Universitat de València
Facultat de Filologia, Traducció i Comunicació
Departament de Filologia Espanyola

VNIVERSITAT DE VALÈNCIA

Tesis doctoral

Bases teórico-metodológicas para la construcción de un corpus multidialectal de conversación coloquial: el corpus Ameresco

Presentada por

Andrea I. Carcelén Guerrero

Dirigida por

Maria Estellés Arguedas

Marta Albelda Marco

Doctorado en Estudios Hispánicos Avanzados

Valencia, octubre de 2023

Esta tesis ha contado con una ayuda predoctoral para Formación del Personal Investigador (FPI) de la Agencia Estatal de Investigación y el Ministerio de Ciencia e Innovación (referencia BES-2017-080287) con subvención de fondos comunitarios FEDER (FSE) y con el apoyo de dos proyectos de investigación, *La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de España y América, Es.Vag.Atenuación* (referencia FFI2016-75249-P) financiado por el Ministerio de Economía y Competitividad y el proyecto 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 el Ministerio de Ciencia e Innovación y la Agencia Estatal de Investigación, ambos proyectos dirigidos por las investigadoras principales Marta Albelda Marco y Maria Estellés Arguedas.

A Alberto, in memoriam

AGRADECIMIENTOS

La lingüística de corpus se puso en mi camino allá por 2010, cuando a través de la Universitat de València entré a formar parte del equipo de trabajo del corpus académico, CORPES XXI, como codificadora y más tarde también validadora; contaron conmigo para transcribir PRESEEA y otros corpus de aprendices de español. Quién me iba a decir que una década después iba a volver a esta casa como doctoranda y que mi camino como lingüista de corpus se iba a materializar en esta tesis. Y es gracias a mis directoras, las doctoras Marta Albelda y Maria Estellés, investigadoras responsables tanto del proyecto que me concede esta ayuda, como del corpus Ameresco, a quien debo mis primeros agradecimientos.

Marta, Maria, en lo profesional, gracias por vuestra labor como directoras, esta supera con creces las expectativas y así lo demuestra vuestro asesoramiento durante estos años, la revisión paciente que habéis hecho de este trabajo, las ideas que me habéis sugerido y el regalo de vuestro tiempo en las tutorías a cualquier hora, cualquier día de la semana. Vuestra excelencia como investigadoras y directoras es insuperable, sois mis referentes académicos y profesionales. En lo personal, gracias por no dejar que descarrilara, por haberme ayudado en tantos sentidos, por cogerme de la mano cuando hacía falta parar, pero también por darme impulso cuando había un poco de luz. Gracias por no haber perdido vuestra confianza en mí y por conseguir que yo misma confiara. ¡Qué suerte teneros en mi vida! Quiero señalar también a Adrián Cabedo cuya ayuda y apoyo han sido fundamentales en esta etapa; aunque esta *padawan* se haya alejado en algún momento del camino de la Fuerza, sus enseñanzas jedi me han llevado siempre por el buen camino.

Al Departamento de Filología Española, por haber sido mi casa estos años, en especial a Antonio Briz por su acogida en el grupo Val.Es.Co. y por haber podido trabajar a su lado; a las administrativas Elena Plano, Miriam Izquierdo y Pilar Almor, por su ayuda en tantos trámites, siempre amables y dispuestas a hacer el papeleo más fácil; a mis compañeras de Lingualitarias, Mercedes Quilis y Marta Pilar Montañez, por su enorme generosidad durante este trayecto y por haber aprendido de ellas y con ellas a activar la lupa violeta.

A mis hermanas académicas, Dorota Kotwica, Cristina Villalba, Gloria Uclés y Lissette Mondaca. A Dorota por los consejos de supervivencia en la fase final de tesis, por su compañerismo y por los hilos e historias sobre Pedro Pascal que atesoramos. A Cristina, por su paciencia cuando hemos compartido docencia y he tenido mil y una dudas, por las recomendaciones literarias y cinematográficas, por crear juntas un club de bordado que solo

celebró una sesión. A Gloria, mi *gemelier*, con la que he compartido salseos, recetas y compadreos académicos, por derribar muros y por reconducirnos por los campos de naranjos. A Lissette, porque me ha ayudado a encontrar paz y calma cada vez que lo he necesitado, por dejar que ocupara su casa en este verano de tesis y disfrutara del aire acondicionado en las olas de calor, por ser la Mónica de Rachel, porque somos un equipo. A las cuatro: habéis demostrado que el trabajo bien hecho puede ir de la mano de la amistad y de la solidaridad entre compañeras. La exquisitez de vuestra carrera investigadora es toda una inspiración para mí.

Todas vosotras, junto a Marta, Maria y Adrián, sois mi familia de otra sangre; intentar plasmar por escrito lo que supone teneros en mi vida no reflejaría ni un porcentaje mínimo lo importantes que sois para mí.

Gracias también a todos los equipos que forman Ameresco, porque sin su colaboración no habría sido posible el desarrollo de este trabajo: a Juliana de la Mora y Ricardo Maldonado, por las enchiladas y chilaquiles caseros y las conversaciones trascendentales al son de los mariachis; a Lety Colin y Yohana Beatriz Martínez, por encontrarnos a un lado y a otro del Atlántico espero que muchas veces más; a María Eugenia Flores, Armando González, Manuel Santiago Herrera y al resto de profesorado de la UANL de Monterrey, que me enseñaron lo mejor de la vida regiomontana; a Tetsuji Miyahara e Ana Isabel García Tesoro, quienes me acogieron como una más de su familia en mis días en Medellín y donde disfruté mucho a pesar de paros armados y huelgas estudiantiles; a Marta Samper y, con un cariño especial, a Clara Eugenia Hernández, quien nos enseñó el Teide desde Las Palmas, aunque no conseguimos verlo; a Fulvia de Morales y al equipo de Panamá, con quienes disfruté de los contrastes de la ciudad y de la vida a otro ritmo; a Silvana Guerrero y Javier Riffo del equipo de Santiago de Chile, y a Claudia Borzi, de Buenos Aires, por su gran compromiso e implicación en la construcción de este corpus, aunque queda pendiente una visita por esos lares; a Ana María González y Yohana Beatriz por el corpus de La Habana, a Yolanda Rodríguez y Yasmín Torres por el de Barranquilla, a Silvina Douglas por Tucumán, a Lissette Mondaca por Ameresco-Temuco y a Danny Murillo por el corpus de Tegucigalpa. A todos vosotros, gracias.

A la profesora Miriam Bouzouita, que amablemente me acogió en mi estancia en la Humboldt-Universität de Berlin, en tiempos aún convulsos por la pandemia; también al resto de los miembros de su grupo, por su ayuda con la burocracia alemana.

A mis padres, José Antonio y María Eugenia, por haber respetado mis tiempos, aunque no los entendieran, por haberme educado en la responsabilidad y el esfuerzo, en la bondad, por su amor infinito. A mi hermana Amaya, porque a pesar de los kilómetros que nos separan me siento más cerca de ella que nunca. A mí tía Encarni, por regalarme sus risas contagiosas, y a toda mi familia, por querernos con calidad.

A mi cable a tierra: Patricia, Noelia, María José, Elena y Marga. Por los mensajes y audios interminables en Whatsapp y Telegram, en los que hemos hablado de lo humano y lo divino, por los viernes mochi, porque habéis estado ahí siempre, para las risas, para los llantos y para todo lo demás.

A Leticia, porque gracias a ella he aprendido a entenderme, así como a comprender mejor cómo funciona el mundo. Tu ayuda ha sido (y es) imprescindible para vencer los obstáculos que van apareciendo y para superar mis miedos.

Y unas últimas líneas para agradecer a mis compañeritas peludas de vida, Cleo y Buffy, aunque nunca entiendan lo que significa su compañía para mí. En definitiva, gracias a quienes de una manera u otra habéis estado ahí, mandándome cariño y memes de gatetes durante este proceso turbulento.

Por todos ellos y por ti, Alberto, estoy hoy aquí.

RESUMEN

La presente investigación se centra en el estudio de los fundamento teórico-metodológicos para el diseño y construcción de corpus orales, en particular, de un corpus multidialectal de conversación coloquial en América y España: el corpus Ameresco.

El estudio se divide en dos partes principales: una fase exploratoria y una fase de aplicación técnica. La fase exploratoria se centra en entender el proceso de construcción de corpus y en establecer estándares uniformes para Ameresco, basándose en un estudio detallado del panorama de corpus existentes en el ámbito hispánico. La fase de aplicación técnica conlleva la exploración de herramientas informáticas para el procesamiento de corpus orales, como la transcripción y el alineado, y las implicadas para la difusión del corpus entre la comunidad científica a través de plataformas de búsqueda en línea que permitan la extracción selectiva y el análisis lingüístico de los materiales compilados.

La importancia de esta investigación radica en su enfoque en la metodología de construcción de corpus lingüísticos, en lugar de realizar estudios de caso particulares a partir de los datos específicos obtenidos a través de un corpus. Los objetivos de la investigación se dividen en dos bloques: la reflexión metodológica y la aplicación técnica. En primer lugar, se busca proporcionar las bases teórico-metodológicas para el diseño de un corpus oral conversacional a partir del análisis de corpus orales previos, analizando sus fortalezas y sus debilidades; identificar los problemas en las diferentes fases de diseño de corpus y proponer soluciones en cada una de ellas. En segundo lugar, desde la perspectiva aplicada, se pretende desarrollar el corpus Ameresco; hacerlo accesible a la comunidad científica desde un punto de vista amplio, esto es, compartiendo no solo el trabajo final del corpus, sino también, las decisiones metodológicas aplicadas en su construcción; como último objetivo se pretende reflexionar sobre la importancia del trabajo de diseño y compilación de corpus, a menudo minusvalorada.

Como cristalización de este trabajo se presentan las bases para establecer un modelo de referencia para la creación de otros corpus orales, maximizando la replicabilidad, a partir del corpus Ameresco.

ABSTRACT

The present research focuses on the study of the theoretical and methodological foundations for the design and construction of oral corpora, in particular, of a multidialectal corpus of colloquial conversation in America and Spain: the Ameresco corpus.

The study is divided into two main parts: an exploratory phase and a technical application phase. The exploratory phase focuses on understanding the corpus construction process and establishing uniform standards for Ameresco, based on a detailed study of the existing Hispanic corpus landscape. The technical application phase involves the exploration of computer tools for processing oral corpora, such as transcription and alignment, and those involved in disseminating the corpus to the scientific community through online search platforms that allow selective extraction and linguistic analysis of the compiled materials.

The importance of this research lies on its focus on the methodology of constructing linguistic corpora, rather than conducting particular case studies based on specific data obtained through a corpus. The research objectives are divided into two blocks: methodological reflection and technical application. Firstly, the aim is to provide the theoretical and methodological bases for the design of a conversational oral corpus based on the analysis of previous oral corpora, analyzing their strengths and weaknesses; to identify the problems in the different phases of corpus design and to propose solutions in each of them. Secondly, from the applied perspective, the aim is to develop the Ameresco corpus; to make it accessible to the scientific community from a broad point of view, that is, sharing not only the final work of the corpus, but also the methodological decisions applied in its construction; the last objective is to reflect on the importance of the corpus design and compilation work, often underestimated.

As a crystallization of this work, the bases are presented to establish a reference model for the creation of other oral corpora, maximizing replicability, based on the Ameresco corpus.

Capítulo 1

Introducción

1.1. Motivación y justificación del estudio

La tesis doctoral que se presenta a continuación es fruto del trabajo desarrollado gracias a un contrato predoctoral FPI para la formación de personal investigador llevado a cabo durante casi cinco años y al que le preceden años de interés particular por la lingüística de corpus. En el marco de este contrato, se han desarrollado tareas que iban de la mano de los primeros pasos en la construcción del corpus Ameresco, iniciado apenas unos años antes. Por tanto, desde el comienzo, las inquietudes de esta investigación tenían que ver con los retos y planteamientos metodológicos exigidos a la hora de construir un corpus lingüístico, particularmente, un corpus oral de conversación coloquial grabada secretamente.

Ante el reto de crear un corpus, especialmente uno de la envergadura del que se nos había encargado, hay dos pasos previos fundamentales. El primero consiste en adquirir una formación técnica lo más sólida posible dentro de las posibilidades de un lingüista; el segundo, en aprender de la experiencia de otros corpus y construir sobre el conocimiento, replicando los aciertos y tratando de enmendar los errores, de quienes nos han precedido. En consecuencia, la investigación realizada ha sido doble: primero, ha conllevado una fase exploratoria y, en segundo lugar, una fase de aplicación técnica.

La base exploratoria de este trabajo se concibe como medio para entender el proceso de construcción de corpus y para establecer los cimientos del corpus Ameresco. Una mirada rápida a la bibliografía permite observar una llamativa falta de homogeneidad terminológica, metodológica y tipológica dentro del panorama de corpus orales disponibles, en las diferentes fases de su diseño, desde su concepción hasta su conclusión. De ahí que uno de los primeros intereses perseguidos en este sentido haya sido buscar para Ameresco un estándar de construcción uniforme, eficiente y replicable.

Para ello, se ha partido de un estudio pormenorizado de los corpus existentes en español, prestando especial atención a los corpus orales. Este estudio no solo ha permitido cartografiar el estado actual de dichos corpus en el ámbito hispánico, sino que nos ha permitido analizar críticamente este mapa de corpus y así profundizar en las fortalezas y debilidades de cada uno de ellos en sus diferentes etapas de construcción. Estas son la fase 1, de concepción y recogida de los datos; la fase 2, de tratamiento de los datos recogidos, y la fase 3, de archivo, distribución y acceso a dichos datos por parte de la comunidad científica.

La base técnica aplicada de esta tesis ha pasado por explorar y explotar múltiples opciones informáticas para el procesamiento computacional de corpus orales, sobre todo aquellas destinadas a la transcripción y el alineado de corpus orales, así como otros medios para la síntesis y recuperación de los materiales por medio de motores de búsqueda en línea. Estas herramientas y sistemas han sido valorados en función de sus ventajas e inconvenientes atendiendo a los intereses propios de la construcción del corpus Ameresco.

Esta revisión metodológica se ha concretado en la propuesta justificada de una metodología de trabajo en cuanto al diseño y construcción del corpus Ameresco en las fases 1, 2 y 3 materializada en esta tesis.

El valor de esta investigación reside, por tanto, en que se presenta como una novedad de estudio, ya que se articula en torno a la metarreflexión metodológica de la lingüística de corpus. Es decir, no se realiza un trabajo de análisis de corpus o de análisis con corpus para una variable lingüística, concepciones estas más clásicas de la lingüística de corpus, sino que se atiende a la propia metodología de construcción de los corpus para establecer un modelo de trabajo. Se ha observado cierta opacidad en la bibliografía a este respecto, posiblemente derivada de una falsa creencia en la que se valoran más a nivel científico los análisis realizados con los datos recogidos por el corpus que el propio trabajo realizado a la hora de diseñar y protocolizar la compilación de este.

En este sentido, se aspira a establecer un modelo que pueda servir de referencia para la elaboración de otros corpus orales, fundamentado en la puesta a disposición de la comunidad científica de todos los materiales y protocolos resultantes de manera que se maximice la replicabilidad. Si bien, los estudios basados en corpus son numerosos, no ocurre así con los que ofrecen una reflexión metodológica sobre la construcción de dichos corpus. El hecho de que esta investigación haya estado ligada laboralmente a las tareas desarrolladas en la etapa predoctoral, como se ha señalado al comienzo, hace que sea una tesis experimentada en cuanto a que bebe directamente de la experiencia personal, del trabajo de campo en la gestión y construcción de un corpus real, que puede ser consultado y explotado por la comunidad científica, gracias a estas tareas: el corpus multidialectal de conversación coloquial Ameresco.

Por último, este trabajo nace desde la reivindicación de la necesidad de poner en valor el trabajo de compilación de corpus, tanto en sus aspectos técnicos como en la labor, invisible y poco reconocida, de concepción y planificación previas.

1.2. Objetivos de investigación

La investigación que se presenta a continuación se divide en dos grandes bloques: la investigación y reflexión metodológica, por un lado, y su aplicación técnica, por otro.

Desde el punto de vista de la investigación y la reflexión previa, este trabajo se propone:

1. Ofrecer las bases metodológicas fundamentales para el diseño y la creación de un corpus de lengua oral conversacional, desde su concepción hasta su distribución en redes públicas; en particular, de un corpus que pretende recoger diversas variedades del español y que aspira a ser un macrocorpus panhispánico confeccionado por varios equipos.
2. Analizar críticamente las ventajas e inconvenientes, las fortalezas y debilidades de otros diseños de corpus orales previos, como base y aprovechamiento para la propuesta que aquí se pretende realizar.
3. Identificar y señalar todos los problemas asociados a las diversas fases de diseño de un corpus oral a partir de la exploración de los distintos procesos de diseño, desde la recogida hasta su incorporación en motores de búsqueda en la red.
4. Estudiar las diversas posibilidades técnicas de la solución de los problemas en el caso del corpus Ameresco, en cada una de sus fases de diseño y creación.
5. Proponer soluciones en cada una de estas fases, justificadas de acuerdo con criterios de rentabilidad para su empleo y facilitación de acceso al usuario de corpus, así como de respeto a la naturaleza lingüístico-discursiva de los materiales. Para ello se aprovecharán tanto las alternativas ofrecidas por la bibliografía como las experiencias propias en la creación y elaboración del corpus Ameresco.

Desde el punto de vista aplicado, esta tesis persigue tres objetivos adicionales:

6. A la luz de los objetivos 1 a 5, desarrollar al máximo posible el corpus Ameresco de conversaciones coloquiales en español de América y España.
7. Ofrecer el corpus a la comunidad científica dotándolo de un espacio accesible a través de la red y construyendo una interfaz de búsqueda que logre un equilibrio entre efectividad e intuitividad.
8. Reflexionar sobre la importancia del trabajo de diseño, coordinación, compilación y tratamiento de corpus, así como la necesidad de publicitar los mecanismos y protocolos como parte imprescindible del proceso de investigación.

Tras detallar los objetivos de investigación presentaremos, a continuación, la estructura general de esta tesis, así como indicaciones para su lectura.

1.3. Estructura del trabajo

En las líneas siguientes se presenta la distribución del contenido de esta investigación a lo largo de cinco capítulos. En ellos se abordarán, en primer lugar, una aproximación a la lingüística de corpus desde su origen hasta el momento actual; en segundo lugar, el panorama de corpus orales en español; en tercer lugar, las fases de diseño y construcción de corpus orales; a continuación, se presentará la caracterización del corpus Ameresco y la metodología de trabajo desarrollada; y, por último, las conclusiones de este trabajo, así como las proyecciones de futuro con respecto a la investigación realizada. En este punto, cabe señalar que los fundamentos teóricos en los que se basa este trabajo y sus aplicaciones prácticas se trabajan conjuntamente en cada capítulo, de manera que cada una de las partes que lo componen contienen una sección teórica y una sección aplicada en la que se analizan de manera práctica los contenidos teóricos que se han comentado en primer lugar.

En el Capítulo 2 se exploran los principios esenciales de la lingüística de corpus y se realiza un recorrido por esta disciplina, revisando su origen y evolución en el contexto de la lingüística en general. Esta aproximación sienta los cimientos para los capítulos posteriores en los cuales se proporciona una visión general de la aplicación de la lingüística de corpus en el análisis de corpus orales del español, foco central de esta investigación. En la primera sección (§ 2.1.) exploraremos las posturas adoptadas en la literatura con respecto a la lingüística de corpus, tanto como una disciplina lingüística en sí misma como una metodología para la investigación. También examinaremos otros conceptos relevantes señalados por la bibliografía y, a continuación, haremos un recorrido a lo largo de la evolución histórica de los corpus, desde sus inicios y el desarrollo de los primeros, hasta su estado actual, detallando sus posibles utilidades (§ 2.2). Seguidamente, examinaremos las principales categorizaciones de corpus según las diferentes clasificaciones establecidas por la bibliografía (§ 2.3), tras lo cual analizaremos la relación entre los corpus y los estudios lingüísticos en contextos específicos, con un enfoque especial en los corpus utilizados para investigaciones pragmáticas (§ 2.4). Por último, concluiremos con un resumen y evaluación de este capítulo (§ 2.5).

En el Capítulo 3, por su parte, se ofrece un breve repaso de la evolución de los corpus orales en español (§ 3.1). A continuación (§ 3.2), se lleva a cabo un análisis detallado de los estudios recopilatorios más destacados que, con enfoques variados, han enumerado los corpus orales disponibles en español hasta la fecha. Posteriormente (§ 3.3), se presenta una panorámica general de los corpus orales disponibles en el ámbito hispánico, con el objetivo doble de servir como recurso para la comunidad científica y de ofrecer una lista actualizada a fecha 2023. Para concluir, en la última sección (§ 3.4), resumiremos las ideas principales que se han abordado en este capítulo.

El Capítulo 4 detalla las fases implicadas en la creación de corpus orales, específicamente aquellos de naturaleza conversacional y espontánea. Comenzando con las consideraciones generales que deben tenerse en cuenta al abordar la construcción de un corpus (§ 4.1), se exploran las distintas fases que conforman la ejecución del diseño acordado (§ 4.2), las cuales se dividen en las siguientes partes: la primera (§ 4.2.1) se centra en los aspectos que afectan a la concepción del corpus y a la recopilación de datos; la segunda (§ 4.2.2) examina los temas relacionados con el tratamiento de estos datos; la tercera (§ 4.2.3) se ocupa de la explotación de los datos. Luego, se lleva a cabo un análisis comparativo entre las metodologías utilizadas en la creación de corpus orales en español en cada una de las fases mencionadas anteriormente (§ 4.3). Para concluir, se presentará una síntesis del capítulo (§ 4.4).

El Capítulo 5 se enfoca en las particularidades del corpus Ameresco en términos de su metodología de recopilación y construcción, siguiendo las tres fases establecidas en el Capítulo 4. En primer lugar (§ 5.1), se exponen las características generales del proyecto, incluyendo sus antecedentes, sus objetivos de investigación y los grupos que han participado en él. Luego (§ 5.2), se profundiza en el diseño y la creación del corpus, abordando las diversas etapas de trabajo: en la fase 1, se trata la concepción y recolección de los datos (§ 5.2.1); en la fase 2, se analiza el proceso de tratamiento de los datos (§ 5.2.2); y en la fase 3, se detallan aspectos relacionados con el almacenamiento, la distribución y el acceso de los datos por parte de los usuarios (§ 5.2.3). A continuación, (§ 5.3) se abordan las dificultades encontradas en cada una de las fases y las posibles soluciones que se han propuesto para el corpus Ameresco. Cerramos el capítulo (§5.4), con una síntesis del contenido expuesto en él.

En el Capítulo 6 se recogen las consideraciones finales de esta investigación. Primeramente, (§ 6.1.) se detallan los principales hallazgos obtenidos en este trabajo. A continuación, (§ 6.2.) nos centramos en exponer la relevancia de los resultados obtenidos, concretándolos en la aportación particular del corpus Ameresco al panorama internacional de corpus orales y en la defensa de la explicitación de la metodología de construcción de corpus como punto fundamental en el desarrollo de la lingüística de corpus. Por último, (§ 6.3.) se hace una reivindicación de la figura del lingüista compilador de corpus, papel que suele pasar desapercibido en el reconocimiento académico y científico.

Advertencias para la lectura

Con el fin de facilitar el acceso a determinadas informaciones y materiales, especialmente durante su consulta en la versión impresa, en este trabajo se han incluido códigos QR a través de los cuales puede accederse a dos tipos de datos: por un lado, dan acceso a material disponible en línea de corpus orales mencionados en el texto (en este caso, también se aporta el hipervínculo, por lo que se contemplan dos modos de lectura, en papel y digital); por otro, dan acceso a material sonoro del corpus Ameresco utilizado en los ejemplos, que permitirá a los lectores ilustrar los ejemplos escritos con el material sonoro del que partieron. Además, algunas tablas llevan su propio código QR que remite al listado de corpus referidos en este trabajo en donde se desarrolla el nombre completo, junto a su acrónimo.

Capítulo 2

La lingüística de corpus

En este capítulo se abordan los conceptos fundamentales de la lingüística de corpus. Se hace un repaso de esta disciplina y de sus instrumentos, acudiendo a su génesis y a su desarrollo en el ámbito de la lingüística general. De este modo, se establecen las bases para los capítulos siguientes en los que se ofrece una panorámica de la lingüística de corpus aplicada a los corpus orales del español, que es el tema que nos ocupa.

En el primer apartado se discutirá qué posición se ha adoptado en la bibliografía referida a la consideración de la lingüística de corpus, bien como disciplina de la lingüística, bien como metodología de trabajo e investigación; también se revisarán otros conceptos relevantes señalados por la literatura al respecto (§ 2.1.). Seguidamente, ofreceremos un recorrido a lo largo de la historia de los corpus, desde sus orígenes, con el desarrollo de los primeros corpus, hasta la situación actual, describiendo las posibles funcionalidades de estos (§ 2.2). Luego, reuniremos las principales tipologías de corpus según las clasificaciones establecidas por diversos autores (§ 2.3.). A continuación, veremos qué relación existe entre los corpus y los estudios lingüísticos en contextos determinados, prestando atención especial a los corpus para el estudio pragmático (§ 2.4). Para terminar, se presentará un resumen y valoración del capítulo (§ 2.5.)

2.1. Aproximación a la lingüística de corpus

Sin ánimo de ofrecer una nueva definición de qué es un corpus lingüístico, definiciones hoy ya ampliamente desarrolladas y tratadas en la bibliografía, cabe señalar que la mayoría de los autores coinciden al apuntar que un corpus consiste en reunir una serie de textos naturales, escritos u orales, que se almacenan y son procesados digitalmente con mayor o menor grado de detalle y cuyo objetivo es servir de material de estudio empírico sobre la lengua. Así lo señalan autores como Crystal (1991), Leech (1992), Sinclair (1996), Torruella y Llisterri (1999), McEnery y Wilson (2001), Berber Sardinha (2004), Parodi (2008), Briz y Albelda (2009), Martín Herrero (2009), Baker (2010), McEnery y Hardie (2011), Rojo (2016a), entre otros.

Como señala Parodi (2008), hay tres aspectos relevantes en los que coinciden las diversas definiciones, a saber,

> 1) un corpus debe estar compuesto por textos producidos en situaciones reales, 2) la recolección de estas instancias de lengua en uso debe estar guiada por parámetros explícitos que permitan tener clara la constitución de las mismas, de modo que se apoyen en el análisis y se posibilite la replicabilidad en estudios posteriores, y 3) un corpus (aunque dicho de modo

> implícito) debe estar disponible en formato electrónico con el fin de ser analizado por medio de programas computacionales. (Parodi, 2008, p. 103)

Resulta relevante tratar en este punto la diferenciación entre *archivos de textos*, *bases de datos* y *corpus*. Lo que diferencia un corpus de un archivo es, precisamente, que a menudo los textos en un corpus se han seleccionado para que puedan considerarse representativos de una determinada variedad lingüística o género, por lo que sirven de referencia estándar (Baker, Hardie y McEnery, 2006, pp. 48-49). Un *archivo* es un depósito de textos, generalmente de gran tamaño, recogido de forma oportunista, y normalmente no estructurado (Kennedy 1998, p. 4), a diferencia del *corpus* que estaría construido según criterios de diseño explícitos para un fin específico (Atkins, Clear y Ostler, 1992, p. 1). Por su parte, según Baker, Hardie y McEnery (2006, p. 55), el concepto *base de datos* suele utilizarse para referirse a grandes colecciones de textos que, a diferencia de los corpus, no están formadas por muestras, sino que constituyen toda una población de datos. Kennedy (1998, p. 4) señala que, aunque muchas bases de datos no se someten al análisis de corpus, no hay razón para que no se utilicen para este fin. Por lo tanto, podría decirse que un archivo y una base de datos son simples repositorios textuales, mientras que un corpus sigue unos criterios predefinidos en cuanto a la representatividad y al equilibrio de la muestra. La diferenciación entre estos conceptos, tal y como apunta Baker (2006, p. 27) se difumina en la práctica. Esta indefinición terminológica se analizará en el Capítulo 3.

Precisamente, que sean textos producidos en situaciones reales lleva a considerar la lingüística de corpus como una aproximación teórica y metodológica en los estudios lingüísticos muy alejada de la visión generativista de la lengua, basada esta última en la introspección. La lingüística de corpus, muy al contrario, se focaliza en el estudio empírico de datos objetivos y computables.

En esta línea, las consideraciones en torno a si la lingüística de corpus se constituye como ciencia o, más bien, como una metodología han generado cierta controversia. Es decir, cabe preguntarse si constituye una disciplina de por sí, como lo hacen, por ejemplo, la sociolingüística, la dialectología o la fonética, o si se constituye más bien como una metodología de trabajo aplicable a cualquiera de las disciplinas en las que se desglosan los estudios lingüísticos.

Para Parodi (2008), la lingüística de corpus en su versión actual constituye un enfoque metodológico para el estudio de las lenguas que brinda una base empírica para el desarrollo de materiales. Por tanto, según este autor, la lingüística de corpus

> constituye un conjunto o colección de principios metodológicos para estudiar cualquier dominio lingüístico y que se caracteriza por brindar sustento a la investigación de la lengua en uso a partir de corpus lingüísticos con sustrato en tecnología computacional y programas informáticos *ad hoc.* (Parodi, 2008, pp. 95-96)

Según la propuesta anterior, la lingüística de corpus no constituiría una rama o un área de la lingüística, sino que se vería como un método de investigación aplicado a ella. Por su parte, también McEnery y Wilson (2001) se plantean la cuestión de si estamos ante una rama de la lingüística o ante una aproximación metodológica; a esta pregunta, sin embargo, los autores no responden de forma tajante. Pese a todo, parecen decantarse más por considerarla una metodología:

> Corpus linguistics is not a branch of linguistis in the same sense as syntax, semantics, sociolinguistics and so on. All of these disciplines concentrate on describing/explaining some aspects of language use. Corpus linguistics in contrast is a methodology rather than an aspect of language requiring explanation or description. […] Corpus linguistics is a methodology that may be used in almost any area of linguistics, but it does not truly delimit an area of linguistics itself. (McEnery y Wilson, 2001, p. 2)

También Rojo (2021), señala a la falta de acuerdo sobre si la lingüística de corpus es una nueva teoría o una nueva metodología. Según este autor, no constituye en sí una teoría, ya que los datos recogidos en los corpus pueden ser analizados desde diferentes perspectivas lingüísticas. No obstante, tampoco parece fácil entenderla únicamente como una metodología, a pesar de que ha supuesto una revolución instrumental por la incorporación del procesamiento informático[1]. Concluye Rojo que se trataría, por tanto, "de una aproximación al estudio de los hechos lingüísticos de orientación empírica y basada en el análisis detallado de gran cantidad de datos" (2021, p. 47).

[1] No deben confundirse la lingüística de corpus con la lingüística computacional. Como señala Bolaños (2015, p. 33) se trata de actividades estrechamente ligadas pero que tienen objetivos distintos; la lingüística computacional se encarga de diseñar herramientas informáticas para un óptimo procesamiento automático del lenguaje natural, mientras que la lingüística de corpus se interesa por realizar análisis de carácter lingüístico mediante el uso de herramientas informáticas diseñadas para este fin.

Justamente, la lingüística se considera hoy una ciencia empírica cultural, si bien esta consideración es relativamente reciente. Como tal "está sometida a los requisitos del conocimiento científico, pero sin posibilidad de aspirar a alcanzar las características de fijeza y predictibilidad que poseen las que se ocupan de objetos naturales" (Rojo, 2021, pp. 35-36). A este respecto los corpus se han convertido en material imprescindible en los que basar la investigación empírica por medio de la observación, la inducción, la deducción, la comprobación y la evaluación de los datos (Krug, Schlüter y Rosenbach, 2013, p. 5).

La lingüística de corpus no siempre ha gozado del prestigio que tiene actualmente. Recibió duras críticas de mano de autores de la talla de Chomsky o Abercrombie, lo que llevó a un "desprestigio general de la metodología basada en corpus (empirismo) a favor de una nueva ortodoxia en los estudios lingüísticos: un acercamiento basado en las intuiciones del lingüista (racionalismo)" (Villayandre, 2008, p. 331). Como describe esta autora, las críticas de Chomsky se basaron en dos hechos fundamentales:

I. La apelación al recurso de la intuición, a la introspección del lingüista, como el único criterio válido para el estudio de la lengua.
II. El papel central otorgado a la sintaxis en las primeras versiones del modelo generativista. (Villayandre, 2008, p. 332)

En cambio, por su parte, la crítica de Abercrombie estaba más encaminada a la ejecución práctica de los corpus que a la teórica vertida por Chomsky. En este sentido, la lingüística de corpus fue tildada de pseudo-técnica en la que el procesamiento de datos era lento, propenso al error y caro (Villayandre, 2008, p. 333). Estas críticas fueron rebatidas décadas después por Leech (1992), quien señala, entre otros aspectos, que los corpus ofrecen una serie de ventajas frente a la intuición, que los datos cuantitativos que se obtienen son de gran utilidad y que el procesamiento de datos a través de ordenadores garantiza la fidelidad y disminuye la posibilidad de cometer errores, frente a un procesamiento humano (Villayandre, 2008, p. 336). Treinta años después de estas afirmaciones de Leech, y muy asentado el desarrollo informático de los corpus, así como la madurez en los procesos estadísticos, puede, sin duda, demostrarse que el procesamiento de los datos en la lingüística de corpus, ni es lento, ni propenso al error ni caro. Es, más bien, todo lo contrario, y prescindir de esta ayuda en el estudio de la lengua, es obviar el progreso en el desarrollo de la ciencia.

No cabe duda ya, pues, de que los estudios científicos sobre la lengua deben estar fundamentados en ejemplos reales que corroboren o desmientan las intuiciones de quien investiga, evitando el condicionamiento y la predisposición de nuestro conocimiento de la norma, así como relativizando también nuestras propias intuiciones, que pueden pecar de particulares y subjetivas. La introspección, por tanto, cada vez tiene menos cabida en el estudio del comportamiento lingüístico. Como apunta Tognini-Bonelli (2001, p. 86), "the unexpectedness of the findings derived from corpus evidence leads to the conclusion that intuition is not comprehensively reliable as a source of information about language". Pese a esto, la introspección es crucial para la interpretación y el análisis de los resultados de las colocaciones y para la identificación de las relaciones léxicas (Sinclair, 1997, p. 29), entre muchas otras ventajas.

Para garantizar la naturalidad de las muestras recogidas que conforman un corpus, estas pueden obtenerse por medio de diferentes técnicas que respaldan la idea que venimos comentando aquí, esto es, la observación de los datos nos permite conocer cómo funciona la lengua y reconocer patrones de comportamiento en diversas dimensiones lingüísticas. Estos datos pueden obtenerse directamente de los hablantes, esto es, cuando estos producen de manera natural muestras de habla, como sucede, por ejemplo, con los corpus de conversaciones espontáneas en las que suele aparecer la figura del investigador observador. Sin embargo, también existen otras técnicas de obtención de datos, centradas en técnicas de elicitación, como, por ejemplo, los test de hábitos sociales, las tareas DCT (*Discourse Completion task*), las encuestas, las entrevistas o los *role-plays*.

Como señala Senft (2009, p. 105), estas dos maneras de proceder (material no elicitado, en el caso del investigador observador, frente a material elicitado, en el resto de los casos), marcan los dos extremos de la obtención de los datos lingüísticos. Dependerá de la finalidad de la investigación elegir qué método será más adecuado al propósito particular de cada estudio, de manera que se obtenga una muestra lo más representativa posible (Senft 2009, p. 106).

En la siguiente figura se observa la organización general de los distintos procedimientos de obtención de datos en lingüística según la gradación que sucede entre los dos extremos, el más natural y el más monitorizado (Krug, Schlüter y Rosenbach, 2013, p. 6)[2]:

[2] En Rojo (2021, p. 43) se puede encontrar una traducción al español de esta figura.

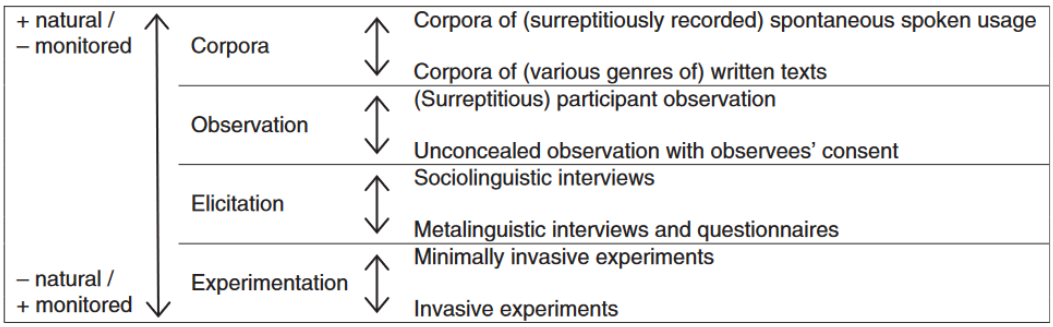


Figura 1. Diversos modos de obtención de datos lingüísticos para la investigación de acuerdo con el gradiente naturalidad/monitorización

La manera de enfocar los estudios con corpus o basados en corpus, responden a la distinción entre *corpus-based* y *corpus-driven*. Esta distinción fue introducida por Tognini-Bonelli (2001) quien saca a colación la oposición entre el método deductivo (*corpus-based*) y el método inductivo (*corpus-driven*).

En un enfoque *corpus-based* se utiliza un corpus existente como fuente de datos para respaldar afirmaciones lingüísticas o hipótesis previamente formuladas (Tognini-Bonelli 2001, p. 65). Sería el caso, por ejemplo, del estudio de los usos de *deber/deber de + infinitivo* con valor de probabilidad en un corpus de lengua (por ejemplo, este es uno de los objetivos de estudio del Proyecto PRESEEA). El reconocimiento, análisis y clasificación de los casos de estas perífrasis vendrían precedidos por un *a priori* lingüístico, por el que ya es sabido que algunos hablantes emplean la perífrasis sin preposición (*deber + infinitivo*) no con valor de obligación (como le es propio), sino de probabilidad (véase, por ejemplo, Blas, 2010, Gómez Molina, 2013, Manjón Cabeza, 2017). Se parte, pues, de un *a priori* teórico, que orienta el análisis de los casos particulares del corpus en cuestión.

No obstante, como señalan McEnery, Xiao y Tono (2006, p. 6), existen detractores sobre esta posición de la lingüística de corpus, puesto que, en ocasiones, ciertas investigaciones basadas en este enfoque no se comprometen plenamente con los datos del corpus en su conjunto, pues descartan las pruebas inconvenientes, es decir, los datos que no se ajustan a la idea concebida antes de obtener el corpus.

En cambio, en un enfoque *corpus-driven* el corpus se utiliza como punto de partida y fuente principal de evidencia, es decir, el corpus revela patrones y regularidades en el lenguaje sin preconcepciones teóricas previas. En este enfoque, los lingüistas están

estrictamente comprometidos con "la integridad de los datos en su conjunto" (Tognini-Bonelli, 2001, p. 84) y, por lo tanto, se afirma que los presupuestos teóricos son totalmente coherentes con las pruebas proporcionadas por el corpus y las reflejan directamente (McEnery, Xiao y Tono, 2006, p. 6). Un ejemplo de este enfoque sería el estudio de una estructura lingüística en la que se observe la generación de nuevos valores en el uso. Pensemos, por ejemplo, en los valores polisémicos que parecen estar emergiendo y consolidándose en algunas variantes del español en estructuras con valor evidencial, como *se ve, por lo que se ve, según parece, por lo visto,* etc. (véase, entre otros, Albelda y Jansegers, 2019, Estellés y Albelda, 2020). El estudio de estas combinaciones en corpus permite detectar qué diversos valores surgen en estas estructuras y qué grado de frecuencia y de fijación tienen en la lengua actual.

Ambos enfoques tienes sus ventajas y sus inconvenientes. La lingüística *corpus-driven* permite un acceso sin sesgo a los datos, lo que implica una apertura a la observación de la lengua desde el descubrimiento, atenta a los cambios y movimientos naturales de esta en su uso. Como inconveniente, cabe destacar principalmente, el coste de trabajo, puesto que llegar a detectar y establecer el comportamiento de determinados elementos de la lengua requerirá el estudio de gran cantidad de datos y de diversos géneros y variedades de la lengua, puesto que, en muchos casos, algunos fenómenos lingüísticos o bien son poco frecuentes, o bien se especializan caprichosamente en algún tipo de variedad.

Presentados ambos tipos de enfoques, sus ventajas y sus inconvenientes, como apuntan McEnery, Xiao y Tono, cuando las circunstancias de la investigación permiten tanto el análisis *corpus-based* como el *corpus-driven*, el investigador puede hibridar ambos enfoques combinando los méritos de la deducción y la inducción al mismo tiempo (McEnery, Xiao y Tono, 2006, pp. 9-10). El resultado de un trabajo triangulado con diversos enfoques enriquecerá siempre la investigación.

Algunas de las ventajas principales de trabajar con corpus han sido recogidas por Gries (2006, pp. 191-192), particularmente desde una perspectiva cuantitativa y estadística. Para este autor, las cuantificaciones basadas en corpus permiten una identificación más objetiva de lo que puede resultar importante y lo que puede considerarse más bien marginal. Así mismo, el apoyo en pruebas estadísticas para la lingüística de corpus permite realizar pruebas fiables, así como comparaciones de diferentes estudios, a diferencia de lo que ocurre si el análisis de los datos se basa exclusivamente en el juicio, más o menos subjetivo del

investigador. Cabe señalar también que los enfoques basados en corpus suelen permitir estudios empíricamente más versátiles que los estudios basados en juicios aislados; y, dado que los corpus consisten en habla y escritura producidas de forma natural, los enfoques basados en corpus permiten una perspectiva más adecuada que la investigación del lenguaje producido de forma aislada y desprovisto de todo contexto.

La lingüística de corpus facilita describir qué ocurre en un determinado contexto discursivo, explicar por qué ocurre, predecir qué ocurrirá y controlar qué influencias puede recibir o qué elementos pueden condicionar el comportamiento de determinados componentes lingüísticos (Gries, 2010, pp. 269-272). Así mismo, permite, además, expandir los estudios lingüísticos particulares, ya que gracias al tratamiento informático de los corpus y al surgimiento de *software* específico para la explotación de los datos podemos realizar búsquedas de palabras o construcciones en un contexto, hacer extracciones de manera exhaustiva (BIG DATA), calcular cuántas veces ocurre (tiene lugar) esa construcción o clasificar los datos de diversas maneras y aplicar sobre ellos técnicas estadísticas. En concreto, en términos de Davies y Parodi (2022), gracias a los corpus es posible

> generar listas de frecuencias de palabras, encontrar la frecuencia de palabras, frases y n-grammas (cadenas con N número de palabras), investigar construcciones sintácticas, realizar otras búsquedas que aborden información semántica y extraer las colocaciones de palabras y frases. (Davies y Parodi, 2022, p.17)

2.2. Origen histórico y desarrollo inicial de los corpus lingüísticos

Como se ha adelantado al comienzo de este capítulo, la historia de la lingüística de corpus no siempre ha gozado del respaldo de la comunidad científica. Este hecho, por tanto, ha afectado a su desarrollo, motivo por el cual, podrían considerarse tres periodos clave desde su origen hasta el momento actual, (Villayandre, 2008): primera generación (hasta mediados del siglo XX), segunda generación (años 60 y 70), y tercera generación o renacimiento de la lingüística de corpus (a partir de los años 80). Las etapas de este desarrollo inicial vienen marcadas, en cualquier caso, por la irrupción de la informática y los avances en tecnología sucedidos a partir de este momento (Berber Sardinha, 2004, McEnery, Xiao y Tono, 2006, Bolaños, 2015, Rojo, 2016a, Hincapié y Bernal, 2018).

