 2016) tiene como objetivo estudiar la variedad coloquial de todos los dialectos del español, tanto de América como de España; en ese sentido, el corpus Ameresco (Albelda y Estellés, en línea) es una de las iniciativas para recopilar el material lingüístico necesario para llevar a cabo el proyecto. El corpus cuenta actualmente con la participación de equipos de 14 ciudades de habla hispana y espera poder contar próximamente con la participación de más equipos locales.

En la clasificación de corpus se encuentra una división fundamental que separa los orales frente a los escritos. En general, hay una mayor presencia de corpus escritos, dado que son más sencillos y rápidos de recopilar (Briz, 2012). Por ejemplo, las técnicas para recoger corpus escritos presentan menos dificultades frente a los orales, ya que basta con reunir archivos en formato electrónico o usar la web como corpus para obtener millones de palabras al instante.

El principal corpus panhispánico oral es el surgido del proyecto PRESEEA (Moreno, 2005; Moreno, 2016; PRESEEA, 2014-) que reúne entrevistas semidirigidas de las principales ciudades de América y España (43 ciudades en 14 países), y que lleva recopilando materiales desde los años 90. Ameresco comparte no solo el carácter panhispánico y oral de PRESEEA, sino también su carácter urbano, y pretende añadir una nueva perspectiva variacionista, de acuerdo con los intereses del Grupo Val.Es.Co., al incluir conversaciones coloquiales como objeto de estudio. Con el corpus Ameresco se pretende, así, completar el panorama panhispánico de corpus orales con un conjunto de conversaciones espontáneas, un género discursivo escasamente representado, como han señalado los trabajos de Barcala et alii (2018), Briz (2012), Briz y Carcelén (2019), Albelda y Briz (2009) o Recalde y Vázquez (2009), entre otros.

En este artículo se describe el proceso de construcción del corpus Ameresco, y se organiza de la siguiente manera: en el apartado 2 se exponen las características generales del corpus; le siguen los criterios metodológicos empleados para su diseño, construcción y transcripción (apartado 3). En el apartado 4 se da cuenta de las opciones de recuperación de información a través del motor de búsqueda. En el apartado 5, consideraciones finales, se exponen de los principales retos que la construcción de este corpus multidialectal ha debido enfrentar.

2 CARACTERÍSTICAS DEL CORPUS AMERESCO

2.1 Origen de la iniciativa Ameresco

Como se mencionaba en la introducción, el corpus que se describe en este artículo se enmarca en la iniciativa Ameresco, que pretende estudiar y caracterizar el español coloquial atendiendo a las diferentes variedades dialectales. En el seno de esta iniciativa

[1] Este trabajo ha sido desarrollado gracias a la financiación recibida en el proyecto "Es.Vag.Atenuación. La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de España y América" (MINECO, FFI2016-75249-P). Agradecemos también la ayuda financiada por la Red de Excelencia (FFI2017-90738-REDT).

se han desarrollado proyectos como ES.VAG.ATENUACIÓN "La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de España y América" (MINECO FFI2016-75249-P) y el proyecto anterior ES.VAR.ATENUACIÓN ("La atenuación pragmática en el español hablado: su variación diafásica y diatópica", MINECO FFI2013-40905-P). Dentro de los objetivos de estos proyectos, se encuentra el estudio de fenómenos pragmáticos como la atenuación y la intensificación en el español europeo y americano, para lo cual es imprescindible la construcción de un corpus que recoja conversaciones coloquiales espontáneas, grabadas de manera secreta, en diferentes puntos del ámbito hispánico.

El corpus Ameresco toma como base los fundamentos metodológicos del corpus Val.Es.Co. (Briz et alii, 1995, 2002). A partir de esta base, se han incluido adaptaciones conducentes a preparar los materiales para su procesamiento informático, como la alineación temporal del archivo de audio con la transcripción o la anotación de fenómenos del habla mediante un etiquetado basado en el lenguaje XML, con variaciones respecto al sistema de transcripción original, como se detalla en el apartado 3.5.

2.2 Grupos participantes

La recopilación de un corpus en el que se incluye material lingüístico recogido tanto en diferentes ciudades de España como de Latinoamérica requiere necesariamente de la colaboración de grupos locales que se encarguen no solo de la recolección de los audios y documentos, sino también de garantizar la correcta transcripción de las grabaciones, algo notablemente complicado de lograr si la tarea la llevan a cabo transcriptores no nativos en la variedad dialectal de la grabación. Por ello, el equipo coordinador de Valencia colabora con una serie de grupos locales, ubicados en los países de recogida y coordinados por investigadores del país. Estos equipos, siguiendo la metodología de recogida y transcripción del corpus Ameresco (ver sección 3), se encargan de las tareas mencionadas y envían los resultados al equipo coordinador, que se ocupa de su posterior procesamiento informático.

A continuación (Tabla 1), se detallan las ciudades en las que se ya se han elaborado o se están elaborando los subcorpus. En total se cuenta con la participación de 14 ciudades situadas en 7 países. A esta lista está prevista la próxima incorporación de nuevos equipos en Honduras (Tegucigalpa), Venezuela (Caracas), Argentina (Buenos Aires) y Chile (Valparaíso). Junto a cada ciudad se proporciona el investigador/a responsable de la coordinación del equipo, cada uno de los cuales está normalmente integrado por un grupo de entre 2 y 10 personas.[2]

[2] La lista completa de miembros se puede consultar en la página web del proyecto http://esvaratenuacion.es/miembros-del-proyecto/.