2.2.1. Primera generación

La primera generación, atendiendo a la periodización de Villayandre (2008), tuvo su base en la lingüística estructural americana de la primera mitad del siglo XX, en donde se estableció una lingüística de corpus basada en métodos empíricos, aunque aún no se la denomina *lingüística de corpus*[3]. Sin embargo, en este primer tiempo, tras el posicionamiento y las críticas de Chomsky, se produjo cierto desprestigio en esta metodología basada en corpus a favor de la intuición lingüística. Como señalan McEnery, Xiao y Tono (2006), "in the late 1950s, however, the corpus methodology was so severely criticized that it became marginalized, if not totally abandoned, in large part because of the alleged 'skewedness' of corpora" (McEnery, Xiao y Tono, 2006, p. 3).

Estos primeros corpus se recogieron en papel; en ellos no se atendió a cuestiones de representatividad, dado a que el análisis debía realizarse de forma manual, lo que imposibilitaba el manejo de un número elevado de datos (Villayandre, 2008, p. 331).

2.2.2. Segunda generación

La segunda generación, situada en las décadas de los años 60 y 70, se caracterizó por el desarrollo tecnológico y la presencia de los ordenadores, hecho que favorece la aparición de los primeros corpus informatizados. Los medios que se emplearon fueron cada vez más potentes y ofrecieron una capacidad de procesamiento cada vez mayor, gracias a un almacenamiento masivo de coste relativamente bajo, lo que llevó a que la explotación de corpus de forma masiva se hiciera factible (McEnery, Xiao y Tono, 2006, p. 3). En este contexto surgieron corpus destacados como el *Survey of English Usage Corpus* (SEU), el *Brown University Corpus of American English* (Brown Corpus), el *Lancaster-Oslo/Bergen Corpus* (LOB) o el *London-Lund Corpus of Spoken English* (LLC).

El primero de estos corpus, el SEU, fue fundado por Randolph Quirk a finales de los 60 en la University College London. Este corpus, según la información que ofrece en su página electrónica (UCL-SEU, en línea) contiene un millón de palabras recogidas del inglés británico escrito y hablado producido entre 1955 y 1985. Consta de 200 textos de 5 000 palabras cada uno. Los textos/discursos orales incluyen diálogos y monólogos, mientras que

[3] El término *lingüística de corpus* no aparece hasta principios de la década de 1980, con la publicación de la obra *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*, editada por Aarts and Meijs en 1984 (Leech, 1992, p. 105).

los textos escritos incluyen no solo material impreso y manuscrito, sino también ejemplos de inglés leído en voz alta, como en los informativos y los discursos guionizados. Aunque se compiló originalmente en papel, en forma de fichas, con anotaciones gramaticales detalladas, actualmente se encuentra informatizado y cada elemento léxico se ha etiquetado automáticamente según la clase de palabra.

Junto al SEU, creado para el estudio del inglés británico, aparece en esta etapa el *Brown Corpus*, primer corpus que nace para el inglés americano. Este corpus fue fundado por W. Nelson Francis and Henry Kučera, de la Brown University Providence. El corpus consta de 1 millón de palabras (500 muestras de más de 2 000 palabras cada una) de texto en prosa obtenidas de publicaciones impresas en Estados Unidos durante el año 1961, si bien fue revisado y ampliado en 1979. Como indica Villayandre (2008, p. 335), "es el primer corpus concebido de forma íntegra para ser informatizado".

Por su parte, el LOB se constituyó como un corpus coordinado entre la University of Lancaster, la University of Oslo, y el Norwegian Computing Centre for the Humanities, de Bergen, dirigido por Leech, Johansson y Hofland. Al igual que el corpus mencionado anteriormente, contiene 500 textos de unas 2 000 palabras extraídas del inglés británico en 1961. Para su elaboración se siguieron los mismos criterios de diseño que el *Brown Corpus* ya que su objetivo era ser su equivalente para el inglés británico (Villayandre, 2008, p. 335). Existen dos versiones principales, la versión original, compilada entre 1970 y 1978, y una versión con etiquetado *part of speech* (POS) de 1981-1986.

En último lugar, cabe destacar el LLC, creado por Svartvik en la Universidad de Lund en 1975. Su construcción derivó de dos proyectos, por un lado, del SEU mencionado arriba, por otro del *Survey of Spoken English* (SEE) iniciado por el propio Svartvik como proyecto paralelo (Svartvik, 1990, p. 11). Su objetivo inicial era poner a disposición, en formato electrónico, el material hablado transcrito en Londres de una muestra de 87 textos con un total de 435 000 palabras. Este corpus original se completó con los 13 discursos orales recogidos por el corpus SEU, utilizando el sistema establecido ya para el propio *London-Lund Corpus*.

En esta segunda generación en la historia de la lingüística de corpus hay una "tendencia a desfavorecer los datos orales por las dificultades técnicas y de transcripción. Predominan los corpus de textos escritos, aunque con notables excepciones" (Villayandre, 2008, p. 334).

2.2.3. Tercera generación

A partir de los años 80, podría decirse que hay un renacimiento de la lingüística de corpus en el sentido en el que la entendemos hoy en día, gracias a las reivindicaciones a favor del uso de corpus realizadas por autores como Leech (1992), como se señaló en el apartado anterior. A este auge contribuyó el desarrollo de la lingüística computacional de la mano del impulso de nuevas tecnologías, además de un cambio en la manera de abordar los estudios lingüísticos que optó por combinar el enfoque intuitivo con la observación empírica de los datos. En términos de Villayandre (2008),

> el uso de corpus no se concibe como incompatible con el recurso a los juicios del lingüista […], se reconoce que los corpus no se pueden analizar válidamente sin la intuición y facultad interpretativa del analista, que usa conocimientos de la lengua (como hablante nativo o no nativo competente) y conocimientos acerca del lenguaje (como lingüista). (Villayandre, 2008, p. 337)

Además, señala esta autora, se hacía patente la necesidad en estos momentos de "contar con vocabularios o diccionarios más extensos para ampliar la cobertura de los sistemas" y "de manejar frecuencias, estadísticas y cálculos de probabilidades para manipular cantidades cada vez más grandes de texto" (Villayandre, 2008, p. 336).

Los corpus surgidos a partir de este momento contaron con la posibilidad de conformarse como macrocorpus, en comparación con los trabajos mencionados arriba que como máximo alcanzaron a recoger un millón de palabras. En este sentido, podrían destacarse corpus como el *Bank of English* (COBUILD Corpus), el *British National Corpus* (BNC) o los corpus de referencia académicos CREA, CORDE y CORPES XXI o el *Corpus del Español* de Mark Davies, para el caso de la lengua española que se detallarán en el capítulo 3[4].

El *Bank of English* (COBUILD Corpus) se inició en 1991 de la mano de Sinclair en el marco del proyecto COBUILD. Está compuesto por textos escritos (75 %) procedentes de diversos medios y de datos procedentes del medio oral (25 %). Contiene 524 millones de palabras de inglés escrito y hablado, con un 70 % dedicado al inglés británico, un 20 % al estadounidense y un 10 % a otras variedades (Xiao, 2008, p. 394).

[4] Para un listado exhaustivo de los corpus más conocidos e influyentes a nivel internacional consúltese Xiao (2008). No obstante, cabe señalar la notablemente escasa representación en este trabajo de corpus de español.

Por su parte, el BNC supone una colección de 100 millones de palabras de muestras de lenguaje escrito y hablado procedente de diversas fuentes, diseñada para representar una amplia muestra representativa del inglés británico de finales del siglo XX, tanto hablado como escrito. Está gestionado por un consorcio liderado por la editorial Oxford University Press, del que forman parte también las grandes editoriales de diccionarios como Addison-Wesley Longman y Larousse Kingfisher Chambers; los centros de investigación académica de Oxford University Computing Services (OUCS), el University Centre for Computer Corpus Research on Language (UCREL) de la Universidad de Lancaster, y el Centro de Investigación e Innovación de la Biblioteca Británica, tal y como refieren en su página electrónica (BNC, en línea). Su diseño y construcción ha servido de referencia para otros corpus como el *American National Corpus*, el *Polish National Corpus* y el *Russian Reference Corpus* (Xiao, 2008, p. 385).

Más allá de la creación y aparición de nuevos corpus en esta tercera generación, como reflejo del buen momento por el que atraviesa la lingüística de corpus, cabe señalar también la existencia de diferentes asociaciones dedicadas a este ámbito, como por ejemplo la *American Association for Corpus Linguistics* (AACL) o la *Asociación Española de Lingüística de Corpus* (Aelinco), así como la creación de centros de investigación especializados como el *Centre for Corpus Linguistics* (Universidad de Portsmouth), el *Centre for Corpus Research* de la Universidad de Birmingham o el *University Centre for Computer Corpus Research on Language* de la Universidad de Lancaster, tal y como recogen Hincapié y Bernal (2018, p. 39). Además, cuenta con un respaldo de revistas científicas especializadas tales como *Research in Corpus Linguistics* (RiCL), *CHIMERA Romance Corpora and Linguistic Studies*, *International Journal of Corpus Linguistics*, *Corpus Linguistics & Linguistics theory* o *Corpus Pragmatics*. En el caso español, contamos, además, con la Red Estratégica INTELE (CLARIAH-ES) financiada por el Ministerio de Ciencia, Innovación y Universidades (RED2022-134527-E), por la que España entra a formar parte oficialmente de las actuales infraestructuras europeas de investigación para el desarrollo de las humanidades y las ciencias sociales CLARIN y DARIAH. El consorcio CLARIAH-ES está constituido por centros de investigación nacionales líderes en las áreas de las ciencias sociales, las artes y las humanidades, así como en áreas específicas tales como la lingüística, la inteligencia artificial, las tecnologías del lenguaje, la informática y la computación de alto rendimiento, entre otras.

Estos son solo algunos ejemplos de corpus surgidos una vez delimitados los límites de la lingüística de corpus y sus posibilidades gracias al desarrollo de las tecnologías. Tal auge y desarrollo de la disciplina ha llevado a la consolidación y al crecimiento del número de corpus existentes con planteamientos, metodologías, características y finalidades diversas, como veremos en el siguiente capítulo centrándonos en el panorama de corpus para el español.

2.3. Tipología de corpus lingüísticos

Como se verá en el capítulo 4, el eje de partida que guía el diseño y la construcción de un corpus es el objetivo que este persigue. Esta finalidad será la que condicione su configuración y adscripción a un tipo u otro de corpus, buscando, en todo caso, la tipología más adecuada a sus intereses. Como señala Rojo (2021, p. 24), "la clasificación de los corpus no admite una organización jerarquizada, sino que tiene que hacerse atendiendo a diferentes perspectivas".

A continuación, en la Tabla 1 se presentan los principales bloques clasificatorios recogidos por la literatura (Torruella y Llisterri, 1999, Berber Sardinha, 2004, Baker, 2006, McEnery, Xiao y Tono, 2006, Lüdeling y Kytö, 2008, Martín Herrero, 2009, Sierra, 2017, Villayandre, 2018, Hincapié y Bernal, 2018, Rojo, 2021).

MEDIO	Escrito Oral Multimodal	
PERIODO TEMPORAL	Diacrónico Sincrónico	
NÚMERO DE LENGUAS	Monolingüe	monodialectal, multidialectal o panhispánico
	Multilingüe Bilingüe	paralelo
ESPECIFICIDAD DE LOS TEXTOS	Generales o de referencia Especializados o técnicos De entrenamiento	
TAMAÑO	Microcorpus Macrocorpus Abierto o monitor Cerrado	

INFORMACIÓN ADICIONAL	Simples o puros Anotados

Tabla 1. Tipología de corpus

Un primer bloque de caracterización y clasificación de los corpus es el referido al medio o canal, en el que se sitúan **los corpus textuales, los corpus orales y los corpus multimodales**. En el primer caso, los materiales proceden del medio escrito, ya sean textos completos o muestrales, y puede tener numerosas fuentes, como por ejemplo prensa, libro, revistas, internet, etc. Es, por ejemplo, el caso del corpus METAPRES (dirigido por Carmen Marimón) formado por columnas de prensa sobre la lengua publicadas en español en periódicos nacionales -en papel y digital- desde finales del siglo XIX hasta la actualidad.

Los corpus orales, por su parte, recogen lengua hablada de diferentes géneros discursivos y, además de para el análisis lingüístico general, pueden recogerse para el desarrollo de aplicaciones para la síntesis y el reconocimiento de voz (ver Capítulo 3). Un ejemplo de corpus oral del inglés es el *Santa Barbara Corpus of Spoken American English* (SBCSAE) (Dubois *et al*. 2000-2005). Este corpus recoge grabaciones de interacciones orales naturales procedentes de diversos lugares de Estados Unidos de diferentes géneros discursivos y contextos de interacción tal y como se indica en su portal electrónico:

> the predominant form of language use represented is face-to-face conversation, but the corpus also documents many other ways that that people use language in their everyday lives: telephone conversations, card games, food preparation, on-the-job talk, classroom lectures, sermons, story-telling, town hall meetings, tour-guide spiels, and more. (En línea, https://www.linguistics.ucsb.edu/research/santa-barbara-corpus#Citing)

Por último, los corpus multimodales se caracterizan por combinar texto, audio e imagen y su recogida, al igual que los corpus orales, presenta un mayor grado de complejidad en cuanto a su diseño y construcción que los corpus escritos como se detallará en el capítulo 4. Un ejemplo de este tipo de corpus es el *Corpus Español Multimodal de Actos de Habla* (COREMAH) (Vaca Matos, 2017) compuesto por 108 vídeos de hablantes nativos y estudiantes de español y sus correspondientes transcripciones, con un total de 18 737 palabras.

Un segundo bloque de clasificación de los corpus es el referido al periodo temporal que pretenden abarcar los datos recogidos, esto es, **corpus sincrónicos o corpus diacrónicos**. Los corpus sincrónicos están compuestos por textos recopilados en un periodo de tiempo específico y son útiles para observar cambios y tendencias en el lenguaje en un momento dado. Los corpus diacrónicos, por su parte, recogen textos a lo largo de un periodo extenso de tiempo, lo que permite el estudio de la lengua en su dinamismo y evolución. De orientación diacrónica es, por ejemplo, el *Corpus histórico del español de México* (CHEM), mientras que el *Corpus de referencia del español actual* (CREA) es un caso de corpus de orientación sincrónica.

También los corpus se pueden clasificar de acuerdo con el número de lenguas que se recogen en sus instancias; así, un **corpus monolingüe** es aquel que solamente recoge una lengua, a diferencia de los **bilingües** que recogen dos y los **multilingües** que recogen más de dos. En esta línea cabe añadir también los **corpus paralelos**, entendidos como corpus que contienen más de una lengua. En este tipo de corpus, a diferencia de los simplemente multilingües, los textos aparecen alineados de forma paralela de manera que se pueden realizar comparaciones entre ellas. Estos corpus son útiles para la traducción automática y el estudio comparado de las relaciones entre lenguas. Un caso que ejemplificaría este tipo de corpus es el *Corpus paralelo alemán-español PaGeS* (Doval, 2017), que además de paralelo es bilingüe, frente a un corpus monolingüe, como el CORPES XXI del español.

Así mismo, dentro de los corpus monolingües, podría hablarse también de corpus que recogen una única variedad dialectal (**monodialectales**), diversas variedades dialectales (**multidialectal**) o que tienen un carácter **panhispánico**, esto es, aspiran a recoger muestras de las variedades dialectales más significativas de una lengua. Sería el caso del corpus Ameresco y del corpus PRESEEA, ambos para el español.

Los corpus pueden ser clasificados también de acuerdo con la especificidad o generalidad de los textos o discursos que se recogen. De esta manera, encontramos los **corpus generales o de referencia** que se constituyen como trabajos de grandes dimensiones cuyo objetivo es proporcionar una visión completa de la lengua. Suelen ser corpus que cuentan con un respaldo institucional, como por ejemplo de las diferentes academias de la lengua, o sirven como base para la escritura de diccionarios y gramáticas. Los **corpus especializados o técnicos** son corpus compuestos por textos de uso específico. Es el caso, por ejemplo, de un corpus de traducción médica inglés-español como Cardiocor (Escarrá y Díaz, 2011), un

corpus comparable, sincrónico y especializado sobre aspectos relacionados con la salud y la enfermedad cardiovascular (Escarrá y Díaz, 2011, p. 134).

Por su parte, un **corpus de entrenamiento** es aquel que "se construye para que las aplicaciones de anotación y lematización adquieran los datos necesarios para poder proceder luego al procesamiento automático de grandes cantidades de textos" (Rojo, 2021, p. 25). El corpus ESLORA (2022), por ejemplo, ha desarrollado su propio corpus de entrenamiento para el desarrollo de su etiquetador gramatical.

Otro de los bloques clasificatorios para la caracterización de los corpus es el tamaño con el que se conciben, así como su posibilidad o no de ser ampliado. A este respecto, se suele diferenciar entre **macrocorpus y microcorpus** según las dimensiones que alcanzan. No se ha establecido una frontera clara en la bibliografía sobre cuándo se considera macro y cuándo micro, sin embargo, autores como Briz (2012a, pp. 116-117) señalan como elemento diferenciador entre unos y otros la posibilidad de contrastar los resultados entre distintas normas regionales.

Relacionado también con la cantidad de material que recoge, podríamos hablar de **corpus abiertos o monitor**; son aquellos que van incorporando nuevos materiales a medida que pasa el tiempo, pues se conciben como instrumentos de actualización de lengua. Por el contrario, un **corpus cerrado** responde a una configuración previa y, una vez se ha alcanzado la muestra prevista, se da por concluido.

Se constituye en macrocorpus el corpus de entrevistas sociolingüísticas PRESEEA que, además, tiene carácter abierto en cuanto a que se siguen recopilando y actualizando materiales. En comparación, el *Corpus Oral Juvenil del Español de Mallorca* (COJEM), de conversaciones espontáneas, es un microcorpus cerrado ya que la recogida de los datos está completa y responde a una muestra pequeña de informantes.

Se puede considerar también como factor de caracterización de los corpus el tipo de procesamiento de las muestras que se lleva a cabo, esto es, si son corpus simples o corpus anotados. Se entienden por corpus anotados, aquellos que añaden una capa de información extra para el análisis lingüístico automatizado (§ 4.2.2.5.), frente a los corpus simples o puros que no ofrecen dicha información y se comparten tal cual, en su puro traslado lingüístico. El corpus MeSA (Fuentes, 2021), de fuentes digitales, es un corpus puro, puesto que ofrece sus materiales sin marcas, más allá de unos signos de transcripción mínimos, mientras que el

corpus ESLORA (2022), de conversaciones y entrevistas, es un ejemplo de corpus anotado morfosintácticamente.

Además, podríamos hablar de otros criterios clasificatorios, por ejemplo, del derivado de usar la web como corpus. En este sentido, se situarían los corpus que Rojo (2021, p. 25) denomina **corpus oportunistas**, aquellos que están "formados mediante la detección y descarga de textos escritos en una lengua determinada que están en la parte pública de la red y que son integrados en un corpus". Este es el caso, por ejemplo, del *Corpus del Español* de Mark Davies (CE), que recoge materiales de blogs, páginas de periódicos, foros, etc. Podría considerarse también la clasificación en cuanto a la lengua utilizada por los hablantes, es decir, si son textos producidos por hablantes nativos o si han sido elaborados por estudiantes de una segunda lengua, **corpus de aprendices**, como es el caso del *Corpus de Aprendices de Español* (CAES) recogido por el Instituto Cervantes.

Según Hincapié y Bernal (2018),

> un corpus no responde a un único criterio […] sino que responde a una característica por criterio, es decir, un corpus puede ser oral, monolingüe, anotado, etc. De esta manera los objetivos de creación se sustentan los unos en los otros y el corpus resultante termina abarcando y definiendo más la variedad o lengua que representa. (Hincapié y Bernal, 2018, p. 34)

Así, teniendo en cuenta los diferentes bloques clasificatorios que se han revisado, si tomamos, por ejemplo, el corpus *Santa Barbara Corpus of Spoken American English*, se puede caracterizar con los siguientes atributos: es un corpus oral, sincrónico, monolingüe, general, de hablantes nativos, y con anotaciones de carácter discursivo. Tomando el ejemplo del *Corpus Diacrónico del Español* (CORDE), podemos definirlo con las siguientes características: es un corpus escrito, diacrónico, de referencia y de carácter panhispánico, de hablantes nativos, y al menos, en su versión de consulta superficial, es un corpus puro, no está anotado.

2.4. Los corpus y la lingüística en contexto. El enfoque pragmático

Con la información aportada hasta este momento, parece fácil prever, como han señalado McEnery, Xiao y Tono (2006, p. 4), que la creación y desarrollo de corpus ha llevado a una revolución en casi todas las ramas de la lingüística en cuanto a que los corpus se han erigido en una herramienta para el estudio y análisis fundamentado de hechos lingüísticos. Así, es

fácil reconocer su utilidad para estudios aplicados en las distintas manifestaciones de variación de una lengua. El análisis de datos lingüísticos realizado sobre diversos tipos de muestras aporta un factor fundamental para el estudio de la lengua real, el contexto de realización del habla. Disponer de muestras de lengua en contexto en sus diversas variedades permite afinar en cualquier subdisciplina lingüística. Se puede estudiar, por tanto, con veracidad y realismo, la variación fonética y fonológica, la variación morfosintáctica y léxica. Así mismo, se puede atender a otro tipo de variaciones incidentes en la lengua, pero externas a ellas, como la variación geográfica y social y la variación situacional. Este sentido, también, los corpus, además, permiten el desarrollo de trabajos sobre la enseñanza y el aprendizaje de una lengua, sobre discurso especializado y terminología, y sobre estudios pragmáticos, apartado en el que centraremos nuestra atención por ser este el objetivo particular de la construcción del corpus Ameresco, eje central de esta investigación.

A pesar del esfuerzo y del coste que supone la recolección de corpus orales, en comparación con el procesamiento de corpus escritos[5], no cabe duda de que el abanico de opciones de estudio que se abre ante el personal investigador se multiplica. En este sentido, trabajar con material oral, sea cual sea su género discursivo, posibilita los estudios pragmáticos (dentro de los que cabe destacar, la posibilidad del estudio de la prosodia) en cuanto que se puede trabajar con la lengua emitida en su contexto natural, el medio oral. En particular, la recolección de corpus orales de conversación coloquial espontánea ha favorecido la investigación y el análisis de la conversación, ya que, según Albelda (2022, pp. 225-226), las mejores posibilidades contextuales para el estudio pragmático las ofrecen los géneros más conversacionales, puesto que justamente la conversación constituye el modo de comunicación más pragmático y libre en cuanto a conducta interaccional (Givon, 1979, Briz, 2001, Briz y García Ramón, 2020).

Como señala Ajmer (2018, p. 557), muchos fenómenos orales, como, por ejemplo, los marcadores del discurso, las interjecciones, los vocativos o los marcadores de vacilación, habían desafiado el análisis gramatical debido a sus propiedades formales y funcionales. Sin embargo, con la aparición y desarrollo de corpus orales con contexto y contextualizados, el estudio y caracterización sobre estos fenómenos se ha visto ampliado, y se ha podido profundizar en su verdadero comportamiento funcional. Así mismo, los corpus orales

[5] Para la recogida de corpus orales se necesita recoger el material oral, transcribirlo y procesarlo para su posterior tratamiento informático, mientras que en un corpus de material escrito el proceso se simplifica (§ 4.1.).

conversacionales permiten el análisis de este género discursivo en sí mismo, en cuanto a su caracterización más prototípica. Este ha sido el punto de partida y eje de los estudios del Análisis de la Conversación, que han supuesto un desarrollo exponencial de las distintas facetas que se ven implicadas en la interacción, tanto de un punto de vista estructural como desde un punto de vista sociológico. El estudio sobre la configuración de los turnos, las diversas estructuras de la conversación, las unidades del discurso, los fenómenos naturales de habla como los solapamientos, los reinicios, las interrupciones, etc., permite no solo conocer cómo es de libre la gestión interaccional entre los participantes de una conversación (Cestero, 1994, Gallardo, 1996, Cortés, 2002, Briz y Grupo Val.Es.Co., 2014, entre muchos otros), sino también, a partir de ella, reconocer cómo se gestionan otro tipo de géneros discursivos (Briz 2010a, 2010b, 2012b, Briz y García Ramón, 2020), así como identificar cómo son las relaciones interpersonales que subyacen a tal uso de la gestión conversacional (Sacks, Schegloff y Jefferson, 1974, Sacks, 1992, Goffman, 1983, Schegloff, 2007, Heritage, 2013, Walsh, 2013, Espinosa y García Ramón, 2019, entre otros).

Cabe señalar, como ya se ha visto, y como recuerda Albelda (2022, p. 227), que, para el estudio de fenómenos pragmáticos, además de corpus naturales de lengua hablada, se han utilizado otros métodos, como los tests de hábitos sociales, los *role-plays*, las encuestas, así como los cuestionarios del *discourse completion task* (Jucker, Schneider y Bublitz, 2018). Ahora bien, "según los intereses específicos de investigación, podría resultar suficiente solicitar las intuiciones o hábitos lingüísticos de los hablantes, o emplear corpus de concordancias". En cambio, los corpus de lengua hablada permiten estudiar los fenómenos pragmáticos de una forma más abarcadora que el resto de los métodos.

Los corpus discursivos de lengua hablada, en definitiva, permiten acceder a los discursos o textos en su totalidad, así como al contexto natural en que fueron realizados natural y espontáneamente, puesto que no fueron producidos originalmente para fines de investigación. Este tipo de corpus ofrece un material auténtico, de carácter puramente empírico, lo que los hace plenamente idóneos para el análisis pragmático. Vienen, además, acompañados de sus datos situacionales y sociolingüísticos. Así, los datos contextuales que acompañan o se deducen de los corpus orales permiten conocer los parámetros situacionales para la interpretación adecuada de los significados implícitos en el habla, las coordenadas que permiten detectar el frecuente desajuste entre formas lingüísticas y funciones comunicativas, el papel de los interlocutores en la comunicación, los valores sociales que rodean a los actos de habla, etc. (Albelda, 2022).

En este sentido, cabe mencionar el reciente desarrollo de una nueva disciplina que aúna los objetivos de la lingüística de corpus (más cuantitativos) y de la pragmática (en su inicio, más cualitativos), que durante mucho tiempo se habían considerado excluyentes (Romero Trillo, 2020; Rühlemann y Ajmer, 2015). A esta se la ha denominado *pragmática de corpus*, y permite el estudio cabal y con garantías de fenómenos pragmáticos que requieren en su análisis tanto de una aproximación cualitativa como cuantitativa. Es el caso del estudio de actos de habla, de estrategias pragmáticas de base retórica como la atenuación e intensificación, la cortesía verbal, los marcadores discursivos, la gestión interaccional, etc.

De acuerdo con los estudios sobre el tema, "las mejores posibilidades contextuales para el estudio pragmático las ofrecen los géneros más conversacionales", pues, como se ha visto, la conversación es el modo de comunicación más pragmático y libre en lo relativo a la conducta interaccional (Albelda, 2022). Así, para la investigación pragmática resulta de gran ayuda la información que aportan los contextos de los corpus.

2.5. Síntesis del capítulo

En este capítulo se ha abordado la definición de la lingüística de corpus, partiendo de la delimitación de los corpus como colecciones digitales de textos naturales, escritos u orales, que se utilizan como material empírico para el estudio del lenguaje. Podemos concluir que las tres características de los corpus más aceptadas por la bibliografía son: los corpus han sido producidos en situaciones reales, han sido recopilados de acuerdo con parámetros explícitos y están disponibles en algún tipo de formato electrónico apto para análisis computacionales. La literatura generalmente diferencia entre archivos de textos, bases de datos y corpus, donde los últimos siguen criterios específicos de representatividad y equilibrio de la muestra.

Se ha debatido vivamente si la lingüística de corpus es una disciplina independiente o, en cambio, se trata de una metodología aplicable a otras disciplinas lingüísticas. Inicialmente, el trabajo de corpus enfrentó críticas por una supuesta falta de rigor y fiabilidad, pero con el tiempo se ha consolidado como un enfoque científico válido y sólido, en primer lugar, por proporcionar datos lingüísticos basados en ejemplos reales y evitar la subjetividad de la introspección, y, en segundo lugar, por los notable avances que ha experimentado de la mano

del *software*, cada vez más preciso y sofisticado, que facilita la búsqueda, extracción y clasificación de los datos.

Respecto a la evolución de la lingüística de corpus, tomando la propuesta de Villayandre (2008), esta se divide en tres fases o generaciones, marcadas por los avances tecnológicos y los cambios en la metodología. Con la llegada de la tercera generación de corpus, a partir de los años 80, llegó el empuje definitivo para este enfoque.

También se ha visto en este capítulo cómo los datos de corpus pueden obtenerse de dos maneras principales: a partir de muestras de habla natural o mediante técnicas de elicitación por parte de los investigadores quienes, a la hora de explotar los datos, pueden acercarse con una perspectiva *corpus-based* (partiendo de hipótesis previas) o *corpus-driven* (sin preconcepciones teóricas). Los criterios de clasificación mediante los que se caracterizan los corpus –que pueden combinarse entre sí– incluyen el medio, el período temporal, el número de lenguas incluidas, la especificidad de los textos, el tamaño y la información adicional proporcionada.

Como se avanzaba en la introducción, el enfoque de este trabajo se centra en un subtipo concreto, el de los corpus orales. Los corpus orales han transformado la lingüística en la medida que han permitido el estudio de la variación del lenguaje en su contexto natural de producción, algo especialmente valioso para el estudio de la pragmática y la conversación. Gracias a ello, la pragmática de corpus surge como una disciplina que combina la lingüística de corpus con la pragmática para estudiar fenómenos interaccionales, discursivos, etc., de manera cuantitativa y cualitativa.

En el capítulo siguiente, se aborda con mayor profundidad la definición y las características diferenciales de los corpus orales, con especial atención a los corpus orales en español.

Capítulo 3

Los corpus orales en español

En este capítulo, se repasa brevemente la historia de los corpus orales en español, desde sus orígenes a la actualidad (§ 3.1); a continuación, se analizan los principales estudios recopilatorios que, hasta hoy y bajo distintas premisas, listan los corpus orales del español existentes (§ 3.2). En la siguiente sección (§ 3.3) se propone una visión general del conjunto de corpus orales disponibles en el ámbito hispánico con el doble objetivo de servir de herramienta para la comunidad científica y de propuesta recopilatoria actualizada; los corpus incluidos se han seleccionado criterios como el de su disponibilidad abierta y en línea. Por último, dedicaremos un apartado (§ 3.4) a sintetizar las principales ideas contenidas en este capítulo.

En las circunstancias actuales, con las facilidades técnicas y la accesibilidad de materiales en línea, todo intento recopilatorio de corpus nace condenado a la obsolescencia casi inmediata. No obstante, en las líneas siguientes se ofrece una descripción actualizada a fecha de 2023 del panorama de corpus orales del español.

3.1. Nacimiento y desarrollo de los corpus orales en español

Los orígenes de los corpus orales se remontan a la labor recopilatoria de la dialectología con los atlas lingüísticos (Moreno Fernández, 2005a, 2005b, Julià, 2021), que deja de lado las investigaciones basadas en la intuición a favor de trabajos de observación empírica de los datos, como hemos visto en el capítulo anterior. Según Rojo (2016a, 2021), la lingüística de corpus en español comenzó con retraso con respecto al ámbito anglosajón, si bien tuvo un desarrollo rápido e intenso.

Para construir el panorama de corpus orales del español hay que remontarse al proyecto del *Programa Interamericano de Lingüística y Enseñanza de Idiomas* (PILEI) en la década de los 60, impulsado por Lope Blanch. Tal proyecto nace con el objetivo de estudiar coordinadamente el habla de las grandes concentraciones urbanas de América (Samper, 2005, p. 105). Bajo este proyecto surgieron corpus de diferentes localizaciones de Hispanoamérica y España, y están representadas ciudades como Ciudad de México, Caracas, Santiago de Chile, Madrid, Sevilla, Bogotá, Buenos Aires, Lima, San Juan de Puerto Rico y La Paz (Briz y Albelda, 2009, p. 6). Sin embargo, dado el momento en el que se recogieron, estos corpus no están disponibles al público más que en algunos casos a través de publicaciones en papel, sin acceso a las transcripciones. Como han señalado Enghels, Vanderschueren y Bouzouita (2015, p. 153), aunque se han construido siguiendo criterios sociolingüísticos, la transcripción y transliteración es ortográfica y no es homogénea en

todos ellos. PILEI recoge principalmente entrevistas, si bien incluye otros materiales como cuestionarios léxicos.

Uno de los primeros corpus que proporciona acceso al audio en formato CD-ROM, es el *Macrocorpus de la Norma Lingüística Culta de las principales ciudades de España y América* (MC-NC) (Samper, Hernández y Troya, 1998). Este macrocorpus recoge una selección de entrevistas individuales del corpus PILEI, revisadas y agrupadas bajo una serie de criterios ya homogéneos y uniformes con respecto a su predecesor, e incorpora dos ciudades más, San José de Costa Rica y Las Palmas de Gran Canaria (Briz y Albelda, 2009, p. 6). En total, ofrece la transliteración de 84 horas de grabación que recogen las voces de 168 hablantes representativos del nivel culto de las ciudades mencionadas arriba (Samper, 2005, p. 107). El género discursivo que recoge es el de la entrevista semidirigida individual y está transcrito de manera ortográfica. Esta transliteración se caracteriza por seguir las reglas normativas aceptadas en el uso escrito del español (Samper, 2005, p. 111). Para una caracterización de los sistemas de transcripción habituales en los corpus orales, véase § 4.2.2.3.

Posteriormente, el MC-NC fue incorporado al *Corpus de Referencia del Español Actual* (CREA) de la Real Academia Española y al *Corpus del Español* de Mark Davies (CE).

Otro de los trabajos destacables en estos primeros años es el macrocorpus para el *Estudio Gramatical del Español Hablado en América* (EGREHA), que surge de un proyecto de investigación dirigido por Hernández Alonso. Contiene materiales procedentes del PILEI y del MC-NC y añade un nuevo conjunto de entrevistas sin transcripción. Según Solís (2018, p. 118), la parte que hay trasliterada no se encuentra publicada, solamente se pueden consultar los resultados de las investigaciones.

Los tres proyectos citados, PILEI, MC-NC y EGREHA, que han derivado en sus respectivos macrocorpus, se han considerado habitualmente tres de los hitos fundamentales en el devenir del desarrollo de los corpus orales del español (Briz y Albelda, 2009, Enghels, Vanderschueren y Bouzouita, 2015, Briz y Carcelén, 2019). Como se ha visto, su germen es de naturaleza dialectológica y sociolingüística. En este sentido, Moreno Fernández (2005a) explica que

> el desarrollo del estudio de la norma culta vino a coincidir en el tiempo con la irrupción de la Sociolingüística en el panorama de la lingüística hispánica. […] se fue haciendo evidente que las bondades del proyecto de la norma culta no cubrían todas las necesidades y exigencias de

> una investigación lingüística básica: se hacía imprescindible conocer también las hablas populares de las ciudades del mundo hispánico y abordar su estudio sociolingüístico, de acuerdo con las posibilidades que la nueva disciplina ponía en manos de los investigadores. (Moreno Fernández, 2005a, p. 123)

De este cruce del análisis de la lengua culta oral con la sociolingüística de orientación laboviana surgieron enseguida una gran cantidad de proyectos de recogida de materiales en un gran número de ciudades hispánicas (Rojo, 2016a, p. 291) (véase § 3.2). Sin duda, el gran proyecto que nace bajo esta premisa será el *Proyecto de Estudio Sociolingüístico del Español de España y América* (PRESEEA).

PRESEEA nace a finales de los 90 bajo la dirección de Moreno Fernández, con el objetivo de aunar en un mismo corpus muestras de habla de las principales ciudades de habla hispana. Este macrocorpus recoge entrevistas semidirigidas en las que participan exclusivamente dos personas, entrevistador/a e informante, y está construido bajo criterios de representatividad sociolingüística. Nace, además, con la finalidad de que sus materiales permitan el estudio de la lengua hablada desde diferentes perspectivas —dialectológicas, sociolingüísticas, fonéticas, gramaticales, de análisis de la conversación, de análisis del discurso, pragmáticas y etnolingüísticas— (Moreno Fernández, 2005a, p. 126).

En la actualidad, son más de 40[6] los grupos que trabajan en la construcción de este corpus, tanto en España, como en América; aunque es un proyecto abierto, por lo que se pretende seguir incorporando equipos y, por tanto, nuevas ciudades de recolección de materiales. Las labores de recogida del corpus y de incorporación de los materiales al motor de búsqueda se encuentran en diferentes estados según cada equipo de trabajo. Equipos como el de Cuenca (Ecuador) se incorporó al proyecto en 2021, y aún se encuentra en la fase de recogida de materiales, mientras otros, como Las Palmas de Gran Canaria, están completos. Solo 29 ciudades están disponibles hasta hoy al público por medio del motor de búsqueda del corpus, si bien, en PRESEEA se integra solamente una muestra representativa y equilibrada de los materiales recogidos de cada ciudad, algo que no ocurre en otros corpus que siguen en gran parte, su metodología de trabajo, como es el corpus Ameresco, que sí publica e incorpora al motor de búsqueda el total de los materiales recogidos como veremos más adelante.

[6] El listado completo de equipos puede consultarse en línea en el portal del proyecto https://preseea.uah.es/equipos.

Todos los equipos del proyecto PRESEEA siguen unas directrices metodológicas homogéneas y comunes. Estas entrevistas se presentan en formato de transcripción ortográfica enriquecida (§ 4.2.2.3.) e incluyen un marcado y etiquetado TEI en XML (§ 4.2.2.4.). La proyección para cada ciudad es de 72 entrevistas con una duración media de 30-40 minutos de entrevista.

Los primeros corpus de habla que aparecen en el ámbito hispánico vienen impulsados por iniciativas privadas, en concreto, por proyectos de investigación en el ámbito de la universidad[7]. Sin embargo, no existían a nivel institucional obras representativas hasta la aparición de los corpus académicos. Como corpus de referencia, cabe mencionar los corpus promovidos por decisión académica, el *Corpus de Referencia del Español Actual* (CREA) y el *Corpus del Español del Siglo XXI* (CORPES XXI), en su sección oral[8].

Como explica Rojo (2016b), la decisión de iniciar los trabajos de CREA nace en 1995 y solo tres años después, en 1998, se publicaría la primera versión. Rojo (2016b, p. 198) señala que, si bien este primer corpus académico surge tres décadas después del primer corpus oral de referencia en inglés, el *Brown Corpus* (iniciado en la década de los 60), durante esos años no se construyeron muchos corpus textuales en la lingüística universal, y los que surgieron no obtuvieron, en general, el alcance y volumen del CREA.

Las referencias fundamentales para la constitución del CREA fueron del inglés el *Lancaster-Oslo-Bergen* (LOB) dirigido por Francis y Kučera, el *Collins Birmingham University International Language Database* (COBUILD) liderado por Sinclair y el *British National Corpus* (BNC); en el ámbito hispánico, el *Corpus de Lovaina*, corpus escrito del español dirigido por de Kock, el corpus *ENTREVIS*, construido por Kjær Jensen (Universidad de Århus) que recoge 725 mil formas procedentes de entrevistas publicadas en las revistas *Tiempo* y *Cambio 16*; recogidos con propósitos lexicográficos *el Vox-Bibliograf* de Alvar Ezquerra, el *CUMBRE* de Aquilino Sánchez y el *Corpus del Español Mexicano Contemporáneo* (CEMC) dirigido por Lara. Otros proyectos europeos que sirvieron de base para la creación de CREA fueron el *Corpus Resources and Terminology Extraction* (CRATER), el proyecto *Network of European Reference Corpus* (NERC), el

[7] No obstante, desde el ámbito empresarial (educativo y editorial) surgieron otros corpus con fines específicos, como, por ejemplo, aquellos con fines lexicográficos como el corpus CUMBRE, propiedad de la editorial SGEL.

[8] Como es sabido, además de estos, contamos con el *Corpus Diacrónico del Español* (CORDE) que no se menciona puesto que tiene carácter diacrónico y está compuesto por documentos escritos exclusivamente.

PARallèle Oral en Langue Etrangère (PAROLE) y *Léxico informatizado del español* (LEXESP), y otros corpus de carácter general como el *Corpus de Referencia de la Lengua Española Contemporánea* (CORLEC), *el Corpus Lingüístico de Referencia de la Lengua Española en Argentina* y el *Corpus Lingüístico de Referencia de la Lengua Española en Chile*, dirigidos los tres por Marcos Marín (Rojo 2016a, Rojo 2016b).

Estos últimos mencionados, aunque recogen principalmente materiales escritos, ya conciben la presencia de lengua hablada. Así, por ejemplo, CREA recogió mayoritariamente material de medios escrito (un 90 %), junto con una porción de material oral (un 10 %). La última versión de los materiales del CREA, la 3.2, es de 2008, pero en 2021 se publicó la versión anotada (0.3) que permite realizar búsquedas por forma, lema y categoría gramatical.

Los materiales orales proceden de tres fuentes: en primer lugar, de convenios con otras instituciones como la cadena SER y RTVE; en segundo lugar, de internet; y, en tercer lugar, de la cesión de materiales procedentes de otros corpus orales (Sánchez, 2005, pp. 39-40). Si bien, aun siendo corpus grabados de origen oral, el CREA no permite el acceso a sus audios, pues algunos de estos corpus no cedieron a la RAE su material sonoro, y los que sí lo hicieron, no fueron incorporados a los motores de búsqueda del CREA, como sí se logró más tarde, con el último corpus de la Academia, el CORPES XXI. De acuerdo con Sánchez (2005), los corpus cedidos que se integraron en CREA fueron los siguientes:

(i) con cesión -aunque no acceso- de soporte de audio:

- *Análisis de la Conversación de Alcalá de Henares* (ACUAH), construido por Ana M.ª Cestero.
- *Corpus Oral de la Variedad Juvenil universitaria del español hablado en Alicante* (COVJA), dirigido por Dolores Azorín.
- *Corpus para el estudio del español hablado en Santiago de Compostela* (CSC).
- *Corpus Oral de Referencia del Español Contemporáneo* (UAM), dirigido por Francisco Marcos Marín.

(ii) sin cesión del soporte de audio:

- *Macrocorpus de la Norma Lingüística Culta de las principales ciudades del mundo hispánico* de la Asociación de Lingüística y Filología de la América Latina (ALFAL), dirigido por José Antonio Samper.

- *Estudio sociolingüístico de Caracas 1977* y *1978* (CARACAS-77 y CARACAS-78), coordinados por Paola Bentivoglio y Mercedes Sedano.
- *Corpus de Encuestas en Asunción de Paraguay* (CEAP).
- *Corpus sociolingüístico de Mérida-Venezuela* (CSMV) coordinado por Carmen Luisa Domínguez y Elsa Mora.

Si bien, como hemos mencionado arriba, Rojo (2016a, 2016b) señala que los trabajos de corpus para el español comenzaron tarde con respecto de otras lenguas como el inglés, Rojo (2016b) alude a que

> la visión dominante de la historia de la lingüística de corpus se refiere sistemáticamente a un periodo inicial muy difícil, en un contexto hostil dominado por la pujante y novedosa orientación chomskiana, lo cierto es que esa caracterización es válida solo para los Estados Unidos, mientras que en países como Inglaterra, Noruega, Suecia y, en menor medida, Francia, Alemania o Italia la lingüística de corpus tuvo en esa época un desarrollo creciente y progresivo desde sus arranques respectivos. (Rojo, 2016b, p.199)

Añade, no obstante, que ese retraso en los trabajos de recolección de corpus, refiriéndose a los corpus académicos, conllevó ciertas ventajas. En primer lugar, hubo una gran evolución en el ámbito de la tecnología informática, hecho que mejoró la capacidad y velocidad de procesamiento de los textos. En segundo lugar, fue muy beneficiosa la aparición de la *Text Encoding Initiative* (TEI) y el sistema estándar de codificación SGML, que establecía un modelo de codificación replicable a cualquier proyecto de corpus. Por último, el desarrollo de internet empezaba a permitir la consulta cómoda y sencilla de los corpus. Por tanto, los primeros trabajos académicos de corpus (con el CREA y el CORDE) nacieron en un contexto más favorable en cuanto al sistema de codificación, estructuración y recuperación de datos que propiciaría su desarrollo (Rojo, 2016b, pp. 200-201).