Ciudad	Coordinación
Equipo central: Valencia	Marta Albelda Marco y Maria Estellés Arguedas
Tucumán (Argentina)	Silvina Douglas de Sirgo
Iquique (Chile)	Corpus cedido por Renata Enghels y Kristina Helinks
Santiago de Chile (Chile)	Silvana Guerrero González
Barranquilla (Colombia)	Yolanda Rodríguez Cadena
Medellín (Colombia)	Ji Son Jang y Ana I. García Tesoro
La Habana (Cuba)	Ana María González Mafud y Yohana Martínez
Santiago de Cuba (Cuba)	Celia Pérez Márquez
Las Palmas de Gran Canaria (España)	Marta Samper Hernández
Ciudad de México (México)	Corpus cedido por Katharina Pater Nuevos materiales Ricardo Maldonado
Monterrey (México)	María Eugenia Flores Treviño
Querétaro (México)	Juliana de la Mora
Ciudad de Panamá (Panamá)	Fulvia Morales de Castillo
Tegucigalpa (Honduras)	Rosario Buezo Velásquez

Tabla 1. Equipos de trabajo Corpus Ameresco

Actualmente los materiales de Ameresco se pueden consultar en línea por medio de dos métodos:

1) A través del motor de búsqueda (http://corpusameresco.com/) Permite realizar consultas de formas concretas (ver apartado 4). A esta plataforma se incorporan las conversaciones que cumplen con todas las características de calidad acústica, espontaneidad y prototipicidad conversacional (ver apartado 3).
2) En forma de repositorio (http://esvaratenuacion.es/corpus-discursivo-propio/). Permite la descarga del archivo alineado completo de cada conversación, así como del audio, la transcripción por separado y la ficha técnica. En este catálogo se pueden encontrar grabaciones que no cumplen con todos los requisitos establecidos en la metodología (ver apartado 3).

En el futuro está previsto que los archivos que forman parte del repositorio tengan también su propia aplicación de consulta.

3 DISEÑO DEL CORPUS[3]

En esta sección se detallan los pasos que se han seguido para la construcción del corpus Ameresco (Briz et alii, 2019). En primer lugar, se exponen los parámetros seguidos para seleccionar las muestras de habla que constituyen el corpus (sección 3.1.). En segundo lugar, se presenta el proceso de grabación secreta de conversaciones

[3] El protocolo de trabajo (Briz et alii, 2019) puede consultarse en http://esvaratenuacion.es/material/

y todos los factores que deben tenerse en cuenta para su realización (sección 3.2.). A continuación, se detalla la estructura de las fichas técnicas que acompañan a las conversaciones (sección 3.3.), el proceso de transcripción y etiquetado (sección 3.4.) y la validación de las etiquetas (sección 3.5.). Por último, se tratan las cuestiones éticas y legales en cuanto a la anonimización y protección de datos (sección 3.6.) y se expone el sistema de identificación de los archivos que conforman el corpus (sección 3.7.).

3.1 Selección de hablantes

El corpus Ameresco mantiene los criterios de representatividad sociolingüística establecidos por el proyecto PRESEEA, en combinación con la metodología del Grupo Val.Es.Co. De acuerdo con esto, se han seguido parámetros referidos a sexo, edad y nivel sociocultural (Labov, 1972). Se pretende que la muestra tenga un número equilibrado de hablantes siguiendo las variables que se recogen en la Tabla 2.

Variante	Variable
Sexo	Mujer Varón
Grupo etario	18-25 26-55 ≥56
Nivel sociocultural	Nivel bajo: estudios primarios o sin estudios Nivel medio: estudios de secundaria y formación profesional Nivel alto: estudios superiores

Tabla 2. Criterios de selección de la muestra

Una vez establecidos estos criterios, en base a los cuales se va a recoger la muestra, se estipulan las cuotas que hay que obtener para garantizar la homogeneidad de la selección (Tabla 3). Para ello, se ha seguido la ficha desarrollada por el Grupo Val.Es.Co. para la construcción del corpus (Briz et alii, 2002).

Edad	Nivel sociocultural									Total		
	Alto			Medio			Bajo					
		Partic.	Conv.		Partic.	Conv.		Partic.	Conv.		Partic.	Conv.
18-25	V* 4 M** 4	8	3	V 4 M 4	8	3	V 4 M 4	8	3	V 12 M 12	24	9
26-55	V 4 M 4	8	3	V 4 M 4	8	3	V 4 M 4	8	3	V 12 M 12	24	9
≥56	V 4 M 4	8	3	V 4 M 4	8	3	V 4 M 4	8	3	V 12 M 12	24	9
Total	V 12 M 12	24	9	V 12 M 12	24	9	V 12 M 12	24	9	V 36 M 36	72	27

* Varón
** Mujer

Tabla 3. Resumen de la muestra extraída (Grupo Val.Es.Co.)

Según los totales del cuadro, por cada ciudad se deben recoger al menos 27 conversaciones con las que, sumando los participantes necesarios para cubrir las diferentes celdas (los estratos sociolingüísticos), se obtienen un total de 72 hablantes, 8 por cada grupo etario y nivel de estudios (4 mujeres y 4 varones).

Esta muestra puede ampliarse con más conversaciones. Así, el plan de grabaciones no es rígido, puesto que no siempre es posible –ni recomendable– forzar las circunstancias de grabación para rellenar huecos en las celdas de forma exacta; es, por tanto, una tabla de mínimos.

3.2 Pasos previos a la grabación. Requisitos legales

Dado que se recogen grabaciones de forma secreta, para poder cumplir con la normativa legal de protección de datos se ha diseñado un sistema de autorización en tres pasos, que se refleja en un documento de consentimiento legal con tres secciones que deben firmar todos los participantes en la conversación. En un primer apartado, el responsable de la grabación solicita el permiso a los hablantes para grabarlos en algún momento futuro, no especificado. Esta primera firma manifiesta el consentimiento del hablante a ser grabado y recoge la fecha en que lo autorizó. En el segundo apartado, una vez realizada la grabación y habiendo sido informado el hablante de que acaba de ser grabado, se le da la posibilidad de escuchar la conversación y, si consiente en cederla para su análisis lingüístico, deberá firmar y fechar el segundo apartado de la autorización. Con esta firma, por tanto, el hablante manifiesta que ha escuchado la conversación y que está de acuerdo con que se haga pública para fines de investigación, previa anonimización de nombres y lugares. En último lugar, los hablantes deben firmar la sección para el tratamiento de datos personales, de acuerdo con la normativa vigente, y aceptar los términos. En el caso de no obtener dicha autorización o de que esta esté incompleta, el archivo no podrá utilizarse y deberá ser destruido.