El *Corpus del Español del Siglo XXI* (CORPES XXI) continúa los trabajos iniciados por CREA como corpus de referencia. Podría decirse que este nuevo corpus académico surge con propósito de enmienda, es decir, partiendo de la base de los defectos que tenía su predecesor, en CORPES se intentan solventar y ofrecer mejoras en determinados aspectos. Es lo que ocurre, por ejemplo, con las cuestiones relacionadas con la representatividad geolingüística, ya que se pasa de recopilar materiales procedentes de España y América en un porcentaje del 50 % para cada continente en CREA, a un cambio sustancial en CORPES

donde para España se contempla un 30 % y para América un 70 %. Respecto de la procedencia de los textos, esto es, si se obtienen de medio escrito u oral, los porcentajes no han cambiado, se mantiene en 90 % para material escrito y 10 % para material oral. Sin embargo, sí que se han enriquecido las búsquedas ya que, a diferencia de CREA, en CORPES puede accederse al registro sonoro de la forma buscada en la mayoría de los casos.

En 2007 la RAE y ASALE toman la decisión de emprender esta nueva tarea y en 2013 se publica la primera versión beta (0.6) del CORPES XXI. En 2023 disponemos de la versión 1.0, que contiene más de 395 millones de formas en total. Con respecto al anterior CREA, también presenta la novedad de que se puede recuperar el sonido alineado con textos orales y la consulta por categoría gramatical, ya que ha sido codificado en XML.

Los materiales orales que componen este corpus proceden de convenios con radio y televisión, de transcripciones y codificaciones de vídeos extraídos de Youtube y de la adaptación de textos procedentes de otros corpus orales, como son el corpus CORALES, construido por la RAE al final de la etapa de CREA con textos orales de todos los países hispánicos, materiales que fueron producidos entre 2001 y 2004; y otros procedentes de PRESEEA (Rojo, 2016b, p. 210).

A modo de recapitulación, en la Figura 2 pueden observarse de manera esquemática los corpus existentes hasta este momento, con indicaciones de las diferentes agrupaciones que se han realizado al integrarse algunos de estos corpus en otros proyectos de mayor tamaño.

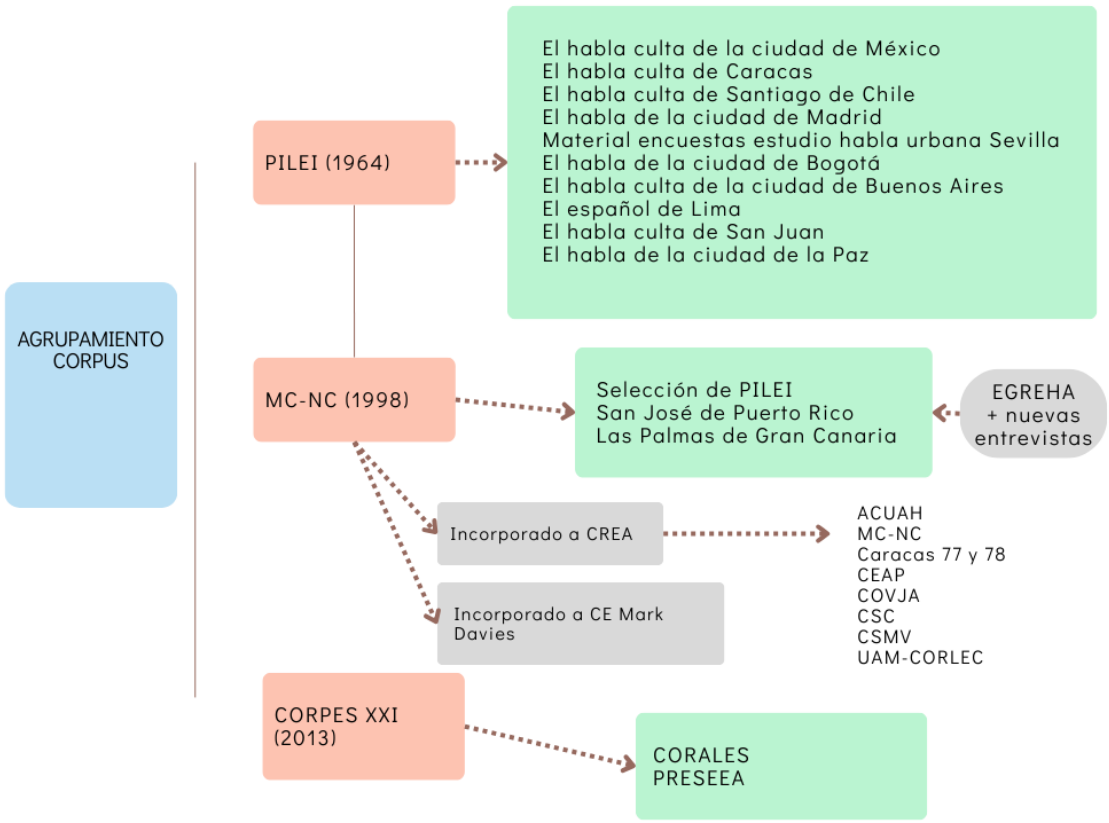


Figura 2. Recapitulación de primeros corpus orales del español y sus agrupamientos

3.2. Panorama de corpus orales en español

Podemos considerar que hoy en día el español es una de las lenguas con mayor desarrollo de corpus orales. En esta sección se da cuenta del panorama actual de corpus orales del español con el objetivo de determinar cuáles son los corpus más relevantes según la bibliografía especializada y cuáles son sus principales características.

Si bien existen plataformas en línea que reúnen información sobre corpus existentes a modo de indexador como, por ejemplo, para el caso de corpus de aprendices de español (Díaz, 2018, en línea) o para los corpus históricos iberorrománicos (Torruella y Kabatek, 2018, en línea), no existe por el momento ninguna plataforma que detalle el panorama de corpus orales del español más allá de pequeños repositorios ofrecidos a nivel particular por

diferentes grupos de investigación[9], ni tampoco es posible la búsqueda combinada en diferentes corpus al mismo tiempo. Sí que encontramos, en cambio, bibliografía específica sobre estudios de corpus existentes, como detallamos a continuación. Estas investigaciones recogen con mayor o menor grado de detalle, entre otras variables, cuáles son estos corpus, qué información ofrecen o dónde se pueden consultar, partiendo siempre 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.

Las recopilaciones más relevantes de corpus orales del español hasta la actualidad son, hasta donde llega nuestro conocimiento, y por orden de aparición, los trabajos de Moreno (2005c), Briz y Albelda (2009), Enghels, Vanderschueren y Bouzouita (2015), Rojo (2016a), Solís (2018), Parodi y Burdiles (2019), Briz y Carcelén (2019) y Llisterri (2021). No obstante, para el análisis pormenorizado de los corpus orales que mostramos a continuación, no se ha seguido un parámetro cronológico, sino que se comentarán atendiendo a su objetivo inicial, esto es, si son trabajos concebidos para mostrar un panorama de corpus orales del español, en un sentido abarcador y general, o si bien se han recopilado atendiendo a un criterio de selección concreto y específico y, por tanto, más restringido.

En el primer bloque analizaremos las recopilaciones realizadas por Moreno (2005c), Briz y Albelda (2009), Enghels, Vanderschueren y Bouzouita (2015) y Briz y Carcelén (2019); en el segundo bloque estudiaremos los trabajos de Solís (2018), Parodi y Burdiles (2019) y Llisterri (2021). Entre ambos bloques, situaremos el estudio de Rojo (2016a) cuya inclusión, si bien no constituye *per se* una panorámica, resulta relevante porque recoge desde una perspectiva histórica la aparición de corpus orales en español más significativos. Téngase en cuenta que estos recopilatorios muestran una visión del panorama en el momento en el que se publicaron y, por tanto, no pueden recoger todos los corpus existentes hasta el momento.

[9] La Universidad de Gante 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, entre ellas, del español, que permite filtrar por idioma o tipo de corpus, entre otros (accesible en https://www.corpusfinder.ugent.be/corpora#edit-language-collapsible--2--content).

Más allá de esto, lo que encontramos son pequeños recopilatorios ofrecidos por diferentes grupos de investigación en lingüística en sus respectivas páginas web. Es el caso de los proyectos Ameresco y Val.Es.Co. de la Universitat de València, el Instituto de Lingüística Aplicada de la Universidad de Cádiz, el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid, el Instituto Caro y Cuervo o la página web personal del lingüista Manuel Alcántara Plà, entre otros.

3.2.1. Recopilatorios generales de corpus orales en español

La primera recopilación de carácter general que encontramos es la de Moreno Fernández (2005c). Este artículo bebe de la sociolingüística, de ahí que se destaquen estas variables para su objetivo de presentar los corpus orales más importantes desde la década de los 90. En este trabajo, se presta especial atención a los corpus para el desarrollo de tecnologías del habla y a los corpus que tienen como objetivo el estudio de la propia lengua, poniendo atención especial en los corpus que ofrecen variedades geolingüísticas del español (Moreno Fernández, 2005c, p. 2), es decir, los primeros estarían enfocados en la lingüística aplicada, mientras que los segundos se centrarían en el componente puramente lingüístico sobre la propia estructura de la lengua. Además, reflexiona sobre uno de los aspectos que considera más difíciles a la hora de obtener este tipo de muestras, la representatividad de los materiales recogidos. Para lo cual, en este artículo se analiza si los criterios utilizados para la selección de hablantes son adecuados o no para el estudio de la lengua hablada y sus aplicaciones.

Este autor señala críticamente que, en los corpus que recoge en su trabajo, si bien es posible distinguir tipos de representatividad según la variación en el uso de la lengua y la variación en el usuario, se puede apreciar que, mientras la primera ha sido tratada de manera correcta en la mayoría de las circunstancias, la segunda ha recibido una atención desigual (Moreno Fernández, 2005c, p. 15).

Según Moreno Fernández, es importante que los trabajos de recolección de corpus se ajusten estrictamente a una tipología establecida de forma previa, atendiendo al *registro*, es decir, a la variedad de lengua según su uso. Sin embargo, para el desarrollo de tecnologías del habla, la representatividad no es un problema porque se trata de textos producidos específicamente para la elaboración del corpus y no tienen significado fuera de él. No obstante, además de la variedad según el uso, hay que prestar atención a las variedades de la lengua según el usuario, denominado *dialectos* según Halliday (Moreno Fernández, 2005c, p. 6). No conviene elaborar o trabajar con corpus cuyo tipo de hablante no se conozca bien, o cuyo origen no esté bien diseñado en función de los cuatro parámetros que determinan la variación lingüística: tiempo, espacio, sociedad y situación. De acuerdo con este autor, “if suitable care is not taken with these criteria, the oral *corpora* achieved would not be truly representative of the spoken language, though it would be valid for application to certain scopes” (Moreno Fernández, 2005c, p. 6).

En cuanto a la tipología que establece Moreno Fernández para su análisis, encontramos que distingue dos bloques: por un lado, los corpus que han sido creados para el desarrollo de tecnologías de reconocimiento del habla, a su vez, subdivididos en corpus generales y corpus especializados; por otro, los corpus creados para el estudio lingüístico de la lengua oral, que también responden a la subdivisión mencionada anteriormente. Cabe señalar que por *corpus generales* se entienden aquellos que no se centran en un nivel o aspecto específico de la lengua, mientras que los *especializados* sí atienden a un objetivo específico (Moreno Fernández, 2005c, p. 2). Además, incorpora un tercer bloque en el que se mencionan otros corpus, que o bien no están publicados o no son accesibles, o bien son de uso interno del grupo de investigación responsable.

Respecto de la clasificación realizada atendiendo a criterios de representatividad, destaca la muestra recogida según sean (a) corpus de referencia o (b) corpus de variedades del español y, a su vez, clasificados según la variedad dialectal que recogen, las variedades sociolingüísticas que se han tenido en cuenta para su recopilación y el género discursivo. Este segundo criterio, ya avanzando anteriormente, muestra la preocupación de los primeros estudiosos de corpus de lengua hablada por las variedades dialectales. El propio Moreno Fernández (2005c) explica que la variación dialectal, presenta una estrecha relación con las variaciones diastrática y diafásica, en el sentido propuesto por Halliday (1990).

En resumen, la propuesta de clasificación y recopilatorio de corpus orales del español de Moreno Fernández (2005c), es la siguiente, y se puede subagrupar en tres bloques:

1. Corpus orales para el desarrollo de tecnologías del habla
    a. Corpus orales generales para el desarrollo de tecnologías del habla
    b. Corpus orales especializados para el desarrollo de tecnologías del habla
2. Corpus orales para estudios lingüísticos
    a. Corpus orales generales para estudios lingüísticos
    b. Corpus orales especializados para estudios lingüísticos
3. Corpus de español con información geolingüística, sociolingüística y parámetros situacionales

En el caso de los corpus de tipo 1 y 2, se ofrece el nombre de cada corpus, la institución a la que está adscrito, la fecha de publicación o comienzo del corpus, si bien este último ítem no aparece en todos los casos. Por último, encontramos el enlace a la página electrónica de cada corpus que posee portal en línea, independientemente de que estén a disponibilidad

pública o no; sin embargo, sobre todo en los corpus descritos en el bloque 1, la mayoría de los enlaces han dejado de funcionar, algo lógico teniendo en cuenta que este trabajo se publicó en 2005.

En el caso del bloque 3, la información que se recoge es diferente de la mencionada en los anteriores. En este caso se proporciona el nombre del corpus, la variedad dialectal (ya sean ciudades concretas o diversas variedades), la variedad sociolingüística, es decir, si se ha aplicado algún filtro a la selección de hablantes como grado de instrucción, edad o varios a la vez; y, por último, parámetros situacionales referidos al material recogido (si es formal o informal, el género discursivo, entre otros). De los mencionados aquí, desarrolla cuatro corpus que tienen en común su carácter panhispánico, y cuatro más considerados corpus de referencia. Estas variables no son tan relevantes para los casos recogidos en el bloque 1 por tratarse de corpus específicos para el desarrollo de tecnologías del habla, en los que los criterios de representatividad no rigen la selección de la muestra de hablantes.

La segunda muestra recopilatoria de corpus de lengua hablada del español la constituye el trabajo de Briz y Albelda (2009), que surge por iniciativa del Instituto Cervantes como herramienta imprescindible para la comunidad investigadora del ámbito hispánico. Este trabajo tiene como objetivo dar cuenta de los corpus de lengua hablada y escrita en español, ya elaborados o en avanzada fase de elaboración (Briz y Albelda 2009, p. 1), al cierre de la primera década del siglo XXI. A diferencia del trabajo anterior, en este caso se recogen también corpus escritos, que dejaremos al margen de nuestro análisis; sin embargo, como señalan los propios autores,

> predominan los corpus de lengua hablada tomados en contextos naturales y de producción espontánea, en algunos casos más dirigida que en otros, donde la 'autoría' del habla se debe más bien a informantes anónimos, y donde el protagonismo del valor del corpus recae en el lingüista que selecciona unas muestras de habla o de escritura con un plan previo de representatividad, de objetivos y de selección de dichas muestras. (Briz y Albelda, 2009, p. 2)

Los criterios en los que han basado su selección son los siguientes: son corpus textuales o discursivos que (a) permiten acceso directo a los datos, aunque este sea mínimo[10]; (b) de cualquier variedad geográfica del español y cualquier nivel sociocultural; (c) que son sincrónicos, esto es, que se han recogido desde 1970 aproximadamente; (d) de acceso público, bien en formato impreso, bien en formato digital (incluyendo acceso en línea); y (e)

[10] Consideramos este criterio como fundamental para realizar nuestra clasificación posterior (§ 3.3).

preferentemente recogidos por grupos de trabajo y no por investigadores particulares o para tesis doctorales (Briz y Albelda, 2009, pp. 2-3).

Por último, los autores establecen distintos bloques para su clasificación. Se distingue el medio (si es escrito o hablado), el fin (general o concreto), el tamaño (macrocorpus o microcorpus) y el formato (textual o motor electrónico de búsqueda), así como a su modo de acceso, que distingue aquellos con acceso completo al texto o con acceso a través de concordancias. Al igual que el trabajo anterior de Moreno Fernández (2005c), incluyen otros tipos de corpus como los recogidos para el desarrollo de tecnologías de habla, de lenguajes técnicos y de adquisición y desarrollo del lenguaje.

Los bloques de información sobre la panorámica de corpus se presentan estructurados de la siguiente manera:

1. Corpus discursivos orales de acceso completo al texto
    a. De diferentes áreas geográficas agrupados en el mismo proyecto
    b. De áreas geográficas concretas con predominio de entrevistas
        · De España
        · De Hispanoamérica
    c. De conversaciones
    d. Con variedad de géneros discursivos orales
2. Corpus de acceso a través de concordancias
    a. Con fines generales
    b. Con fines específicos
3. Otros tipos
    a. De lenguajes técnicos
    b. De adquisición y desarrollo del lenguaje
    c. Para el reconocimiento del habla

La forma de presentar cada corpus varía a lo largo del trabajo. En el bloque 1 (corpus discursivos orales de acceso completo al texto) se ofrece la información más completa. Para

cada corpus incluido en este bloque, se proporciona la información general del material, así como un cuadro esquemático con los datos más relevantes, esto es, el nombre del corpus, su tamaño, el tipo de transcripción utilizada, los tipos de texto que contienen (incluyen aquí el género discursivo y las variables sociolingüísticas aplicadas), la variedad geográfica y los datos referidos a su publicación: en el caso de estar publicados en papel, se incluye la referencia, y en el caso de estar disponibles en línea, su URL. Se añaden, por último, otros datos sobre si se ofrece información contextual, si es posible el acceso al material y de qué manera, si dicho material ha sido sometido a algún tipo de tratamiento informático y, por último, información relativa a la financiación de cada proyecto.

Este recopilatorio también recoge corpus que no han sido publicados, al menos en el momento en el que se realizó este trabajo, como, por ejemplo, el *Corpus del Español Oral en Bilbao y su área metropolitana* (dirigido por Maitena Etxebarria) y el *Corpus del Español Oral en Bilbao* (dirigido por Maitena Etxebarria) que aún hoy se encuentran sin publicar, junto con otros en esa misma situación como el *Corpus de Puerto Cabello* (dirigido por Manuel Navarro) o el *Corpus de Valencia (Venezuela)* (dirigido por Manuel Navarro). Así mismo, se incluyen otros corpus que, en la fecha de publicación del citado trabajo, aún se encontraban en proceso de construcción, pero que a fecha de hoy ya se puede acceder a los materiales. Es el caso del *Corpus Sociolingüístico de la Ciudad de México* (CSCM) (dirigido por Pedro Martín Butragueño y Yolanda Lastra), cuyas transcripciones están disponibles en línea, o el *Corpus del Grupo de Investigación Lingüística Aplicada* (COGILA) (dirigido por Pedro Barros), que ha sido publicado en formato libro (tanto en papel como en formato de libro electrónico).

Por su parte, los corpus incluidos en los bloques 2 y 3, en cambio, no incorporan exactamente la misma información y el modo de exponerla que los corpus del bloque 1. En el bloque 2 se incluyen corpus de referencia que son considerados mayormente bancos de datos digitalizados (§ 2.1.), con acceso abierto en línea (Briz y Albelda, 2009, p. 32). Se presentan siguiendo un orden de mayor a menor volumen de número de palabras por corpus. En este punto, cabe mencionar que los autores de este recopilatorio denominan *corpus* a aquellas compilaciones de muestras de habla o de escritura recogidas en su contexto natural de enunciación; mientras que consideran *bases de datos textuales* los materiales extraídos de publicaciones (artículos científicos, periódicos, novelas, entre otros, incluyendo grabaciones de programas de televisión o radio, sesiones de juicios orales, etc.) agrupados de acuerdo con criterios homogéneos y que han sido digitalizados, es decir, sometidos a un

proceso de informatización para facilitar búsquedas o realizar cualquier tipo de análisis lingüístico siempre y cuando se hayan recogido siguiendo un proceso de selección de la muestra que sea representativo según sus objetivos (Briz y Albelda, 2009, pp. 1-2).

En cuanto a los corpus del bloque 3, se consideran un grupo aparte por ser corpus orientados a lenguajes de especialidad, adquisición de la lengua, o por recoger muestras producidas en entornos artificiales para el desarrollo de tecnologías de habla. Al ser corpus aplicados que no cumplen con las condiciones de los recogidos en el primer bloque, solo se da cuenta de algunos de ellos, ya que no es el objetivo fundamental de los autores (Briz y Albelda, 2009, p. 39).

El trabajo de Briz y Albelda, concebido, como se indicaba arriba, por iniciativa del Instituto Cervantes en 2009, ha sido tomado como referencia para otras panorámicas publicadas posteriormente, ya que es uno de los artículos recopilatorios más completos y detallados hasta el momento. La inclusión de los cuadros informativos de cada corpus facilita al personal investigador el acceso a todos los datos de una vez de manera clara y concisa y destaca, así mismo, la incorporación de remisiones internas que enlazan corpus que han sido englobados dentro de otros proyectos. Más abajo presentamos la continuación de este trabajo, con una actualización diez años más tarde (Briz y Carcelén, 2019).

Seis años después del recopilatorio del Instituto Cervantes, encontramos el panorama de corpus realizado por Enghels, Vanderschueren y Bouzouita (2015), cuyo objetivo es el de proporcionar una herramienta que permita determinar fácil y rápidamente los corpus disponibles del español europeo contemporáneo (Enghels, Vanderschueren y Bouzouita, 2015, p. 147). De nuevo, no se atenderá aquí a la selección de corpus recogidos por estas tres autoras en su trabajo que sean exclusivamente escritos.

Los criterios de selección de los corpus recogidos por estas autoras resultan en un conjunto de corpus (a) procedentes de proyectos finalizados o en gran medida disponibles al público; (b) de dimensiones razonablemente grandes o con una importancia particular; (c) de libre y fácil acceso, y (d) tanto corpus orales como bases de datos (Enghels, Vanderschueren y Bouzouita, 2015, p. 147). Los corpus orales recogidos pueden estar en soporte electrónico o en papel y atienden, además, a criterios geolingüísticos. Así pues, distinguen corpus panhispánicos, corpus de español peninsular y estos, a su vez, aparecen subclasificados según su variedad diatópica (variedad andaluza frente al resto de variedades

dialectales). A modo de recapitulación, los bloques de contenido quedarían dispuestos como se muestra a continuación:

1. Corpus del español europeo actual considerados bases de datos
    a. Sin finalidad de búsqueda particular
    b. Con finalidad de búsqueda específica
2. Corpus orales del español europeo actual
    a. Proyectos panhispánicos y peninsulares
    b. Representativos de determinades variedades peninsulares
        · Región andaluza
        · Otras regiones

Estas autoras parten de la diferenciación terminológica entre *corpus orales* y *bases de* datos establecida por Briz y Albelda (2009) y comentada anteriormente. En el primer bloque se recogen las bases de datos conformadas por grandes colecciones de textos, generalmente de diversos géneros que no suelen dar acceso al texto en su totalidad, sino a través de un soporte lógico de búsqueda en la red (Enghels, Vanderschueren y Bouzouita, 2015, p. 147). Se incluyen tanto las que no tienen una finalidad de búsqueda particular como las que han sido diseñadas para búsquedas de esquemas sintáctico-semánticos verbales predefinidos. En primer lugar, aparecen en forma de lista los datos más relevantes del corpus o base de datos descrito (URL, número de palabras, institución a la que está vinculado, variedad del español) y su estado actual, es decir, si está finalizado o no. Tras esta información esquemática, se detallan otras informaciones como la fecha en que se iniciaron los trabajos de recopilación, el género discursivo que recoge, y algunas características del funcionamiento del motor de búsqueda. En el segundo bloque, aparecen aquellos materiales considerados *corpus*, esto es, textos completos en soporte electrónico o en versión en papel (Enghels, Vanderschueren y Bouzouita, 2015, p. 147). La información está dispuesta de igual manera, y en el caso de PRESEEA incorporan un cuadro que recopila los diferentes grupos de trabajo en España, indicando la ciudad a la que pertenece, los datos referidos a la coordinación del subcorpus, su extensión y su publicación.

La labor de recopilación de un panorama de corpus es una tarea extensa y minuciosa que, además, en este caso, de antemano se encuentra condicionada por las limitaciones de espacio

que rigen la publicación, razón por la cual los datos ofrecidos en este último artículo que estamos presentando son menos abarcadores. Con este fin, Enghels, Vanderschueren y Bouzouita (2015) restringen la selección de corpus a aquellos que recogen español europeo, dejando fuera, en principio, las diferentes variedades hispanoamericanas. Sin embargo, en su selección aparecen representados corpus y bases de datos que sí que incluyen dichas variedades y, además, se ofrece la información de las variedades dialectales que recogen, como sucede en el caso de CORPES XXI, CREA, CE, COLA (*Corpus Oral de Lenguaje Adolescente*), PILEI y MC-NC. Se observa, así mismo, cierta vacilación terminológica en la denominación de los subapartados del bloque 1: sin finalidad de búsqueda particular y con finalidad de búsqueda específica. Según parece, *sin finalidad de búsqueda particular* hace alusión a aquellas bases de datos de referencia, construidas con propósitos generales, más que por finalidad de búsqueda, mientras que con *finalidad de búsqueda específica* se refiere a que su construcción responde a un objetivo concreto y particular, en este caso, el análisis sintáctico-semántico como se ha mencionado anteriormente. Por último, no se especifica a qué responde el orden de aparición de los distintos corpus. En el recopilatorio previo de Moreno Fernández (2005c) estos aparecen por orden alfabético, en Briz y Albelda (2009) se organizan internamente bien por su variedad dialectal, distinguiendo España e Hispanoamérica, bien por su tamaño, para el caso de otros corpus. En este último recopilatorio no se observa, ni se menciona, ningún criterio ordenador más allá de los diferentes bloques de contenido.

Por último, como se mencionaba anteriormente, diez años después de la panorámica realizada por Briz y Albelda (2009) para el Instituto Cervantes surge la actualización de Briz y Carcelén (2019), también en el *Anuario* del Instituto, cuyo objetivo es ofrecer una puesta al día sobre el estado de corpus orales del español (Briz y Carcelén, 2019, p. 193).

Los corpus recogidos se han clasificado según los siguientes criterios: por un lado, corpus de carácter panhispánico; por otro, corpus de variedades dialectales concretas, en ambos casos se puede acceder a ellos por medio de plataformas en línea con motor de búsqueda. Tras ello, se muestran otros corpus discursivos que ponen a disposición del usuario material transcrito sin motor de búsqueda. Por último, se ofrecen de manera esquemática todos los corpus orales de los que se ha tenido noticia hasta la fecha, ya tengan acceso abierto en línea o no, atendiendo a las siguientes variables: género discursivo, tipo de acceso, publicación y estado (finalizado o en desarrollo). Cabe mencionar que en este trabajo no se recopilan

corpus cuya finalidad sea el desarrollo de tecnologías de habla (Briz y Carcelén, 2019, p. 206).

En resumen, los bloques de contenido quedan dispuestos de la siguiente manera:

1. Corpus orales panhispánicos, entendidos como aquellos "que recogen materiales orales procedentes tanto de España como de Hispanoamérica" (Briz y Carcelén, 2019, p. 194).

2. Corpus de variedades geográficas concretas

3. Otros corpus discursivos

4. Esquema del panorama general de corpus orales del español

En el bloque 1, se ofrecen diversos cuadros informativos sobre cada uno de los corpus orales de carácter panhispánico recogidos. Estos muestran la descripción de cada uno a través de los conceptos (a) nombre, (b) clase de corpus, (c) objetivo por el que empezó a recopilarse, (d) tamaño en tiempo o en formas, (e) el tipo de transcripción y/o etiquetado, (f) el género discursivo que recoge, (g) la variedad geográfica, (h) si ofrece información contextual sobre el material que contiene, (i) si facilita el acceso a los audios y sus transcripciones y de qué manera, (j) a qué tipo de tratamiento informático ha sido sometido, (k) si ha sido publicado y dónde, (l) qué financiación ha recibido, (m) quién es el equipo o entidad responsable y (n) su estado actual. Para otro tipo de corpus construidos sin un objetivo concreto de investigación (Briz y Carcelén, 2019, p. 199), como son los corpus de referencia, solo aparece la información descriptiva en párrafo.

En el bloque 2, los corpus de variedades dialectales concretas siguen el mismo esquema de información recogido en el bloque anterior. No obstante, se recogen otros corpus discursivos en texto que ofrecen sus transcripciones en línea, pero que, en cambio, no facilitan acceso a los audios.

En el bloque 3 se recogen otros corpus discursivos que ofrecen sus trascripciones a modo de repositorio, sin acceso a los audios y sin herramientas para filtrar las búsquedas (Briz y Carcelén, 2019, p. 206).

Por último, aparecen de forma esquemática todos aquellos corpus orales del español localizados por los autores, bien por medio de otros trabajos publicados, bien por la búsqueda en profundidad y la colaboración de los autores de corpus que han facilitado la información (Briz y Carcelén, 2019, p. 194). A diferencia de las tablas descriptivas de los bloques

anteriores, en este listado la información que se muestra es menos exhaustiva y se centra en recoger el género discursivo, la presencia o no de acceso abierto en línea, la fecha de publicación y el estado (finalizado o no). Aparecen clasificados, a su vez, según sean corpus panhispánicos, de variedades dialectales concretas o de adquisición y desarrollo del lenguaje.

El valor de este trabajo radica en que amplía y actualiza la colección de corpus recopilados por Briz y Albelda (2009); ahora bien, la descripción de cada uno de ellos en los bloques 1 y 2 se ha restringido a aquellos corpus (a) con acceso directo al material en línea en el momento de publicación del trabajo o que lo tienen previsto en un futuro y (b) con acceso tanto al material oral como a las transcripciones. En el panorama general del tercer bloque aparecen listados todos ellos, con acceso abierto en línea o no. Esta recopilación añade a los corpus recogidos en Briz y Albelda (2009), también ofrecidos aquí, aquellos corpus surgidos en los diez años que han transcurrido entre ambas publicaciones. Sin embargo, el orden de aparición no responde a ningún criterio explicitado por los autores, ni alfabético ni cronológico.

3.2.2. Otros recopilatorios de corpus orales en español

Como hemos mencionado al principio de esta sección, para completar este apartado, analizaremos la publicación de otras obras que incluyen recopilaciones, pero no como propósito fundamental del trabajo, o bien con un alcance mucho más estrecho que las anteriores. La primera de ellas es la obra de Rojo (2016a) que, si bien no es en sí mismo un trabajo recopilatorio de corpus, al tratar sobre los orígenes y situación actual de los corpus del español, sí que aparecen mencionados los principales corpus del ámbito hispánico. Por esta razón no podemos hablar de criterios de clasificación ni tipología, como hemos venido haciendo. Sin embargo, nos ha parecido necesario incluirlo en este apartado por la visión panorámica que ofrece. De nuevo, para los propósitos de este trabajo, nos centraremos en aquellos que son totalmente orales.

El recorrido de Rojo (2016a) parte de los orígenes de la lingüística de corpus en español, remitiendo al predominio de los corpus de referencia, académicos o no, así como a los primeros macroproyectos que surgen de la mano de la dialectología y la sociolingüística como son el PILEI y PRESEEA. Tras este repaso al origen, destacan los primeros corpus que nacen entre 1980 y 1995, corpus de tamaño reducido resultado de proyectos individuales

o de grupos de investigación y otros corpus de tamaño pequeño construidos en el marco de proyectos europeos. Frente a los que acabamos de mencionar de carácter sincrónico, Rojo ofrece también unas líneas a corpus diacrónicos (Rojo, 2016a, pp. 286-289).

En último lugar, este trabajo dedica una sección a describir la situación actual de los corpus de español, incluyendo aquí los corpus más relevantes, ya sean corpus de referencia, corpus para estudios lingüísticos generales, de español como lengua extranjera, lenguaje infantil o para el desarrollo de aplicaciones de análisis y síntesis de voz (Rojo, 2016a, p. 289-292). Resulta muy útil el listado que incluye al finalizar con la relación de corpus y otros recursos electrónicos mencionados en el texto ya que, si bien aparecen referidos por orden alfabético, sin que haya ningún otro criterio clasificatorio, permite el acceso directo a las páginas electrónicas correspondientes.

Mostramos a continuación la descripción de otros trabajos organizados bajo un criterio de selección concreto y específico y, por tanto, más restringido que los mencionados en el apartado anterior. Desde el ámbito del análisis de la conversación, el trabajo de Solís (2018) tiene como objetivo señalar aquellos corpus que puedan usarse para este fin. Por tanto, el criterio elegido para su clasificación prioriza el género discursivo conversacional frente a otras variables, bien sean corpus de conversación espontánea, de debates, mesas redondas o tertulias en medios de comunicación (Solís, 2018, p. 118). No obstante, Solís recoge a modo de lista los principales corpus de entrevistas y conversación semidirigida existentes sin ofrecer más información sobre ellos.

Los bloques de contenido en el estudio de Solís (2018) se presentan en dos ejes:

1. Corpus conversacionales del español

2. Corpus conversacionales y análisis contrastivo de la conversación

En primer lugar, la autora recoge un bloque denominado *macrocorpus de conversaciones del español* que incluye, de manera listada, ejemplos de corpus orales de dicho género discursivo, así como otras colecciones de textos conversacionales para el desarrollo de tecnologías del habla que, como señala la autora, recogen diálogo entre personas y ordenadores. Este listado se acompaña de un breve párrafo descriptivo por cada corpus o colección mencionado, que incluye, de manera general, datos sobre quién lo dirige, qué zona geográfica recoge y el modo de consulta.

En segundo lugar, se agrupan los *corpus conversacionales y de análisis contrastivo de la conversación*, microcorpus, en este caso, paralelos, multilingües y monolingües que nacen en diferentes contextos como la enseñanza de la lengua o la traducción automática (Solís, 2018, p. 123), ya sean espontáneos o semiespontáneos. Al igual que en el bloque anterior, cada corpus incluye información general sobre sus características.

Tras este listado, se ofrece una tabla conjunta donde aparecen referidos los corpus de ambos bloques y en la que se detallan las siguientes características: si ofrecen las transcripciones, si se puede acceder al audio, si el audio y la transcripción se encuentran alineados y etiquetados, si se encuentran disponibles en línea gratuitamente o *en comercio*.[11]

Como hemos mencionado, se establece desde el principio que la atención del trabajo recae en corpus de conversaciones espontáneas, de debates, mesas redondas o tertulias en medios de comunicación. Solís justifica esta decisión porque, aunque las entrevistas puedan llegar a ser espontáneas, siempre habrá una persona que dirija el diálogo y un informante que ofrezca su testimonio lingüístico. Por tanto, de acuerdo con la autora, en ellas no sucede la misma dinámica comunicativa que en las conversaciones de tipo espontáneo (Solís, 2018, p. 118). Sin embargo, a lo dicho por la autora conviene objetar que el nivel de espontaneidad de una entrevista semidirigida, en muchas ocasiones, puede equipararse a la espontaneidad de una conversación informal, un debate, una mesa redonda o una tertulia televisiva (Briz, 2010, 2012b). De hecho, si tomamos como referencia las grabaciones que conforman el corpus PRESEEA, según cómo actúe la persona que dirija la conversación, encontramos casos donde hay mayor fluidez y espontaneidad, mucho más, al menos, que, si lo comparamos con un debate, género que se caracteriza por su grado de formalidad y unas respuestas ajustadas a unos tiempos marcados previamente. Por lo tanto, el criterio clasificatorio elegido no se encontraría totalmente justificado desde el punto de vista de la variación situacional.

El siguiente trabajo que analizamos el de Parodi y Burdiles (2019) sobre corpus y bases de datos. Su objetivo es ofrecer un análisis crítico del panorama de corpus y bases de datos disponibles para los procesos de enseñanza y aprendizaje de español como L2 (Parodi y Burdiles, 2019, p. 596). En cuanto a los criterios de selección de los corpus y bases de datos, los propios autores destacan que son (a) herramientas disponibles en línea; (b) de y para aprendientes de español, sea como L1 o L2; (c) declaran registrar usos del español; (d)

[11] Entendemos que por *comercio* la autora se refiere a 'comerciales', que pueden adquirirse por medio de su compra, sin embargo, no aparece especificado en el trabajo.

ofrecen muestras tanto orales como escritas; (e) presentan soporte tecnológico de diversa índole y grado de desarrollo, y (f) están en funcionamiento en la actualidad. Las variables elegidas para su clasificación son dos, atienden al modo (oral o escrito) y a la procedencia (L1 o L2).

Este trabajo recoge una primera tabla donde se listan los corpus disponibles en español según sean corpus orales de L1 o L2 o corpus escritos también de L1 y L2, pero no se ofrece más información que el nombre de cada uno de ellos. En cuerpo de texto se expone una breve descripción de algunos de ellos, pero no de todos y no con la misma información para cada caso. A modo de anexo, recoge corpus de español como L2 y L1 desarrollados por y para aprendices, incluyendo su nombre, su dirección electrónica, si es oral o escrito y si está anotado.

Este recopilatorio focaliza claramente su objetivo en la utilidad de los corpus para los procesos de enseñanza y aprendizaje del español. Sin embargo, es menos exhaustivo con las descripciones de los corpus que recoge y se centra más en la parte de teorización sobre los procesos de enseñanza-aprendizaje que sobre la propia construcción de los corpus.

En último lugar se encuentra la extensa recopilación propuesta por Llisterri (2021). Este autor realiza un acercamiento al panorama de corpus en español para el estudio del componente fónico en LE/L2. Entre sus objetivos destaca presentar las posibilidades y limitaciones de los corpus orales disponibles en línea, así como ofrecer las características esenciales de dichos corpus, especialmente referidas a su acceso. Este autor diferencia entre *corpus de lengua oral*, es decir, aquellos que ofrecen una transcripción ortográfica de las grabaciones originales y que, por tanto, considera *textos* orales (se estudian desde su forma escrita transcrita), de los *corpus orales* propiamente dichos, que son aquellos que incorporan la transcripción fonética del material sincronizada con la grabación y acompañada o no de algún tipo de anotación, en mayor o menor detalle (Llisterri, 2021, p. 165).

Como criterios clasificatorios encontramos que (a) considera únicamente los recursos en los que se puede trabajar con la señal sonora (corpus orales), bien de nativos, bien de aprendices; (b) aquellos recursos no comerciales, esto es, de uso gratuito, aunque se requiera la creación de una cuenta de usuario; y (c) el grado de utilidad de corpus orales de hablantes nativos y de los que recogen datos de aprendices para la investigación sobre el componente fónico.

En este trabajo, los bloques sobre corpus se presentan de la siguiente manera:

1. Corpus orales de hablantes nativos de español

2. Corpus orales de hablantes no nativos de español

En este caso, la información se organiza en torno a dos ejes: 1) corpus orales de hablantes nativos de español, de dominio público y con acceso a las grabaciones; y 2) corpus orales de estudiantes de español como LE/L2, de dominio público y de acceso a las grabaciones. En ambos bloques la estructura es la siguiente: primero, encontramos un cuadro en el que se registran los corpus de cada eje y su enlace a la página electrónica. Segundo, a cada uno de estos cuadros le siguen dos subtablas: una que describe las posibilidades de acceso a los datos y otra que detalla las posibilidades de búsqueda de dichos corpus.

La subtabla 1 registra por cada corpus información sobre el acceso a la transcripción ortográfica, a la anotación del corpus y a las grabaciones; y cada una de estas variables informan también de si se tiene acceso en pantalla o si se pueden descargar. La subtabla 2, por otro lado, ofrece información sobre las posibilidades de búsqueda según las variables de consulta a partir de los metadatos del corpus, búsqueda de fenómenos anotados, búsqueda en la transcripción ortográfica o concordancias. En el caso de los corpus de estudiantes de español no aparece la última variable (concordancias) posiblemente porque son corpus de menor tamaño con menos posibilidades de desarrollar un motor de búsqueda específico o porque no hay necesidad de restringir el contexto comunicativo como sí ocurre en corpus de referencia, debido a condicionantes por derechos de autoría de los materiales que han servido para su construcción (§ 4.2.1.1.).

La recopilación de este trabajo resulta muy útil a aquellos investigadores/as o usuarios/as de corpus que necesiten trabajar con material oral, ya que solamente da cuenta de los proyectos que recogen audio y a cuyos materiales se puede acceder en línea. Además, la información que se recoge va más allá de la que aparece en los recopilatorios comentados anteriormente, puesto que añade notas de carácter técnico de una manera detallada, en particular, sobre aspectos relacionados con la transcripción y la anotación, que en los trabajos previos solo se mencionaba en algunos casos de manera somera.

3.2.3. Revisión crítica de los recopilatorios sobre corpus orales del español

Con el objetivo de realizar una propuesta actualizada del panorama de corpus orales del español, que parta de las virtudes de los trabajos comentados anteriormente, pero supliendo

las principales deficiencias encontradas, ofrecemos a continuación un análisis pormenorizado del grado de consenso en la recolección de corpus orales, comparativa que ayuda a establecer cuáles son los corpus más importantes según la bibliografía.

3.2.3.1. Coincidencias y divergencias: los corpus orales más importantes del español según la bibliografía

Con la finalidad mencionada arriba, se han revisado y recopilado en una misma base de datos todos los corpus recogidos en los trabajos señalados anteriormente, Moreno Fernández (2005c), Briz y Albelda (2009), Enghels, Vanderschueren y Bouzouita (2015), Rojo (2016a), Solís (2018), Parodi y Burdiles (2019), Briz y Carcelén (2019) y Llisterri, (2021). A partir de aquí, nos proponemos establecer, por orden de consenso, cuáles considera la literatura que son los corpus más importantes del panorama hispánico.

Para tal fin, se ha clasificado el total de corpus compilados en cuatro bloques:

(a) corpus con fines lingüísticos generales;

(b) corpus para la enseñanza de español;

(c) corpus para el desarrollo de aplicaciones de tecnologías del habla, y

(d) otros corpus, donde se incluyen, por ejemplo, corpus para la adquisición de la lengua, corpus de habla infantil o de patologías del lenguaje.

Debe tenerse en cuenta que, como hemos visto detalladamente en la sección anterior, encontramos recopilatorios de carácter generalista, frente a otros centrados en recoger corpus que responden a determinadas especificidades, como el caso de Solís (2018), focalizado en corpus dialógicos, o Parodi y Burdiles (2019), cuyo centro de atención es la enseñanza y aprendizaje de español. Además, no conviene olvidar que estas panorámicas son un reflejo del momento en el que se escribieron y que, a ojos de una lectora o lector actual, pueden notarse algunas ausencias significativas.

En la Tabla 2 se da cuenta de aquellos corpus que tienen como mínimo el respaldo de cuatro de los trabajos revisados en el apartado anterior, esto es, aquellos en los que coinciden al menos la mitad de los autores. Los corpus recogidos corresponden al bloque denominado *corpus lingüísticos generales*. Para el resto de los bloques, es decir, tecnologías del habla, español como lengua extranjera y corpus con otros fines, el grado de consenso encontrado es de tres o menos, motivo por el cual no aparecen aquí mostrados.