3.3 Las grabaciones

Una vez establecida la muestra de hablantes de la que se parte, son los equipos externos los encargados de recoger las grabaciones necesarias para cumplir con los objetivos de representatividad fijados. Las grabaciones se realizan de forma secreta, con petición de autorización previa y posterior (véase sección 3.2.) a los participantes,

por medio de grabadoras o bolígrafos espía –situados estratégicamente para no ser descubiertos– o bien con las aplicaciones de grabación de sonido de teléfonos móviles, si realizan grabaciones con una calidad de sonido adecuada. Estos últimos presentan la ventaja de que son más fáciles de integrar en el contexto que una grabadora.

Para ajustarse a los objetivos del corpus, es decir, la compilación de conversaciones coloquiales espontáneas, el investigador/a responsable de la grabación no debe intervenir en la elicitación de información ni desvirtuar la espontaneidad de la interacción entre hablantes, pues ello podría alterar el grado de prototipicidad de la conversación coloquial (Briz, 2001). En este sentido, se recomienda que la persona responsable de la grabación intervenga lo mínimo imprescindible para que su silencio no resulte sospechoso y, si fuera posible, se sugiere que deje la grabación en marcha y salga o se retire de la escena. En cualquier caso, es esencial que los participantes no sean conscientes de que están siendo grabados.

Además de las mencionadas características genéricas, las grabaciones recopiladas deben reunir una serie de requisitos acústicos que las hagan aptas para el análisis fónico posterior. En cuanto a los requisitos técnicos, las grabaciones deben tener idealmente una duración aproximada de entre 20 y 60 minutos. Sin embargo, se sigue un criterio flexible que prioriza la recolección de una conversación completa ante la necesidad de ajustarse a unos tiempos preestablecidos. Desde el punto de vista acústico, la grabación secreta plantea una serie de problemas externos que pueden afectar a la calidad del sonido. Grabar en exteriores o en sitios públicos como cafeterías o parques puede producir archivos de audio en los que aparecen ruidos de fondo tales como el tráfico o el viento. Por lo tanto, se recomienda que las grabaciones se realicen en momentos o localizaciones que presenten una baja incidencia de ruido externo. De lo contrario, no solo puede generar grabaciones altamente ininteligibles o difíciles de transcribir, sino que este material también quedará inhabilitado para el desarrollo de estudios fónicos.

El número de hablantes constituye otro factor condicionante de la calidad acústica de las grabaciones y, por tanto, de su validez para el análisis fónico y de la inteligibilidad para la transcripción. Las conversaciones con un elevado número de hablantes (mayor a 4 participantes) tienden a tener una mayor cantidad de solapamientos entre dos o incluso más hablantes. El habla simultánea presenta mayores dificultades a la hora de transcribir y, en muchas ocasiones resulta imposible recuperar lo que los hablantes dijeron, por lo que se da un porcentaje más alto de fragmentos ininteligibles. Otra consecuencia de la multiplicidad de interlocutores es la presencia de diálogos laterales, desgajados, en los que solo participa una parte de los hablantes; ello puede producir fragmentos de conversaciones en paralelo, fónicamente solapados y, por tanto, a menudo indescifrables y no aptos para el análisis acústico. Sin embargo, el criterio de la espontaneidad opera también en este caso y cabe destacar que se cuenta con que el número de hablantes no siempre será fijo, debido al menor control del contexto de las conversaciones coloquiales. Así, si en ocasiones aparecen espontáneamente más interlocutores y sus solapamientos son puntuales (no duran toda la conversación) o no invalidan la muestra para el análisis por solaparse poco, las grabaciones pueden considerarse válidas y entrar a formar parte del corpus.

Tras el proceso de grabación, los responsables deben rellenar los datos técnicos de la conversación según el modelo especificado en 3.4. A continuación, se procede a la doble transcripción de las muestras (3.5): una ancha (3.5.1), llevada a cabo por los equipos locales, y una estrecha (3.5.2.), que incluye etiquetado pragmático, realizada en la mayor parte de los casos por el equipo central. Finalmente, de acuerdo con la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales, así como la normativa propia de la Universitat de València, se procede a la anonimización de todas las referencias personales o locales que pudieran llevar al reconocimiento de los hablantes, como detallaremos en el apartado 3.7.

3.4 Ficha técnica

Cada archivo va acompañado de una ficha técnica[4] que rellena la persona responsable de la grabación siguiendo el modelo que se muestra a continuación (Tabla 4). La ficha contiene información general sobre la grabación (fecha, duración, lugar), así como de su contenido temático. La información proporcionada acerca de los hablantes permite extraer los datos necesarios para ajustarse a los criterios de representatividad del corpus previamente expuesto (sección 3.1.). Un aspecto destacable de esta ficha reside en la información sobre las características de la situación comunicativa, especialmente su propósito o tenor funcional predominante. Se distinguen dos: interpersonal –cuando el objetivo es la comunicación *per se*– o transaccional –la conversación persigue un fin concreto– (Briz, 2001).

FICHA TÉCNICA

a) Investigador:
b) Datos identificadores de la grabación:
- Fecha de la grabación:
- Tiempo de la grabación:
- Lugar de grabación:
c) Situación comunicativa:
- Tema:
- Propósito o tenor funcional predominante:
Interpersonal:
Transaccional:
- Tono:
- Modo o canal:
d) Tipo de discurso:
e) Técnica de grabación:
- Conversación libre:
Observador participante/Observador no participante
Grabación secreta/grabación ordinaria
- Conversación semidirigida (grabación ordinaria):
f) Descripción de los participantes:
Número de participantes:
Clave (identificar como A, B, C, D…)
Activos: Pasivos:
- Tipo de relación que los une:

[4] Se toma como modelo la ficha técnica del grupo Val.Es.Co. (Briz et alii, 2002)

- Sexo: varón: mujer: - Edad: 18-25 26-55 ≥56 - Nivel de estudios: Bajos Medios Superiores - Profesiones: - Residencia o domicilio habitual (indicar solo el municipio): - Lengua habitual: monoling. cast.: biling.: g) Grado de prototipicidad coloquial: conversación coloquial prototípica: conversación coloquial periférica:

Tabla 4. Ficha técnica del Grupo Val.Es.Co.

3.5 Transcripción y alineado

Todas las grabaciones están transcritas partiendo de la base de las convenciones del sistema de transcripción del Grupo Val.Es.Co. (Briz et alii, 2002). Para ello, la transcripción clásica llevada a cabo en formato Word fue adaptada en la primera fase del proyecto (Es.Var.Atenuación) a las características del corpus y a los propósitos de investigación, modificaciones que se han recogido en el protocolo de trabajo (Briz et alii, 2019). La necesidad de informatizar el corpus para hacerlo públicamente accesible y consultable ha obligado a transformar estas convenciones en etiquetas, lo que permite procesar los datos con una mayor sistematicidad y eficacia a la hora de recoger y buscar fenómenos propios de la oralidad. Para establecer este sistema de etiquetado, el corpus Ameresco recoge la propuesta de PRESEEA (2008). Así pues, dentro del proceso de transcripción se distinguen dos fases y se aplica un estándar doble de transcripción: por una parte, la transcripción ancha en Word, transcrita en el sistema clásico adaptado y llevada a cabo por los equipos locales (3.5.1); por otra, la revisión, el alineado y el etiquetado por parte del equipo central (3.5.2.).

3.5.1 Transcripción ancha del equipo local

Cada equipo local realiza una transcripción utilizando un procesador de textos, esto es, no alineada con el tiempo. Este primer paso resulta crucial dado que, al trabajar con hablantes de ubicaciones tan diversas, son los equipos locales quienes mejor conocen las variedades dialectales propias, así como el contexto cultural. Por lo tanto, se garantiza una transcripción mucho más precisa que la que se podría realizar desde el equipo central. Como se mencionaba anteriormente, para esta primera transcripción, se siguen las convenciones del Grupo Val.Es.Co. (Briz et alii, 2002), utilizando una selección de signos imprescindibles para una transcripción ancha (Briz et alii, 2019)

(Tabla 5). Por razones de legibilidad, esta transcripción es la recomendable a la hora de utilizar ejemplos extraídos del corpus.

A:	Intervención de hablante identificado como A
?:	Intervención de un hablante no identificado
[ ]	Inicio y final de habla simultánea
-	Reinicios y autointerrupciones sin pausa
/	Pausa
(2")	Silencio (lapso o intervalo de segundos)
((siempre))	Transcripción dudosa
((…))	Interrupción de la grabación o transcripción
(en)tonces	Reconstrucción de una unidad léxica pronunciada incompleta
pa'l	Fenómenos de fonética sintáctica entre palabras
°( )°	Fragmento pronunciado con intensidad baja, próxima al susurro
h	Aspiración de "s" implosiva
PESADO	Pronunciación marcada o enfática
(( ))	Fragmento indescifrable
(RISAS) (TOSES) (GRITOS)	Aparecen antes o después de los enunciados
aa nn	Alargamientos vocálicos
¿?	Alargamientos consonánticos
¡!	Interrogaciones
Cursiva	Exclamaciones
Notas a pie de página	Reproducción e imitación de emisiones en estilo directo Aclaraciones u observaciones sobre algún aspecto de la transcripción

Tabla 5. Signos de transcripción del Grupo Val.Es.Co. para transcripción ancha

A grandes rasgos, este sistema de transcripción va más allá de la mera transliteración ortográfica ya que sí se incorporan fenómenos discursivos propios de la oralidad, esto es, se representan casos de fonética sintáctica, aspiraciones, alargamientos y solapamientos, entre otros. Así, por ejemplo, el uso de signos de puntuación queda relegado exclusivamente a marcar preguntas o sorpresa mediante signos de interrogación y de exclamación respectivamente, para no someter a lo oral a un sistema de codificación e interpretación propio de lo escrito. En cuanto al uso de mayúsculas, solo se emplean aquellas que corresponden a nombres propios y siglas, y las que marcan una pronunciación enfática. Las cifras y símbolos se transcriben con letra.

3.5.2 Transcripción alineada por parte del equipo central

Las transcripciones anchas en formato Word se alinean con los archivos de audio correspondientes utilizando ELAN, un programa para la anotación lingüística que permite asociar pequeños fragmentos transcritos con un código de tiempo. Se distribuye como una herramienta libre y es de código abierto (https://tla.mpi.nl/tools/tla-tools/elan/download/). Entre las ventajas que presenta ELAN se encuentra que permite establecer una línea (*tier*) para cada hablante, lo que facilita la visualización de solapamientos y pausas.

El corpus Ameresco incorpora en esta fase una serie de etiquetas para marcar parte de los fenómenos discursivos que el sistema de transcripción de Val.Es.Co. reflejaba en la transcripción presentada en 3.5.1. El sistema de etiquetado traduce las marcas clásicas a un lenguaje XML, basado –como se indicó más arriba– en el sistema utilizado por PRESEEA (2008), que regula la codificación de textos siguiendo estándares internacionales.

En el sistema de transcripción estrecha se diferencian dos tipos de etiquetas:

- etiquetas simples: compuestas por un solo elemento (<ininteligible/>)
- etiquetas dobles: compuestas por dos elementos, uno de apertura y otro de cierre (<cita></cita>).

Además, algunas etiquetas contienen atributos, esto es, información adicional que no forma parte de la transcripción y que debe aparecer dentro de las comillas (es el caso de <énfasis t=“ ”></énfasis>, <obs t=“ ”/> y <siglas t=“ ”></siglas>, entre otras). Las etiquetas simples se colocan entre espacios, a excepción de la etiqueta <alargamiento/>, que se escribe pegada a la letra que afecta (es decir, también puede aparecer en mitad de palabra).