En los puestos más altos de la tabla, en color naranja, se observan aquellos corpus que han sido recogidos mayoritariamente, que cuentan con siete u ocho coincidencias, de las ocho posibles; en segundo lugar, le siguen en amarillo aquellos que cuentan con seis concurrencias; en tercer lugar, en verde, se recogen los corpus mencionados por cinco autores y, por último, en azul, aquellos que son recogidos por la mitad de los recopilatorios descritos.[12]

CORPUS	MO	B&A	E, V & B	B&C	RO	SO	PA	LLIS
COLA	✗	✓	✓	✓	✓	✓	✓	✓
C-ORAL-ROM	✓	✓	✓	✓	✓	✓	✓	✗
CORLEC	✓	✓	✓	✓	✓	✓	✓	✗
PRESEEA	✓	✓	✓	✓	✓	✗	✓	✓
Val.Es.Co. (2002)	✓	✓	✓	✓	✓	✓	✓	✓
CE (Mark Davies)	✓	✓	✓	✓	✓	✓	✗	✗
CECBNA	✓	✓	✓	✓	✓	✓	✗	✗
CORDIAL	✗	✗	✓	✓	✓	✓	✓	✓
COSER	✗	✓	✓	✓	✓	✗	✓	✓
CREA	✓	✓	✓	✓	✓	✓	✗	✗
ALCORE	✓	✓	✗	✓	✓	✓	✗	✗
CLH de Almería	✓	✓	✓	✓	✗	✗	✗	✓
CORPES XXI	✗	✗	✓	✓	✓	✓	✓	✗
COVJUA	✓	✓	✗	✓	✓	✓	✗	✗
VUM	✓	✓	✓	✓	✓	✗	✗	✗
COGILA	✗	✓	✓	✓	✗	✓	✗	✗
CUMBRE	✓	✓	✗	✓	✓	✗	✗	✗
GRIAL	✗	✓	✗	✗	✓	✓	✓	✗
MC-NC	✗	✓	✓	✓	✓	✗	✗	✗
PILEI	✓	✓	✓	✗	✓	✗	✗	✗

Consenso en 7-8 artículos | Consenso en 6 artículos | Consenso en 5 artículos | Consenso en 4 artículos

Mo (Moreno Fernández, 2005c), B&A (Briz y Albelda, 2009), E, V & B (Enghels, Vandershueren y Bouzouita, 2015), B&C (Briz y Carcelén, 2019), RO (Rojo, 2016a), SO (Solís, 2018), P&B (Parodi y Burdiles, 2019), LLI (Llisterri, 2021)

Tabla 2. Consenso entre recopilatorios

[12] El desarrollo de los acrónimos de los corpus puede consultarse en el Anexo 1 y accediendo a los siguientes enlaces: https://nuvol.uv.es/owncloud/index.php/s/IiJ0bjHxYfGKoUc

Como mencionábamos arriba, primeramente, en color naranja, aparecen los corpus que han sido recopilados por unanimidad en la totalidad de trabajos, si bien cabe destacar que no hay ningún corpus que aparezca en todos los recopilatorios estudiados, lo cual nos da una idea de la dispersión y relativa falta de consenso entre las obras. Una excepción la constituiría, curiosamente, el corpus Val.Es.Co., si se considera que Fonocortesía, citado por Llisterri (2021), es en realidad un subcorpus de este. Cabe señalar que en los recopilatorios cuyo objetivo es ofrecer una panorámica general hay mayor grado de coincidencia entre ellos, mientras que en los trabajos con fines específicos hay mayor divergencia.

El *Corpus Oral de Lenguaje Adolescente* (COLA), de Annette Myre Jørgensen, si bien, no ha contado con un entramado de grupos de trabajo colaboradores como sí ocurre en otros proyectos, destaca por ser un corpus de lenguaje adolescente con representación de diversas variedades dialectales que se encuentra etiquetado y alineado y que cuenta con motor de búsqueda. El *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC) y el *Corpus Oral de las Lenguas Romances* (C-ORAL-ROM) han sido realizados por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid. El primero se incorporó a CREA y el segundo puede adquirirse en versión DVD, si bien hay una muestra disponible en línea.

No es sorprendente que el corpus PRESEEA se encuentre en esta franja ya que, a diferencia de otros corpus de carácter panhispánico construidos en el pasado, este no solo cuenta con una ingente infraestructura de equipos de trabajo detrás, sino que, además, facilita el acceso a los materiales en línea a través de un motor de búsqueda que permite el filtrado por criterios sociolingüísticos y lingüísticos.

Es significativo el alto grado de consenso con respecto al corpus Val.Es.Co. Cabe señalar que la versión mencionada en estos artículos se corresponde con el primer corpus Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002), que fue publicado en papel[13]. Este corpus, en comparación con los anteriores, tiene un tamaño pequeño y se centra en recoger una variedad del español, el español de Valencia, que corresponde a la zona septentrional, en la que además convive con otra lengua, el valenciano. Con estas características, en principio, no era esperable que estuviera en los primeros puestos de la tabla. Sin embargo, el valor de este trabajo, y de ahí el reconocimiento que ha recibido en los recopilatorios descritos, reside en que se constituye

[13] Actualmente, el corpus Val.Es.Co. en su versión 3.0 permite el acceso a una muestra de nuevos materiales a través de su página electrónica. Para la consulta de las versiones anteriores ha de acudirse a Briz y Grupo Val.Es.Co. (2002).

como el primer corpus de conversaciones coloquiales de lengua hablada en español, conversación coloquial espontánea grabada secretamente, género discursivo con escasa representación en el panorama de corpus por las dificultades metodológicas que conllevan su recolección (§ 4.1) y, además, por establecer un sistema de trabajo y un modelo de transcripción que ha sido adoptado por trabajos posteriores.

En cuanto al segundo bloque, en amarillo, los corpus recogidos son el *Corpus del español* de Mark Davies (CE), el *Corpus del Español Conversacional de Barcelona y su área metropolitana* (CEBNA), el *Corpus Oral Didáctico Anotado Lingüísticamente* (CORDIAL), el *Corpus Oral y Sonoro del Español Rural* (COSER) y el *Corpus de Referencia del Español Actual* (CREA). La razón por la que creemos que COSER no se encuentra recogido en el total de los repertorios revisados en esta tesis, a pesar de que sus características estructurales son bastante parecidas a las de los corpus que aparecen en primer lugar, es que es un corpus reciente, si tenemos en cuenta la horquilla temporal existente desde la aparición de estos trabajos. Si bien los materiales de COSER se empezaron a recoger en la década de los 90 del siglo pasado, su publicación en un portal electrónico y motor de búsqueda, y, por tanto, su posibilidad de acceso público es de la década posterior.

Sorprende que CREA, siendo el primer corpus académico que recoge oralidad, no cuente con mayor grado de consenso. El hecho de que Llisterri (2021) no lo mencione puede deberse a que no responde a su objetivo de recopilación, esto es, no es un corpus válido para el estudio del componente fónico y CREA, aunque recoja material oral, no permite el acceso al audio. Parodi y Burdiles (2019) lo mencionan como corpus escrito, por tanto, no hemos considerado que lo incluya. Igual sucede con el CE de Mark Davies ya que, aunque vuelca en su motor de búsqueda materiales procedentes del medio oral, no hay recogidos segmentos orales[14].

Pasando al tercer bloque de consenso, en color verde, los corpus aquí recogidos son el *Alicante Corpus Oral del Español* (ALCORE), el *Corpus Lingüístico del Habla de Almería* (CLHA), el *Corpus Oral de la Variedad Juvenil Universitaria de Alicante* (COVJUA), el *Vernáculo Urbano Malagueño* (VUM) y CORPES XXI. El rasgo común de todos ellos, exceptuando el CORPES XXI, y que los diferencia de los corpus de los bloques anteriores

[14] Si se accede a la página web, la interfaz de búsqueda puede hacernos creer que sí que se dispone de este material ya que, junto a las concordancias resultantes de nuestra consulta aparece un icono que refiere a la reproducción del fragmento. Sin embargo, este enlaza con la opción de Google Translator de reproducción automática a través de su propio asistente.

es su reducido tamaño, que en parte se debe a que recogen una variedad dialectal concreta. Que el corpus académico aparezca en esta franja se debe, previsiblemente, a que se publicó su primera versión en 2013, varios años después de los primeros recopilatorios generales. Por otro lado, el VUM acabará formando parte del corpus PRESEEA.

En último lugar, en azul, los corpus que presentan menos consenso en su cita son: el *Corpus del Grupo de Investigación Lingüística de la Universidad de Granada* (COGILA), el *Corpus para fines lexicográficos y de análisis gramatical* CUMBRE, y el *Macrocorpus para el Estudio de la Norma Lingüística Culta* (MC-NC), derivado de los trabajos realizados por el *Proyecto del Programa Interamericano de Lingüística y Enseñanza de Idiomas* (PILEI). Son corpus más antiguos, en el mejor de los casos publicados en versión comercial, como sucede con COGILA, o integrados en el CREA, como el MC-NC. Mientras que el corpus CUMBRE surge por iniciativa privada y no se encuentra a disposición de la comunidad investigadora. Respecto al corpus GRIAL[15], dirigido por Parodi, si bien autores como Briz y Albelda (2009) y Solís (2018) se refieren a él como base de datos de grandes dimensiones, formada por diversos corpus, el propio Parodi (2006) lo cataloga como interfaz computacional, más que como corpus en sí.

A modo de recapitulación, puede observarse que existe poco consenso en la mención de los corpus por parte de la historiografía de corpus. De los más de 70 corpus listados, exclusivamente de la categoría *fines lingüísticos generales*, solo en 19 corpus se da una coincidencia en cuatro o más artículos. Esto puede deberse probablemente a dos factores: en primer lugar, el tiempo, ya que hay una horquilla de casi 20 años entre la publicación del primer y último repertorio aquí recogido. En segundo lugar, puede tener incidencia el foco de investigación, ya que estas panorámicas responden a diferentes objetivos así, contamos con panorámicas centradas en ofrecer una visión general de los corpus orales de español, frente a otras focalizadas en objetivos particulares, como la recogida de corpus de carácter dialógico en exclusiva o de aquellos útiles para los procesos de enseñanza y aprendizaje del español. El mayor grado de consenso, como se ha podido observar, sucede entre los recopilatorios de carácter general.

[15] Cabe señalar la imposibilidad actual de acceder a la página web, por lo que no podemos ofrecer una descripción más detallada.

3.2.3.2. Aspectos de mejora

No cabe duda de que los trabajos analizados en las líneas anteriores constituyen una gran aportación a la comunidad investigadora. Como se mencionó al principio de este capítulo, ante la ausencia de portales en línea configurados para ser un repositorio de corpus orales del español, la información reunida en estos trabajos facilita la visión en conjunto de los corpus orales del español más conocidos y permite obtener una revisión actualizada de los corpus disponibles a fecha actual. Ahora bien, partimos de la base de que los casos estudiados se enfrentan a la imposibilidad de recoger todos los corpus existentes y no son más que una fotografía del momento en el que se publicaron estos trabajos. Las posibilidades de conocer los corpus orales de menor tamaño o realizados por personal investigador de manera individual, sin el respaldo de grandes instituciones, es uno de los hándicaps a los que nos enfrentamos cuando se realizan estudios recopilatorios.

Los trabajos de Briz y Albelda (2009) y Llisterri (2021) son, a nuestro parecer, los más completos. En el primer caso destaca por la exhaustividad de la búsqueda de corpus orales existentes, así como por la descripción extensa y detallada que ofrecen de las características de cada uno de los corpus recogidos. En el segundo caso, junto a esta minuciosidad es relevante la descripción de los aspectos técnicos referidos a la transcripción, la anotación y el material oral en sí, información que apenas se ofrece en los otros trabajos.

Uno de los principales problemas de los que adolece la mayoría de los recopilatorios es la presencia de cierta vacilación terminológica que no siempre se aclara, ya que se deja en manos del lector/a la interpretación de los criterios clasificatorios. Uno de los ejemplos más claros es la indeterminación de *micro* y *macrocorpus*, cuya definición solo se aporta brevemente en Briz y Albelda (2009), de manera que queda a criterio del lector/a determinar qué se considera un corpus grande, qué un corpus conformado por otros subcorpus, etc. De manera particular, se observan puntos de indefinición en Enghels, Vanderschueren y Bouzouita (2015), como, por ejemplo, no establecer qué se entiende por *búsqueda no específica* frente a *búsqueda particular*, algo que no se explica en el propio trabajo, pero tampoco en la remisión a otras obras de contenido teórico en las que se hayan basado para realizar cada panorámica. Tampoco quedan claros otros criterios clasificatorios como la extensión de los corpus denominados *panhispánicos* frente a los que no lo son. Así, bajo este epígrafe deberían englobarse, a grandes rasgos, los corpus que incluyen tanto variedades del español peninsular como del español americano. Sin embargo, no aparece ninguna

matización a este respecto, salvo en Briz y Carcelén (2019, p. 124), quienes especifican que son aquellos que recogen materiales orales procedentes tanto de España como de Hispanoamérica, aunque no detallan en qué porcentaje. Es decir, cabría preguntarse qué requisitos deben reunirse para considerar que un corpus es panhispánico y no un corpus que recoge diversas variedades dialectales.

Hay que considerar, por otro lado, que los objetos de estudio de cada recopilatorio son diferentes, algo motivado en parte por el distinto foco de interés y, en parte, por el uso distinto de los términos. En esta última situación se encuentra la propia definición de *corpus oral* en Llisterri (2021) frente a los otros recopilatorios, ya que, como hemos visto, distingue entre corpus oral y corpus de lengua hablada, y matiza que un corpus oral es aquel que incluye una transcripción fonética de los materiales sincronizada con la grabación y, en ocasiones, acompañada de algún tipo de anotación, mientras que un corpus de lengua oral ofrecería la transcripción ortográfica de las grabaciones originales (Llisterri, 2021, p. 165). Si tenemos en cuenta esta diferenciación, la mayoría de los corpus recogidos por los trabajos aquí presentados no se considerarían corpus orales, sino corpus de lengua oral, puesto que, por lo general, no incluyen la transcripción fonética. Sin embargo, a excepción de Llisterri, en el resto de los autores no aparece esta distinción y se habla de corpus orales para englobar cualquier tipología de corpus en los que hay un trabajo de recogida de material oral, independientemente del tipo de transcripción y anotación que se añada sobre dicho material.

Debemos tener en cuenta también que todos los corpus recogidos responden a fines lingüísticos, ya que la confección de un corpus responde a objetivos de investigación concretos de cada grupo. Sin embargo, parece habitual diferenciar trabajos con fines lingüísticos específicos, como los corpus para el desarrollo de tecnologías del habla, de adquisición del lenguaje y corpus de español como lengua extranjera, de otros trabajos para el análisis lingüístico del español en general, aunque estos respondan a sus propios objetivos de recolección. Es decir, cualquier corpus, aunque responda a objetivos específicos como puede ser la lexicología, la sociolingüística, la pragmática o el análisis de la conversación, puede extrapolarse a cualquier investigación que responda a otros más generales, sobre todo si para su construcción se han seguido sistemas internacionales de estandarización que permitan la reutilización de los datos (§ 4.2.2), siempre dependiendo del tipo de acceso que permitan a los datos.

En general, se da por supuesto que los investigadores que se acercan a los recopilatorios estudiados tienen clara esta información; sin embargo, la vacilación terminológica es notable y, por tanto, se necesitarían indicaciones concretas sobre la elección de cada uno de los criterios y sus implicaciones.

Atendiendo a estos aspectos problemáticos mencionados anteriormente y con propósito de mejora, en el siguiente apartado se presenta una revisión actualizada del panorama de corpus oral del español actual y se propone una clasificación integradora y una delimitación del objeto de estudio concreto.

3.3. Los corpus orales en español en 2023: revisión y propuesta clasificatoria

La revisión previa sobre la historiografía de corpus orales del español nos ha permitido conocer el estado de la cuestión sobre los corpus hasta la fecha de la última publicación sobre repertorios de corpus, Llisterri (2021). A los materiales recogidos en los trabajos anteriores hay que añadir otros corpus que no se han citado en tales repertorios, por ser de creación más reciente, o por no haber sido considerados bajo el abanico de criterios de inclusión de dichas obras.

Así, este apartado se propone presentar una recopilación de corpus orales del español con el objetivo de ofrecer una actualización del panorama de corpus a fecha de hoy. Para la selección de los trabajos aquí recogidos se han tenido en cuenta los siguientes criterios de selección:

(a) son corpus orales, entendidos en sentido amplio, esto es recogen muestras de habla para su análisis posterior, independientemente de que estas incluyan o no transcripción fonética;

(b) sincrónicos;

(c) de acceso libre y no comercial, aunque se necesite cuenta de usuario para poder consultarlos;

(d) recogen cualquier variedad del español;

(e) responden a fines lingüísticos generales; por tanto, quedan fuera de esta recopilación aquellos construidos para el desarrollo de tecnologías de habla (generalmente privados y de

acceso restringido), corpus de aprendices de español y otros corpus como los construidos para adquisición y desarrollo del lenguaje o patologías del lenguaje;

(f) permiten el acceso en línea a los datos, bien sea a través de un motor de búsqueda, 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;

(g) de cualquier género discursivo; y

(h) que se encuentren tanto finalizados como en construcción.

En cuanto a los criterios clasificatorios, nos centraremos en describir aquellos que, como señalan Briz y Albelda (2009) y Briz y Carcelén (2019), están disponibles en formato digital, aunque sea en grado mínimo y que permiten su accesibilidad en formato electrónico, siguiendo también los criterios de clasificación adoptados por Llisterri (2021) sobre aquella información técnica que resulte pertinente ofrecer sobre estos corpus en cuanto a sus sistemas de transcripción y codificación.

Para su descripción, se han seguido estas consideraciones previas, 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, así mismo, el modo de acceso y consulta de los datos, esto es, que sea en línea, con o sin motor de búsqueda disponible. Aquellos corpus que no estén abiertos al público o que estén publicados en papel no serán considerados en esta recopilación.
- 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 tienen voluntad de recogerlos. Si solo recogen muestras de alguna de

estas variedades debería hablarse de *corpus multidialectal* o de *una sola variedad*, o de *corpus de variedades del español*, como el caso del corpus COLA, que reúne muestras del español de Madrid, Buenos Aires y Santiago de Chile.

- Se explica qué materiales ofrece (transcripción, audio, información contextual) cada corpus y si se permite la descarga.
- En cuanto al tamaño, se tiene en cuenta por igual que constituyan macrocorpus, que ofrecen la posibilidad de contrastar resultados entre distintas normas regionales y que pretenden constituirse en grandes bases de datos, o microcorpus, cuyo ámbito de acción y sus dimensiones son más reducidas, así como sus objetivos, que son más concretos (Briz, 2012a, p. 116-117). Más particularmente, en este punto, seguimos la propuesta de Briz (2012a), que considera que un material es un macrocorpus si recoge, al menos, tres variedades de habla. De manera inversa, un microcorpus será aquel material que recoja dos variedades o una. Conviene tener en cuenta esta precisión, pues ha sido lugar común en la bibliografía sobre el tema, asociar los términos *macrocorpus* y *microcorpus* al número de palabras o formas que componen un corpus. La propuesta de Briz (2012a) 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, 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 (§ 3.3.1); y, en segundo lugar, la descripción detallada de cada uno de los corpus (§ 3.3.2).

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, o los repositorios con las transcripciones del Parlamento Europeo, que no tienen un propósito lingüístico por sí mismos, para poder realizar cualquier investigación. Sin embargo, y tal y como ya ha

planteado la literatura sobre el tema (Torruella y Llisterri, 1999, McEnery, Xiao, Tono, 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 (§2.1.). Así mismo, a nuestro juicio, y teniendo en cuenta las obras sobre repertorios de corpus revisadas en el apartado anterior, dicha muestra debe haber recibido un tratamiento informático que permita el trabajo con el audio/vídeo y la transcripción, unidos por un código de tiempo.

3.3.1. Listado de corpus orales del español disponibles en línea

A continuación, en la Tabla 3, se presentan por orden alfabético los corpus que conforman la versión actualizada del panorama de corpus orales del español, reunidos aquí siguiendo los criterios clasificatorios explicados en el apartado anterior, si bien, el primer corpus que aparece, el corpus Ameresco, no será desarrollado por medio de la ficha descriptiva ya que, su caracterización y particularidades serán detalladas en el capítulo 5. Como se ha mencionado en el Capítulo 1, el diseño y construcción de este corpus responde al objetivo principal de esta tesis y se abordara de manera extensa en su propio capítulo.

Una vez listados, se ofrece la ficha técnica 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, o bien poseen unas características diferentes. Por esta razón no siguen la misma línea estructural que los anteriores[16].

Como hemos comentado a lo largo de este capítulo, 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.

[16] No se han considerado, así mismo, 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, si bien cuentan con sus propios portales electrónicos donde se puede acceder a los materiales.

CORPUS[17]	ACCESO EN LÍNEA
AMERESCO	http://corpusameresco.com/
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	https://ebuah.uah.es/dspace/bitstream/handle/10017/25298/Mendez_Corpus_LR__2015_13.PDF?sequence=1&isAllowed=y
COLA	https://blogg.hiof.no/colam-esp/el-corpus-cola/
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 LINGÜÍSTICO 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/
VHW	https://dialectos.osu.edu/

Tabla 3. Corpus orales del español disponibles en línea

[17] El desarrollo de los acrónimos de los corpus puede consultarse en el Anexo 1 y accediendo a los siguientes enlaces: https://nuvol.uv.es/owncloud/index.php/s/IiJ0bjHxYfGKoUc

3.3.2. Descripción de los corpus orales del español disponibles en línea

Se describen, a continuación, las principales características de los corpus recogidos en la Tabla 3. De cada uno de ellos, a excepción de los corpus señalados en la sección anterior, se detalla su nombre, el acceso a su página electrónica mediante enlace y código QR, el nombre de la persona o grupo responsable, su variedad dialectal, el género o géneros discursivos que recoge, el objeto de estudio, la manera en que se permite la consulta en línea y los materiales que se ponen a disposición de la comunidad de usuarios.

3.3.2.1. 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)[18] 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 electrónica (CE, en línea).

Corpus del Español de Mark Davies https://www.corpusdelespanol.org/	
Responsable	Mark Davies
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 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.

3.3.2.2. 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

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

1921 a 1974 y de 1975 a 2018. Esta recopilación está orientada, sobre todo, a documentar vocabulario, tanto del más reciente, como del tradicional, sin perjuicio para los análisis sintácticos que también son posibles con estos materiales.

CORPUS DEL ESPAÑOL MEXICANO CONTEMPORÁNEO (CEMC) I y II https://repositorio-cell.colmex.mx/corpus.html	
Responsable	Colegio de México
Variedad dialectal	México
Género discursivo	Entrevista
Objeto de estudio	Estudio del léxico
Consulta en línea	Sí, con motor de búsqueda. Pueden realizarse búsquedas combinadas en el CEMC I y II, así como en otros corpus procedentes del medio escrito. Permite filtrar por año y género (tipología de la muestra). Ofrece 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*.

3.3.2.3. 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 vídeo, archivos audio, transcripciones completas y anotaciones de categoría gramatical.

3.3.2.4. 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 muy 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) https://ebuah.uah.es/dspace/bitstream/handle/10017/25298/Mendez_Corpus_LR__2015_13.PDF?sequence=1&isAllowed=y	
Responsable	Beatriz Méndez Guerrero, Universidad Complutense de Madrid
Variedad dialectal	Peninsular, 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.

3.3.2.5. Corpus Oral de Lenguaje Adolescente (COLA)

El Corpus COLA contiene el habla juvenil y conversaciones espontáneas e informales, recogidas en Madrid (COLAM), Santiago de Chile (COLAS) y Buenos Aires (COLABA). Contiene un total de 700 000 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.

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.

3.3.2.6. 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 2023, por el momento cuenta con 165 informantes (97 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 172 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. Antes de la actualización de la página electrónica con fecha 28/08/2023 había una muestra disponible en línea, ahora no aparece.

3.3.2.7. 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).

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.
Materiales disponibles	A falta de que activen la consulta en abierto en línea, pueden solicitarse las transcripciones en formato PDF.

3.3.2.8. 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 que 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).

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 y 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.

3.3.2.9. 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.

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.

3.3.2.10. 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.

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
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.

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

Desde el Observatorio de la lengua española y las culturas hispánicas del Instituto Cervantes en la Universidad de Harvard, con la colaboración de la Academia Norteamericana de la Lengua Española (ANLE), bajo la dirección de Francisco Moreno Fernández, se inician los trabajos para la construcción del CORPEEU. 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 la Universidad de Harvard
Variedad dialectal	Multidialectal, Español de los Estados Unidos
Género discursivo	Miscelánea, las entrevistas corresponden a PRESEEA Nueva York
Objeto de estudio	Documentar el español en los Estados Unidos
Consulta	Sí, 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.

3.3.2.12. Corpus del Habla de Almería

Nace en el seno del 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 (CLHA) 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
Consulta	En línea
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.

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

Dirigido por Inés Fernández Ordóñez, este corpus está formado por grabaciones de la lengua hablada en enclaves 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, pero puede oscilar desde solo media hora hasta más de dos horas y media, y, como se indica en su metodología de trabajo, aunque se han registrado unos 2 910 informantes, la inmensa mayoría de las veces solo ha sido encuestado uno informante por enclave con el detenimiento deseable

Corpus Oral y Sonoro del Español Rural (COSER) http://www.corpusrural.es/	
Responsable	Inés Fernández Ordóñez, Universidad Autónoma de Madrid
Variedad dialectal	Multidialectal, español 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.

3.3.2.14. 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. El porcentaje de textos orales que incluye con respecto a los escritos es del 10 % frente al 90 %, 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.

Corpus de Referencia del Español Actual (CREA) https://www.rae.es/banco-de-datos/crea	
Responsable	RAE
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, dos versiones, 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.

3.3.2.15. El español hablado en Bogotá

Dirigido por José Joaquín Montes Giraldo y coordinado por Jennie Figueroa Lorza, del departamento de Dialectología 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 investigativos 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).

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	Dialectológicos y sociolingüísticos
Consulta	Sí
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.

3.3.2.16. ESLORA

Este corpus ha sido elaborado por los miembros del 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 de la Universidad de Santiago de Compostela
Variedad dialectal	Peninsular, Santiago de Compostela
Género discursivo	Entrevistas (integradas en PRESEEA) y conversaciones
Objeto de estudio	Ofrecer material oral de referencia para el estudio de la lengua española en su variedad propia de un área bilingüe con el gallego
Consulta	Sí
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.

3.3.2.17. 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, español de España
Género discursivo	Miscelánea

Objeto de estudio	Análisis y el estudio de la sintaxis del discurso
Consulta	Sí, no cuenta con motor de búsqueda
Materiales disponibles	Pueden descargarse las transcripciones en formato PDF, sin acceso al audio

3.3.2.18. 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).

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	Entrevista semidirigida
Objeto de estudio	Investigación 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.

3.3.2.19. 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).

Valencia Español Coloquial (Val.Es.Co.) versión 3.0[19] https://www.valesco.es/	
Responsable	Salvador Pons Bordería, Universitat de València
Variedad dialectal	Peninsular, Valencia
Género discursivo	Conversación coloquial espontánea grabada secretamente
Objeto de estudio	Análisis pragmático del discurso
Consulta	Sí, motor de búsqueda
Materiales disponibles	Búsqueda por concordancias sin acceso al audio. Permite la descarga de los resultados, pero no de los materiales.

3.3.2.20. Voices of Hispanic World

Este corpus, desarrollado desde la Ohio State University, se constituye como un recurso audiovisual de muestras dialectales del español. Las muestras de habla están clasificadas por geografía, rasgo lingüístico y tema de conversación, además, aparecen con transcripciones escritas, anotaciones dialectológicas y funciones de navegación fáciles de usar (Voices of Hispanic World, en línea).

Voices of Hispanic World https://dialectos.osu.edu/	
Responsable	Terrell A. Morgan, Ohio State University
Variedad dialectal	Panhispánico
Género discursivo	Miscelánea
Objeto de estudio	Dialectología
Consulta	Sí, sin motor de búsqueda para formas ortográficas
Materiales disponibles	Vídeos y transcripción en PDF, no alineados. No se puede descargar.

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

3.3.2.21. Otros

Cabe señalar, en último lugar, otros corpus dignos de mención que, si bien trabajan con material oral, no cumplen con todos los requisitos clasificatorios establecidos. Por ejemplo, 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 que ARTHUS (Rojo), aunque nace como un trabajo de corpus, sus materiales se han incorporado a bases de datos para el estudio sintáctico y verbal como ADESSE (García Miguel) que solo permite búsquedas en estos términos; o el *Corpus Oral del Español de México* (COEM) (Martín Butragueño, Mendoza y Orozco), cuyo motor de búsqueda solo permite recuperar enunciados aseverativos e interrogativos, sin poder recuperar otras informaciones. Los casos de C-ORAL-ROM*, Corpus Oral de Referencia del Español en Contacto* (COREC) (Palacios) y *El habla popular* solo permiten acceso a una muestra de los materiales, el material completo debe comprarse.

3.4. Síntesis del capítulo

El Capítulo 3 se ha centrado en el desarrollo de los corpus orales en español, que tienen sus raíces en la dialectología y los atlas lingüísticos.

Aunque la lingüística de corpus en español comenzó con cierto retraso en comparación con el mundo anglosajón, experimentó un rápido desarrollo. Uno de los primeros hitos en el desarrollo de corpus orales en español fue el *Programa Interamericano de Lingüística y Enseñanza de Idiomas* (PILEI) en la década de 1960, seguido del *Estudio gramatical del español hablado en América* (EGREHA), que contiene materiales del PILEI y el MC-NC, el *Macrocorpus de la norma lingüística culta de las principales ciudades de España y América* (MC-NC), uno de los primeros que publicó en CD-ROM y que se incorporó posteriormente al *Corpus de Referencia del Español Actual* (CREA) de la Real Academia Española y al *Corpus del Español* de Mark Davies (CE).

A finales de los años 90, nació el proyecto más importante de estudio sociolingüístico del español de España y América (PRESEEA) bajo la dirección de Francisco Moreno Fernández y posteriormente surgieron corpus académicos como el *Corpus de Referencia del Español Actual* (CREA) y el *Corpus del Español del Siglo XXI* (CORPES XXI), en 1998 y 2013,

respectivamente. El último mejoró la representatividad geolingüística y permitió el acceso al registro sonoro, marcando un avance en la disponibilidad de material oral en español para la investigación lingüística.

A pesar de la creciente disponibilidad de corpus orales en español, resulta imposible recoger todos los corpus existentes, y cualquier recopilación está destinada a quedarse anticuada rápidamente. En este capítulo se han enumerado algunas de las principales obras recopilatorias de corpus orales, trabajos que llenan un vacío importante al proporcionar una panorámica del estado de la investigación en corpus orales del español y entre los cuales destacan Briz y Albelda (2009) y Llisterri (2021).

A la luz de los datos aportados, queda patente la variedad terminológica, que resulta en una ambigüedad, y se pone de manifiesto la importancia de definir claramente la terminología y los criterios para la clasificación de corpus orales. Para ello, a partir de una selección de corpus, se describen desde varios puntos de vista y se ofrece un nuevo panorama, actualizado a fecha de 2023, de los principales corpus orales en español con acceso abierto y disponibles en línea.

Capítulo 4

Diseño y construcción de corpus orales

En este capítulo se describen las fases de diseño de la construcción de corpus orales, particularmente conversacionales y espontáneos. Partiendo de las consideraciones generales que deben tenerse en cuenta cuando se plantea la construcción de un corpus (§ 4.1), se abordan las diferentes fases en las que se concreta la ejecución del diseño pactado (§ 4.2) divididas en las siguientes secciones: la primera (§ 4.2.1) trata sobre aquellos aspectos que afectan a la concepción del corpus y a la recogida de los datos; la segunda (§ 4.2.2) se dedica a revisar las cuestiones referidas al tratamiento de dichos datos; en tercer lugar, (§ 4.2.3) se aborda la explotación de los datos; a continuación, se realiza un análisis contrastivo entre las metodologías de construcción de corpus orales del español (§ 4.3). En último lugar, se ofrece una síntesis como cierre de este capítulo (§ 4.4).

4.1. Consideraciones generales para el diseño de corpus orales: planteamientos previos

La construcción de corpus de lengua oral requiere de un proceso minucioso de planificación y reflexión previa para asegurar que los datos recopilados sean representativos y útiles para su análisis. La literatura existente (Torruella y Llisterri, 1999, Sinclair, 2004, McEnery, Xiao y Tono, 2006, Adolph y Knight, 2010, Reppen, 2010, Rojo, 2021, Pons, 2022, Egbert, Biber y Gray, 2022, entre otros) aborda los principios generales que deben regir un buen diseño de corpus, ya sean escritos u orales. Como señalan Hincapié y Bernal (2018, p. 53), dentro de la lingüística de corpus no existe un protocolo que determine paso a paso cómo crear un corpus, si bien, autores como Kennedy (1998) o Atkins, Clear y Ostler (1992) han señalado cinco estadios fundamentales en la construcción de un corpus que podrían resumirse en los siguientes pasos: diseño de corpus, obtención de permisos y captura de los datos, planificación y preparación del sistema de almacenamiento, procesamiento del corpus y opciones de uso. Por su parte, Leech, Myers y Thomas (1995) concretan también cinco fases, englobadas en torno a las tareas de 1) grabación, 2) transcripción, 3) representación o marcado, 4) codificación o anotación y 5) aplicación.

Cuando se plantea el diseño de un corpus, el primer principio básico tiene que ver con el objetivo de investigación, a saber, cuál es la finalidad de su creación, qué tipo de análisis lingüístico se pretende realizar y cuáles son las preguntas de investigación que guían la recogida de corpus. En segundo lugar, se debe establecer de qué manera el corpus va a ser representativo, es decir, qué criterios de selección se seguirán para obtener una muestra que refleje la diversidad en términos de variación diatópica, diastrática o diafásica. El tamaño

que tendrá el corpus atendiendo a su objetivo y a las características de representatividad escogidas es otra de las reflexiones que deben plantearse. Por ejemplo, no tendrá las mismas implicaciones construir un corpus para una investigación particular, como puede ser una tesis doctoral, con escasa o nula financiación y con una persona responsable de todo el proceso, que construir un corpus bajo el amparo de un grupo de investigación o de una institución que, generalmente, va a contar con más medios técnicos, personales y económicos para abordar su construcción.

Otro de los ejes fundamentales tiene que ver con los materiales que se van a recopilar y qué implicaciones llevan aparejados dichos materiales. En el caso de corpus escritos, por ejemplo, deberán considerarse cuestiones relacionadas con los derechos de autor de los archivos que conformarán la muestra; en el caso de los corpus orales, este punto se hace más complejo, especialmente cuando se van a recopilar grabaciones realizadas a particulares, ya que implica el conocimiento y cumplimiento de la legislación en materia de privacidad y protección de datos, como se verá detalladamente en la sección 4.2.1. No obstante, en ambos casos deben operar siempre los principios de ética y buenas prácticas en la investigación.

Deben plantearse, por último, aspectos relacionados con la conservación y almacenamiento de los datos, así como decidir de qué manera se va a realizar el acceso y la distribución del corpus, es decir, se deberá establecer la manera en que el público va a acceder al corpus y en qué condiciones, si los materiales recopilados se pueden compartir de manera abierta o si, por el contrario, se debe restringir el acceso según sea necesario. Esto sucede, por ejemplo, si el corpus recogido contiene grabaciones realizadas en ámbitos de mayor privacidad o confidencialidad, como sería el caso de un entorno médico-sanitario. Tales grabaciones, aun cuando se hayan sometido a un proceso de anonimización, no pueden por ley estar accesibles al público general, sino que su uso queda restringido al grupo de investigación (§ 4.2.1.2.). Por tanto, este último aspecto dependerá del contexto en el que surja cada corpus, por lo general, a mayores medios técnicos, personales y económicos disponibles, mejores condiciones de acceso y distribución del corpus.

Centrándonos exclusivamente en los corpus orales, hay que dar un paso más. Mientras que, en la recopilación de materiales escritos, su incorporación a un corpus es más ágil —por ejemplo, bastaría con recopilar las noticias de prensa en formato digital, codificarlas para

su procesamiento informático y volcar este material en un motor de búsqueda en línea[20]—, en el caso de corpus orales nos encontramos con que, para llegar al paso final de incorporación de materiales a la página electrónica, debe superarse primero una fase de transcripción y etiquetado del material. Por tanto, cabe plantearse la manera en que, por medio de lo escrito, va a representarse la lengua hablada, cómo de exhaustiva va a ser la transcripción y la codificación, cómo se van a recoger los metadatos e informaciones contextuales sobre las circunstancias de la grabación y los participantes, incluyendo además los permisos y consentimientos informados, de las personas que participen en las grabaciones.

Cierto es que, en la actualidad, gracias a los rápidos avances en tecnología y en inteligencia artificial, han surgido numerosas herramientas que deberían facilitar la transcripción automática, como pueden ser Whisper, Otter.ai, Google Cloud Speech-to-Text o Microsoft Azure Speech Service, entre muchos otros. Estas herramientas son realmente efectivas en casos determinados, como en grabaciones dirigidas donde hay un mayor control de la situación, en entornos de reconocimiento y síntesis de habla, en discurso monologal formal o en otros en los que la calidad de la grabación es esperablemente óptima, como en televisión, radio o determinados creadores de contenido profesionales en redes. Sin embargo, estas herramientas presentan aún limitaciones si las comparamos con la capacidad humana[21]. Además, la transcripción de la lengua hablada, larga y costosa de por sí, es más difícil cuanto más detallado sea el nivel de transcripción (Adolphs y Knight, 2010) y, por lo general, los grandes corpus se siguen transcribiendo de forma manual (Gadet *et al*., 2012, Niemants, 2018) o al menos se completan de forma manual (Rufino, 2020, p. 132).

4.2. El diseño de corpus orales. Fases de la construcción

A continuación, se abordarán las fases de la construcción de un corpus. Partiendo de la ya mencionada clasificación de Atkins, Clear y Ostler (1992) en cinco estadios fundamentales, se han establecido tres grandes bloques: una fase previa de concepción y recogida, que incluye el diseño del corpus, la obtención de los permisos y la captación física de los datos (§ 4.2.1.); una segunda fase de tratamiento de los materiales recogidos, que

[20] Puede consultarse bibliografía específica sobre el proceso de diseño de corpus escritos en McEnery y Wilson (2001), Biber, Conrad y Reppen (1998), Berber Sardinha (2004), McEnery y Hardie (2011), Stefanowitsch (2020), entre otros.

[21] O si se plantea la transcripción automática de determinados géneros discursivos, como el caso de la conversación coloquial espontánea.

incluye la planificación y preparación del sistema de transcripción y codificación, incluyendo la fase de anonimización de los materiales, el procesamiento del corpus y las opciones de anotación (§ 4.2.2.); y, finalmente, una fase posterior (que puede darse o no) que implica la difusión de los datos del corpus a través de plataformas para su explotación (§ 4.2.3.).

4.2.1. Fase 1: Concepción del corpus y recogida de los datos

4.2.1.1. Cuestiones generales de la recogida de datos

En este punto del proceso de diseño debe reflexionarse sobre las características que ha de presentar el corpus según el objetivo de creación al que responde. Según la literatura (Meyer, 2002, Berber Sardinha, 2004, McEnery, Xiao y Tono, 2006, Rojo, 2021) en la concepción del corpus deben tenerse en cuenta aspectos como el tamaño, la representatividad o el equilibrio en el tamaño de la muestra, factores que determinarán la selección de los materiales que se van a recoger.

La reflexión propia del diseño y concepción de un corpus consiste en estudiar y calcular *a priori* los costes de tiempo, de personal y de financiación, así como prever la disponibilidad suficiente de los medios, instrumentos y conocimientos técnicos para llevar a cabo este trabajo. Esta reflexión anticipada resulta necesaria en tres sentidos. En primer lugar, para evaluar el grado de realismo y capacidad para emprender la recogida de datos y creación del corpus. En segundo lugar, lógicamente, para la propia planificación y temporalización del trabajo, así como para la preparación y apropiación de los instrumentos materiales y personales necesarios para su ejecución. En tercer lugar, para verificar que dicho material no existe ya, por lo que sería innecesario plantearse recogerlo.

El criterio fundamental y general que guía el diseño de un corpus es, como hemos visto, su finalidad de uso en la investigación. Establecer este criterio con sus particulares características es la primera fase del diseño. A partir de ello, surgen otras preguntas que ha de hacerse el investigador, siempre vinculadas a la finalidad general. Como guía para abordar esta tarea, puede resultar útil evaluar las siguientes dimensiones al proyectar la recopilación de los datos. La finalidad investigadora implica determinar el fenómeno que se pretende estudiar y evaluar sus diversas dimensiones en relación con el tipo de materiales que resultarán más adecuados para su investigación. La Tabla 4 recoge las principales dimensiones que han de valorarse respecto a los datos del corpus que se quiere recoger:

Naturaleza lingüística de los datos	• Fónicos, prosódicos, morfosintácticos, léxicos, discursivos, etc.
Naturaleza variacionista de los datos	• Diastrática: rasgos sociolectales de las personas informantes (edad, sexo, nivel de instrucción) • Diafásica: rasgos de la situación comunicativa (tipo de relación entre los hablantes, grado de familiaridad del lugar físico de grabación, temas de habla, finalidad interpersonal o transaccional) • Diatópica: circunscripción a áreas dialectales particulares • Jergas o lenguas de especialidad o lengua general: académico, administrativo, deportivo, médico, empresarial, etc.
Género discursivo	• Monológico (charla, clase, conferencia, monólogo de humor, etc.) • Dialógico (conversación, entrevista, debate, reunión de trabajo, transacción comercial, tutoría, consulta médica, rueda de prensa, etc.) • DMO: discurso mediado por ordenador (videoconferencia, audios de mensajería espontánea, etc.)
Necesidad de contexto discursivo	• Innecesario • Necesidad de cotexto y contexto inmediato local de la interacción • Necesidad de contexto global de la interacción
Impacto de la interacción y tipo de destinatario	• Privada • Pública no grabada previamente • Pública grabada previamente (tv, radio, cine, etc.)
Previsión del índice de frecuencia del fenómeno de estudio	• Muy frecuente • Frecuencia intermedia • Esporádico
Tipo de investigación	• Exploratoria • Descriptiva • Predictiva
Grado de naturalidad de los datos	• Materiales naturales, espontáneos e inconscientes (secretos) • Materiales naturales, pero conscientes (por ejemplo, se sabe que se está grabando una conversación, se está grabando en TV) • Materiales elicitados
Necesidad de imagen/vídeo	• Sí • No

Tabla 4. Factores incidentes en la toma de decisiones inicial en la concepción de un corpus

Tomando como base los parámetros establecidos en el cuadro anterior, si, por ejemplo, una investigación se propone estudiar las funciones de la interjección *ay*, en cuanto a la

finalidad investigadora, habrá que determinar si el estudio prosódico, gramatical y pragmático de esta interjección se quiere realizar en una particular comunidad lingüística (como, por ejemplo, jóvenes chilenos) o en la lengua general. Tratándose de un fenómeno más propio de la dialogicidad, será interesante decidirse por la construcción de un corpus dialógico. La conversación puede ser el tipo de género más adecuado, si solo se proyecta estudiar un género, pero quizás cabe también plantearse la posibilidad de observar diversos géneros, puesto que puede presentar comportamientos y funciones diferentes. Así mismo, hay que valorar en qué grado se necesita disponer de datos contextuales ya que, si pensamos en la naturaleza de las interjecciones, estas pueden formularse ante diversos estímulos externos (una reacción ante algo que está sucediendo frente al hablante, por ejemplo), no solo por factores de índole lingüística o discursiva. Si, por ejemplo, el investigador se planteara obtener el índice de frecuencia de la interjección *ay,* en consonancia con el género discursivo elegido, habría que seleccionar bien el género discursivo para los datos, pues habría géneros en los que la interjección se registraría poco, como sucedería si se analizara un corpus de discursos parlamentarios.

Por otro lado, el acercamiento al estudio de esta interjección podría realizarse de diferentes modos: de carácter puramente exploratorio, por ejemplo, para confirmar la presencia o ausencia de *ay*; también podría ser de carácter descriptivo, en este sentido, solamente se identificarían los casos aparecidos y su significado en el empleo idiomático, o podría ser de tipo predictivo, estudio en el que se deberían valorar los contextos de uso más favorables a que esta aparezca. Al haber planteado previamente que quizá sea una partícula más frecuente en interacción, este hecho nos podría decantar hacia la recogida de materiales que se hayan obtenido de una manera espontánea y natural, ya que quizá el hecho de saber que se está siendo grabado podría implicar una distorsión del discurso, tendente a un grado de formalidad en el que la aparición de la interjección esté más limitada. Se podría, incluso, plantear la posibilidad de trabajar con materiales elicitados obtenidos en un contexto de grabación de tipo *role-play* en el que se le plantearan diferentes situaciones al hablante con el objetivo de propiciar el uso de esta interjección. Además, podría ser interesante obtener muestras de vídeo, material que permitiría estudiar la kinésica del hablante en cuanto a los gestos corporales, los movimientos y las expresiones faciales que aparecen cuando se hace empleo de *ay*.