A: me dijo tu primo que<alargamiento/> bue<alargamiento/>no que no sabía
B: bueno es que él nunca sabe <risas/> mi primo nunca sabe

Las etiquetas de apertura y cierre de las etiquetas dobles aparecen junto al fragmento al que acompañan.

A: ayer vi una película que <énfasis t=“silabeo”>madre mía</énfasis> era buenísima

A continuación, en la Tabla 6 se describen las etiquetas usadas en la transcripción de conversaciones del corpus Ameresco y el símbolo de Val.Es.Co. al que sustituyen. La Tabla 7 recoge las nuevas etiquetas que se han introducido durante el desarrollo del corpus Ameresco.

Cabe también señalar que algunas etiquetas, mediante el atributo, presentan la posibilidad de distinguir entre el caso y tipo. Concretamente, se ha incluido esta diferenciación en los fenómenos discursivos en los que la realización oral y su representación normativa es variable, esto es, en cuestiones relacionadas con la fonética sintáctica (<fsr t=“ ”></fsr>), extranjerismos (<extranjero t=“ ”></extranjero>) y siglas (<siglas t=“ ”></siglas>). Esta doble representación, si bien es posible a través del etiquetado, comporta una decisión metodológica: se debe establecer si es la producción oral la que se transcribe y se indica en el atributo el referente al que alude o, de lo contrario, se transcribe la forma canónica y se introduce en el atributo la forma oralizada. Ante esta disyuntiva, se ha optado por introducir el caso dentro del atributo de la etiqueta y en el cuerpo de la transcripción, entre las etiquetas dobles, el tipo. De esta forma, un caso en el que se dice *voy pa casa* se representaría de la siguiente manera: *voy <fsr t=”pa”>para</fsr> casa*.

En esta manera de representar los datos, se favorece, por tanto, el tipo sobre el caso como la opción preferida. Ello conlleva una serie de ventajas en relación con el procesamiento de datos y su posterior consulta por parte de los usuarios del corpus,

puesto que las diferentes manifestaciones fónicas se pueden recoger bajo una misma forma (obsérvese, por ejemplo, el caso de *pues* que, en el corpus Val.Es.Co. 2.0. podía encontrarse como *pues*, *pos*, *puees* y *pueh*), lo que permite una gestión más eficiente de los datos y, por tanto, que las búsquedas sean sencillas y completas. En definitiva, se consiguen unos resultados más sistemáticos sin perder información relevante para la oralidad (diferentes manifestaciones de fonética sintáctica y de pronunciación de extranjerismos o siglas).

Etiqueta	Símbolo al que sustituye
<alargamiento/>	aa nn alargamiento vocálico y consonántico
Se utiliza para marcar tanto alargamientos vocálicos como consonánticos Ejemplo: camió<alargamiento/>n en<alargamiento/>	
Etiqueta	Símbolo al que sustituye
<ininteligible/>	(( ))
Marca fragmentos indescifrables Ejemplo: parece que <ininteligible/> mañana	
Etiqueta	Símbolo al que sustituye
<énfasis t=" "></énfasis>	PESADO
Se utiliza para pronunciación marcada y silabeo. Esta información aparece dentro del atributo. Ejemplo: eres un <énfasis t="silabeo">pesado</énfasis> eres un <énfasis t="pronunciación_marcada">pesado</énfasis>	
Etiqueta	Símbolo al que sustituye
<susurro></susurro>	º( )º
Marca voz baja o susurros. Ejemplo: mañana vamos a <susurro>comprar el regalo</susurro>	
Etiqueta	Símbolo al que sustituye
<cita></cita>	*Estilo directo*
Indica que se está reproduciendo el estilo directo. Ejemplo: y luego me dijo <cita>mañana no iré a trabajar</cita>	
Etiqueta	Símbolo al que sustituye
<fsr t=" "></fsr>	(en)tonces pa'l
Recoge fenómenos de fonética sintáctica y en general aquellos casos en los que la ortografía y la pronunciación de una palabra no coinciden. En el atributo se indica cómo se ha pronunciado y entre las etiquetas la forma ortográfica. Ejemplos: voy <fsr t="pa'l">para el</fsr> supermercado <fsr t="tonces">entonces</fsr> le dijo que no lo hiciera <fsr t="toa">toda</fsr> la noche <fsr t="po">pues</fsr> entonces no sé	
Etiqueta	Símbolo al que sustituye
<obs t=" "></obs>	Nota al pie

Recoge cualquier tipo de información relevante para la transcripción que no se recoge en el resto de las etiquetas. Se utiliza con fenómenos que solo afecten a un hablante. Ejemplos: <obs t="da palmas">madre mía</obs> veníamos de la <obs t="lugar de trabajo de la hablante B">cueva</obs>	
Etiqueta	Símbolo al que sustituye
<obs t=" "/>	Nota al pie
Recoge cualquier tipo de información relevante para la transcripción que no se recoge en el resto de las etiquetas. Puede aparecer en la línea de ELAN dedicada a observaciones o en la propia línea de hablante. Se utiliza con fenómenos que afectan a más de una intervención. Ejemplo: <obs t="hablan mientras ven un vídeo en el móvil"/>	
Etiqueta	Símbolo al que sustituye
<entre_risas></entre_risas>	Nota al pie
Delimita fragmentos entre risas. Ejemplo: mañana <entre_risas>vamos a flipar con el examen</entre_risas>	
Etiqueta	Símbolo al que sustituye
<risas/>	(RISAS)
Señala que alguno de los hablantes ríe. Ejemplo: no te estás enterando de nada <risas/> de nada	
Etiqueta	Símbolo al que sustituye
<tos/>	(TOS)
Señala que alguno de los hablantes tose. Ejemplo: esto<alargamiento/> <tos/>	
Etiqueta	Símbolo al que sustituye
<gritos/>	(GRITOS)
Señala que alguno de los hablantes chilla. Ejemplo: <gritos/> no me lo digas más veces	

Tabla 6. Sistema de etiquetas que sustituyen a signos de transcripción Val.Es.Co.

En la Tabla 7 se recogen aquellas etiquetas creadas para señalar –y, por tanto, poder recuperar posteriormente– información relativa a la anonimización, las siglas o los extranjerismos que en el sistema tradicional Val.Es.Co. no se señalaban o se hacía a través de notas al pie.

Etiqueta <anónimo></anónimo>
Marca las antropónimos y topónimos que han sido sustituidos por otros ficticios. Ejemplo: Sin anonimizar: ayer vi a María y a Pedro Anonimizado: ayer vi a <anónimo>Marta</anónimo> y a <anónimo>Pablo</anónimo>
Etiqueta <sic></sic>
Marca pronunciación errónea que no es un error de transcripción. Ejemplo: cómete las <sic>almóndigas</sic>
Etiqueta <siglas t=“ ”></siglas>
Marca el uso de siglas. Ejemplo: en <siglas t= “erre te uve e”>RTVE</siglas> dijeron ayer que el <siglas t=“pesoe”>PSOE</siglas> necesitaba más votos
Etiqueta <extranjero t=“ ”></extranjero>
Señala el uso de un extranjerismo. Entre las dos etiquetas se introduce la voz extranjera a la que se hace referencia y en el atributo la pronunciación de lo que se oye. Ejemplo: <extranjero t="pílic">peeling</extranjero>

Tabla 7. Etiquetas no existentes en el sistema Val.Es.Co.

En último lugar, como vemos en la Tabla 8, hay una serie de símbolos del sistema Val.Es.Co. que se han mantenido, ya que no entran en conflicto con el lenguaje XML y que, además, facilitan la lectura de la transcripción por parte del usuario no familiarizado con el entorno ELAN.

[	lugar donde se inicia un solapamiento o superposición.
]	final del habla simultánea.
-	reinicios y autointerrupciones sin pausa.
((siempre))	transcripción dudosa.
¿ ?	interrogaciones.
¿¡ !?	interrogaciones exclamativas.
¡ !	exclamaciones.
?	cuando no se reconozca el interlocutor se creará una línea en ELAN

Tabla 8. Signos de transcripción Val.Es.Co. que se mantienen

El alineado del audio y la transcripción a través de ELAN ofrece una serie de ventajas a la hora de representar la conversación. Además de incluir la visualización del oscilograma del fragmento, ELAN permite consultar las anotaciones realizadas en las líneas de los hablantes. Estas ventajas han permitido prescindir de algunos de los signos usados en el sistema Val.Es.Co., que pasan a ser innecesarios. Es el caso de las pausas, por ser visibles en el oscilograma cuando aparecen, o de la marca de sucesión inmediata:

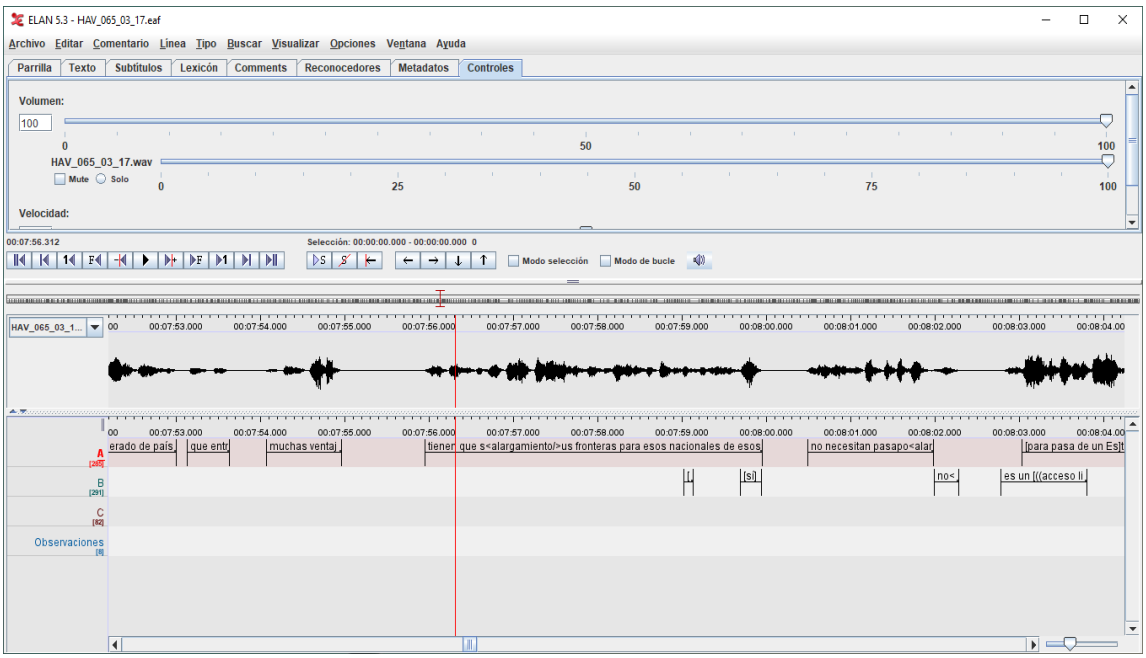

Imagen 1. Captura de pantalla del entorno de trabajo ELAN

Los símbolos utilizados en el sistema Val.Es.Co. que se han suprimido se recogen en la Tabla 9.

Pausas Puesto que se aprecian claramente en la ventana de trabajo de ELAN no es necesario marcarlas. Tonemas No se tienen que marcar los tonemas, aunque no coincidan con la pronunciación habitual. Sucesión inmediata entre interlocutores y mantenimiento del turno El sistema de transcripción Val.Es.Co. utiliza el símbolo § para indicar una sucesión inmediata, sin pausas, entre interlocutores. ELAN permite la visualización de este fenómeno sin necesidad de marcarlo. Lo mismo sucede con el símbolo = utilizado para indicar el mantenimiento del turno de un participante en un solapamiento.