Una vez valorados y decididos los rasgos anteriores, el diseño requiere pasar a una segunda fase de decisiones previas, y que implica hacerse preguntas en las siguientes

dimensiones (Torruella y Llisterri, 1999, Wynne, 2005, O'Keefe y McCarthy, 2010, Pons, 2022, Rojo, 2021):

Tamaño	• Número de hablantes • Número de interacciones (conversaciones, por ejemplo) • Duración de las grabaciones de cada hablante y/o interacciones • Número/cantidad de fragmentos de habla necesarios por cada hablante: relación con el equilibrio y la variedad de muestras
Representatividad	• Número de muestras necesario, de acuerdo con la población de origen: proporcional o estratificado
Equilibrio (variedad de las muestras)	• Diversidad de muestras • Proporción y equilibrio de las muestras
Grabaciones	• Circunscripción de las grabaciones a una fase/marco temporal • Aspectos técnicos de la recogida: dispositivos de grabación, ruidos ambientales, capacidad de la grabadora, etc. • Información contextual de la grabación (metadatos)
Cuestiones éticas y legales	• *Copyright* de los materiales • Confidencialidad y/o privacidad • Consentimiento de los hablantes • Cuestiones legales de cada país, cuestiones éticas • Posibilidades de contacto y acceso a los hablantes

Tabla 5. Criterios definidores del diseño de corpus orales

Respecto al **tamaño** que debe tener un corpus, esta es una de las principales cuestiones que afectan a la construcción de un corpus y la decisión final dependerá de diversos factores (Baker, Hardie y McEnery, 2006, p. 146). Por ejemplo, como señalan McEnery, Xiao y Tono (2006),

> First, the size of the corpus needed to explore a research question is dependent on the frequency and distribution of the linguistic features under consideration in that corpus (cf. McEnery and Wilson 2001: 80). As Leech (1991: 8-29) observes, size is not all-important. Small corpora may contain sufficient examples of frequent linguistic features. […] Second, small specialized corpora serve a very different yet important purpose from large multi-million-word corpora (Shimazumi and Berber-Sardinha 1996) […] Third, corpora that need extensive manual

> annotation (e.g. semantic annotation and pragmatic annotation, are necessarily small. Fourth, many corpus tools set a ceiling on the number of concordances that can be extracted, e.g. WordSmith version 3 can extract a maximum of 16,868 concordances (version 4 does not has this limit) (McEnery, Xiao y Tono, 2006, p. 78)

Es decir, consideran que hay cuatro factores fundamentales para determinar el tamaño. En primer lugar, dependerá de la frecuencia y la distribución de los rasgos lingüísticos considerados en dicho corpus; en segundo lugar, de la finalidad en sí, ya que esta será diferente si se construye un corpus pequeño y/o especializado o un macrocorpus; en tercer lugar, de la propia capacidad del investigador ya que, si el corpus va a necesitar una anotación manual exhaustiva, esto determinará el tamaño puesto que necesariamente tendrá que ser pequeño para poder abarcar esta tarea; y, por último, de la capacidad de las herramientas de corpus que se emplearán ya que, por ejemplo, existen herramientas que fijan un límite al número de concordancias que se pueden extraer. En este sentido, en Kennedy (1998) se establecen diferentes tamaños en cuanto a número de palabras según sea la finalidad del corpus:

> Kennedy (1998:68) suggests that for the study of prosody 100,000 words of spontaneous speech is adequate, whereas an analysis of verb-form morphology would require half a million words. For lexicography, a million words is unlikely to be large enough, as up to half the words will only occur once (and many of these may be polysemous). However, Biber (1993) suggests that a million words would be enough for grammatical studies. (Baker, Hardie y McEnery, 2006, p. 146)

Las consideraciones entorno al **equilibrio y la representatividad** son mencionadas sistemáticamente en las definiciones de corpus. Una muestra se considera representativa si reproduce la configuración de la población de la que ha sido extraída en los parámetros que se consideran relevantes (Rojo, 2021, p. 68), en cuanto a proporción y estratificación. Por su parte, el equilibrio *(balance)* tiene que ver con la proporcionalidad de los materiales, como plantean Baker, Hardie y McEnery (2016, p. 18), un corpus estará equilibrado cuando "contains texts from a wide range of different language, genres and text domains, so that, for example, it may include both spoken and written, and public and private texts", esto es, establece en qué medida va a recoger textos de diferentes tipologías, medio de producción, procedencia geográfica, etc.

Según Rojo (2021),

> el planteamiento más realista de la representatividad consiste en la garantía de que el corpus está equilibrado *(balanced* en inglés*),* lo cual implica que contiene, en cada uno de los subcorpus que se pueden establecer en función de su diseño, un número de textos y volumen suficiente para que la información específica que se puede extraer de ese subcorpus no esté sesgada y resulte fiable (Rojo, 2021, pp. 68-69).

De acuerdo con McEnery, Xiao y Tono (2006, p. 72), el equilibrio y la representatividad son consideraciones importantes a la hora de diseñar un corpus, estas dependen de la pregunta de investigación y de la facilidad con la que se puedan capturar los datos, por lo que deben interpretarse en términos relativos, es decir, un corpus debe ser lo más representativo posible de la variedad lingüística que se esté estudiando (McEnery, Xiao y Tono, 2006, p. 72). Si el objetivo de estudio es eminentemente gramatical, como por ejemplo el análisis de los verbos transitivos, no será tan relevante el equilibrio entre géneros discursivos recogidos; mientras que, si el análisis va a versar sobre el uso en interacciones dialógicas de dichos verbos, sí que cobra relevancia.

Por tanto, puede observarse cierta problemática. Como señala Rojo (2021, p. 69), "representatividad y equilibrio son, pues, nociones de difícil fijación en factores concretos. Se trata, más bien, de valores de imposible consecución, pero a los que hay que tender". Esto es, como también tratan McEnery y Hardie (2011), existen ciertas controversias en cuanto a qué debe considerarse por *representativo* y *equilibrado* y, realmente, no existe unanimidad:

> V´aradi (2001) has been critical of the failure of corpus linguists to fully define and realise a balanced and representative corpus. Even proposals, such as those of Biber (1993), to produce empirically determined representative corpora have not actually been pursued. Biber's proposal for representativeness to be realised by measuring internal variation within a corpus – i.e. a corpus is representative if it fully captures the variability of a language – has yet to be adopted in practice. It is also only one of many potential definitions of representativeness, as Leech (2007) points out. (McEnery y Hardie, 2011, p. 10)

Por tanto, si llevamos al extremo la caracterización de lo que significa que un corpus sea representativo y que esté equilibrado, nos vamos a encontrar con obstáculos difíciles de salvar. Por ejemplo, si se pretende construir un corpus de español que recoja las diferentes variedades diatópicas, para que este sea equilibrado debería tener el mismo número de palabras de cada variedad, así se recopilaría una cantidad de palabras fija e igual para cada

una de dichas variedades (español de España, de México, de Argentina, de Panamá, etc.); pero si bien, así se garantizaría que el corpus esté equilibrado, este equilibrio no conlleva que el corpus sea representativo. En el ejemplo señalado anteriormente, dadas las diferencias en cuanto al volumen de población de cada uno de los países de habla hispana, las proporciones de las muestras de México frente a las de Panamá deberían ser diferentes, más extensa en el caso mexicano y menos en el panameño.

Podría decirse, además, que el hecho de considerar este tipo de factores (equilibrio y representatividad) hace que estemos hablando de corpus y no de un archivo (McEnery, Xiao y Tono, 2006, p. 13), ya que un archivo recopila materiales por puro criterio cuantitativo, sin incluir parámetros de selección en este sentido, dicho de otra manera, un material no es representativo si se contempla solamente como un conjunto de textos, ya que será, entonces, solo un conjunto de datos arbitrario.

Según Torruella y Llisterri (1999, p. 2) "la función de un corpus es establecer la relación entre la teoría y los datos", por tanto, "(…) el corpus tiene que mostrar a pequeña escala cómo funciona la lengua natural; (…) pero para ello es necesario que esté diseñado correctamente sobre unas bases estadísticas apropiadas que aseguren que el resultado sea efectivamente un modelo de la realidad." No obstante, según estos autores, todas las muestras son, de algún modo, tendenciosas ya que cabe cuestionarse en todo momento cómo se obtuvieron las muestras y hasta qué punto son válidas para extraer las conclusiones de su análisis (Torruella y Llisterri, 1999, p. 19).

En resumen, como bien apuntan McEnery, Xiao y Tono (2006, pp. 21-22), se considera que un corpus es representativo si lo que encontramos en él también es válido para la lengua o variedad lingüística que se supone que representa. Normalmente, la representatividad se consigue mediante el equilibrio, es decir, cubriendo una amplia variedad de categorías de texto frecuentes e importantes que se muestrean proporcionalmente de la población objetivo. Sin embargo, la caracterización de la representatividad y el equilibrio de un corpus debe interpretarse en términos relativos y considerarse más como una declaración de fe que como un hecho, ya que actualmente no existe una forma objetiva de equilibrar un corpus o de medir su representatividad, como hemos señalado arriba.

Estrechamente vinculada con los conceptos de representatividad y equilibrio, otra de las dimensiones que hay que tener en cuenta y que se ha ido anticipando con anterioridad, es la del **muestreo**. Es imposible, en este sentido, recoger todo el material de todos los géneros

necesarios para un objetivo de investigación, así como obtener grabaciones de todos los sujetos que conforman la comunidad de habla que se quiere estudiar. Por este motivo, es esencial recurrir a una selección o muestreo. En un sentido estadístico el objetivo del muestreo es conseguir una muestra que, con sujeción a las limitaciones de tamaño, reproduzca lo más fielmente posible las características de la población, especialmente las de interés inmediato (Yates, 1965, p. 9, citado en McEnery, Xiao y Tono, 2006, p.18).

Labov (1966, p. 170), por ejemplo, sostiene que un número ideal de individuos para la muestra es de 25 por cada 100 000. No obstante, como señalan Torruella y Llisterri (1999),

> en algunas ocasiones es difícil poder aplicar las fórmulas de extracción de muestras porque es muy complejo (a veces imposible) delimitar el total de la población y, además, en el caso de que ésta pueda ser delimitada, siempre habrá alguna característica de la población que no se habrá tenido en cuenta o no estará representada adecuadamente por las muestras. (Torruella y Llisterri, 1999, p. 19)

Sin embargo, es necesario apuntar que no siempre tiene por qué ser así. Por ejemplo, sería el caso de un estudio en el que se pretendiera analizar algún fenómeno sobre toda la muestra de las sesiones parlamentarias de un determinado gobierno. En este caso, podría ser posible hacer coincidir la población con la configuración de la muestra, pues hay un registro de todas las intervenciones sucedidas en sede parlamentaria y no hay más que estas. Si este estudio se expandiera a las intervenciones sucedidas en los parlamentos de las diferentes comunidades autónomas, la población y la muestra se incrementarían. Si bien, es un caso bastante particular y no suele ser lo habitual, ya que generalmente se opta por el muestro precisamente porque no puede analizarse todo el material posible si antes no se han establecido unos mínimos.

Respecto de los requisitos sobre **la recogida de las grabaciones**, son varios los aspectos que se han de contemplar. En primer lugar, se debe considerar el marco temporal en el que estas se recogerán, es decir, cabe plantearse si estas van a obtenerse en un periodo concreto, con una motivación sincrónica o si, por el contrario, se precisaría de grabaciones recogidas en diferentes momentos para así realizar estudios comparados. En este caso, además, debe plantearse la relación entre coste y beneficio. Como señalan Adolphs y Knight (2010, p. 41), dado que la construcción de corpus hablados es muy costosa hay que plantearse cuestiones como la rentabilidad, esto es, valorar las ventajas e inconveniente de capturar grandes cantidades de datos (en términos de tiempo, número de encuentros o contextos discursivos),

decidir la cantidad de detalles que se van a añadir durante la fase de transcripción y anotación, y la naturaleza de los análisis que podrían generarse a partir de las grabaciones.

En segundo lugar, cabe plantearse el grado de calidad de las grabaciones que, como se puede entender, dependerá de nuevo del objetivo de estudio y de los parámetros contemplados en el proceso de concepción del corpus y la toma de decisiones. Así, para estudios de carácter fonético y prosódico, idealmente los materiales deberán recogerse en entornos de laboratorio, o una localización lo más parecida posible; no obstante, este tipo de análisis entraría en conflicto si desea estudiarse en el marco de interacción de la conversación coloquial ya que esta no sucedería de manera natural en un entorno como el descrito. Es más, para su caracterización es fundamental que se desarrolle en entornos de familiaridad para los interlocutores (Briz, 2001). Además, en el caso de la conversación espontánea, deben valorarse también las características del entorno de grabación para poder tener un control mínimo sobre la calidad de la muestra; a saber, deberán elegirse preferiblemente lugares que sean cerrados, para evitar la aparición de ruidos de fondo procedentes del tráfico, de una cafetería, por ejemplo.

En tercer lugar, la consecución de las grabaciones para la construcción de un corpus oral conlleva la recopilación de una amplia variedad de materiales orales que, como hemos venido observando desde el Capítulo 2, pasaría por grabaciones de muy diversa índole: entrevistas, conversación espontánea, llamadas telefónicas, discursos y debates políticos y parlamentarios, transmisiones en medios de comunicación y redes sociales, etc. Todos estos son materiales susceptibles de formar parte de un corpus oral, siempre y cuando se sometan a una metodología de trabajo que cumpla con el diseño y los objetivos de un corpus, como hemos comentado en 2.1.

La particularidad de la recogida de conversación coloquial radica en que las condiciones de grabación no son controladas ya que, para garantizar la naturalidad de los datos, estas muestras se recogen sin que los hablantes sepan que están siendo grabados[22].

Como se apuntaba en la sección 2.1., hoy día para recoger materiales orales se cuenta con la facilidad derivada de los avances informáticos, sin embargo, sin un procesamiento que conlleve la transcripción, la codificación y, a ser posible, la alineación de audio y texto, no podríamos hablar de corpus orales.

[22] No obstante, para la obtención de los consentimientos informados de grabación véase § 4.2.1.2.

La consecución de los materiales orales conlleva, así mismo, la recogida de otro tipo de materiales asociados, como son la información contextual de la grabación y los permisos de uso de dichas grabaciones.

En cuanto a **la información contextual**, si bien dependerá del objeto de estudio particular que se pretenda realizar con cada corpus, es muy recomendable que, junto a la recolección de las grabaciones, se recopile además la mayor cantidad de datos sobre las muestras de habla grabadas en cuanto a las circunstancias de la grabación y los metadatos de los participantes. En términos de Adolphs y Knight (2010),

> It is therefore advisable to strive to collect data which is as accurate and exhaustive as it can be, capturing as much information from the content and context of the discursive environment as possible (Strassel and Cole 2006: 3). This involves documenting information about the participants, the location and the overall context in which the event takes place, as well as about the type of recording equipment that is being used, and the technical and physical specifications that are being applied to the recording itself. This is because the loss or omission of data cannot be easily rectified at a later date, since real-life communication cannot be authentically rehearsed and replicated. (Adolphs y Knight, 2010, p. 41)

A pesar del valor que ofrece la información contextual que acompaña a los datos orales, no siempre esta se recopila y se ofrece a los usuarios de los corpus. Especialmente, se hace difícil recuperar esta información cuando se trata de corpus que se proponen recoger grandes muestras de datos, más frecuentemente proveniente de datos en línea. Si pensamos, por ejemplo, en el *Corpus del Español* de Mark Davies, este material no ofrece acceso a los datos sobre los hablantes (piénsese en información sobre la edad o el nivel de estudios) o sobre el origen geográfico del material, ya que con este método de recolección automatizado solamente puede saberse, en todo caso, el lugar donde se publicó ese texto, no así el lugar de procedencia de la persona que lo produjo. Lo mismo sucede con el material obtenido de grabaciones en redes sociales, salvo que pertenezca a personalidades públicas cuyas circunstancias vitales sean conocidas, no podremos recuperar esa información contextual.

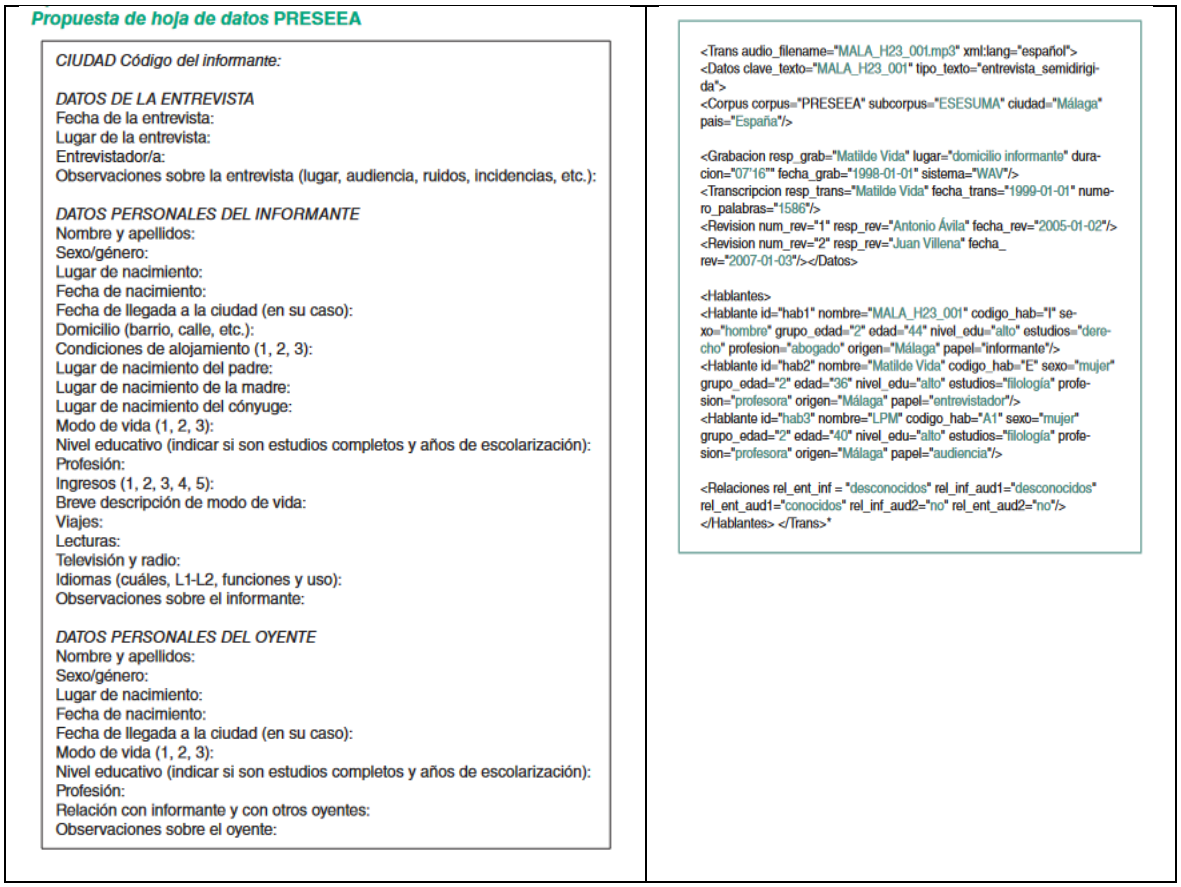
En esta misma línea, los metadatos recogidos para los archivos que forman el CORPES XXI, también presentan muchas veces limitaciones, esto es, los materiales orales recogidos para el corpus académico se seleccionan según criterios como el origen geográfico de los datos y según su fuente de procedencia, sin embargo, cuando aparecen hablantes que no pertenecen a la esfera pública, −volvemos a encontrarnos con la misma casuística con respecto a la dificultad de recoger información sobre los participantes. Cierto es que como

corpus de referencia no incluye criterios como la representatividad sociolingüística, aunque en su esquema de codificación sí estén contemplados estos datos, como puede verse en la Figura 3 más abajo. Por tanto, el investigador debe tener estos condicionantes en cuenta, o bien, mencionarlos explícitamente como limitaciones de su estudio.