Tabla 9. Signos de transcripción Val.Es.Co. no necesarios en ELAN

3.6 Validación de etiquetas

El equipo central se encarga de alinear la primera transcripción ancha hecha por cada equipo local e introducir las etiquetas. Ello implica, por tanto, una tarea doble para la cual se emplea el *software* especializado en la anotación lingüística (ELAN) introducido en el apartado 3.5.2. Este procedimiento cuenta con un inconveniente: mientras que el programa está diseñado para producir transcripciones alineadas con el tiempo, no está pensado para que la introducción de lenguaje basado en XML se pueda realizar de forma semiautomática. Ello quiere decir que el etiquetado se debe llevar a cabo de manera manual dentro del entorno de ELAN. Dicha limitación técnica implica que la posibilidad de error sea más alta y, por tanto, la necesidad de revisión del etiquetado adquiere mayor importancia.

Para asegurar que las etiquetas se han introducido correctamente, se ha incorporado un paso adicional en el proceso de construcción del corpus, esto es, la validación de las etiquetas. Para poder llevar a cabo esta labor se emplea el editor de XML Oxygen (https://www.oxygenxml.com/xml_editor.html). Entre muchas otras funciones, este programa está diseñado para poder detectar cualquier tipo de error que no permita que la etiqueta se pueda procesar correctamente, ya sea porque esté incompleta, existan errores de escritura o cualquier otro problema que entre en conflicto con la sintaxis del lenguaje XML.

La validación de archivos procedentes del programa ELAN supone un reto adicional, precisamente porque el formato de los archivos que produce emplea también XML. Por tanto, entran en conflicto las etiquetas con las que ELAN codifica la alineación temporal de las anotaciones con las empleadas en la transcripción de conversaciones. Así, es necesario importar solamente la transcripción producida con este programa y comprobar que los errores encontrados en la interfaz de Oxygen se corrigen en el archivo de ELAN.

3.7 La anonimización de los datos

Para garantizar el cumplimiento de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales, de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (LOPD) y su reglamento de desarrollo (RLOPD) y de la normativa propia de la Universitat de València, por una parte, los datos recogidos han sido sometidos a un proceso de anonimización y, por otra, se cuenta con una autorización en varias fases que refleja el consentimiento de los participantes con la grabación y el uso de sus datos, como se indica en el apartado 3.2.

Los nombres y apellidos de las personas explícitamente mencionadas, así como los nombres de calles, lugares de trabajo o cualquier otro elemento que pueda ser sensible a la identificación de hablantes han sido omitidos en el audio y sustituidos por nombres ficticios en la transcripción. En las fichas técnicas, los hablantes (sean nombrados o no en la conversación) se identifican mediante claves. Además, los datos se encuentran disociados, es decir, los archivos de audio, las transcripciones y las fichas técnicas están albergadas en ubicaciones diferentes a las de las autorizaciones, donde sí aparece el nombre real de los participantes de las conversaciones. Esto permite que no se puede rastrear la relación entre ellos y contribuye a garantizar la imposibilidad de identificar por ningún medio a los hablantes que participan en las conversaciones, incluso aunque se accediera a la base de datos interna del corpus.

3.8 Identificación de los archivos

Cada conversación posee un identificador único que designa los archivos asociados con cada conversación, esto es, el archivo de audio, la transcripción y la ficha técnica. Este código no solo tiene como cometido identificar la conversación en general y los diferentes archivos que la componen, sino que también refleja información básica sobre esta, tal como la ciudad de origen, el número de hablantes y el año en el que fue registrada. El equipo central se encarga de adjudicar un identificador a cada archivo según el sistema que se expone a continuación.

Si se toma como ejemplo el siguiente identificador, se puede observar que aparecen cuatro bloques de información separados por guiones bajos. El primero de ellos contiene caracteres alfabéticos, mientras que los tres siguientes se componen de datos numéricos.

VLC_001_03_18

El primer bloque de información está formado por tres o cuatro letras que representan el nombre de la ciudad en el que se ha grabado la conversación. Estas letras se corresponden con las abreviaturas de los códigos oficiales de sus aeropuertos. En caso de que no se pueda tomar un aeropuerto de referencia, se toman las tres primeras letras de la localidad. En el caso del ejemplo, corresponde a Valencia, cuyo aeropuerto tiene asignado el código VLC.

Los tres dígitos siguientes representan el código de grabación interno y se establece por orden correlativo a medida que las conversaciones se van grabando. Cabe mencionar que este número concierne exclusivamente a cada ciudad. Es decir, tiene en cuenta la cantidad de grabaciones en cada una de ellas y no del corpus en general. Si en un identificador encontramos que es el número 5 de grabación de Ciudad de México, ello significa que es la quinta grabación que se ha recibido de esta ciudad y no la quinta grabación en general que se ha registrado en el corpus. En el ejemplo anterior 001 indica que es la primera grabación.

El tercer bloque de información lo componen dos dígitos que representan el número de participantes en la conversación. Por tanto, en el ejemplo 03 quiere decir que en la conversación participan 3 hablantes. Finalmente, los dos últimos dígitos indican el año de recogida de la grabación. Se emplean únicamente las dos últimas cifras del año teniendo en cuenta que se graba después del año 2000. El número 18 indica pues que la grabación se produjo en el 2018.

4 LA RECUPERACIÓN DE LA INFORMACIÓN: EL MOTOR DE BÚSQUEDA

En la actualidad se está trabajando en la construcción de un motor de búsqueda en línea que facilite la recuperación de información al usuario de corpus. La versión beta pueda consultarse en http://corpusameresco.com/.

Las tareas de transcripción y anotación llevadas a cabo previamente facilitan las posibilidades de recuperación de la información en esta plataforma que no solo ofrece diferentes opciones de búsqueda, sino que también nos permite acceder a los fragmentos de audio y transcripción que le corresponden y descargar los resultados.

En esta aplicación de consulta podemos filtrar las búsquedas a través de diferentes opciones:

1. Búsquedas. Desde esta primera ventana de consulta se pueden realizar:
   a) Búsquedas por palabras (ya sea token, lema, categoría morfológica o forma).
   b) Búsquedas por opciones (palabras afectadas por etiquetas de cita, extranjero o alargamiento).

   c) Búsquedas por criterios prosódicos y de orden (orden de la palabra en la anotación, orden de la anotación en la intervención, orden de la intervención en la conversación, duración y pausa). Esta opción ha sido establecida dada la naturaleza del Grupo Val.Es.Co. especializado en el análisis de la conversación coloquial.
   d) Búsquedas por criterios sociolingüísticos: sexo, edad y nivel sociocultural del hablante.
   e) Búsquedas por origen de la conversación (esto es, ciudad, país y tema).
   f) Búsquedas por proximidad (bien por categoría morfológica y lema, bien por token y forma)
2. Acceso a conversaciones completas (tanto al audio como a su transcripción alineada).
3. Acceso a las intervenciones de cada hablante que conforma el corpus.
4. Acceso a las estadísticas: frecuencias léxicas por país, inventario de formas léxicas, frecuencias por hablantes y frecuencias de bigramas y trigramas (N-Grams)

La anotación morfológica y la lematización se ha realizado de manera automática mediante el etiquetador *Treetagger*.

Como se ha visto en el apartado 2, todos los materiales que se han recogido en el corpus también pueden consultarse por medio de la sección repositorio http://esvaratenuacion.es/corpus-discursivo-propio/. El usuario podrá encontrar allí los audios y transcripciones acompañados de su respectiva ficha técnica. Además, se ofrece la transcripción en formato de texto tabulado junto con el audio alineado.

5 CONSIDERACIONES FINALES

La recopilación de este corpus panhispánico es compleja desde diversos puntos de vista. Desde el punto de vista administrativo, existe dificultad, en ocasiones, para establecer convenios interuniversitarios. Desde el económico, en la mayoría de los casos, los equipos locales carecen de financiación para llevar a cabo las tareas de grabación y transcripción, y las labores de formación de los equipos deben realizarse necesariamente a distancia, lo cual ralentiza el proceso notablemente. Aunque ya se cuenta con la participación de 14 equipos en 7 naciones, no se considera esta tarea como finalizada. Se prevé que el corpus crezca en el futuro y se incluyan muestras de habla de países todavía no cubiertos, así como de otras zonas de España.

Por otro lado, como se ha señalado en las secciones 3 y 4, hay cuestiones técnicas referidas al tratamiento informático de los datos para las que aún no se han encontrado soluciones definitivas. Muestra de ello es el conflicto que se produce por ejemplo con el etiquetado XML propio del *software* ELAN en combinación con el sistema propio de etiquetado del corpus y la validación del etiquetado. Además, no se puede perder de vista que transcribir la oralidad supone un sometimiento a la norma escrita en la que el transcriptor tiene que decidir cómo mantener la idiosincrasia de este canal y a la vez ser fiel a la representación de los datos.

A pesar de estos obstáculos, la incorporación del corpus Ameresco al panorama de corpus del español existente rellena una parcela que estaba vacía en cuanto al género discursivo –esto es, conversación coloquial grabada secretamente– y en cuanto a su aspiración por ofrecer a la comunidad científica muestras de habla coloquial de toda Hispanoamérica.

6 BIBLIOGRAFÍA

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

Albelda, M. y M. Estellés (coords.) [en línea]: Corpus Ameresco, Universitat de València, ISSN: 2659-8337, www.corpusameresco.com

Barcala et alii (2018): «El corpus ESLORA de español oral», *CHIMERA: Romance Corpora and Linguistic Studies*, N.º 5, 2, 217-237.

Briz Gómez, A. y Grupo Val.Es.Co. (2002): «Corpus de conversaciones coloquiales», *Anejo 1 Oralia*, Madrid: Arco Libros.

Briz et alii (2019): *Protocolo de trabajo equipos Ameresco* [en línea].

Briz Gómez A. y A. Carcelén Guerrero (2019): «El futuro iberoamericano del español. La investigación del español oral y en español». *El español en el mundo:* Anuario del Instituto Cervantes.

Briz Gómez, A. (coord.) (1995): «La conversación coloquial (Materiales para su estudio)». Universidad de Valencia. *Anejo XVI de la Revista Cuadernos de Filología*, Valencia, Cuadernos de Filología.

Briz Gómez, A. (2001): *El español coloquial en la conversación*, Ariel, Madrid.

Briz Gómez, A. (2012): «Los déficits de los corpus orales del español (y de algunos análisis)». *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Servizo de Publicacións e Intercambio Científico da Universidade de Santiago de Compostela, 115-137.

Briz Gómez, A. (2016): «El proyecto AMERESCO. La idea de un corpus de conversaciones coloquiales del español de América», en *Oralidad y análisis del discurso. Homenaje a Luis Cortés*, Bañón Hernández, M. et alii (eds.), Almería: Editorial Universidad de Almería.

Cabedo, Adrián y Pons, Salvador (eds.): *Corpus Val.Es.Co 2.0.* Consultado online en http://www.valesco.es

Labov, W. (1972): *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Moreno Fernández, F. (2005): «Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA», *Oralia: Análisis del discurso oral*, Nº 8, 123-140.

Moreno Fernández, F. (2016): «En torno a preseea: Notas de investigación y de sociología de la ciencia», *Boletín de filología*: Universidad de Chile, Vol. 51, Nº. 2, (Ejemplar dedicado a: Estudios sobre la lengua española hablada en el mundo hispánico en su variedad geográfica y social con materiales del PRESEEA), 369-376.

Recalde Fernández, M. y V. Vázquez Rozas (2009): «Problemas metodológicos en la formación de corpus orales», en Pascual Cantos Gómez (ed. lit.), Aquilino Sánchez Pérez (ed. lit.) *A survey of corpus-based research*, 51-64.

Pons, S. (2019): *Corpus Val.Es.Co 2.1*, http:/www.valesco.es/corpus.

PRESEEA (2008): "Marcas y etiquetas mínimas obligatorias". Vers. 1.2. 17-02-2008. [http://www.linguas.net/preseea]

PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. [http://preseea.linguas.net]