En los casos en que esta información contextual sí pueda obtenerse, será posible recogerla en distintos formatos, ya sea a través de la elaboración de una ficha técnica, como sucede en los corpus Val.Es.Co. y Ameresco (§ 5.2.1.4.), que después se procesa para filtrar las búsquedas en el motor electrónico, bien directamente, con la recogida de dichos metadatos durante el proceso de codificación, como sucede, por ejemplo, en CORPES XXI. Como puede verse en la Figura 3, la cabecera electrónica en formato XML (ver 4.2.2.4.) de un archivo oral contiene aquella información relevante sobre el vídeo que se ha transcrito y codificado como, por ejemplo, la fecha y el lugar de emisión, el lugar de grabación, el título del archivo, su duración y, en último lugar, la información que se ha podido reunir de la hablante que participa en el vídeo. En el caso de la Figura 4 puede verse, a la izquierda, un modelo de ficha técnica, en particular del corpus PRESEEA, que recoge el investigador cuando realiza una grabación; y, a la derecha, el formato final que adquieren estos metadatos en lenguaje XML.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CORPES SYSTEM "file:/C:/DTD/CORPESXXI_ORAL.dtd">
<CORPES id="OR2012_0014">
    <cabecera fecha_electrónica="2017-08-15">
        <título_principal autor_título_principal="lainformacion.com">'Cazadores' de murciélagos en
Lavapiés</título_principal>
        <edición procedencia="Transcripción_y_codificación_CORPES" subcorpus=""
archivo_fuente_tipo="vídeo" archivo_fuente_localización="https://youtu.be/LuU7SLTGfIc?
list=PLCW3TCaPXbl-KT2rd6DhDKg36tfYepFDj" lugar_grabación="Madrid" fecha_de_grabación=""
fecha_de_emisión="2012-08-13" fecha_de_transcripción="2017-08-15" sonido_alineado="No"/>
        <numpal n="831"/>
        <duración minutos="05" segundos="26"/>
        <criterio_clasificación_CORPES criterio="Fecha_de_emisión" año="2012"/>
        <clasificación_textual medio_difusión="Internet" tipología="Reportajes_y_documentales"/>
        <hablante hb="001" nombre="Barboza Márquez, Kathrin" sexo="mujer" grupo_edad="20-34"
edad="29" nivel_edu="superior" estudios="Biología" profesión="bióloga" ciudad_origen="Cochabamba"
país="Bolivia" zona="Andina" origen="A" otros_datos="" papel="presentadora"/>
```

Figura 3. Ejemplo de cabecera con los metadatos de un archivo de CORPES



Propuesta de hoja de datos PRESEEA

CIUDAD Código del informante:

DATOS DE LA ENTREVISTA
Fecha de la entrevista:
Lugar de la entrevista:
Entrevistador/a:
Observaciones sobre la entrevista (lugar, audiencia, ruidos, incidencias, etc.):

DATOS PERSONALES DEL INFORMANTE
Nombre y apellidos:
Sexo/género:
Lugar de nacimiento:
Fecha de nacimiento:
Fecha de llegada a la ciudad (en su caso):
Domicilio (barrio, calle, etc.):
Condiciones de alojamiento (1, 2, 3):
Lugar de nacimiento del padre:
Lugar de nacimiento de la madre:
Lugar de nacimiento del cónyuge:
Modo de vida (1, 2, 3):
Nivel educativo (indicar si son estudios completos y años de escolarización):
Profesión:
Ingresos (1, 2, 3, 4, 5):
Breve descripción de modo de vida:
Viajes:
Lecturas:
Televisión y radio:
Idiomas (cuáles, L1-L2, funciones y uso):
Observaciones sobre el informante:

DATOS PERSONALES DEL OYENTE
Nombre y apellidos:
Sexo/género:
Lugar de nacimiento:
Fecha de nacimiento:
Fecha de llegada a la ciudad (en su caso):
Modo de vida (1, 2, 3):
Nivel educativo (indicar si son estudios completos y años de escolarización):
Profesión:
Relación con informante y con otros oyentes:
Observaciones sobre el oyente:

<Trans audio_filename="MALA_H23_001.mp3" xml:lang="español">
<Datos clave_texto="MALA_H23_001" tipo_texto="entrevista_semidirigida">
<Corpus corpus="PRESEEA" subcorpus="ESESUMA" ciudad="Málaga" pais="España"/>

<Grabacion resp_grab="Matilde Vida" lugar="domicilio informante" duracion="07'16"" fecha_grab="1998-01-01" sistema="WAV"/>
<Transcripcion resp_trans="Matilde Vida" fecha_trans="1999-01-01" numero_palabras="1586"/>
<Revision num_rev="1" resp_rev="Antonio Ávila" fecha_rev="2005-01-02"/>
<Revision num_rev="2" resp_rev="Juan Villena" fecha_rev="2007-01-03"/></Datos>

<Hablantes>
<Hablante id="hab1" nombre="MALA_H23_001" codigo_hab="I" sexo="hombre" grupo_edad="2" edad="44" nivel_edu="alto" estudios="derecho" profesion="abogado" origen="Málaga" papel="informante"/>
<Hablante id="hab2" nombre="Matilde Vida" codigo_hab="E" sexo="mujer" grupo_edad="2" edad="36" nivel_edu="alto" estudios="filología" profesion="profesora" origen="Málaga" papel="entrevistador"/>
<Hablante id="hab3" nombre="LPM" codigo_hab="A1" sexo="mujer" grupo_edad="2" edad="40" nivel_edu="alto" estudios="filología" profesion="profesora" origen="Málaga" papel="audiencia"/>

<Relaciones rel_ent_inf = "desconocidos" rel_inf_aud1="desconocidos" rel_ent_aud1="conocidos" rel_inf_aud2="no" rel_ent_aud2="no"/>
</Hablantes> </Trans>"

Figura 4. Ficha técnica y cabecera electrónica de corpus PRESEEA (Moreno Fernández, 2021b)

En el caso de las conversaciones coloquiales grabadas secretamente, conviene recoger esta información justo después de terminar la grabación, ya que, debido a las premisas establecidas por la ley de protección de datos y el derecho a la intimidad, se descarta la posibilidad de volver a contactar con los informantes para obtener o rectificar sus datos (Carcelén, en prensa, § 4.2.1.2.). Además, para los estudios de carácter pragmático, en los que es habitual el análisis de la conversación, se hace necesario contar con el marco interaccional y la información del resto de datos situacionales para determinar el grado de coloquialidad de una interacción. Por ejemplo, para analizar fenómenos de atenuación en una conversación conflictiva entre dos personas será fundamental conocer el tipo de relación que une a los participantes.

Respecto al parámetro referido a **las cuestiones éticas y legales**, en la fase de recogida de datos se abordan cuestiones como la recopilación del material oral en sí, de las fichas técnicas que acompañan a cada grabación y/o a la recogida de la información contextual que

sea relevante, como se ha comentado arriba. Así mismo, también se requiere la obtención de los permisos necesarios en cada caso, bien sean por la obtención de la cesión de los derechos de autor —*copyright*— de los materiales, o bien a través del consentimiento informado de los hablantes.

Como señalan Torruella y Llisterri (1999, p. 24), la transcripción de textos orales registrados de un medio de comunicación (radio, televisión) está sujeta a la normativa vigente en materia de derechos de autor. Dado que el trabajo de Torruella y Llisterri cuenta ya con casi 25 años, en la actualidad deben considerarse en este punto también aquellos materiales procedentes de redes sociales en las que se publica material audiovisual, como Youtube, Tik Tok, Instagram o Twitch.

Por su parte, Baker, Hardie y McEnery (2006) se refieren al *copyright* como

> the right to publish and sell literary, musical or artistic work. Corpus compilers need to observe copyright law by ensuring that they seek permission from the relevant copyright holders to include particular texts. This can often be a difficult and time-consuming process as copyright ownership is not always clear –some texts are owned by the publisher, while others are owned by the author. If the corpus is likely to be made publicly available, copyright holders may require a fee for allowing their text(s) to be included, particularly if the corpus is believed to hold commercial value. Kennedy (1998:77) notes that if permission is sought, many copyright holders are willing to facilitate genuine research by allowing their texts to appear in a corpus. (Baker, Hardie y McEnery, 2006, p. 48)

Es decir, de acuerdo con los autores, los compiladores de corpus deben respetar la legislación sobre derechos de autor y solicitar permiso a los titulares de los derechos para incluir determinados textos.

Así mismo, de acuerdo con Llisterri (en línea), si se constituye un corpus para la difusión pública o para la explotación comercial es preciso tener en cuenta que los materiales originales pueden estar sujetos a derechos de autor, regulados por las leyes relativas a la propiedad intelectual. En el caso de España, opera la *Ley de Propiedad Intelectual* 1/1996, de 12 de abril, y sus posteriores modificaciones. Como señala esta ley en su artículo 2:

> A efectos de la presente Ley, y sin perjuicio de lo dispuesto en el apartado anterior, se consideran bases de datos las colecciones de obras, de datos, o de otros elementos independientes dispuestos de manera sistemática o metódica y accesibles individualmente por medios electrónicos o de otra forma.

Si nos fijamos en los corpus académicos como el CREA o el CORPES XXI en el caso del español, para salvaguardar los derechos de autor se han firmado convenios con plataformas de radio y televisión, como se detalló en la sección 3.1.

En el caso de la difusión de grabaciones que no proceden de medios de comunicación, o de redes sociales, como en las conversaciones coloquiales, sin embargo, se requiere el permiso por escrito de los hablantes, "obtenido en general con posterioridad a la realización de las mismas para no restar espontaneidad al intercambio comunicativo. Es necesario también proteger la intimidad de las personas, cambiando, por ejemplo, sus nombres por iniciales" (Torruella y Llisterri, 1999, p. 24).

Por lo tanto, si se recogen grabaciones realizadas *ex profeso* para su posterior análisis, no hablaríamos de derechos de autor, sino que las cuestiones ético-legales que entran en juego aquí tienen que ver con el derecho a la intimidad y a la protección de datos. Esto sucede, como veremos a continuación, en el caso de los corpus orales de conversación espontánea grabados secretamente, ya que el hecho de solicitar el consentimiento de las personas seleccionadas para la muestra conlleva una serie de implicaciones ético-legales que no se presentan, o se dan en menor medida, cuando se recopila otro género discursivo. De hecho, con la legislación actual, hay una obligatoriedad de obtener el consentimiento de manera previa, y no posterior, como indican los autores previamente mencionados.

Dado que el eje principal de esta investigación lo constituye la metodología para la construcción del corpus Ameresco, centrado en la recolección de conversación coloquial espontánea grabada secretamente, se hace necesario un desarrollo exhaustivo de las implicaciones éticas derivadas de la recogida de este género discursivo. Como se verá en la siguiente sección, se ha detectado una falta sistemática de explicitación metodológica sobre esta cuestión en los corpus existentes (Carcelén, en prensa) y, por lo tanto, ha sido una de las principales preocupaciones metodológicas en la fase de recogida del corpus Ameresco.

En las líneas que siguen veremos cuál es la particularidad de este género discursivo y cómo debe aplicarse la legislación vigente para garantizar una investigación ética y responsable.

4.2.1.2. La particularidad de los corpus de conversación espontánea grabados secretamente: el consentimiento informado

La singularidad de este tipo de discurso radica en que, para llevar a cabo una investigación lingüística lo más precisa e imparcial posible, los hablantes no deben tener conocimiento de que están siendo grabados. Esto se debe a la *paradoja del observador laboviana* (Labov, 1983), que plantea que el investigador debe encontrar formas de observar cómo una persona habla cuando no está siendo sistemáticamente vigilada. No obstante, para obtener estos datos, la observación sistemática es necesaria. Lo ideal sería disponer de muestras de habla naturales en un entorno de grabación donde los hablantes no fueran conscientes de que están siendo grabados. Sin embargo, desde una perspectiva ética, esto podría plantear problemas legales, ya que entraría en conflicto con la legislación que protege el derecho a la intimidad, la propia imagen y la protección de datos personales y derechos digitales.

Si nos centramos en los corpus orales construidos en España, estas directivas son la Ley Orgánica 1/1982, de 5 de mayo, de *Protección civil al derecho al honor, la intimidad personal y familiar y a la propia imagen*, la Ley Orgánica 3/2018, de 5 de diciembre, de *Protección de datos personales y garantía de los derechos digitales*, y el Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 sobre la protección de datos personales y la libre circulación de estos datos, que deroga la Directiva 95/46/CE sobre la protección de datos. No obstante, cada territorio deberá estudiar la situación legislativa particular que debe contemplarse a la hora de diseñar cada corpus[23].

Por lo tanto, antes de emprender cualquier investigación que involucre la utilización de grabaciones de conversaciones espontáneas, es esencial establecer una metodología que defina cómo se llevarán a cabo las tareas de recopilación y cómo se realizará el proceso de anonimización de los datos para evitar la identificación de los interlocutores. También es necesario diseñar un modelo de consentimiento informado que los participantes deberán firmar para poder participar. Los investigadores e investigadoras deberán garantizar el respeto de los principios éticos que promuevan un manejo responsable de estos datos durante la recolección de corpus (Adolphs y Knight, 2010, McEnery y Hardie 2011, Schneider, 2018, Childs *et al*., 2011, Carcelén, en prensa).

[23] Para esta investigación hemos tenido que acotar al territorio español, ya que un estudio de la legislación a nivel internacional se escaparía de los límites de este trabajo.

Para lograr este último fin, existen dos posibilidades. La primera opción, como señala Schneider (2018, p. 53), consiste en que los participantes sepan que están siendo grabados, ya que "in some studies, it is, however, reported that participants tend to forget about being recorded and behave increasingly naturally the longer the recording takes and the speech event lasts, feeling particularly at ease in familiar situations and among friends".

Sin embargo, no existen investigaciones sistemáticas que respalden de manera objetiva esta impresión. Se ha supuesto que los hablantes eventualmente olvidan la presencia de los dispositivos de grabación, pero no se ha verificado mediante corpus de grabaciones secretas si esto es cierto, y en caso de serlo, en qué momentos de la grabación ocurre que dejen de ser conscientes.

La segunda opción implica el uso de un modelo de consentimiento informado que permita obtener una muestra lingüística natural y al mismo tiempo cumpla con la normativa legal. No existe ningún modelo *a priori* ni están establecidas formalmente las características de este tipo de documento. En la sección 5.2.1.2., se expone y explica el consentimiento informado que se diseñó en el año 2015 para el proyecto Ameresco (Briz *et. al*., 2019, Carcelén y Uclés, 2019, Carcelén, en prensa), y que se ha ido actualizando, de acuerdo con las nuevas necesidades y en función de la legislación.

Como ya se ha dicho, recopilar un corpus es una tarea que se ha vuelto más accesible gracias a los avances tecnológicos y a las nuevas herramientas informáticas. Estas herramientas proporcionan a los investigadores una abundante fuente de datos que pueden utilizar como base para construir su propio corpus lingüístico. La tarea resulta relativamente sencilla cuando se trata de investigar géneros discursivos que son principalmente textuales y no orales. Además, si se trata de géneros orales, el proceso es más productivo si las grabaciones provienen de medios públicos como la televisión, la radio o las redes sociales, o cuando se trabaja con géneros discursivos que implican un mayor grado de planificación, como entrevistas o conversaciones semidirigidas ya que el control sobre las condiciones de grabación es prácticamente completo.

Sin embargo, cuando se plantea la creación de corpus orales a partir de conversaciones espontáneas grabadas secretamente, la situación se vuelve más compleja. Esto se debe a que no es posible planificar previamente las condiciones de la grabación y, además, la grabación secreta requiere la implementación de un protocolo de consentimiento informado diferente al utilizado en otros géneros discursivos. El acto de grabar a alguien sin su conocimiento y

divulgar el contenido de la grabación, incluso con fines de investigación, puede conllevar, como se ha visto, problemas legales que resulten en sanciones penales.

Rock (2001) recopila las opiniones de diversos investigadores, creadores y usuarios de corpus a nivel internacional sobre este tema, con el objetivo de examinar las diversas actitudes y prácticas que existen en la actualidad en relación con cuestiones éticas relacionadas con el tratamiento de datos y la anonimización. A partir de sus investigaciones, se destaca que, si bien en cualquier campo de investigación o práctica profesional existen métodos de anonimización válidos, en el ámbito de la lingüística no parece haber un estándar establecido (Rock, 2001, p. 2).

Por otro lado, McEnery y Hardie (2011, p. 60) hacen referencia al hecho de que, si bien hay información disponible sobre aspectos legales relacionados con la creación de corpus, no se ha prestado suficiente atención a las cuestiones éticas involucradas en estos trabajos. Esto es evidente en los primeros corpus orales, cronológicamente hablando, que fueron creados en un momento en el que no existían los requisitos legales actuales y no se consideraban aspectos como la invasión de la privacidad de los participantes. Por ejemplo, en estos primeros corpus orales en el caso del español, se solicitaba el consentimiento de los hablantes, pero la forma en que se obtenía variaba según el género discursivo. Por su parte, en el género de la entrevista semidirigida, como sucede en el corpus PRESEEA, obtener autorización previa a la grabación no suponía un problema ya que los participantes son conscientes de que están siendo grabados. Sin embargo, en otros casos, como las conversaciones coloquiales del corpus Val.Es.Co. (2002), de carácter espontáneo y, en principio, con grabación secreta, solicitar autorización previa podía afectar a las condiciones de producción y, por ende, a las características fundamentales y a la autenticidad de la conversación. En estos casos, solo se buscaba la autorización después de realizar la grabación.

Por tanto, aunque en la recogida de los primeros corpus se tomaron medidas éticas para la recopilación y el manejo de la información en los primeros corpus orales, estas decisiones se tomaron sin una reflexión más profunda sobre las motivaciones detrás de las mismas. Hoy en día puede observarse, sin embargo, que como también reconocen McEnery y Hardie (2011, p. 68), muchos de los errores cometidos en el pasado en la lingüística de corpus en relación con estas cuestiones éticas se han corregido.

Así pues, en la actualidad existe un consenso generalizado en la literatura respecto al hecho de que la recopilación de corpus orales debe incluir necesariamente la obtención de la autorización y la firma del consentimiento informado por parte de los hablantes. Además, es obligatorio proteger tanto a los participantes como los datos que proporcionan a la investigación mediante un sistema de anonimización que puede ser más o menos complejo, dependiendo de las características del corpus y la naturaleza de la investigación. En consecuencia, se deben establecer diferentes estándares en función del tipo de datos recolectados, ya sean exclusivamente auditivos, incluyan contenido multimedia, etc. (Rock, 2001, Adolphs y Knight, 2010, Childs *et al*., 2011, McEnery y Hardie, 2011, Schneider, 2018, D'Arcy y Bender, 2023, Carcelén, en prensa).

Centrándonos exclusivamente en el caso de la normativa que opera en España, citada anteriormente, el primer documento que afecta a la recogida de datos es la Ley Orgánica 1/1982, de 5 de mayo, que establece en su artículo 1 los derechos fundamentales del honor, la intimidad personal y familiar, así como la propia imagen, los cuales están reconocidos por la Constitución. Por lo tanto, cualquier intrusión no autorizada en este ámbito que no cuente con una autorización explícita de la ley o no haya obtenido un consentimiento claro por parte de la persona afectada sería considerada como un acto punible (artículo 2.2). Conforme al artículo 7 de esta legislación de 1982, se define una *intromisión ilegítima* como:

> 1. El emplazamiento en cualquier lugar de aparatos de escucha, de filmación, de dispositivos ópticos o de cualquier otro medio apto para grabar o reproducir la vida íntima de las personas.
>
> 2. La utilización de aparatos de escucha, dispositivos ópticos, o de cualquier otro medio para el conocimiento de la vida íntima de las personas o de manifestaciones o cartas privadas no destinadas a quien haga uso de tales medios, así como su grabación, registro o reproducción. […].

Es importante destacar que la recolección de muestras de conversaciones informales espontáneas, preferiblemente obtenidas de manera secreta para evitar influir en el comportamiento lingüístico de los participantes, implica necesariamente el uso de dispositivos de grabación en contextos físicos. De acuerdo con lo mencionado anteriormente, esta práctica podría potencialmente violar la privacidad de las personas. Además, es relevante mencionar que la grabación de una conversación sin el consentimiento de las personas involucradas puede constituir un delito grave contra la privacidad según el Código Penal (artículo 197). Si bien, puede considerarse legal en situaciones específicas, como cuando la persona que realiza la grabación es una participante activa en la conversación y

esta grabación podría utilizarse como evidencia en un proceso judicial, en cualquier otro escenario, como la grabación de conversaciones ajenas o conversaciones en las que la persona que graba apenas participa, y especialmente si se comparte el contenido con terceros, se considera un delito grave contra la privacidad. Si lo trasladamos a la recogida de conversaciones del corpus Ameresco, por ejemplo, en la que 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 (Carcelén y Uclés, 2019, p. 23).

se estaría atentando contra las disposiciones del Código Penal señaladas más arriba y, por tanto, de no contar con el consentimiento previo, la recogida del corpus contravendría los principios éticos y legales que se deben cumplir.

Por otro lado, la recogida de estas grabaciones lleva consigo la recolección de datos personales que indefectiblemente van a aparecer en las conversaciones. La conversación coloquial sucede en un marco de interacción familiar, entre íntimos y conocidos, con una temática cotidiana (Briz, 1995) y, por tanto, la aparición de datos sensibles o que pudieran dar lugar a la identificación de los hablantes es esperable. El segundo texto mencionado arriba, el Reglamento Europeo 2016/679 y su aplicación en España mediante la Ley Orgánica 3/2018 dan cuenta de cómo debe gestionarse el tratamiento de datos personales.

Según ambas directrices, los principios de la protección de datos deben aplicarse a toda información relativa a una persona física identificada o identificable, ya estén los datos pseudonimizados o anonimizados[24]. Por tanto, la clave está en adoptar medidas para que no se pueda llegar a saber de qué persona se trata:

> Los datos personales seudonimizados, que cabría atribuir a una persona física mediante la utilización de información adicional, deben considerarse información sobre una persona física

[24] Cabe señalar que el término *anonimización de datos* en el ámbito legal difiere en cierta medida del concepto de anonimización en el campo de la lingüística. En la lingüística de corpus, se utiliza el término *anonimización de datos* para describir el proceso mediante el cual se elimina o se reemplaza con un término ficticio o una codificación cualquier dato que pueda llevar a la identificación del hablante (Rock, 2001, Childs *et al*., 2011, Adolphs y Knight, 2010, McEnery y Hardie, 2011, Schneider, 2018).

Por otro lado, desde una perspectiva legal, el proceso de *anonimización* implica la eliminación completa de los datos personales, lo que impide cualquier posibilidad de reidentificación. En cambio, la *pseudonimización* implica el uso de un sistema de encriptación basado en archivos que, en casos específicos contemplados por la Ley 3/2018, podría permitir la reidentificación del usuario. Un ejemplo de esto se da en una investigación orientada a encontrar una cura para una enfermedad, donde, después de realizar los ensayos necesarios, los resultados podrían mejorar la calidad de vida del paciente involucrado en el estudio.

> identificable. Para determinar si una persona física es identificable, deben tenerse en cuenta todos los medios, como la singularización, que razonablemente pueda utilizar el responsable del tratamiento o cualquier otra persona para identificar directa o indirectamente a la persona física. Para determinar si existe una probabilidad razonable de que se utilicen medios para identificar a una persona física, deben tenerse en cuenta todos los factores objetivos, como los costes y el tiempo necesarios para la identificación, teniendo en cuenta tanto la tecnología disponible en el momento del tratamiento como los avances tecnológicos. Por lo tanto, los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación. (Reglamento UE 2016/679, apartado 26).

Respecto de los datos anonimizados, se entiende que

> son anonimizados si se han eliminado todos los elementos identificativos de un conjunto de datos personales. No puede dejarse en la información elementos que podrían, ejerciendo un esfuerzo razonable, servir para volver a identificar a la(s) persona(s) de que se trate. (Agencia de los Derechos Fundamentales de la Unión Europea 2014)

Aunque el Reglamento establece que los datos que han sido efectivamente anonimizados ya no se consideran datos personales y, por lo tanto, no se les aplican los principios de protección de datos (lo que significa que podrían divulgarse públicamente), en el caso de los corpus orales, se requiere la implementación de medidas restrictivas adicionales para evitar problemas legales. Esto se debe a que la anonimización completa no es factible, especialmente si el corpus se planea como un recurso de acceso público. En este caso, se tendría acceso al audio de las grabaciones, lo que implica la voz, un atributo directamente identificable de cada individuo.

Además, es esencial prestar atención a que se deben considerar *todos los factores objetivos* como se menciona en el Reglamento Europeo arriba, teniendo en cuenta tanto la tecnología disponible en el momento del procesamiento como los avances tecnológicos. En otras palabras, dado que el desarrollo de herramientas informáticas para el tratamiento de sonido y voz está avanzando rápidamente y se está volviendo más accesible incluso para equipos domésticos, se debe ejercer una precaución adicional de cara a proyecciones futuras en las que se deberán implementar medidas paliativas, como por ejemplo, el establecimiento

de un sistema de anonimización del audio, por medio del borrado del segmento o la inclusión de una señal sonora que sustituya al fragmento original.

Por otra parte, retomando el concepto de *consentimiento informado*, es fundamental que los informantes otorguen su autorización de manera voluntaria, informada y sin ambigüedades para el uso de los datos recopilados, incluyendo los detalles específicos en los casos que involucren a menores de edad (Reglamento 2016/679 apartados 32 y 42 y Ley 3/2018 artículos 6 y 7). En la creación de este consentimiento, la regulación establece el *principio de transparencia* (Reglamento 2016/679, 39, Ley 3/2018 art. 11). Esto significa que se debe informar a los informantes de manera clara y comprensible, que la información debe ser de fácil acceso y estar redactada en un lenguaje claro y sencillo acerca de los propósitos específicos y legítimos para los cuales se utilizarán sus datos. Además, se designa a un responsable del tratamiento de los datos que "estará obligado a informar al afectado sobre los medios a su disposición para ejercer los derechos que le corresponden. Los medios deberán ser fácilmente accesibles para el afectado" (Ley 3/2018, artículo 12). A saber, la persona afectada debe estar al tanto de que tiene la opción de retirar su consentimiento en cualquier momento y debe conocer a quién o a qué entidad puede dirigirse para ejercer su derecho a hacerlo.

La normativa mencionada es aplicable en todos los casos que involucran el manejo de datos personales, abarcando desde grabaciones de cámaras de seguridad hasta archivos comerciales. Sin embargo, surge la pregunta de si la ley contempla alguna especificidad cuando se trata de la investigación científica. En relación con esto, el Reglamento Europeo 2016/679, en su apartado 159, establece que su alcance también se extiende al tratamiento de datos personales con fines de investigación científica en un sentido amplio. Esto engloba el desarrollo tecnológico y la demostración, así como la investigación básica, la investigación aplicada y la investigación financiada por el sector privado. Se presta especial atención a las investigaciones en el ámbito de la salud, especialmente en el campo biomédico. De hecho, la Ley 3/2018 contiene disposiciones adicionales que detallan específicamente las características del tratamiento de datos para la investigación en el ámbito de la salud, como se establece en las disposiciones adicionales sexta y decimoséptima.

A partir de lo anterior, se derivan dos aspectos importantes. En primer lugar, carecemos hoy por hoy de legislación específica que regule las condiciones para la recopilación de datos en la investigación en general, posiblemente debido a la diversidad de objetivos que puede

tener este tipo de investigación. La regulación se enfoca principalmente en el ámbito de las ciencias de la salud, donde se manejan datos altamente confidenciales de los pacientes y se requiere un proceso riguroso de anonimización y almacenamiento de datos. Sin embargo, este proceso no es necesario en el caso de los corpus orales espontáneos.

En segundo lugar, se enfatiza la importancia de que el consentimiento firmado por los sujetos sea claro y transparente en cuanto a la finalidad de la recopilación de datos. Esto implica que se deben especificar con precisión los aspectos que son de particular interés para el grupo de investigación, de modo que los informantes tengan un conocimiento completo de la razón por la cual se están recopilando sus datos.

Sin embargo, nos encontramos en una paradoja en este punto: detallar el objeto de investigación, que en nuestro caso se relaciona con aspectos lingüísticos, puede inducir al hablante a modificar considerablemente su forma de expresarse, ya que se vuelve consciente de que su manera de hablar es el enfoque de interés en la investigación y, por lo tanto, su propio discurso pasa a ocupar el centro de su atención. En este contexto, el protocolo que se ha desarrollado para el corpus Ameresco (ver § 5.2.1.2., Briz *et al*. 2019, Carcelén y Uclés, 2019, Carcelén, en prensa) busca mitigar el impacto de esta situación al introducir un periodo de tiempo entre la obtención del consentimiento y la grabación efectiva. Este lapso de tiempo permite que el hablante olvide la grabación y, con ello, que deje de concentrarse en la forma en que se comunica.

En resumen, al trabajar con datos personales, ya sea en el ámbito de la investigación científica o en cualquier otro contexto (como el comercial, administrativo, financiero, publicitario, etc.), es esencial, en primer lugar, (a) obtener el consentimiento informado previo por escrito de los usuarios, informantes o hablantes en el caso de la recopilación de corpus orales, y (b) someter estos datos a un proceso de anonimización que evite su reidentificación.

El proceso de anonimización aquí mencionado se tratará en la sección 5.2.2.3. ya que consideramos que forma parte del tratamiento de los datos al requerir una intervención técnica e informática sobre el material.

4.2.2. Fase 2. Tratamiento de los datos: Transcripción, codificación y anotación

En este apartado veremos que, para el procesamiento informático de los materiales de un corpus, es indispensable el tratamiento de los datos de manera que estos puedan ser aprovechados al máximo de cara a su explotación posterior. Este tratamiento del material sonoro pasa por diferentes capas, desde la transcripción y la codificación al anotado; no obstante, la terminología empleada en la bibliografía no establece claramente las fronteras entre ellas, como analizamos a continuación.

4.2.2.1. Transliteración, transcripción, codificación, anotación, etiquetado…: un mapa de fronteras difusas

En la revisión bibliográfica realizada en esta investigación se ha observado la utilización de una variedad de términos para referirse a las distintas capas necesarias en el tratamiento de los datos recopilados, materiales que en un primer momento se encuentran en formato de audio y que, para su posterior análisis y explotación pasan por su transformación al medio escrito para hacer posible el tratamiento informático. Las fronteras entre transliteración, transcripción, codificación, etiquetado, marcado y anotación no terminan de estar claras.

Si bien, el hecho de pasar un material obtenido del medio oral, al medio escrito puede considerarse *codificar* en un sentido general, tal y como recoge el *Diccionario de la Lengua Española,* implica "registrar algo siguiendo un código (‖ combinación de letras, números u otros caracteres)"; en términos de lingüística de corpus adquiere un sentido específico, como veremos más abajo. Podría pensarse que el simple hecho de transcribir ya supone una codificación por el cambio de código que opera, sin embargo, *transcripción* y *codificación* son considerados procesos diferentes en la bibliografía.

Partiendo del primer nivel de representación, parece que existe cierta unanimidad al diferenciar la *transliteración* de la *transcripción* —que podría, por ejemplo, ser también fonética (Hidalgo y Sanmartín, 2005, Llisterri, 2021)—. Además, Llisterri (2021, p. 166) señala el concepto de *transcripción ortográfica enriquecida* para referirse a una representación ortográfica en la que se señalan algunos elementos relacionados con la oralidad. Torruella y Llisterri (1999, p. 27) exponen que para la gestión de proyectos como los corpus de referencia se han creado "una serie de convenciones para la transcripción ortográfica —a veces denominada transliteración—de la lengua oral", lo que viene a

demostrar la vacilación terminológica. En otros casos (Hidalgo y Sanmartín, 2005), se señala la transliteración ortográfica como método diferenciado de la transcripción fonética y la codificación, y que consiste en la representación de la oralidad siguiendo las convenciones ortográficas normativas.

No obstante, no siempre las fronteras están claras ya que, si nos fijamos en cómo definen diversos corpus actuales sus sistemas de representación de la oralidad, podemos ver que, por ejemplo, PRESEEA (Moreno Fernández, 2021b, p. 5) se autodenomina *transcripción ortográfica enriquecida,* pero esta sigue una ortografía normativa convencional, incluida la acentuación, sin intentar reproducir fenómenos orales, como elisiones, aspiraciones, diptongaciones orales o sinalefas y además, se reconstruyen las palabras pronunciadas incompletas (Moreno Fernández, 2021b, p. 12); o el corpus COSER, que también la denomina *transcripción enriquecida*, sin embargo, siguen las convenciones de la ortografía habitual, incluyendo puntuación, pero también otro tipo de marcas (COSER, en línea, sección *Transcripción ortográfica*), por tanto, podemos decir que este camino se encuentra a medio camino entre la transliteración y la transcripción enriquecida.

No parece existir consenso con los términos *codificar*, *etiquetar* o *marcar*. Por mencionar algunos ejemplos, Adolphs y Knight (2010, p. 40) y Thompson (2004, p. 74) diferencian *transcription*, *coding* y *mark-up* (transcripción, codificación y marcado). De acuerdo con Hincapié y Bernal (2018, p. 57), el proceso de codificación implica la conversión de caracteres del lenguaje natural a un lenguaje que se pueda procesar por medio de máquinas o sistemas que se valen de programas computacionales. Así lo señalan también Baker, Hardie y McEnery (2006), si bien puede comprobarse la indefinición terminológica cuando refieren que a veces se denomina *anotación*, *etiquetado* o *marcado*:

> Encoding is usually the last of the five stages of corpus compilation, and is sometimes referred to as annotation, tagging or markup. Encoding is a way of representing elements in texts such as paragraph breaks, utterance boundaries etc. in a standardised way across a corpus, so that they can be more easily recognised by computer software and by corpus users. (Baker, Hardie y McEnery, 2006, p. 66)

Para Adolphs y Knight (2010), la *codificación*

> is essentially a development of the transcription stage, providing further detail to the basic systems of annotation and mark-up applied through the use of transcription notation. The coding stage thus operates at a higher level of abstraction compared to the transcription stage,

> and may include, among others, annotation of grammatical, semantic, pragmatic or discoursal features or categories. Coding is a key part of the process of annotating language resources. (Adolphs y Knight, 2010, p. 47)

Para estos últimos autores, la codificación implica un escalón más en la transcripción y en ella podrían incluirse también tareas de anotación.

Respecto de la anotación, parece haber mayor consenso sobre su definición, ya que la mayoría de los autores (entre otros McEnery, Xiao y Tono, 2006, Baker, Hardie y McEnery, 2006, Rojo, 2021) coinciden en señalar que este proceso constituye una fase lingüística, frente a la codificación que está relacionada con lo computacional y que, así mismo, añade información adicional. De hecho, Hincapié y Bernal (2018, p. 20) opinan que la anotación no es una característica primordial de un corpus, puesto que existen corpus no anotados o planos; no obstante, reconocen que esta información adicional es muy útil ya que permite hacer búsquedas más específicas dentro de los corpus. Estos autores definen la *anotación* como "la inclusión de datos que buscan enriquecer el corpus con información lingüística adicional." Es uno de los tres tipos de información, junto con los metadatos y la codificación, que pueden ayudar a llevar a cabo distintas investigaciones con los datos contenidos en el corpus (McEnery y Hardie, 2011, p. 29).

Baker, Hardie y McEnery (2006, p. 154) refieren, además, el término *tagging* (etiquetado) como "a more informal term for the act of applying additional levels of annotation to corpus data. A tag usually consists of a code, which can be attached to a phoneme, morpheme, word, phrase or longer stretch of text in a number of ways".

De acuerdo con lo señalado, entendemos que *etiquetado* sería un sinónimo de *anotación*, hecho que podría entrar en conflicto con la parcela de la *codificación* ya que, si atendemos a la definición propuesta por estos autores más arriba para *encoding* o *codificación*, también la incluyen como equivalente para *codificación* y *marcado*.

Tras haber dado cuenta de esta variedad terminológica, hemos optado, como se verá en el siguiente apartado, por adoptar una definición operativa con el único propósito de servir a la precisión léxica de este trabajo; no se pretende, por tanto, establecer una denominación específica para cada una de estas fases.

4.2.2.2. Propuesta de definición operativa: transliteración, transcripción, codificación y anotación

Para esta propuesta de definición operativa de los conceptos señalados hemos partido de tres aspectos. El primero tiene que ver con el cambio de medio o canal, es decir, del paso de la oralidad a la escritura; el segundo pone el foco en el tratamiento informatizado de los datos para su procesamiento; por último, el aspecto referido al añadido de información lingüística al material base.

Así, hablaremos de *transliteración* cuando se opte por reflejar el material oral de manera escrita siguiendo las convenciones gramaticales y ortográficas normativas, incluyendo la puntuación; mientras que la *transcripción* implicaría reflejar por escrito características propias de la oralidad y, si bien puede ser ortográfica (frente a la fonética), no suele utilizar un sistema de puntuación por considerarse que su uso puede transmitir subjetividad por parte de quien transcribe (López Morales, 1997).

Por tanto, en este sentido, podremos hablar de *transcripción fonético-fonológica, transliteración* y *transcripción enriquecida*, esta última utilizada para describir sistemas de transcripción (no fonética), que incorporan ciertos símbolos y convenciones y que no se somete a las normas de la escritura. Tomando como ejemplo un fragmento de audio al que se puede acceder con el código QR que aparece a continuación, en la Figura 5 pueden observarse los tres tipos de representación comentados.

Transliteración: Pero sí, ¿quién sabe? Si yo estuviera en esa posición podría darte otra opinión, pero no. Y lo único que hago con el pelo, porque sí me han dicho que me crece muy rápido. Este… lo tengo muy lacio.

Transcripción fonética (AFI): pe ro 'si 'kjen 'sa βe si yo es tu 'βje ra en 'e sa po si 'θjom po 'ðri a 'ðar te 'o tra o pi 'njom 'pe ro no i lo 'u ni ko ke 'a ɣo kon el 'pe lo 'por ke 'si me an 'di ʧo ke me 'kre θe 'mu i 'ra pi ðo 'es te lo 'teŋ go 'mu i 'la θjo

Transcripción enriquecida (marcas de codificación): pero<alargamiento/>/ sí quién sabe si yo estuviera en esa posición/ podría darte otra opinión/ pero no y<alargamiento/>/ lo único que hago con el<alargamiento/>// pelo porque sí me han dicho que me/ crece muy rápido// este<alargamiento/>/ lo tengo muy lacio

https://nuvol.uv.es/owncloud/index.php/s/kxOZEPUFOu2bcpQ

Figura 5. Diferencias entre transliteración, transcripción fonética y transcripción enriquecida

Esta fase, ya se opte por la transliteración ya por alguna clase de transcripción de las enumeradas anteriormente, es necesaria en cualquier diseño de corpus orales, ya que es el paso mínimo imprescindible que permite trabajar con el material recogido.

Consideramos que el término *codificación* se aplica al hecho de asignar etiquetas, marcas o códigos a elementos específicos en un corpus para facilitar su búsqueda informática y análisis posterior. Por tanto, quedan incluidos en esta definición los términos *etiquetado* y *marcado*. Esta fase será necesaria especialmente si el corpus va a estar disponible en línea y va a contar con un motor de búsqueda que permita filtrar búsquedas atendiendo a parámetros tanto textuales como extratextuales. Esta codificación se realiza de manera semiautomática por medio de diferentes herramientas y lenguajes de marcado.

Por último, con respecto de la *anotación*, se considerará aquí el proceso por el cual se añade información adicional que enriquece los datos, siendo esta de carácter lingüístico. Nos referimos, por lo general, a información de carácter gramatical −morfológica y sintáctica−, semántica o léxica, así como a otro tipo de información pertinente con la finalidad del corpus y de la línea de investigación a la que este responde. Esta fase es opcional; suele realizarse de manera automática a través de *software* específico y de lenguajes de programación. Como puede verse en la Figura 6, tomando como base un fragmento del ejemplo anterior, se ha realizado la tokenización y la anotación morfológica (§ 4.2.2.5.) de manera automática[25]:

[25] Realizada a través del analizador morfológico automático del Laboratorio de herramientas de la Biblioteca Virtual Miguel de Cervantes, disponible en línea: https://data.cervantesvirtual.com/analizador

Palabra	Categoría gramatical	Descripción
Pero	cc	Conjunción (coordinación). Ejemplo: y, o, pero
sí	rg	Adverbio (general). Ejemplos: siempre, más, personalmente
,	fc	Coma (,)
¿	fia	Signo interrogación invertido (¿)
quién	pt000000	Pronombre interrogativo. Ejemplos: cómo, cuánto, qué
sabe	vmip000	Verbo (principal, indicativo, presente). Ejemplos: da, trabajamos
?	fit	Signo interrogación (?)

Figura 6. Ejemplo de tokenización y anotación morfológica automática

4.2.2.3. La transcripción de corpus orales

En lo relativo a la decisión sobre qué sistema de transcripción se va a adoptar para trasladar los datos orales a la escritura, habrá que determinar, en primer lugar, los eventos de la oralidad que se quieren reflejar y en qué grado de granularidad se va a realizar, decidir hasta qué punto es pertinente ser más o menos exhaustivo de acuerdo con la finalidad del corpus, etc.

A este respecto, en la literatura (Ochs, 1979, Du Bois, 1991, Payrató, 1995, Torruella y Llisterri, 1999, Du Bois, 2015, Du Bois *et al*., 2015) se señalan una serie de requisitos que se deben cumplir para obtener un sistema de transcripción idóneo. Así, Du Bois (1991, pp. 78-80) establece una serie de máximas agrupadas en cinco categorías: (i) *category definition*, se deben definir categorías suficientemente explícitas y generales; esto tiene que ver no tanto con qué simbología se elige para transcribir el discurso, sino a qué categorías analíticas van a representar dichos símbolos, qué fenómenos discursivos se están representando (ii) *accesibility*, son preferibles los símbolos familiares, motivados, fáciles de aprender, que maximicen el acceso a los datos; (iii) *robustness*, esto es, se deben emplear caracteres disponibles ampliamente; (iv) *economy*, se debe procurar hacer representaciones económicas, evitando las etiquetas largas, y permitiendo que estas sean discriminables, es decir, suprimibles y que no entorpezcan la lectura; y (v) *adaptability*, eso es, ya que es imposible para el diseño de un sistema de transcripción de uso general anticipar todo lo que los usuarios querrán hacer con él, hay que considerar que el sistema debe diseñarse con anticipación para permitir a los usuarios introducir sus propias categorías de transcripción.

Retomando los presupuestos de Du Bois, Payrató (1995, p. 52) establece los requisitos que teórica e idealmente deberían exigirse a una transcripción, a saber: *neutralidad o fidelidad* (que la transcripción no sea interpretativa), *globalidad o complejidad* (que recoja todos los fenómenos del discurso), *omnifuncionalidad* (que permita diversos usos y aplicaciones), *claridad* (para el aprendizaje y legibilidad de lo representado), y *universalidad y compatibilidad* entre sistemas informáticos. Sin embargo, señala que estos principios no dejan de ser un ideal y que, en realidad, un sistema de transcripción debe preocuparse de ser fácil de usar y, por tanto, será interpretativo de los datos —si bien, toda transcripción conlleva la subjetividad del propio transcriptor[26]—, selectivo en cuanto a los fenómenos que transcribe, pertinente para el objeto de investigación, coherente con la base teórica adoptada por el investigador. Que sea fiel en cuanto a la representación de los datos, flexible para que sea posible su utilización en diversos estudios, que emplee una simbología clara, económica, sencilla, sin ambigüedad y compatible con sistemas internacionales estandarizados (Payrató 1995, p. 53-54) son máximas a tener en cuenta, en cualquier caso. Por tanto, el corpus ideal será aquel que apueste por la estandarización, permita el intercambio de los datos, permita la reutilización de información, favorezca la legibilidad y apueste por la universalidad.

Sin embargo, hay que contemplar que la transcripción de la oralidad es un proceso detallado y complejo que implica convertir el habla en forma escrita y que, por tanto, conlleva el sometimiento de la gramática de la oralidad a la gramática de la escritura. A saber, plasmar fenómenos propios de la oralidad en la escritura se convierte en una tarea ardua y difícil, más si se pretende que no se pierda información. Adolphs y Knight (2010) reflejan esta problemática cuando dicen que

> one of the biggest challenges in corpus linguistic research is probably the representation of spoken data. There is no doubt that the collection of spoken language is far more laborious than the collection of written samples, but the richness of this type of data can make the extra effort worthwhile. […] However, the representation of spoken data is a major issue in this context as the recorded conversations have to undergo a transition from the spoken mode to the written before they can be included in a corpus. In transcribing spoken discourse we have to make various choices as to the amount of detail we wish to include in the written record. (Adolphs y Knight, 2010, p. 44)

[26] En las tareas de revisión del corpus Ameresco se ha podido comprobar esta afirmación ya que, ante la misma grabación, a pesar de que las directrices y el sistema de transcripción es común, se han registrado transcripciones que diferían en algunos segmentos.

Cuando se plantea cómo se va a transcribir un corpus oral, hay que tener en cuenta diferentes factores que influirán indefectiblemente en el modelo elegido, partiendo de las decisiones tomadas en los planteamientos previos de diseño mencionados al comienzo de este capítulo. Cabe señalar, entre otros, qué género discursivo se va a transcribir, qué variedad dialectal, qué nivel de detalle se quiere reflejar en la transcripción, con qué materiales se va a contar, esto es, si estará alineado con el audio, si no lo estará, si el audio estará disponible en abierto, qué tipo de corpus estamos construyendo, etc., de acuerdo con la toma de decisiones que se realiza en la primera fase de diseño de un corpus y que se ha comentado en la sección 4.2.1. La reflexión sobre cada uno de estos aspectos conllevará la toma de determinadas decisiones que cristalizarán en la metodología concreta de trabajo de un corpus oral. Sin duda, como han señalado diversos autores a este respecto (Ochs, 1979, Payrató, 1995, Hidalgo y Sanmartín, 2005, Vázquez y Recalde, 2009, Adolphs y Knight, 2010, Briz, 2012, Rojo 2021, Llisterri, 2021, entre otros), el condicionante principal será el objetivo o finalidad para la que se construye el corpus.

Respecto del sistema de transcripción, Rojo (2021, p. 65) señala cómo "sin necesidad de optar directamente por una transcripción fonética o fonológica, el uso de un sistema basado en la ortografía convencional plantea todos los problemas relacionados con qué tratamiento debe darse a la pronunciación". Se refiere el autor a las distintas realizaciones que suceden en la lengua oral, como por ejemplo al caso de la perdida de la *-d-* intervocálica y la manera en qué debería representarse. Es decir,

> la fidelidad a la pronunciación supone el problema de la fijación de límites en el detalle fonético y las dificultades de reflejar diferentes fonéticas en un sistema que no está diseñado para ese fin, además de complicar la recuperación de la información. (Rojo, 2021, p. 65)

Por tanto, además de la configuración inicial del corpus y de sus objetivos, se deben contemplar otros problemas posibles, como por ejemplo "cómo reflejar la entonación, así como en fenómenos del tipo de las palabras cortadas, las repeticiones de alargamientos o los solapamientos en las intervenciones de distintos participantes" (Rojo, 2021, p. 93). Para este autor, sería interesante dejar constancia de todo aquello que es propio de la lengua hablada y que no se refleja o lo hace de forma muy imperfecta en la lengua escrita; sin embargo, esta fidelidad a la oralidad chocaría con la economía del sistema, señalada por Payrató (1995) y Du Bois (1991, 2015), y con el procesamiento de los datos textuales, ya que cuantas más indicaciones sobre estos se introduzcan en la transcripción, más complicado será el procesamiento del texto y su posterior recuperación por medio de un motor de búsqueda.

Para Rojo (2021, p. 93) también sería ideal que se pudieran vincular sonido y transcripción, y más concretamente, si se alinearan secuencias cortas de transcripción con el fragmento de audio correspondiente.

Estos planteamientos previos a la construcción de un corpus oral han llevado a que autores como Briz (2012a) señalen los déficits encontrados en los corpus orales debido a la dificultad de recoger género oral de manera general, y de manera particular, a lo que sucede cuando se trata de género conversacional espontáneo. Según este autor,

> en principio, cualquier sistema de transcripción es adecuado siempre que se ajuste al objeto de estudio y a la finalidad para la que se emplee y, por supuesto, cumpla los principios de exhaustividad y pertinencia de los signos […]. Conviene, no obstante, que el sistema presente otra característica, la adaptabilidad, esto es, que sea flexible en cuanto a su capacidad de estrecharse (introducir otros signos) o de ensancharse (eliminar algunos) en función de los objetivos más o menos concretos que puedan surgir. (Briz, 2012a, p. 128).

Por tanto, una vez que se establece el objetivo y el género discursivo con el que se va a trabajar, el tercer condicionante tendría que ver con el tipo de transcripción que se va a adoptar atendiendo, justamente, a la finalidad y al material oral seleccionado (§ 4.2.2.3.):

> Así, la transcripción, por ejemplo, de una conversación no puede ser solo ortográfica (o solo podría serlo para ciertos fines), ya que los signos ortográficos no pueden dar cuenta de muchos fenómenos de la lengua hablada (pensemos solo en el caso del habla simultánea, en las anotaciones prosódicas). Por otro lado, los datos obtenidos de ciertos corpus orales no se entienden en ocasiones por la falta de enriquecimiento contextual, lo que puede inducir a errores de interpretación. Como afirmábamos, un corpus oral transcrito sin prosodia, sin al menos algunas indicaciones prosódicas, es como un texto sin voz, lo cual es una paradoja difícil de superar. Claro que ello no puede suplir ni el audio ni el vídeo. Es obvio que el mejor sistema en este sentido será el que pueda ofrecer alineados, transcripción con audio (e, incluso, vídeo). (Briz, 2012a, p. 129)

Y, es más, como señalan Vázquez y Recalde (2009),

> incluso usando un sistema de transcripción que trate de reflejar ciertas peculiaridades del discurso oral, especialmente del conversacional, la fijación gráfica del habla implica transformar un proceso dinámico en un producto textual estático, implica atribuir secuencialidad a lo simultáneo (proxémico-gestual, paraverbal, suprasegmental), e inevitablemente conlleva perder de vista muchos de los elementos comunicativos presentes en el habla. (Vázquez y Recalde, 2009, p. 60)

Briz (2012a, p. 130) aboga por "una doble transcripción, que permitiera un tratamiento informático a través de marcas y que tendiera también a la posibilidad de su lectura y fácil comprensión". Adolphs y Knight (2010) señalan la dificultad de alcanzar este equilibrio entre exhaustividad y comprensibilidad:

> Once decisions have been taken as to the features which are to be transcribed, and the level of granularity and detail of the information to be included, the next step is to decide on an appropriate layout of the transcription. There are many different possibilities for laying out a transcript, but it is important to acknowledge that 'there will always be something of a tension between validity and ease of reading' (Graddol et al. 1994: 185). The most commonly used format is still a linear representation of turns with varying degrees of detail in terms of overlapping speech, prosody and extra-linguistic information. (Adolphs y Knight. 2010, p. 45)

A partir de las dificultades señaladas arriba, se han propuesto múltiples modelos de transcripción de materiales orales. Como venimos señalando en este capítulo, uno de los condicionantes a la hora de seleccionar el tipo de transcripción que se practicará en la construcción de corpus orales tiene que ver con el objetivo y finalidad del corpus en sí. Atendiendo a esto, la bibliografía recoge tres tipos principales de transcripción utilizados mayoritariamente, como se señaló anteriormente: (a) transcripción fonética y fonológica; (b) transliteración y (c) transcripción enriquecida.

(a) Transcripción fonética y fonológica

La transcripción fonética de un corpus consiste en representar las unidades sonoras del habla (fonemas) a través de un sistema de símbolos. Esta puede realizarse atendiendo a dos niveles: segmental y suprasegmental. El propósito de la transcripción fonética es capturar de manera precisa cómo se pronuncia un segmento de habla, como puede observarse en el ejemplo de la Figura 7. A este tipo de transcripción suele recurrirse para la construcción de aquellos corpus orales centrados en la investigación del habla, en estudios de adquisición de lenguaje o el desarrollo de tecnologías de síntesis y reconocimiento de voz, principalmente.

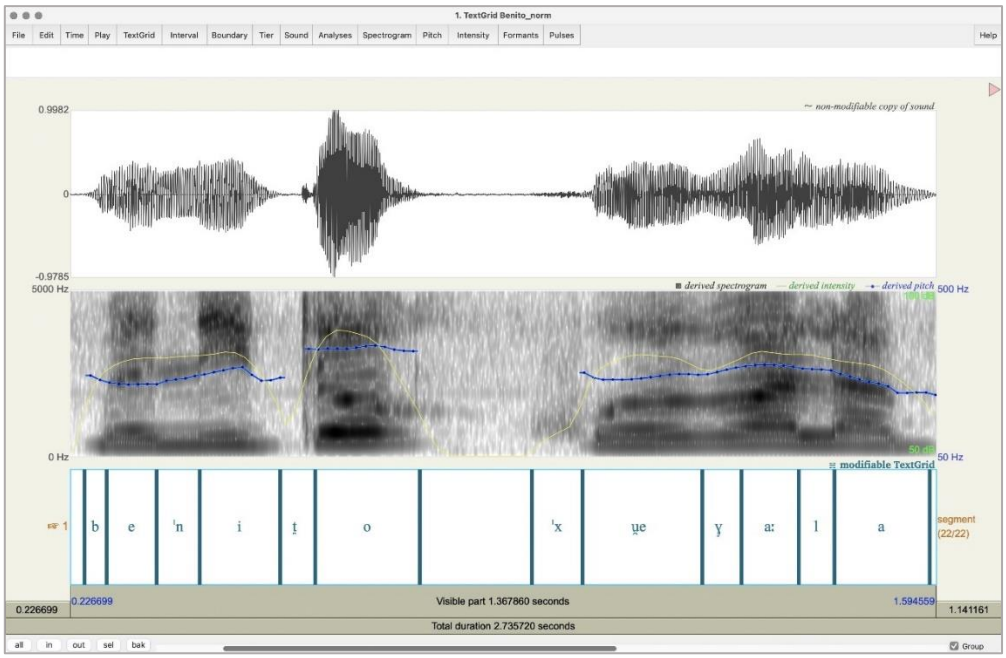
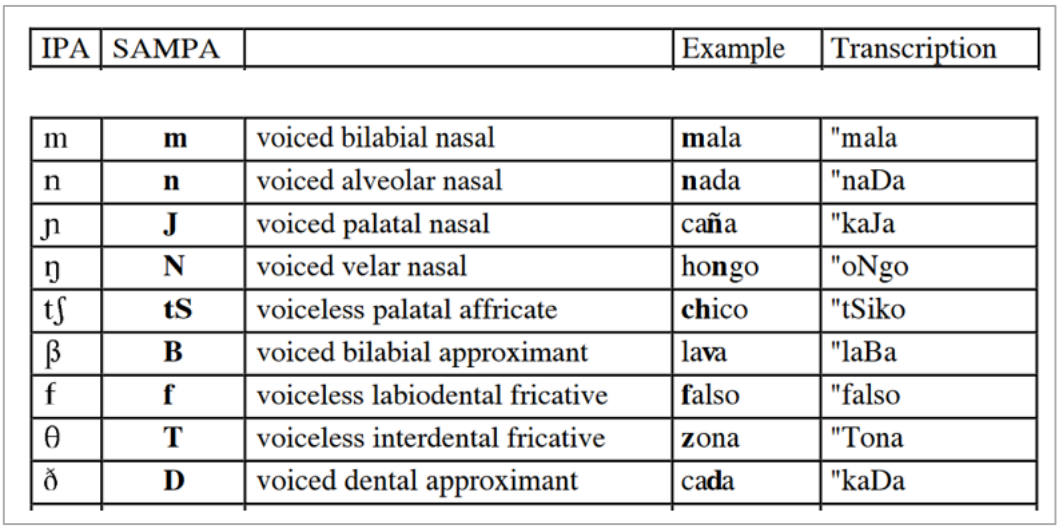


Figura 7. Ejemplo de segmentación y etiquetado fonético segmental del principio del enunciado «Benito juega a la petanca». Fuente: Llisterri, (en línea)

Para la transcripción de la fonética segmental, según Torruella y Llisterri (1999, p. 27) suele recomendarse el uso del Alfabeto Fonético Internacional (AFI). Este se construye utilizando un conjunto de caracteres tomados de diferentes alfabetos, como el griego, el latino o el árabe, entre otros. La ventaja del AFI reside en que posibilita representar gráficamente cualquier lengua y constituye un estándar de transcripción. Si bien es un sistema muy preciso para representar todos los sonidos del habla, la simbología empleada, esto es, los caracteres, no es legible en un lenguaje informatizado.

Debido a estos problemas de compatibilidad para la representación del nivel segmental surge SAMPA, *Sam Phonetic Alphabet* (Wells, 1997), que utiliza los símbolos presentes en un teclado convencional y ha sido adaptado a buena parte de las lenguas europeas (Torruella y Llisterri, 1999, p. 27). Este sistema constituye un intento muy significativo y extendido de desarrollo de un modelo de transcripción fonética informatizado. Se basa en un conjunto de equivalencias de los símbolos AFI con códigos ASCII (*American Standard Code for Information Interchange*) que opera sobre el alfabeto latino facilitando la lectura de las transcripciones a nivel computacional. Este código combina caracteres alfanuméricos y caracteres especiales que permiten su procesamiento informático, como puede verse en la

Figura 8. SAMPA ha tenido una gran difusión entre los desarrollos de tecnologías del habla realizados en Europa[27].



IPA	SAMPA		Example	Transcription
m	**m**	voiced bilabial nasal	**m**ala	"mala
n	**n**	voiced alveolar nasal	**n**ada	"naDa
ɲ	**J**	voiced palatal nasal	ca**ñ**a	"kaJa
ŋ	**N**	voiced velar nasal	ho**n**go	"oNgo
tʃ	**tS**	voiceless palatal affricate	**ch**ico	"tSiko
β	**B**	voiced bilabial approximant	la**v**a	"laBa
f	**f**	voiceless labiodental fricative	**f**also	"falso
θ	**T**	voiceless interdental fricative	**z**ona	"Tona
ð	**D**	voiced dental approximant	ca**d**a	"kaDa

Figura 8. Adaptación española del sistema internacional de transcripción fonética y del SAMPA. Fuente: Llisterri y Mariño (1993)

Lo mismo sucede en lo que respecta a la transcripción de los elementos suprasegmentales, para la cual surgen sistemas de codificación de la entonación, como los señalados a continuación, según Torruella y Llisterri:

> además del conjunto de símbolos del AFI, se dispone también de varios sistemas que pueden ser utilizados en corpus en soporte electrónico […]. Entre los más difundidos cabe citar ToBI (*Tone and Break Index*) (Silverman et al., 1991), SAMPROSA (*SAM Prosodic Alphabet*), desarrollado en el marco del proyecto SAM (Gibbon, 1989) e INTSINT (*International Transcription System for Intonation*) (Hirst et al., 1994). (Torruella y Llisterri, 1999, p. 27)

No obstante, como veremos más adelante,

> se han desarrollado también sistemas de anotación prosódica de corpus orales ortográficamente transcritos en los que suelen marcarse pausas, unidades tonales, cambios de intensidad, de rango melódico o de velocidad de elocución o bien sílabas acentuadas, sílabas prominentes no acentuadas y movimientos tonales. (Torruella y Llisterri, 1999, p. 27)

Este tipo de ampliaciones y especificaciones en los sistemas fonéticos y prosódicos también ha sido realizado para otro tipo de símbolos y sistemas de representación, como sucede con las convenciones TEI y las propuestas por el proyecto NERC, así como con

[27] X-SAMPA es una aplicación más reciente de SAMPA.

determinadas propuestas derivadas del Análisis Conversacional. Se refieren estos autores a los sistemas de transcripción enriquecida, como veremos más adelante.

El *Tone and Break Index* o ToBi (Silverman *et al*., 1991), al que hacen referencia Torruella y Llisterri (1999), es una notación utilizada en lingüística y análisis del discurso para representar y analizar la prosodia y la estructura entonativa del habla; se utiliza principalmente en la transcripción fonética de discursos orales para resaltar elementos como la entonación, los tonos, las pausas y las relaciones entre las unidades de discurso. Hidalgo y Sanmartín (2005, p. 17) señalan su mérito por permitir la representación por separado de las unidades prosódicas y del fenómeno entonativo. Por su parte, SAMPROSA, *SAM Prosodic Alphabet*, (Wells, 1995), la versión suprasegmental de SAMPA, conforma uno de los sistemas más extendidos para la transcripción prosódica (Hidalgo y Sanmartín, 2005, p. 16, Torruella y Llisterri, 1999, p. 27) precisamente por su base ASCII que garantiza una máxima compatibilidad y legibilidad por las computadoras, lo mismo que el menos extendido WORLDBET (Hieronymus, 1994 y 1997), también con código ASCII, de los laboratorios Bell, que goza de relativo prestigio en los Estados Unidos o el método INTSINT*, International Transcription System for Intonation* (Hirst *et al*., 1994), que tiene como objetivo proporcionar un sistema para la intercomparación lingüística de sistemas prosódicos y que permite la transcripción de la curva melódica de forma automática (Llisterri, 1999, p. 73, Llisterri *et al*., 2005, p. 310).

(b) Transliteración

En este caso, se realiza un volcado a la escritura sin incluir ninguna clase de marca o codificación, generalmente ajustándola a las normas ortográficas y gramaticales de la lengua, incluyendo puntuación. En principio, un corpus que utilice este sistema estaría más interesado en estudiar el contenido de la palabra, sus concordancias, por tanto, su finalidad tendría más que ver con el análisis léxico, y no tanto con fenómenos fonéticos o pragmáticos. Por tanto, la transliteración ortográfica puede ser suficiente según los objetivos de cada corpus. Como señalan Hidalgo y Sanmartín (2005, pp. 18-19), si atendemos a la finalidad particular de cada corpus, puede darse el caso de corpus que no pretendan en origen representar fenómenos lingüísticos (fonéticos, prosódicos, morfológicos ni léxicos). Es el caso, por ejemplo, del *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*.

Este tipo de transcripción presenta una serie de inconvenientes, tal y como señalan Torruella y Llisterri (1999):

> plantea diversos problemas entre los que se cuentan las variaciones en la pronunciación no recogidas en los diccionarios normativos, el uso de los signos de puntuación y la representación de siglas, abreviaturas, palabras deletreadas o secuencias numéricas. (Torruella y Llisterri, 1999, p. 27).

Además de estas cuestiones, añaden otras como la falta de representación de

> los elementos propios de la lengua oral, las pausas, la delimitación de los enunciados y de las unidades tonales, las variaciones en los elementos suprasegmentales, los elementos vocales tanto semi-léxicos (por ejemplo, las denominadas 'pausas llenas') como no léxicos (por ejemplo, risas o toses), los cambios de turno de palabra, las intervenciones simultáneas de varios hablantes o las dudas, palabras truncadas, repeticiones y errores de producción corregidos o no por el propio hablante (Torruella y Llisterri, 1999, p. 27).

Por tanto, esa estandarización de la lengua que sucede al seguir un sistema ortográfico normativo deja de lado la realidad de la lengua oral y el conjunto de realizaciones posibles propios de este medio, que, además, responden a cuestiones de variación dialectal, diastrática y diafásica. Además, se debe tener en cuenta que, la adaptación de un texto oral al medio escrito supone la puesta en entredicho de la objetividad de la muestra ya que supone la interpretación por parte del transcriptor (López Morales, 1997) como decíamos en el apartado 4.2.2.2.

La principal ventaja de este tipo de transcripciones es la facilidad de lectura, como puede observarse en la Figura 9.

MA-1. Hombre de 25 años. Estudiante de Ingeniería

Enc.- ¿Me puedes hablar un poco de tus estudios?

Inf.- Pues sí, que estudio tercero actualmente de ingeniero técnico de Obras Públicas, como se ha venido llamando, pero que actualmente no sabemos cómo... cómo lo llamamos. Termino este año, si es que se aclara esta situación, ya que actualmente estamos en huelga pendientes de unas reuniones de todos los colegios de las... escuelas técnicas y de arquitectos técnicos.

Enc.- ¿Me puedes hablar de lo que piensas hacer en el futuro con tu carrera?

Inf.- Eso actualmente no lo puedo, no lo puedo señalar. Creo que depende de muchísimos factores. Mi intención es trabajar en España, pero si existen... grandes dificultades para ello pues... procuraré buscar otros horizontes; también mi intención es hacer el acceso a la escuela superior como es Caminos en este caso. Para ello, pues necesito encontrar un trabajo apropiado para... para realizar este acceso, al mismo tiempo que me permita una situación económica desahogada, ¿no? Entonces pues... mi trabajo pues es relacionado con Obras Públicas, pero al mismo tiempo puede desarrollarse también en una empresa privada, no solamente pública, ya que existen varias empresas en España hoy día que se dedican a obras públicas privadamente. (Cantarero y Esgueva 1981: 87)

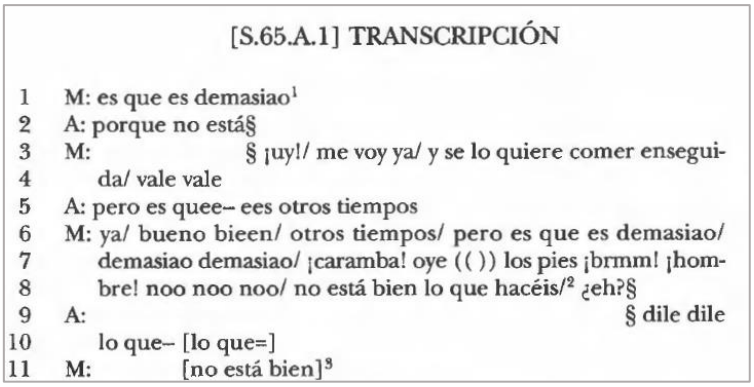
Figura 9. Ejemplo de transliteración ortográfica del MC-NC. Fuente: Samper, Hernández y Troya, (1998)

En la línea de Adolphs y Knight (2010) y Briz (2012a), se privilegia la legibilidad sobre la exhaustividad. Sin embargo, el sometimiento de la escritura a la gramática no deja paso al reflejo de las características propias de la oralidad como hemos mencionado arriba.

(c) Transcripción enriquecida

Llisterri (2021, p. 166) utiliza esta denominación para referirse a una representación en la que se señalan algunos elementos relacionados con la oralidad. Una transliteración como la vista más arriba no da cuenta de la verdadera naturaleza de una secuencia fónica, por esta razón, pueden darse casos en los que se añadan marcas semiortográficas que representen gráficamente aspectos como la entonación, los solapamientos o los alargamientos, así como la introducción de marcas de tiempo (Rojo, 2021, p. 74).

Un ejemplo de corpus que optó por este sistema de transcripción es el del corpus Val.Es.Co. (Briz y Grupo Val.Es.Co., 1995 y 2002), publicados los dos primeros volúmenes en papel originariamente. El corpus Val.Es.Co. adoptó un sistema que refleja en el escrito, niveles propios de la oralidad, como el nivel suprasegmental y el nivel discursivo conversacional (Hidalgo y Sanmartín, 2005, p. 26). Esto se realizó incluyendo una marca para señalar los tonemas, así como otra serie de signos ortográficos convencionales que representan fenómenos como los alargamientos, los reinicios o el habla simultánea, como puede observarse en la Figura 10. En este caso, como señalan Pons y Gurillo (2005, p. 253) no se adoptó un método de transcripción informatizado —o informatizable— de acuerdo con otros modelos existentes como las normas TEI y el lenguaje de marcado SGML, en congruencia con los objetivos de investigación iniciales de este corpus. De acuerdo con Pons y Gurillo (2005, p. 253), "animados por el principio de comprensibilidad de la transcripción", se optó "por un método que, siendo máximamente informativo, dificultara menos la comprensión del texto". Es, por tanto, una simbología clara, económica, sencilla y exenta de ambigüedad que combina recursos ortográficos y tipográficos.



[S.65.A.1] TRANSCRIPCIÓN

1 M: es que es demasiao[1]
2 A: porque no está§
3 M: § ¡uy!/ me voy ya/ y se lo quiere comer enseguí-
4 da/ vale vale
5 A: pero es quee-- ees otros tiempos
6 M: ya/ bueno bieen/ otros tiempos/ pero es que es demasiao/
7 demasiao demasiao/ ¡caramba! oye (( )) los pies ¡brmm! ¡hom-
8 bre! noo noo noo/ no está bien lo que hacéis/[2] ¿eh?§
9 A: § dile dile
10 lo que– [lo que=]
11 M: [no está bien][3]

Figura 10. Ejemplo de transcripción Val.Es.Co. Fuente: Briz y Grupo Val.Es.Co., (2002)

El inconveniente que presenta este sistema, y que se ha tratado de subsanar en el corpus Ameresco como veremos en la sección 5.2.2., es que su apuesta por la legibilidad ha jugado en contra de las posibilidades para su tratamiento informatizado.[28] No obstante, se contemplan dos estándares, como veremos en el Capítulo 5, por un lado el relativo a la transcripción y codificación del material oral, necesariamente codificado para su procesamiento informático; por otro, un sistema de transcripción ancha que prime la legibilidad de cara a la difusión de resultados de investigación.

Hoy en día, de cara al procesado informático de los corpus, un procedimiento como la transliteración resultaría insuficiente ya que no permite su tratamiento informático, si bien es un sistema válido si así se necesita para los objetivos de investigación particulares. Aunque sería ideal poder ofrecer las tres opciones (transcripción fonética-fonológica, transliteración y transcripción enriquecida) a los usuarios en plataformas en línea, tanto para su descarga, como para realizar búsquedas selectivas, esta es una tarea inabarcable, principalmente por la falta de recursos y financiación que suele acompañar a la construcción de los corpus. Por lo tanto, es recomendable utilizar, siempre que sea posible, sistemas internacionales y estandarizados, como comentaremos más adelante, así como ofrecer al usuario/investigador las facilidades para que pueda transportar los materiales de un corpus para su explotación con otras herramientas de análisis, aunque estas no correspondan con los objetivos iniciales del corpus, en línea con lo señalado por Payrató (1995). Estos presupuestos, basados en principios como la reusabilidad de los datos, han guiado la metodología de trabajo del corpus Ameresco, como se detallará en el capítulo siguiente.

[28] En la actualidad, el corpus Val.Es.Co. (versión 3.0) ha sido revisado y reajustado para su tratamiento informático, incorporando un lenguaje de marcado y etiquetado XML.

Tras presentar los principales modelos de transcripción, a continuación, trataremos de mostrar las principales herramientas informáticas utilizadas en esta fase del tratamiento de los datos.

Los primeros corpus se servían de procesadores de texto más o menos sofisticados para plasmar las transcripciones, pero, actualmente existen programas que nos permiten añadir informaciones de manera multicapa, tarea que, además, puede realizarse desde una misma plataforma que permite la reproducción del material sonoro a la vez que se va ejecutando la transcripción.

Uno de los primeros recursos[29] que aparecieron que permitían la alineación del audio y la transcripción fue Transcriber (Boudahmane *et al*., 1998-2008).[30] Supone una herramienta de ayuda a la anotación manual de señales de voz empleada para segmentar grabaciones de voz, transcribirlas y reflejar diferentes condiciones acústicas. Su diseño se creó específicamente para la anotación de grabaciones de noticias, con el fin de crear corpus para el desarrollo de sistemas automáticos de transcripción de este género, pero sus funciones podían resultar útiles en otras áreas de la investigación del habla, como es la lingüística de corpus. Tiene la particularidad de que permite establecer turnos de habla diferentes sobre los cuales pueden reflejarse habla solapada, pero, como señalan Jørgensen y Eguía (2014, p. 10), cuando aparecen más de dos solapamientos, en el caso de más de dos hablantes a la vez, el programa no permite la transcripción de todos los turnos.

El programa ELAN Annotation[31] vino a solucionar este problema. Esta herramienta fue creada por The Language Archive Max Planck (Institute for Psycholinguistics de Nijmegen) y es ampliamente utilizada en lingüística de corpus. Su función principal es permitir la anotación lingüística de archivos de audio o vídeo. Se caracteriza por permitir la anotación de múltiples niveles de datos en un solo archivo, esto es, alineando en el tiempo el material audiovisual o de audio con el material textual, lo que significa que se pueden agregar transcripciones fonéticas, traducciones, anotaciones semánticas, anotaciones kinésicas y marcadores de tiempo, todo dentro de la misma interfaz. Además, incluye un reproductor de medios que permite reproducir el archivo de audio o vídeo mientras se realizan las

[29] Otros de los recursos más conocidos son Transana, desarrollado para ayudar en la transcripción y análisis de datos de vídeo y audio, si bien no es un *software* libre; Atlas.ti, también de pago; y el programa CLAN (*Computerized Language Analysis*) desarrollado en el marco del proyecto CHILDES para la adquisición de lenguaje.

[30] https://trans.sourceforge.net/en/presentation.php

[31] https://archive.mpi.nl/tla/elan

anotaciones, lo que facilita la sincronización precisa de las anotaciones con los segmentos relevantes. Destaca, así mismo, porque puede trabajar con una gran variedad de formatos de archivo, es compatible con WAV, MP3, MP4, AVI, entre otros. Esta compatibilidad se extiende también a otros programas, como por ejemplo Audacity, programa para la edición de audio y vídeo utilizado para realizar las anonimizaciones, entre otras tareas, o Praat, desarrollado para el tratamiento de la señal acústica, ya que ELAN admite la exportación e importación de datos anotados, hecho que facilita la colaboración entre la comunidad científica y el intercambio de datos entre diferentes herramientas y proyectos.

Si bien ELAN se ha convertido en uno de los recursos más importantes, el uso de EXMARaLDA[32] (Schmidt y Wörner, 2014) también es significativo. Esta herramienta agrupa bajo la misma interfaz diferentes aplicaciones. Entre otras, permite la transcripción de audio y vídeo de manera sincronizada y, además, incluye herramientas de edición, de búsqueda y análisis de los datos, incluyendo análisis de frecuencias y patrones lingüísticos. Al igual que ELAN, permite exportar los datos transcritos y codificados en varios formatos.

4.2.2.4. La codificación de corpus orales

Retomando la denominación operativa adoptada en la sección 4.2.2.2., nos referimos con *codificación* al hecho de asignar etiquetas, marcas o códigos a elementos específicos en un corpus para facilitar su búsqueda informática y análisis posterior. Para Rojo, el paso del medio oral al formato escrito implica la toma de una serie de decisiones, pensando en la recuperación posterior de los datos:

> en una conversación habrá que dejar constancia de las características de las personas que participan en ella y suelen ser utilizadas en los estudios sociolingüísticos, edad, sexo y nivel educativo; […] debe tenerse en cuenta también que puede haber hablantes de diferentes procedencias, edades, sexo y nivel educativo, por lo que la aplicación de recuperación de datos tiene que ser capaz de identificar segmentos que correspondan a cada participante en función de sus rasgos. (Rojo, 2021, p. 65)

La codificación, el tratamiento del texto para hacerlo informatizable, puede ser extratextual o intratextual según los datos que se vean afectados. Hablamos de *codificación extratextual* cuando se trata de una codificación no lingüística que incluye datos externos al texto, como son la información bibliográfica del material, de control interno como la fecha

[32] https://exmaralda.org/en/

de incorporación al corpus, la persona o personas responsables de su transcripción, revisión y validación, los datos referentes a las características generales y a las características sociolingüísticas de los hablantes (Rojo, 2021, pp. 97-98); toda esta información extratextual se corresponde con los llamados *metadatos* (ver Figuras 3 y 4 en sección 4.2.1.1.). En cambio, la *codificación intratextual* tiene que ver con la estructura interna del texto: si se trabaja con corpus escritos, se codificarán los capítulos, párrafos, estrofas; si se trata de corpus orales, se etiquetarán las intervenciones.

Retomando los principios establecidos por Du Bois (1991), Payrató (1995) o Torruella y Llisterri (1999), quienes plantean que el diseño del corpus debe garantizar la universalidad y compatibilidad entre sistemas, han surgido, desde diferentes escuelas, propuestas internacionales de estandarización de la codificación. El alto coste que supone codificar y etiquetar, en términos económicos y de tiempo, ha impulsado esta idea de definir estándares para facilitar el intercambio y la reusabilidad de los textos (Torruella y Llisterri, 1999, p. 22).

Una de las iniciativas con mayor impacto en la lingüística de corpus es la *Text Encoding Initiative* (TEI) que nació en un congreso de la Association for Computers and the Humanities (ACH) en el Vassar College en 1987. Es un consorcio que desarrolla colectivamente un estándar para la representación de textos en formato digital. La TEI cuenta con un conjunto de directrices que son usadas para la codificación de textos y su lectura por máquinas y que han sido adaptadas por especialistas en disciplinas como las ciencias humanas, sociales, la lingüística y la informática. Propone un sistema de etiquetas lo suficientemente completo para construir un lenguaje claro de programación. Este sistema establece más de 500 etiquetas en formato XML que, si bien en un alto porcentaje están pensadas para corpus escritos, incluyen una sección dedicada a las etiquetas de lengua hablada. En la Figura 11 se muestra la sección denominada *Elements Unique to Spoken Texts*[33] que da cuenta de fenómenos exclusivos de la oralidad.

<u> (utterance) contains a stretch of speech usually preceded and followed by silence or by a change of speaker.
<pause> (pause) marks a pause either between or within utterances.
<vocal> (vocal) marks any vocalized but not necessarily lexical phenomenon, for example voiced pauses, non-lexical backchannels, etc.

[33] La lista completa de convenciones puede consultarse en https://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html#TSSA

<kinesic> (kinesic) marks any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc.
<incident> (incident) marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication.
<writing> (writing) contains a passage of written text revealed to participants in the course of a spoken text.
<shift> (shift) marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes.

Figura 11. Elementos propios de discursos orales (TEI)

También ofrecen marcas relacionadas con otras particularidades de este medio (Figura 12), como los rasgos prosódicos.

<shift> (shift) marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes.
@feature a paralinguistic feature. Suggested values include: 1] tempo; 2] loud; 3] pitch; 4] tension; 5] rhythm; 6] voice
@new specifies the new state of the paralinguistic feature specified.

Figura 12. Elementos particulares para las marcas prosódicas

Como se aprecia en la Figura 12, estos elementos prosódicos pueden, además, ser subclasificados con otras etiquetas. Así se observa que para el caso de @feature, TEI sugiere una lista de valores susceptibles de ser codificados, entre otros, la velocidad de habla (*tempo*), el volumen (*loud*), los diferentes rangos tonales (*pitch),* el ritmo, o aspectos relacionados con lo que catalogan como *voice (for voice quality)* donde se reflejarían susurros, risas o suspiros, entre otras opciones.

En el ámbito europeo se ha desarrollado la *Network of European Reference Corpora* (NERC), un consorcio integrado por once instituciones representante cada una de ellas de un país miembro de la CEE, cuyo objetivo es la realización de un estudio de viabilidad que proporcione recomendaciones a la CEE sobre el futuro de la provisión de corpus de referencia en Europa (Martín de Santa Olalla Sánchez, 1999, en línea).

Otra propuesta significativa es la del *Expert Advisory Group on Language Engineering Standards* (EAGLES), que propone una transcripción ortográfica de corpus orales agrupando características propias de TEI y NERC. Según Parodi (2008, p. 103), este grupo busca la armonización de los recursos lingüísticos en diferentes lenguas europeas y no pretende, por lo tanto, producir un etiquetario morfosintáctico, sino más bien entregar directrices que ayuden en el desarrollo de uno.

Desde la International Organization for Standardization (ISO) se ha propuesto también el estándar internacional ISO 24624:2016[34] que especifica reglas para representar transcripciones de interacciones habladas grabadas en audio y vídeo en documentos XML basados en las directrices del TEI. Es un recurso de pago.

Estas iniciativas respaldan sus propuestas en determinados lenguajes de marcado, como hemos señalado arriba, también con carácter estándar e internacional. Cabe destacar el *Standard Generalized Markup Language* (SGML) para la definición de la estructura y el contenido de diferentes tipos de documentos que surge en la industria editorial en la década de los 60 para facilitar el intercambio de datos. Aunque SGML sentó las bases para los lenguajes HTML y XML, es un estándar más complejo y menos utilizado en la actualidad. El *HyperText Markup Language* (HTML) es el lenguaje utilizado para crear páginas electrónicas. Este define su estructura y contenido mediante el uso de etiquetas específicas. Se utiliza para marcar el texto, crear enlaces, insertar imágenes, definir encabezados, párrafos y otros elementos estructurales.

Por último, el *eXtensible Markup Language* (XML), se utiliza para estructurar datos de una manera que sea legible tanto para los humanos como para las máquinas. Se utiliza para describir datos y su característica más destacable es que permite un uso flexible, es decir, los mismos usuarios pueden definir sus propias etiquetas y reglas de marcado, lo que lo hace adecuado para una amplia gama de aplicaciones, como pueden ser la representación de datos estructurados en documentos hasta la comunicación entre sistemas de *software* como los empleados para la construcción de corpus orales.

A continuación, explicaremos cómo se estructuran este tipo de documentos utilizando para ello el documento base (Figura 13) para la definición del tipo de documento o DTD extraída de la codificación oral de CORPES XXI (Rojo *et al*., en línea).

[34] https://www.iso.org/obp/ui/en/#iso:std:iso:24624:ed-1:v1:en

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CORPES SYSTEM "file:/C:/DTD/CORPESXXI_ORAL.dtd">
<CORPES id="">
  <cabecera fecha_electrónica="">
    <título_principal autor_título_principal=""></título_principal>
    <edición procedencia="" subcorpus="" archivo_fuente_tipo=""
archivo_fuente_localización="" lugar_grabación="" fecha_de_grabación=""
fecha_de_emisión="" fecha_de_transcripción="" sonido_alineado=""/>
    <numpal n=""/>
    <duración minutos="" segundos=""/>
    <criterio_clasificación_CORPES criterio="" año=""/>
    <clasificación_textual medio_difusión="" tipología=""/>
    <hablante hb="" nombre="" sexo="" grupo_edad="" edad="" nivel_edu="" estudios=""
profesión="" ciudad_origen="" país="" zona="" origen="" otros_datos="" papel=""/>
    <codificación equipo_codificación="" persona_codificación="" fecha_codificación=""/>
    <validación valor_validación="" persona_validación="" fecha_validación=""/>
    <revisión_RAE valor_revisión_RAE="" persona_revisión_RAE=""
fecha_revisión_RAE=""/>
    <notas></notas>
  </cabecera>
  <texto>
    <turno hb="" seg=""></turno>
  </texto>
</CORPES>
```

Figura 13. Esquema de codificación XML CORPES XXI

Este documento organiza la información que se va a codificar en torno a dos ejes mencionados con anterioridad: la codificación extratextual y la codificación intratextual. Para la codificación extratextual, la DTD establece una cabecera (<cabecera></cabecera>) en la que deberán volcarse todos los datos relativos al material que se está codificando (Figura 14). Al archivo concreto que se está codificando, por ejemplo, un fragmento de una entrevista en televisión, deberá otorgársele un código (ID) para la identificación del archivo, se añade la fecha de creación del archivo y, a continuación, se añade la "biografía" o información de carácter técnico del documento, esto es, el título del archivo, el lugar de

procedencia, cuál es el archivo fuente y la localización del mismo, lugar y fecha de grabación, la fecha en que se ha realizado la transcripción y si audio y texto están alineados. Se añaden, además, el número de palabras que contiene, la duración, los criterios de clasificación y textuales[35] de acuerdo con los fijados para este corpus en la fase de concepción del corpus. A continuación, se registran los datos de los participantes que intervienen en esta grabación. El último bloque de la cabecera contiene información relativa a los procesos internos de codificación, revisión y validación por parte de la institución.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CORPES SYSTEM "file:/C:/DTD/CORPESXXI_ORAL.dtd">
<CORPES id="OR2016_0025">
   <cabecera fecha_electrónica="2017-07-02">
      <título_principal autor_título_principal="12tv DoceTV">Café con... Arkano, artista de Rap</título_principal>
      <edición procedencia="Transcripción_y_codificación_CORPES" subcorpus="" archivo_fuente_tipo="vídeo" archivo_fuente_localización="https://youtu.be/YgEdvlF25ng?list=PLCW3TCaPXbl-KT2rd6DhDKg36tfYepFDj" lugar_grabación="Alicante" fecha_de_grabación="" fecha_de_emisión="2016-01-14" fecha_de_transcripción="2017-07-01" sonido_alineado="No"/>
      <numpal n="6692"/>
      <duración minutos="37" segundos="12"/>
      <criterio_clasificación_CORPES criterio="Fecha_de_emisión" año="2016"/>
      <clasificación_textual medio_difusión="Televisión" tipología="Entrevista"/>
      <hablante hb="001" nombre="García, Teddy" sexo="hombre" grupo_edad="no_indicado" edad="No_indicado" nivel_edu="no_indicado" estudios="No_indicado" profesión="presentador y locutor" ciudad_origen="San Pedro de Jujuy" país="Argentina" zona="Río_de_la_Plata" origen="A" otros_datos="reside en Alicante desde el año 2000" papel="entrevistador"/>
      <hablante hb="002" nombre="Rodríguez Godínez, Guillermo" sexo="hombre" grupo_edad="20-34" edad="21" nivel_edu="superior" estudios="Ingeniería Informática" profesión="rapero" ciudad_origen="Alicante" país="España" zona="España" origen="E" otros_datos="conocido como Arkano, todavía no ha terminado Ingeniería Informática" papel="entrevistado"/>
```

[35] Los criterios establecidos por la academia pueden consultarse en el documento *Descripción del sistema de codificación: textos orales*. Disponible en línea: https://www.rae.es/sites/default/files/2020-06/2020_DisyCod_Orales.PDF

```
    <codificación equipo_codificación="" persona_codificación="" fecha_codificación=""/>
    <validación valor_validación="1" persona_validación=" " fecha_validación=""/>
    <revisión_RAE valor_revisión_RAE="" persona_revisión_RAE=""
fecha_revisión_RAE=""/>
    <notas></notas>
  </cabecera>
```

Figura 14. Datos esquema codificación de la cabecera archivo CORPES XXI

Veamos a continuación los datos relativos a la codificación intratextual, marcada por la etiqueta <texto></texto>, como se puede ver en la Figura 15. Este espacio está reservado para la transcripción del audio; en este caso, se establecerán líneas de intervención para los hablantes y podrán marcarse aquellos fenómenos o convenciones que se hayan establecido previamente.

```
<texto>
    <turno hb="001" seg=""><música/> ¿cómo le va / doña Rosa / don Paco? / ¿cómo están? /
¿bien? / hoy estoy así medio <vacilación/> <observación_complementaria desc="hace como que
baila"/> la / la gente joven / la gente joven que uno lo // lo vitaminiza / lo / lo / lo / lo // ¡ah!
<silencio/> gente que tiene ganas de hacer cosas // gente que dice <cita>no me pregunten
cuántos son sino que salgan de a uno</cita> // y van / y van superando obstáculos / y van
mostrándonos nuevos caminos // cosas nuevas // y / y lo hacen tan bien // que inclusive llegan a
ser // eeh / campeones internacionales // de eso nuevo // que nos están mostrando // yo tendría
que comenzar este programa haciendo una especie de rap // pero no soy bueno para eso / pero /
¡ay! <observación_complementaria desc="suspira"/> voy a hacer un papelón / no / no me pidan
que haga un rap porque no / <vacilación/> cuando tome las clases puede ser que lo haga // ¡ah! /
le podría decir / qué se yo // eeh // <cita>estamos en invierno / no esperamos el verano / hoy
tenemos con nosotros / al campeón Arkano</cita> <risas_inicio hb="002"/>¡qué sé yo! /
<transcripción_dudosa>no sé por las dudas</transcripción_dudosa> // ¿cómo le va don
Arkano?<risas_fin hb="002"/> <simultáneo>¿bien?</simultáneo></turno>
    <turno hb="002" seg=""><simultáneo>¡qué</simultáneo> bueno! todo bien / encantado
<simultáneo>de estar aquí / sí</simultáneo></turno>
    <turno hb="001" seg=""><simultáneo>¡qué malo / lo!</simultáneo> ¡qué malo lo mío!
<risa hb="varios"/> ¿no?</turno>
    <turno hb="002" seg="">no / no / oye está bien / ¿eh? / para ser un principio</turno>
    <turno hb="001" seg="">sí / <palabra_cortada>pa-</palabra_cortada> para arrancar /
```

```
digo</turno>
</texto>
</CORPES>
```

Figura 15. Datos esquema codificación del texto CORPES XXI

Si nos fijamos en el sistema de marcas y etiquetas que aparecen la figura anterior, puede observarse que estas se conforman de dos maneras: existen etiquetas simples que están formadas por un solo elemento, como por ejemplo <vacilación/> y etiquetas dobles en las que aparece un elemento de apertura y otro de cierre, como sucede en <transcripción_dudosa> </transcripción_dudosa>. A su vez, algunas etiquetas permiten introducir información adicional o atributos, es el caso de <turno hb="" seg=""></turno> o <observación_complementaria desc=""/>. Esta información no forma parte de la transcripción y por tanto debe añadirse dentro de las comillas. La inclusión de estas etiquetas, tanto extratextuales como intratextuales, permitirá que cuando el archivo se incorpore a la base de datos del corpus, se puedan hacer búsquedas filtradas.

4.2.2.5. Posibilidades de anotación de un corpus oral

Siguiendo la terminología adoptada, la anotación ofrece información lingüística extra sobre el material transcrito y codificado. Existen diferentes tipos de anotación que se pueden aplicar a los corpus de lengua oral, dependiendo de los objetivos de investigación y las necesidades específicas, cada tipo está determinado por el nivel de análisis de la lengua en el que se incluya.

Una vez que se ha transcrito y codificado el material oral, pueden añadírsele capas de anotación que reflejen otras informaciones adicionales. Las más comunes son la tokenización, la lematización, la anotación morfológica o *part of speech tagging* (POS), la anotación sintáctica o *parsing*, y el análisis semántico. Si bien podrían considerarse otras tipologías como la pragmática, la estilística o la anotación de errores en corpus de aprendices.

La tokenización supone “la identificación y caracterización de cada uno de los elementos gramaticales que forman la secuencia analizada” (Rojo, 2021, p. 105); es decir, lo que se

consigue es obtener una lista de todas las palabras que conforman la transcripción. Así, en una oración como *los gatos están jugando en el jardín* se identificarían siete unidades.

La lematización, en cambio, permite al investigador extraer y examinar todas las variantes de un lexema sin tener que introducir todas las opciones posibles, y producir información de frecuencia y distribución del lexema (Baker, Hardie y McEnery, 2006, p. 105), a saber, se extrae un listado de todas las formas lematizadas contenidas en la transcripción. Tomando el mismo ejemplo utilizado para la tokenización, si lematizamos la secuencia *los gatos están jugando en el jardín* obtendríamos los lemas el|gato|estar|jugar|en|el|jardín. La lematización facilita el análisis de texto y la identificación de temas comunes en un conjunto de documentos, así como mejorar la precisión de futuras consultas al reducir las palabras a sus formas base. McEnery, Xiao y Tono (2006) señalan que la lematización es especialmente útil en lenguas muy flexivas, como el español, ya que un lema cubre un gran número de variantes flexivas.

La anotación morfológica o *part of speech tagging* (POS) consiste en un tipo de anotación mediante la cual se asignan categorías gramaticales a palabras (o en algunos casos morfemas o frases), normalmente mediante un etiquetador automático, aunque, como señalan Baker, Hardie y McEnery (2006, p. 128) puede ser necesaria la postedición manual por parte humana en una fase final para corregir errores. Según Leech (2005) es "one common type of annotation is the addition of tags, or labels, indicating the word class to which words in a text belong".

Como se aprecia en la Figura 16, la primera fila corresponde con la búsqueda realizada, la segunda fila incluye la anotación morfológica[36] y la tercera muestra el resultado de la lematización.

que	había	en	la	**casa**	<silencio/>
PLMP	VII3S	X	DAFS	**NCFS**	ETQ_PAUSA
que	haber	en	el	**casa**	

Figura 16. Anotación morfosintáctica del corpus ESLORA. Fuente: Barcala *et al.,* (2018)

[36] El etiquetario morfosintáctico empleado en este corpus puede consultarse en https://eslora.usc.es/guide_tags

Este tipo de anotación es útil para una amplia gama de aplicaciones que van desde la desambiguación de homógrafos hasta usos más sofisticados como, por ejemplo, el cálculo de las ocurrencias de clases de palabras en un corpus (McEnery, Xiao y Tono, 2006, p.37).

Una vez etiquetado un corpus con POS, es posible relacionar estas categorías morfosintácticas en relaciones sintácticas de alto nivel entre sí (McEnery y Wilson, 2001, p. 53, Baker, Hardie y McEnery, 2006, p. 127). Es lo que se conoce como la anotación sintáctica o *parsing* que se basa en añadir etiquetas para indicar la estructura sintáctica de un segmento. Para McEnery, Xiao y Tono (2006):

> Parsing is probably the most common type of annotation after POS tagging. It is important to most natural language processing (NLP) applications – to make sense of a natural language, an NLP system must be able to decode its syntax. Syntactically parsed treebanks are even more useful than POS tagged corpora in linguistic research, as they not only provide part-of-speech information for individual words but also indicate constituent types and membership. For example, it is much easier to study clause types using a parsed corpus. (McEnery, Xiao y Tono, 2006, p. 39)

Por último, mencionaremos la anotación semántica. Esta se encarga de asignar códigos a las palabras según su función semántica. Para McEnery, Xiao y Tono (2006, p. 38) existen al menos dos tipos: el primero marca las relaciones semánticas entre los constituyentes de una frase —tipo que los autores consideran como un tipo de anotación sintáctica—, mientras que el segundo tipo marca las características semánticas de las palabras de un texto. Esta sería la anotación semántica propiamente dicha para estos autores.

Estas capas de información de anotación suelen realizarse mediante programas computacionales que permiten una anotación automática. No obstante, es habitual y recomendable que en casos como la anotación morfológica, sintáctica y semántica se lleve a cabo un proceso de revisión y desambiguación de manera manual, como hemos apuntado arriba. Esta desambiguación puede realizarse, en primer lugar, de manera semiautomática recurriendo a la aproximación basada en reglas o a la aproximación probabilística, es decir, empleando algoritmos de autoaprendizaje que sobreentienden las normas de los corpus de una manera automática y las utilizan para definir otras funciones de palabras. En última instancia conllevaría un proceso de revisión manual, llevado a cabo de manera personal.

Las herramientas informáticas para el procesamiento de lenguaje natural que permiten este tipo de anotación están desarrolladas en lenguajes de programación como Python, R o

C++. Algunas de las más utilizadas son Sketch Engine, NLTK (Natural Language Toolkit), Freeling o Treebank, UDPipe además de otras plataformas en línea que realizan estos procesos como Linguakit, para el español, o Xiada, para el gallego.

4.2.3. Fase 3. Archivo, distribución y acceso al corpus por parte de los usuarios

La última fase que debe considerarse en el diseño de un corpus oral es aquella correspondiente al archivo, la distribución y al acceso al corpus por parte de los usuarios. Esto es, de qué manera se van a almacenar los materiales de forma que se garantice el mantenimiento del corpus en el tiempo, cómo se va a dar a conocer a la comunidad científica, es decir, si se prevé que se incluya dentro de alguna red o iniciativa que actúe de impulsora en la difusión del trabajo; y, por último, de qué manera van a poder acceder los investigadores y usuarios a los materiales recogidos para realizar análisis lingüísticos.

Quizás, el factor más significativo para las decisiones sobre los aspectos comentados para esta fase podría ser el económico, ya que para archivar y alojar un corpus en línea que permita el acceso a los materiales se requiere de una plataforma o *hosting* que necesita una inversión y un mantenimiento técnico por parte de personal especializado. Las implicaciones, en este sentido, variarán si simplemente se suben las transcripciones en línea, o si, además de estas, se facilitan los audios, lo que conllevará que el espacio de almacenamiento aumente considerablemente; si se pretender subir los audios y las transcripciones, se necesitará crear un motor de búsqueda que filtre los datos y que sea capaz de procesar el material previamente codificado y/o anotado. El añadido de cada una de estas facilidades supone una inversión en términos de espacio y económicos, que implica que se multipliquen los gastos. Es decir, cuantas mayores posibilidades se ofrezcan a los usuarios, mayores necesidades económicas y de personal para llevarlo a cabo habrá.

Cabe mencionar que, en muchas ocasiones, las necesidades para almacenar y poder distribuir el material superan los medios y ayudas que un grupo de investigación puede recibir por parte de la institución que lo ampara, porque lo corpus exceden las posibilidades de almacenaje que estas ofrecen y, además, las partidas presupuestarias para el caso de proyectos que cuenten con alguna subvención no suelen incluir ninguna sección dedicada a hacerse cargo de los gastos derivados de estas tareas. En otros casos, el uso de recursos y plataformas de carácter gratuito no siempre es posible debido a que algunos corpus están

obligados a mantener el compromiso adoptado en cuanto a la privacidad de los datos recogidos —no siempre sucede, dependerá del tipo de corpus que se haya construido. Se debe reflexionar, además, sobre el papel que tiene el grupo o el investigador o investigadora que cree y construya el corpus, que deberían ser capaces de garantizar por sí mismos el mantenimiento del corpus a nivel informático para no tener que depender de contratación de personal externo ya que, como hemos mencionado arriba, no siempre se cuenta con financiación suficiente.

Así mismo, deben contemplarse otros factores más allá del económico y es que, como señala Wynne (2005, p. 87), deben tenerse en cuenta los cambios que pueden sucederse en cuanto al personal implicado en la recopilación, como por ejemplo cambios de puesto de los investigadores, finalización de contratos, etc. y deben considerarse también posibles cambios que operarían a nivel informático, es decir, las herramientas utilizadas en la construcción del corpus y en su puesta a disposición pública están sometidas a un proceso de obsolescencia desde que nacen ya que la tecnología no hará más que evolucionar. Por tanto, deben contemplarse acciones que garanticen o posibiliten el manejo del corpus a través de los años y que evite la caducidad de su uso.

Cuando se plantea el acceso público en abierto al corpus se debe atender, así mismo, a varios condicionantes. En primer lugar, hablaríamos del formato en que se van a facilitar los materiales. Según Wynne (2005, p. 94) idealmente debería realizarse a través de formatos no comerciales. De hecho, este autor defiende el uso del formato XML o el texto plano, compatible con sistemas estandarizados internacionales; no obstante, refiere a un posible conflicto entre las necesidades de los lingüistas especializados en corpus y las necesidades derivadas del almacenamiento en donde puede necesitarse otro formato.

Además, en relación con la necesidad de contar con almacenamiento suficiente, comentada anteriormente, debe plantearse en qué formato se almacenarán los audios cuando estos puedan ponerse a disposición pública. Se da aquí un nuevo conflicto entre la decisión de ofrecer la mejor calidad de audio posible frente al tamaño del archivo; esto es, a mayor calidad de audio, mayor peso tiene el archivo, lo que conlleva una necesidad mayor de almacenamiento y un ralentizado del motor de búsqueda. En segundo lugar, se trataría de garantizar el cumplimiento de los acuerdos establecidos previamente en materia de protección de datos o derechos de autor. A saber, hay que asegurarse de que se disponen de los permisos pertinentes por parte de los participantes o de la cesión de los derechos para la

explotación de los datos procedentes de otras fuentes, como radio, televisión o redes sociales antes de difundirlos públicamente.

Una posibilidad de superar estas barreras relacionadas con la capacidad propia de almacenamiento y la gestión económica que venimos señalando, podría encontrarse en la inclusión del corpus en plataformas o repositorios externos que aporten el apoyo necesario para esta fase del corpus en este sentido. En los últimos años han surgido iniciativas como Talk Bank (MacWhinney, en línea), la Common Language Resources and Technology Infrastructure-Clarin (Hinrichs, Erhard y Steven Krauwer, 2014), OLAC: Open Language Archives Community (2010) o Lyneal (Ueda, 2021) que facilitan infraestructuras de almacenamiento para corpus.

Si se opta por la gestión propia del corpus, puede hacerse por medio de servicios de alojamiento web en el mercado que ofrecen la posibilidad de instalar sistemas como Drupal o Wordpress que, para un investigador con formación media en informática, la facilidad de instalación, customización y mantenimiento es prácticamente infinita. Si bien, existen otras opciones más complejas basadas en bases de datos relacionales (como veremos en 5.2.3.) que permiten la creación de corpus masivos en el dominio lingüístico actual, tanto para el español como para otras lenguas como el inglés. Cabe destacar, por ejemplo, los corpus desarrollados por Mark Davies (Davies y Kim 2019; Davies 2009, 2012, 2021). Si bien la mayoría de los corpus desarrollados por Davies proceden de textos escritos, otros desarrolladores también han utilizado bases de datos relacionales, como sería, por ejemplo, el caso del corpus Spokes para el polaco; se trata de una iniciativa desarrollada por Piotr Pezik para el marco de la plataforma de corpus CLARIN e incluye más de 2.2 millones de palabras.

4.3. Análisis contrastivo del diseño y construcción de corpus orales del español

Una vez caracterizadas las diferentes fases que suceden en el diseño y construcción de corpus orales, a continuación, se presenta un análisis contrastivo de las principales decisiones metodológicas que operan en corpus relevantes del español oral.

Para la realización de este análisis, recuperamos aquí la tabla incluida en el Capítulo 2, en la que aparecen en color naranja los corpus que cuentan con mayor respaldo de la literatura revisada. Serán los que nos servirán de base para la revisión crítica de las diferentes

fases del diseño y construcción de corpus orales que veremos a continuación, aunque con algunas modificaciones como se explicará más abajo. Estos corpus son, por orden alfabético el *Corpus Oral de Lenguaje Adolescente* (COLA), el C-Oral-ROM, el *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC), PRESEEA y Val.Es.Co. (2002).

CORPUS	MO	B&A	E, V & B	B&C	RO	SO	PA	LLIS
COLA	×	✓	✓	✓	✓	✓	✓	✓
C-ORAL-ROM	✓	✓	✓	✓	✓	✓	✓	×
CORLEC	✓	✓	✓	✓	✓	✓	✓	×
PRESEEA	✓	✓	✓	✓	✓	×	✓	✓
Val.Es.Co. (2002)	✓	✓	✓	✓	✓	✓	✓	✓
CE (Mark Davies)	✓	✓	✓	✓	✓	✓	×	×
CECBNA	✓	✓	✓	✓	✓	✓	×	×
CORDIAL	×	×	✓	✓	✓	✓	✓	✓
COSER	×	✓	✓	✓	✓	×	✓	✓
CREA	✓	✓	✓	✓	✓	✓	×	×
ALCORE	✓	✓	×	✓	✓	✓	×	×
CLH de Almería	✓	✓	✓	✓	×	×	×	✓
CORPES XXI	×	×	✓	✓	✓	✓	✓	×
COVJUA	✓	✓	×	✓	✓	✓	×	×
VUM	✓	✓	✓	✓	✓	×	×	×
COGILA	×	✓	✓	✓	×	✓	×	×
CUMBRE	✓	✓	×	✓	✓	×	×	×
GRIAL	×	✓	×	×	✓	✓	✓	×
MC-NC	×	✓	✓	✓	✓	×	×	×
PILEI	✓	✓	✓	×	✓	×	×	×

Consenso en 7-8 artículos | Consenso en 6 artículos | Consenso en 5 artículos | Consenso en 4 artículos

Mo (Moreno Fernández, 2005c), B&A (Briz y Albelda, 2009), E, V & B (Enghels, Vandershueren y Bouzouita, 2015), B&C (Briz y Carcelén, 2019), RO (Rojo, 2016a), SO (Solís, 2018), P&B (Parodi y Burdiles, 2019), LLI (Llisterri, 2021)

Tabla 6. Consenso entre recopilatorios

De acuerdo con la tabla anterior son cinco los corpus que presencian su registro en 7 u 8 artículos recopilatorios. De estos cinco corpus, no obstante, para un análisis más completo se ha tomado la decisión de prescindir de C-Oral-Rom y de CORLEC, desarrollados en el Laboratorio de Ingeniería Informática de la Universidad Autónoma de Madrid. En el caso

del primero, la decisión se debe al hecho de no estar disponible en formato abierto, esto es, debe adquirirse la publicación para acceder al corpus completo; en el segundo, no se dispone de una plataforma en línea para la recuperación de la información. En sustitución, se han seleccionados otros dos corpus, el *Corpus Oral y Sonoro del Español Rural* (COSER) y el *Corpus del Español del siglo XXI* (CORPES XXI) que no aparecen en la tabla entre los corpus con más consenso, porque se publicaron posteriormente, y varios de los trabajos recopilatorios analizados (Moreno Fernández, 2005c o Briz y Albelda, 2009) no pudieron, por tanto, recogerlos en sus panorámicas. Su inclusión en nuestro estudio está justificada dado el tamaño de ambos corpus y el respaldo académico e institucional con el que cuentan.

Hemos estructurado esta revisión en tres bloques. Por un lado, se han contemplado los factores externos**,** es decir, aquellos que corresponden con la información que los propios grupos de investigación facilitan del corpus a modo de descriptores: (a) su objetivo o finalidad, (b) el género discursivo, (c) criterios de representatividad, (d) actuación con respecto a aspectos legales (consentimiento), factores que están presentes en la fase general de recogida de datos.

Por otro lado, se han revisado los factores que hemos denominado factores internos, esto es, aquellos que tienen que ver directamente con el tratamiento de los datos, como son los signos y marcas que se han utilizado en su transcripción y codificación, divididos en diferentes categorías: (a) ortografía y puntuación, (b) etiquetado de ruidos, risas y elementos funcionales, (c) etiquetas fonéticas, (d) etiquetas de oralidad, (e) etiquetas léxicas, (f) etiquetas relativas a la transcripción, (g) etiquetas relativas a la anonimización. Se presta especial atención a la enumeración de los principales retos a los que se enfrenta la codificación de los elementos propios de la lengua oral, señalados por Torruella y Llisterri, en el fragmento que retomamos de la sección 4.2.2.3.:

> las pausas, la delimitación de los enunciados y de las unidades tonales, las variaciones en los elementos suprasegmentales, los elementos vocales tanto semi-léxicos (por ejemplo, las denominadas ‘pausas llenas’) como no léxicos (por ejemplo, risas o toses), los cambios de turno de palabra, las intervenciones simultáneas de varios hablantes o las dudas, palabras truncadas, repeticiones y errores de producción corregidos o no por el propio hablante (1999, p. 27).

Y, por último, se han analizado aquellos factores relacionados con acceso a los datos, esto es, de qué manera cada corpus facilita el acceso a sus materiales para su utilización por la comunidad científica.

Cabe señalar que las informaciones relativas a cada uno de los bloques proceden de la propia información que aparece en la página electrónica de cada corpus, así como de bibliografía específica sobre ellos y de la experiencia propia como usuarios en cuanto al acceso a sus plataformas de búsqueda.

4.3.1. Fase 1. Factores externos

A continuación, detallaremos las decisiones adoptadas por los corpus COLA, CORPES XXI, COSER, PRESEEA y Val.Es.Co. (2002), respecto de aquellos aspectos que corresponden con la fase de recogida de los datos, esto es, cuál es el objetivo de investigación y análisis de cada uno, qué género discursivo recogen, si siguen algún criterio en cuanto a la representatividad de la muestra y qué aspectos legales han debido tener en cuenta y cómo los han abordado.

4.3.1.1. Objetivo del corpus

OBJETIVO
COLA: lenguaje adolescente, con intereses diatópicos
CORPES XXI: corpus de referencia, con interés panhispánico
COSER: fenómenos dialectales morfosintácticos, con interés local de España
PRESEEA: estudio sociolingüístico, con interés panhispánico
Val.Es.Co. (2002): estudio diafásico de la conversación coloquial, con interés local en Valencia

Si nos fijamos en los objetivos de recolección de los corpus considerados para este análisis, CORPES XXI nace como corpus de referencia y, por lo tanto, su objetivo es describir, desde una perspectiva panhispánica, el estado de la lengua española en un determinado momento, concretamente a partir del año 2001. De carácter panhispánico es también PRESEEA, que surge con el objetivo de servir como corpus sociolingüísticamente representativo para diversos estudios. Los corpus COLA y COSER son corpus multidialectales, el primero de los cuales tiene como principal finalidad el estudio del lenguaje adolescente, mientras que el segundo se restringe al estudio del habla en zonas rurales para documentar fenómenos relativos a la gramática y para el estudio de la variación morfosintáctica. Por último, el corpus Val.Es.Co. (2002) se constituye como corpus monodialectal para el estudio de la conversación coloquial española.

4.3.1.2. Género discursivo del corpus

GÉNERO DISCURSIVO
COLA: conversación espontánea informal
CORPES XXI: incluye una miscelánea de entrevistas, debates, monólogos, tertulias, entre otros.
COSER: entrevista semidirigida
PRESEEA: entrevista semidirigida
Val.Es.Co. (2002): conversación coloquial espontánea

Con respecto al género discursivo, CORPES XXI, al ser un corpus de referencia, incluye una miscelánea de entrevistas, debates, monólogos y tertulias, entre otros, extraídos de convenios con medios de comunicación y de otros corpus orales; en cambio PRESEEA y COSER se centran en recoger exclusivamente entrevista semidirigida. COLA y Val.Es.Co. (2002) optan por la recogida de conversación coloquial espontánea e informal, en el caso de Val.Es.Co. obtenida de manera secreta, como veremos más abajo.

4.3.1.3. Criterios de representatividad del corpus

CRITERIOS DE REPRESENTATIVIDAD
COLA: sexo y edad
CORPES XXI: no sigue criterios de representatividad sociolingüística
COSER: un hablante por enclave seleccionado, de edad avanzada, oriundos del lugar
PRESEEA: sexo/género, grupos de y nivel de instrucción
Val.Es.Co. (2002): sexo, grupo de edad y nivel de instrucción

En el caso del corpus académico, CORPES XXI, no sigue criterios de representatividad sociolingüística, sino que la selección de materiales se realiza según la procedencia y medio de los materiales (70% español de América, 30 % de España, 90 % material escrito y 10 % material oral); esto puede justificarse porque constituye un corpus de referencia que no pretende ser representativo en ese sentido.

Sin embargo, PRESEEA y Val.Es.Co.[37] sí que se han construido teniendo en cuenta la estratificación en sexo, grupo etario y nivel de instrucción. Comparten ambos corpus la

[37] No obstante, en el volumen de 2002 se recogen conversaciones anteriores que no fueron seleccionadas siguiendo criterios de estratificación sociolingüística, sino que se clasificaron según su adscripción al género conversación coloquial prototípica o periférica.

división en sexo (mujer u hombre), si bien PRESEEA (Moreno Fernández, 2021a) ha incorporado la variable género a la recogida de entrevistas, y la variable nivel de instrucción (bajo, medio, alto). El nivel de bajo se corresponde con analfabetos, sin estudios y Enseñanza Primaria; el nivel medio con Enseñanza Secundaria; y el nivel alto con Enseñanza Superior (universitaria, técnica superior). Para los grupos de edad estos grupos difieren en los rangos: PRESEEA contempla las franjas de entre 18-34, 25-54, +55, mientras que Val.Es.Co. (2002) establece los grupos de 18-25, 26-55, >55.

Para el caso del corpus COLA, que recoge exclusivamente muestras de habla producidas por adolescentes, los criterios de representatividad que operan tienen que ver con el sexo y la edad, recogen grabaciones para la franja de edad de 13 a 19 años. No obstante, los centros escolares desde los que se reclutó a los informantes fueron seleccionados según ingresos económicos, zona residencial y formación académica, como indican Jørgensen y Eguía (2014, p. 8).

En cuanto al corpus COSER, solamente se recoge un hablante por enclave seleccionado, de edad avanzada y oriundo del lugar, de acuerdo con sus fines de investigación. Téngase en cuenta, como hemos visto anteriormente, que su objetivo es el estudio de fenómenos dialectales, no sociolingüísticos.

4.3.1.4. Aspectos legales del corpus

ASPECTOS LEGALES
COLA: recogida de consentimiento informado
CORPES XXI: convenios
COSER: no se especifica
PRESEEA: no se especifica
Val.Es.Co. (2002): no se especifica

Tal y como se mencionó en el apartado 4.2.1.1., en el caso de los corpus académicos conformados como corpus de referencia, los aspectos legales están relacionados con los derechos de autor, motivo por el cual, este tipo de corpus suele operar en base a convenios con editoriales y plataformas audiovisuales que ceden estos derechos. En este tipo de corpus la recuperación de información se realiza por medio de concordancias, es decir, solamente se recupera el contexto inmediatamente anterior y posterior de la forma buscada y no se tiene acceso de ninguna manera al material cedido en su totalidad. En el caso de CORPES XXI,

se han firmado convenios con medios de comunicación, salvaguardando por tanto los derechos de autor, pero no se necesita consentimiento de los propios hablantes. El material que se recoge en este corpus procedente de otros corpus orales se incorpora bajo las premisas de recolección del corpus original y, por tanto, los permisos ya fueron obtenidos en su momento por el grupo que lo gestionaba en primer lugar.

En el caso de COLA (Jørgensen y Eguía, 2014, p. 8), los participantes han firmado un consentimiento informado, si bien, para aquellos informantes menores de edad, estos fueron firmados por su progenitores o tutores.

Para los corpus COSER, PRESEEA y Val.Es.Co. (2002), de su metodología general de trabajo se deduce que los informantes han dado su consentimiento en algún momento anterior o posterior a la grabación, sin embargo, no se menciona expresamente.

4.3.2. Fase 2. Factores internos

A continuación, se abordan los aspectos que hemos denominado factores internos y que están relacionados con la fase de tratamiento de los datos recogidos. Se incluye aquí la descripción de la metodología adoptada por los corpus seleccionados en cuanto a sus sistemas de transcripción y codificación, el uso de ortografía convencional o no, marcas[38] fonéticas, marcas que reflejen ruidos, risas y elementos funcionales, marcas de oralidad, léxicas, de transcripción y anonimización.

4.3.2.1. Transcripción y codificación

TRANSCRIPCIÓN Y CODIFICACIÓN
COLA: HTML y TEI
CORPES XXI: XML y TEI
COSER: Sistema propio
PRESEEA: XML y TEI
Val.Es.Co. (2002): Sistema propio

[38] De manera general hablaremos de *marcas*, si bien, puede encontrarse de manera sinónima el término *etiqueta*. A este respecto, cabe matizar que una etiqueta se caracteriza por el uso de < > y toda etiqueta es una marca, mientras que no toda marca tiene forma de etiqueta. Una marca puede ser también un signo o símbolo, por ejemplo.

El estándar internacional TEI es la base sobre la que se establecen sus sistemas de transcripción y codificación los corpus COLA, CORPES XXI y PRESEEA. Mientras que COSER y Val.Es.Co. desarrollan un sistema propio.

COSER utiliza un conjunto de marcas que sirven para reflejar los turnos de palabra y las diversas circunstancias que concurren en la conversación, sin pretensión alguna de exhaustividad, pues el análisis de la conversación no forma parte de sus objetivos (COSER, en línea). Val.Es.Co. (2002) prefiere una transcripción enriquecida que, dado que su publicación iba a ser en formato papel, no adoptó ninguna metodología de transcripción informatizada siguiendo el modelo TEI y XML.

COLA, además, opta por realizar una transcripción alineada con Transcriber, utilizando un lenguaje de codificación HTML, mientras que CORPES XXI y PRESEEA optan por el lenguaje XML, que en la actualidad se presenta como un lenguaje más sencillo e intuitivo en su uso, así como su versatilidad a la hora de modificar o componer nuevas etiquetas.

4.3.2.2. Ortografía y puntuación

Se engloban bajo este epígrafe las decisiones que adoptan cada uno de estos corpus sobre el uso de criterios ortográficos en la transcripción.

ORTOGRAFÍA Y PUNTUACIÓN[39]
COLA: transcripción ortográfica sin puntuación
CORPES XXI: transcripción ortográfica sin puntuación
COSER: transcripción ortográfica con signos de puntuación
PRESEEA: transcripción ortográfica sin puntación
Val.Es.Co. (2002): transcripción ortográfica sin puntuación

Como tendencia general, a excepción de COSER, la transcripción de estos corpus tiende a romper con las reglas de la escritura. Sin embargo, cabe mencionar que, si bien en varios de estos sistemas sí que se utilizan convenciones ortográficas, en ningún caso se refieren a tales sistemas como *transliteración*, sino como *transcripción*. Esto puede deberse a que,

[39] Cabe señalar aquí que se han recogido las denominaciones utilizadas por cada uno de estos corpus para el tipo de transcripción que realizan, no responden a la propuesta de denominación operativa propuesta en este trabajo en la sección 4.2.2.2.

siguiendo unos estándares internacionales, u otros propios, como hemos visto arriba, en todos los casos hay un enriquecimiento del texto por medio de etiquetas y marcas.

El corpus COLA está transcrito de manera ortográfica, no obstante, no se utilizan signos de puntuación (puntos, comas, exclamaciones, interrogaciones) por considerarse esta una interpretación del texto (COLA, en línea). Igual sucede en el caso de CORPES XXI, aunque en este corpus sí que se señalan las estructuras interrogativas y exclamativas. Se indica también que no se usan las mayúsculas salvo para nombres propios y siglas (Rojo *et al.*, en línea).

PRESEEA (Moreno Fernández, 2021b, p. 12) señala que emplea una ortografía normativa convencional, incluida la acentuación, sin signos de puntuación, excepto para interrogativas y exclamativas y, al igual que CORPES XXI, no se usan las mayúsculas salvo en nombres propios y siglas. Esta decisión tiene sentido, ya que el corpus PRESEEA ha cedido algunos de sus materiales al corpus académico y, por tanto, busca la compatibilidad. Además, Moreno Fernández aclara que no intentan reproducir fenómenos orales, como elisiones, aspiraciones, diptongaciones orales o sinalefas y que las palabras que presenten elisiones se completan en su escritura. No obstante, el texto transcrito también debe ajustarse a ciertos criterios gráficos que no son los habituales en la escritura ordinaria, sino que responden a unas necesidades propias de la representación escrita de la lengua oral. Estos criterios afectan especialmente a la puntuación y a la forma de representar las pausas, pero también a otros aspectos (Moreno Fernández, 2021b, p. 12). Se refiere al sistema de marcas mínimas obligatorias en formato XML que veremos más adelante.

Val.Es.Co. (2002) opta por una ortografía convencional con excepciones, ya que sí reflejan fenómenos de la oralidad como fenómenos de fonética sintáctica, sinalefas o elisiones en la escritura, entre otros. No se emplean signos de puntuación a excepción de interrogativas y exclamativas, no se usan las mayúsculas más que para nombres propios, si bien, también se emplean como marca de una pronunciación enfática (Briz y Grupo Val.Es.Co., 2002, p. 28).

COSER es un corpus orientado al estudio de la variación morfosintáctica, por lo que la transcripción de los materiales se lleva a cabo siguiendo las convenciones de la ortografía habitual. Aunque se realizan algunas concesiones a la pronunciación dialectal — en especial, en el reflejo de la supresión, adición y metátesis de sonidos—, en ningún caso pueden estimarse las transcripciones del COSER un reflejo de la variación fonética o fonológica de

las variedades del español peninsular. Para establecer conclusiones o comparaciones de ese tipo, es necesario recurrir directamente a los archivos sonoros que se ponen a disposición de los usuarios en la web junto a las correspondientes transcripciones. Los signos de puntuación son utilizados de acuerdo con las normas generales de la puntuación española. (COSER, en línea).

4.3.2.3. Marcas fonéticas

Se utiliza la denominación de marcas fonéticas para aquellas marcas que reflejen fenómenos propios del medio oral como son las pausas y silencios, así como vacilaciones, autointerrupciones y pronunciaciones marcadas como el énfasis o el susurro, por ejemplo.

MARCAS FONÉTICAS
COLA: pausas, entonación e interrupciones
CORPES XXI: pausas, silencios, palabras cortadas y vacilaciones
COSER: pausas, silencios y tipos de pronunciación
PRESEEA: pausas, silencios, palabras cortadas, vacilaciones, alargamientos y énfasis
Val.Es.Co. (2002): pausas, silencios, entonación y énfasis

COLA registra fenómenos como el habla poco clara, las palabras interrumpidas, la entonación y las pausas. Recordemos que este corpus utiliza un lenguaje de codificación HTML al que han incorporado las siguientes marcas: XXX señala habla poco clara, % se utiliza para palabra interrumpida, / refleja entonación ascendente de pregunta, \ entonación descendiente de pregunta y . para pausa de un segundo, junto con .. para pausa de dos segundos y … para pausa de tres segundos (COLA, en línea).

CORPES XXI y PRESEEA comparten varias de estas etiquetas, en formato XML: <palabra_cortada></palabra_cortada>, <sic></sic> se emplea para señalar incorrecciones que no son descuido del transcriptor, <vacilación/>, <silencio/>, pausas / //, diferenciando la duración entre pausa mínima / y pausa //. Si bien, PRESEEA añade <alargamiento/> y <énfasis></énfasis> para pronunciaciones claramente marcadas.

COSER establece un sistema propio de etiquetado en el que las diferentes marcas se recogen entre corchetes. Así, para estas etiquetas fónicas registra silencio [SLNC], pausa [PS], pausa tiempo [PS:] para pausas largas se indica su duración, e incluyen cinco marcas de pronunciación: enfática [P-Enf:], relajada [P-Rel:], susurrada [P-Ssr:], silabeante [P-Slb:], otras [P-Otr:].

El otro corpus que utiliza su propio método de etiquetado es Val.Es.Co. (2002). En este caso, han considerado pertinente marcar la entonación descendente ↓, ascendente ↑ y suspendida →. Se marcan pausas cortas de menos de medio segundo con /, pausa entre medio segundo y un segundo con // y para pausas de más de un segundo ///. Los silencios, se señalan mediante la indicación del número exacto de segundos entre paréntesis (5"). Para pronunciación enfática se utilizan las mayúsculas y para el silabeo la representación de la palabra separada por sílabas. Los fragmentos pronunciados como susurros se encierran entre °( )°. Los alargamientos vocálicos y consonánticos se registran mediante el duplicado de la letra afectada.

Nótese que este corpus no se pensó en un primer momento para ser digitalizado, de ahí que se privilegie la facilidad de lectura mediante un sistema que utiliza signos y símbolos tipográficos de fácil comprensión.

4.3.2.4. Marcas de ruidos, risas y elementos funcionales

Se engloban aquí las marcas o etiquetas que los distintos corpus analizados utilizan para reflejar ruidos, risas y elementos funcionales, también llamados disfluencias, pausas oralizadas no léxicas y los sonidos o signos paralingüísticos[40]. Nos referimos a elementos como *mm, mhm, eemm*, carraspeos, siseos, chasquidos de lengua, entre otros, que pueden transmitir o no conductas verbales como manifestar atención/desatención, interés/desinterés, aprobación/desaprobación, etc. (Poyatos, 1994).

MARCAS DE RUIDOS, RISAS Y ELEMENTOS FUNCIONALES
COLA: no existen convenciones
CORPES XXI: música, risa, aplausos, elementos funcionales y fáticos
COSER: ruidos, elementos funcionales y fáticos
PRESEEA: ruidos, risas y elementos cuasi-léxicos y funcionales
Val.Es.Co. (2002): risas, toses y gritos

[40] Llisterri (en línea) utiliza la denominación de *disfluencias* para caracterizar a aquellos fenómenos que tienen que ver con la fluidez debido a la planificación (falta de ella), en conversación coloquial especialmente, del discurso. Poyatos (1994) hace una clasificación de *elementos paralingüísticos* entre los que se incluyen disfluencias, elementos funcionales y sonidos. Cestero (1999) recoge un listado muy completo de *signos paralingüísticos*. Para una clasificación exhaustiva de los tipos de disfluencias véase Pascual Aliaga (2019, pp. 94-101).

Para el corpus COLA, señalan Jørgensen y Eguía (2014, p. 11) que no se añade esta información extralingüística ya que esta información se obtiene escuchando la transcripción con sonido en la red, no obstante, si se consultan las transcripciones puede observarse la representación de la risa por medio de diversas convenciones como *jajajaja* o *jijijiji*.

CORPES XXI, contempla etiquetas para las risas según se realicen de manera esporádica <risa hb=""/> o bien que afecten a una intervención particular <risas_inicio hb=""/><risas_fin hb=""/>; sucede igual para documentar los ruidos <ruido desc=""/> <ruido_de_fondo> </ruido_de_fondo>. No obstante, añaden otros elementos como el etiquetado de música <música/>, música de fondo <música_de_fondo> </música_de_fondo> y aplausos <aplausos/> motivado porque este corpus recoge material procedente de medios de comunicación, como programas de radio y televisión, donde es frecuente la aparición de estos recursos. En cuanto a los elementos funcionales, en su metodología de trabajo se señala que se transcriben, aunque no aparezcan en el *DLE*, y señalan los casos de *eeh, mmm, pff, sshh, eemm, buah, bueh, uff*. Indican que, para permitir su recuperación, se irá elaborando una lista que, dentro de la variedad existente, estandarice su modo de representación, sin embargo, no se ha encontrado más información al respecto (Rojo *et al*., en línea).

En el corpus COSER se observa una gran exhaustividad a la hora de enfrentarse a estos elementos. Se señalan las risas [RISAS] y las risas mientras se habla [Rndo:], el llanto mientras se habla [Llndo:], exclamaciones [EXCL], toses [TOS], carraspeos [CARRASP], chasquidos [CHASQ], onomatopeyas [ONOM], respiraciones [RESPIR], sonidos de asentimiento o fáticos (*ajá, hum, mm*, etc.) que llevan marca Asentimiento [Asent] y elementos funcionales recogidos bajo la marca Otras [OTRAS-EM] como, por ejemplo, *bah, puf, prr, brr, pss, chss*. Además, para ruido encontramos quince marcas diferentes, entre otros, se señalan con esta etiqueta los ruidos de vehículos [R-Vhc], los ruidos de animales [R-Anm], ruido de campanas [R-Cmp] o ruido indeterminado [R-Ind].

PRESEEA solamente detalla ruidos de manera general <ruido=""/> y ruidos que afectan a una intervención <ruido_fondo></ruido_fondo>, también señala <risas=""/> y fragmentos pronunciados entre risas <entre_risas></entre_risas>. En este caso, se refieren a los elementos funcionales como elementos cuasi-léxicos o paralingüísticos y se representan en ortografía ordinaria, de acuerdo con las convenciones más habituales en español, cuando existan (*uhum, ah, eeh, bah*).

El sistema Val.Es.Co. (2002) recoge (RISAS) (TOSES) (GRITOS) y remite al uso de las notas al pie de página para otras observaciones pertinentes. En cuanto a los elementos funcionales, no se proporcionan convenciones.

Podemos observar los diferentes acercamientos a la transcripción de estos elementos, si bien COLA descarta su representación apelando a que se puede consultar el audio para localizarlos si es necesario, en el caso de COSER se aprecia una sobrerrepresentación, sobre todo en el reflejo de ruidos.

4.3.2.5. Marcas de oralidad

Bajo este epígrafe se han englobado aquellas marcas correspondientes con la dinámica discursiva, a saber, solapamientos, reproducción de estilo directo o cita, elisiones y fenómenos de fonética sintáctica.

MARCAS DE ORALIDAD
COLA: solapamientos
CORPES XXI: simultáneos, cita, no reflejan elisiones
COSER: simultáneos, elisiones
PRESEEA: simultáneos, cita
Val.Es.Co. (2002): simultáneos, estilo directo, sucesión inmediata, elisiones, fenómenos de fonética sintáctica, reinicios y autointerrupciones

El corpus Val.Es.Co. (2002) implementa todo un sistema de marcas en este sentido. Estas van desde la reconstrucción de una unidad léxica que se ha pronunciado incompleta (en)tonces, al reflejo de fragmentos de habla solapada que se enmarcan entre corchetes []. Señalan, así mismo, la sucesión inmediata sin pausas entre dos emisiones de distintos interlocutores por medio del signo § y el mantenimiento de la intervención de un participante en un solapamiento con =, fenómenos que no se recogen en los otros corpus estudiados. Aparecen reflejados también reinicios y autointerrupciones sin pausa por medio del guion -, fenómenos de fonética sintáctica entre palabras (pa'l) y la aspiración de la *s* implosiva (h). Por último, aquellos fragmentos que se corresponda con estilo directo se marcan en cursiva. Esta exhaustividad se justifica por la finalidad del corpus, que se propone el análisis de la conversación.

Como hemos señalado anteriormente, CORPES XXI sigue las convenciones ortográficas del español, por tanto, no es de extrañar que no se reflejen fenómenos como seseos,

aspiraciones, rehilamientos, etc., ni formas reducidas, como [pa] para, [na] nada, [to] todo, [amás] además, [alante] adelante, etc. u otras contracciones no contempladas en la ortografía convencional: [pal] para el, [deste] de este, [patrás] para atrás. En cambio, sí representan elementos léxicos con acortamientos cuando se consideran de uso general, estén o no recogidos en el *DLE*: cole, boli, disco. También reflejan los fragmentos de habla simultánea <simultáneo> </simultáneo> y la reproducción del estilo directo <cita> </cita>.

En este sentido, PRESEEA se comporta de manera similar al corpus académico, ya que, como señalan, no les interesa reproducir fenómenos orales, como elisiones, aspiraciones, diptongaciones orales o sinalefas y las palabras que presenten elisiones se completan en su escritura (Moreno Fernández, 2021b, p. 12). Más allá de esto, se señalan simultáneos <simultáneo> </simultáneo> y citas <cita> </cita>.

COSER sigue la línea de los anteriores para el marcado de los solapamientos por medio de las marcas para habla simultánea E1 [HS:E1], conversación cruzada [HCruz:], otros solapamientos [HSim-O:], y voces simultáneas [V-Sml]. No establecen ninguna etiqueta para contracciones o elisiones, pero sí se transcriben, en cambio, no se reflejan aspiraciones. Como señalan en su metodología de trabajo disponible en su página electrónica (COSER, en línea), en todas estas marcas que reflejan solapamientos entre participantes de la conversación o del ambiente solo se indica el principio del solapamiento. Se ha renunciado, por tanto, a marcar el final de las intervenciones que se solapan dado que el análisis de la conversación no es el objetivo primordial en la elaboración del COSER.

En este aspecto, la única convención que recoge COLA es para los solapamientos, que se marcan mediante el uso de corchetes [ ].

4.3.2.6. Marcas léxicas

En este caso, se recogen elementos que tienen que ver con el léxico como, por ejemplo, si interesa señalar las siglas y otros elementos que tienen que ver con el cambio de código idiomático.

MARCAS LÉXICAS
COLA: no utiliza
CORPES XXI: siglas
COSER: no utiliza

PRESEEA: siglas, extranjero, lengua y término
Val.Es.Co. (2002): notas a pie de página

Tanto CORPES XXI como PRESEEA señalan las siglas por medio de la etiqueta <siglas desc=""> </siglas>. PRESEEA, además, incrementa este etiquetado por medio de las etiquetas <término> </término> para formas de uso claramente especializado, <extranjero></extranjero> para la aparición de extranjerismos y para el cambio de lengua <lengua=""></lengua>. En el corpus Val.Es.Co. (2002), recogido en una zona bilingüe donde se habla español y valenciano, cuando aparece un fragmento en valenciano, este se acompaña de una nota al pie en la que se especifica su traducción. Este corpus también aporta información sobre algunas palabras, como los extranjerismos que son transcritos como se pronuncia, o las siglas, en iguales términos.

4.3.2.7. Marcas de transcripción

En este caso, las marcas registran anotaciones que tienen que ver directamente con la labor del transcriptor. Se señalan, por ejemplo, cuestiones relacionadas con la inteligibilidad del audio, ya que hay casos en los que no se puede averiguar lo dicho por el hablante, o se averigua solo parcialmente y, por tanto, solo puede ofrecerse una transcripción aproximada o dudosa. Se incluyen también otras observaciones sobre el material u otras informaciones que el transcriptor considere relevantes incluir para garantizar la comprensibilidad del texto.

MARCAS DE TRANSCRIPCIÓN
COLA: no aparecen
CORPES XXI: marcas sobre la inteligibilidad, observaciones
COSER: marcas sobre la inteligibilidad y la grabación
PRESEEA: marcas sobre la inteligibilidad, observaciones
Val.Es.Co. (2002): marcas sobre la inteligibilidad, observaciones

En esta sección, los corpus analizados presentan un algo grado de coincidencia, así, está extendido el uso de marcas para la transcripción dudosa y los fragmentos indescifrables o ininteligibles. CORPES XXI y PRESEEA utilizan la misma etiqueta <transcripción_dudosa> </transcripción_dudosa> e <ininteligible/>, además de compartir la marca para observaciones y el tiempo: <tiempo seg=""/>, <observación_complementaria

desc=""/>. CORPES XXI señala otros aspectos como elementos que no se han representado con <nrp/> −por ejemplo, si se diera el caso de que apareciera publicidad en la grabación−, o intervenciones que proceden de segmentos traducidos <traducción></traducción>.

COSER detalla fenómenos relacionados con la grabación, como cortes [A-Crt] y errores de grabación [A-Err], así como los fragmentos que resultan ininteligibles [A-Inn] o inteligibles [A-PIn:]. En la misma línea actúa Val.Es.Co. (2002) cuando marca entre dobles paréntesis los fragmentos de transcripción dudosa ((casa)) y entre dobles paréntesis vacíos los segmentos indescifrables (( )). Para otras observaciones se utilizan las notas a pie de página En el corpus COLA no aparecen marcas relacionadas con este aspecto.

4.3.2.8. Marcas de anonimización

En general, los distintos corpus han anonimizado las transcripciones cuando, por las características del material recogido, se ha hecho necesario. En general, además de los antropónimos, también se anonimizan los topónimos y otros nombres propios, como por ejemplo accidentes geográficos, marcas, locales públicos, etc. Estas son las marcas que ha utilizado cada corpus:

MARCAS DE ANONIMIZACIÓN
COLA: sustitución del nombre original por etiqueta
CORPES XXI: sustitución del nombre por otro ficticio marcado con ~
COSER: sustitución del nombre original por etiqueta
PRESEEA: reducción del nombre a su inicial
Val.Es.Co. (2002): sustitución del nombre por otro ficticio sin marca

En el caso de COLA, las transcripciones y los audios han sido anonimizados tal y como exige la legislación europea y noruega. En este caso se sustituye el nombre original por la etiqueta @nombre como puede verse en la Figura 17.

T	L	10	MABPE2G01:	oeeeeeeh chachoooooo
T	L	10		dame un beso ven aquí dame un beso borracha
T	L	10	MABPE2J02:	qué música oyes @nombre
T	L	10		de qué estás escuchando ahora la radio
T	L	10	MABPE2G01:	ven
T	L	10		estoy escuchando el telediario tía a ver que dice

Figura 17. Captura de un fragmento de transcripción del corpus COLA

En la metodología de trabajo de CORPES XXI no aparece ninguna convención explícita para realizar la anonimización. Sin embargo, hay que tener en cuenta que CORPES XXI contiene material oral procedente en su gran mayoría de dos tipos de fuentes. Por un lado, se obtiene de medios de comunicación (radio, televisión e internet), y archivos públicos que no requieren anonimización (Figura 18), muchos de los cuales proceden de acuerdos o convenios con las propias plataformas de comunicación (RTVE, por ejemplo). Por otro, recoge material oral procedente de otros corpus (Figura 19) y, por tanto, adopta los respectivos protocolos de anonimización a los que se sometieron los corpus originales. Como hemos comentado arriba con respecto a la recogida de permisos, cuando CORPES XXI incluye materiales procedentes de otros corpus, se mantiene el procedimiento de anonimización particular del corpus original (en el caso mostrado aquí pertenece al corpus PRESEGAL). En este corpus se sustituye el nombre por uno ficticio que va precedido del signo ~ y en el audio puede escucharse un indicador sonoro.

Figura 18. Captura de resultado de una búsqueda en CORPES XXI

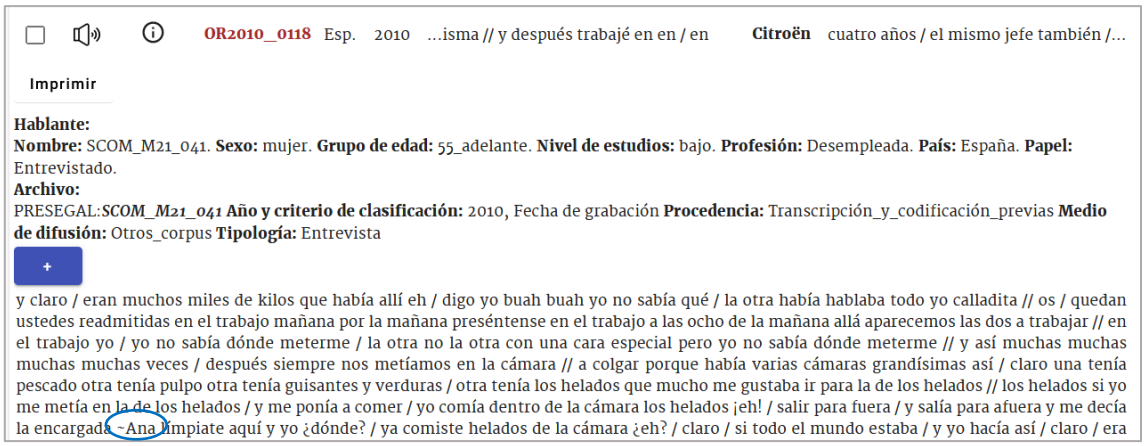


Figura 19. Captura del resultado de una búsqueda en CORPES XXI

En el corpus COSER la anonimización se realiza por medio de la marca Nombre propio [NP] a nivel textual (Figura 20). Esta convención oculta el nombre propio de los informantes, para evitar su divulgación en la publicación de los materiales.

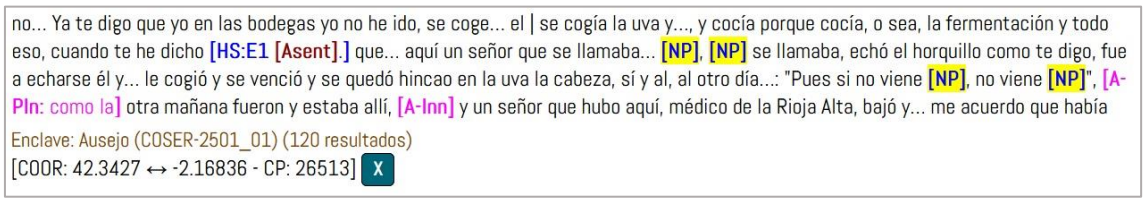


Figura 20. Captura del resultado de una búsqueda en el corpus COSER

Esta marca presenta cierta problemática ya que, aun teniendo acceso al audio, el hecho de tipificar todas las anonimizaciones como [NP] produce la pérdida de referencias, es decir, cuando dice “si no viene [NP], no viene [NP]” no podemos saber muy bien si se está refiriendo a una sola persona, como una repetición (“si no viene Pepe, no viene Pepe), o si que no venga una persona implica que tampoco venga otra, como una consecuencia (“si no viene Lola, no viene Pepe”). Esto podría solucionarse si apareciera un nombre ficticio o una numeración.

Para el siguiente corpus analizado, el corpus PRESEEA, si bien no existe una marca específica para señalar dicha anonimización, al realizar consultas en el corpus, a través de su página general, puede observarse la sustitución de nombres propios por la inicial correspondiente (Figura 21) cuando se trata de nombres propios o nombres de lugares en los que, por su relación con el informante, se ha considerado que debían anonimizarse. Sin embargo, no todos los nombres de lugares se anonimizan, ya que hay casos en los que no

existe dicha relación y por tanto no pueden considerarse identificadores indirectos. En este último caso, se percibe de manera general cierta subjetividad en la decisión de anonimizar o no estos identificadores, ya que queda a criterio de la persona encargada de la transcripción la pertinencia de realizarla o no.

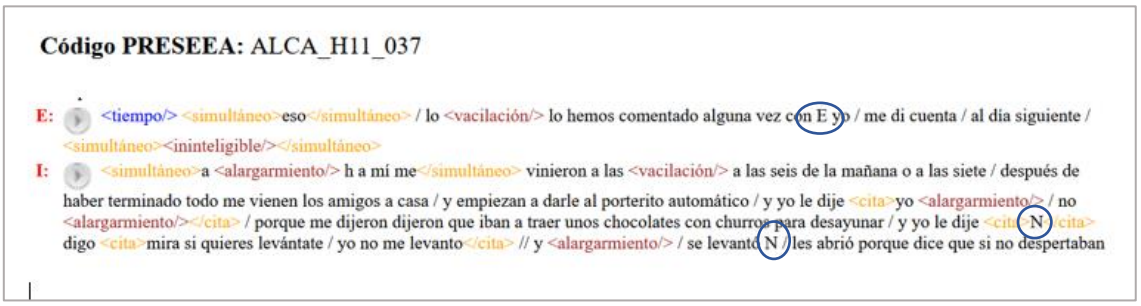
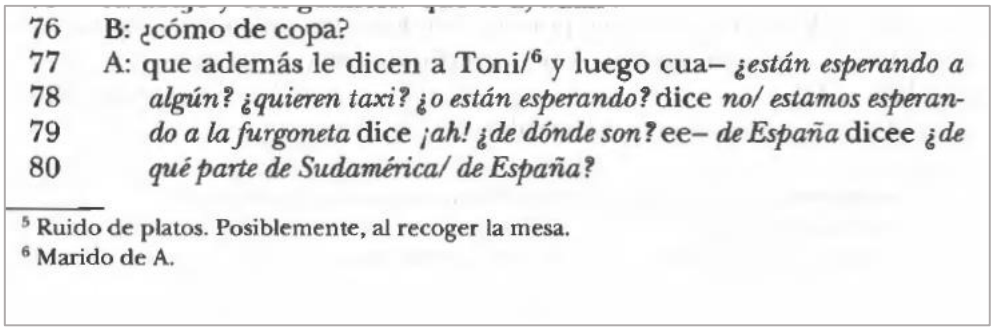


Figura 21. Captura de resultado de búsqueda del corpus PRESEEA

En Val.Es.Co. (2002), según indican los autores, suele sustituirse el nombre propio original por otro ficticio, respetando el patrón fónico siempre que sea posible, como puede verse en la Figura 22.


76 B: ¿cómo de copa?
77 A: que además le dicen a Toni/[6] y luego cua– *¿están esperando a*
78 *algún? ¿quieren taxi? ¿o están esperando?* dice *no/ estamos esperan-*
79 *do a la furgoneta* dice *¡ah! ¿de dónde son?* ee– *de España* dicee *¿de*
80 *qué parte de Sudamérica/ de España?*

[5] Ruido de platos. Posiblemente, al recoger la mesa.
[6] Marido de A.

Figura 22. Captura del corpus Val.Es.Co. (2002), conversación IM339.B.1

En todos los corpus aquí mencionados, además, se ha procedido a la anonimización del fragmento de audio correspondiente con la etiqueta de anonimización, bien se ha silenciado dicho fragmento, bien se ha introducido un indicador sonoro o pitido. El único corpus que no proporciona al acceso público al audio es el corpus Val.Es.Co. (2002) ya que, como hemos señalado, fue publicado en papel y, por tanto, sin acceso al audio. No obstante, estos materiales debidamente anonimizados se han facilitado a aquellos investigadores que lo han solicitado.

De esta breve descripción de cómo se manejan los datos en los corpus orales del español, es posible identificar diversas recomendaciones éticas y legales. Algunas de estas

sugerencias se han integrado en la metodología del corpus Ameresco, junto con otras mejoras, para desarrollar un proceso de consentimiento informado en tres etapas. Como se explicará más adelante (capítulo 5), este proceso busca equilibrar la autenticidad del material con el cumplimiento de las leyes de privacidad y protección de datos.

4.3.3. Fase 3. Acceso al corpus por parte de los usuarios

Bajo este epígrafe se compara la manera de acceso a los datos de cada corpus por parte de los usuarios. Si bien, en la sección 4.2.3. tratábamos también de las condiciones de almacenamiento y distribución, para este análisis nos hemos centrado en la manera en que tienen acceso los usuarios a los corpus ya que los otros dos factores se integran dentro de los planteamientos previos de cada corpus y es una información que no se encuentra a disposición pública.

ACCESO A LOS DATOS
COLA: consulta en línea de transcripciones y fragmentos de audio y motor de búsqueda, resultado por concordancias
CORPES XXI: motor de búsqueda, resultado por concordancias
COSER: motor de búsqueda, resultado por concordancias
PRESEEA: motor de búsqueda con acceso a la muestra por ciudades
Val.Es.Co. (2002): no disponible electrónicamente

A excepción del corpus Val.Es.Co. (2002) que fue publicado en papel, por lo que su sistema de acceso a los datos se restringe a la lectura de las transcripciones, el resto de los corpus analizados aquí disponen de un motor de búsqueda en línea abierto al público.

En el caso del corpus COLA, se pone a disposición del usuario la consulta en línea de las transcripciones junto con el fragmento de audio correspondiente a cada intervención; además, cuenta con un motor de búsqueda que recupera la información a través de concordancias, también con acceso al fragmento de audio. Los materiales no se pueden descargar en su totalidad, pero sí los resultados de la búsqueda (solo texto).

CORPES XXI cuenta con un motor de búsqueda que filtra los resultados por concordancias, en algunos casos, no siempre, se puede ver u oír el audio enlazado con la búsqueda. Por cuestiones de derechos de autor, estos materiales no se pueden descargar, más

allá del resultado de las búsquedas. El corpus COSER recupera la información también por concordancias, sin embargo, se ofrece la posibilidad de escuchar los audios completos, no así, la descarga. Sí permite descargar el documento con la transcripción completa y los metadatos. PRESEEA recupera las búsquedas por concordancias a través de su plataforma de búsqueda y permite, además, la descarga de audio y transcripción, incluidos los metadatos, en su totalidad.

4.4. Síntesis del capítulo

En el Capítulo 4 se ha abordado todos los aspectos referidos a las tres fases de diseño y construcción de corpus orales. Como contextualización del capítulo se ha realizado un repaso al marco teórico con respecto a la literatura que expone y discute los planteamientos de corpus orales desde su concepción, pasando por las fases de recogida de los datos, de tratamiento de los datos y de archivo, distribución y acceso al corpus por parte de los usuarios.

En la primera fase de construcción de corpus se ha observado la pertinencia de atender a criterios de tamaño, representatividad, equilibro de la muestra, recogida de las grabaciones, así como a los condicionamientos éticos y legales que operan sobre corpus de material oral (desde derechos de autor a derechos relacionados con la intimidad y la protección de los datos). Resulta relevante, en este sentido, la particularidad de los corpus de conversación coloquial espontánea grabados de forma secreta ya que, los materiales recogidos para este género discursivo están sometidos a una aplicación de la legislación más estricta en cuanto a garantizar el anonimato de los participantes. Otro de los resultados del análisis realizado en esta sección ha sido la determinación de la existencia de una indefinición terminológica respecto a los diferentes procesos por los que deben tratarse los datos en la segunda fase; así, ha resultado en el establecimiento de una definición operativa para dichos procesos, que implican desde la transcripción, a la codificación y anotación, entre otros. En cuanto a la última fase, la fase 3, se han señalado las necesidades de mantener el corpus en el tiempo, esto es, a través de su almacenado y distribución, garantizando su accesibilidad por parte de la comunidad de usuarios y usuarias.

La segunda parte del capítulo presenta un análisis contrastivo de los diferentes diseños de corpus orales del español en relación con las tres fases señaladas. Este análisis ha permitido

establecer las guías de actuación más prestigiosas a la hora de abordar el diseño del corpus oral Ameresco que se detalla en el siguiente capítulo.

Capítulo 5

El corpus Ameresco

A continuación, se describen las particularidades del corpus Ameresco en cuanto a su metodología de recogida y construcción según las tres fases establecidas en el capítulo 4. En primer lugar (§ 5.1) se presentan las características generales del proyecto, esto es sus orígenes, sus objetivos de investigación y los grupos participantes. En segundo lugar (§ 5.2) se aborda el diseño y construcción del corpus atendiendo a las diferentes fases de trabajo; en la fase 1 se trata la concepción y recogida de los datos (§ 5.2.1), en la fase 2 se atiende al tratamiento de los datos (§ 5.2.2) y en la fase 3, se describe el archivo, la distribución y el acceso a los datos por parte de los usuarios (§ 5.2.3). En la sección 5.3 se da cuenta de las dificultades encontradas en cada una de las fases y las propuestas de solución que se han planteado para cada una de ellas. Por último (§ 5.4), se ofrecerá una síntesis del capítulo en el que se recogen las consideraciones más importantes respecto a la construcción del corpus Ameresco.

5.1. Caracterización del corpus Ameresco

El corpus Ameresco (Albelda y Estellés, en línea) nace con el objetivo de recoger muestras de habla reales para el estudio y la caracterización del español coloquial en sus distintas variedades dialectales (Briz 1995, 1998 [2001], 2016). Se centra en conversaciones coloquiales prototípicas grabadas secretamente en las principales ciudades de Hispanoamérica y España. Como se señalaba en el objetivo de investigación número 6 (§ Capítulo 1), fruto del trabajo de esta tesis, el corpus se encuentra en un estado avanzado en cuanto a su construcción: a fecha de octubre de 2023, cuenta con representación de 9 países, con un total de 14 ciudades y 742 170 palabras incluidas en el motor de búsqueda, si bien se dispone de más grabaciones que aún no han sido procesadas. La recogida de materiales no se ha cerrado ya que está prevista la incorporación de nuevos equipos de trabajo de países que aún no cuentan con representación en este corpus. El corpus Ameresco nace, en este sentido, con voluntad de ser un corpus panhispánico.

Su metodología de trabajo en cuanto a la representatividad sociolingüística parte de los principios establecidos por PRESEEA (Moreno Fernández, 2005a, 2016, 2021a, 2021b) y por Val.Es.Co. (Briz y Grupo Val.Es.Co., 1995, 2002). Recoge las prácticas de PRESEEA en cuanto a los puntos de recogida de América y España, la gestión de los grupos −independientes y coordinados desde un grupo central− y el estándar de transcripción; de Val.Es.Co. toma el género discursivo conversacional como objeto de estudio, el protocolo de recogida de las muestras y los fenómenos pragmáticos estudiados, como veremos en la

sección 5.2.1. Respecto al tratamiento informático que recibe, como se verá detenidamente en la sección 5.2.2., el corpus se encuentra anotado y etiquetado en XML, así como alineado audio y texto por medio del programa de anotación ELAN (Carcelén y Uclés, 2019).

El corpus es de acceso abierto y cuenta con un motor de búsqueda que permite diferentes métodos de consulta, detallados en la sección 5.2.3., y tanto las conversaciones como los documentos de trabajo pueden consultarse en línea. Desde su concepción, uno de los presupuestos del corpus ha sido que se constituya como una herramienta útil para cualquier investigador, de ahí que los materiales completos estén a disposición de la comunidad científica.

5.1.1. Orígenes del proyecto

El corpus Ameresco (Albelda y Estellés, en línea) es uno de los frutos principales surgidos de la iniciativa AMERESCO[41], cuyo objetivo "es proporcionar material de referencia que contribuya al estudio del español hablado en su modalidad prototípica: la conversación coloquial en distintas ciudades de América, y favorecer los contrastes y comparaciones posteriores" (Briz, 2016, p. 82). Es por tanto un corpus de carácter panhispánico, carácter que hereda y comparte con el principal corpus panhispánico oral, el corpus PRESEEA, que reúne entrevistas semidirigidas de las principales ciudades de América y España.

Ameresco surge como extensión natural del corpus Val.Es.Co., concebido originalmente por Antonio Briz para estudiar la conversación coloquial espontánea en el español hablado en Valencia (Briz, 1995, Briz y Grupo Val.Es.Co., 2002). Sin embargo, resulta difícil ampliar el estudio de la conversación coloquial en otras variedades dialectales del español, ya que se trata de un género discursivo escasamente representado, como hemos podido observar e la sección sobre panorama de corpus actuales que se incluye en el Capítulo 3. Por tanto, para avanzar en el estudio del género y el registro mismo, era necesaria la construcción y recopilación de un corpus multidialectal de conversaciones coloquiales espontáneas.

Las dificultades metodológicas que el género oral presenta a la hora de ser trabajado desde la lingüística de corpus son muchas si lo comparamos con la construcción de corpus escritos (Briz y Albelda, 2009, Recalde y Vázquez, 2009, Briz, 2012a). Esto es, para la recolección de un corpus escrito basta, *grosso modo*, con recoger material textual que o bien hay que digitalizar usando generalmente *software* de reconocimiento de caracteres OCR, o bien se

[41] Dentro de esta iniciativa se han desarrollado otros proyectos, como se detallará en § 5.1.2.

encuentra ya procesado informáticamente, como es el caso de los libros electrónicos, la prensa digital, etc. Este material requiere una codificación más sencilla, mientras que un corpus oral conlleva una metodología más compleja, empezando por establecer un sistema de transcripción adecuado a los objetivos de investigación, que refleje en mayor o menor medida las características de la oralidad sin someterse a las normas de la escritura; y llevando a cabo el alineado de audio y texto. En este sentido, el corpus Ameresco bebe de los fundamentos metodológicos de los corpus PRESEEA y Val.Es.Co. para su diseño y construcción.

5.1.2. Objetivo de investigación

Como se ha adelantado, el objetivo de la construcción de este corpus es obtener material para el estudio de la conversación coloquial en español (Briz, 2016, p. 82). En particular, su construcción se plantea inicialmente desde un primer proyecto, *Metodología y bases teóricas para el estudio de la atenuación lingüística en España y América* (INV_PRECOMP12-80407), al que siguió el proyecto Es.Var.Atenuación. *La atenuación pragmática en el español hablado: su variación diafásica y diatópica,* (MINECO FFI2013-40905-P) dedicado al estudio de la variación de la atenuación pragmática en el español hablado y cuyo objetivo principal era la obtención de patrones de comportamiento formales y funcionales de este fenómeno en diversas variedades diatópicas y diafásicas del español, así como su contraste. Como continuación del proyecto anterior surge 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), proyecto que aborda la definición teórica y metodológica de la atenuación pragmática y su manifestación en las principales variedades del español, tanto diatópicas como diafásicas. Desde el punto de vista teórico, recoge los principales aspectos señalados en la bibliografía previa a la hora de definir la atenuación y consigna las principales dificultades, proceso tras el cual propone una definición operativa del concepto basada en una visión tripartita (cognitiva/ social/ lingüística). Además, desde este proyecto se ha observado la aparición de diferencias en las funciones, mecanismos y frecuencias de la atenuación en diversas variantes geográficas del español. Una de las aportaciones metodológicas más importantes, tanto para el desarrollo del proyecto mencionado como para ponerlo a disposición de la comunidad científica, es justamente el corpus Ameresco.

Actualmente, los trabajos de recolección del corpus se enmarcan dentro de un tercer proyecto, ESPRINT. *Estrategias pragmático-retóricas en la interacción conversacional conflictiva entre íntimos y conocidos: intensificación, atenuación y gestión interaccional* (PID2020-114805GB-100), que profundiza teóricamente en los fenómenos pragmático-retóricos de la interacción conversacional espontánea y en la recopilación de materiales para su estudio, además de incluir un nuevo corpus de conversaciones orales cotidianas y espontáneas entre íntimos y conocidos en situaciones de conflicto y problematicidad comunicativa.

5.1.3. Grupos participantes

El corpus Ameresco está formado por el equipo central, con sede en Valencia, y por diferentes equipos de trabajo de distintas ciudades hispanoamericanas y españolas, si bien esta última localización está menos explotada, pero está prevista su ampliación en fases de trabajo futuras. En este sentido, no es un corpus cerrado, es decir, podrán seguir incorporándose nuevos equipos de trabajo hasta conseguir representación al menos de cada país hispanohablante.

La confección de un corpus multidialectal necesita indefectiblemente de la colaboración de grupos localizados en las ciudades cuya variedad del español va a ser objeto de estudio. Un trabajo de tal magnitud conformado únicamente por el equipo central, con sede en España, se plantea imposible debido a la escasa financiación a la que se tiene acceso y otros condicionantes metodológicos como la interferencia de la variedad dialectal peninsular septentrional sobre los materiales obtenidos. Este último punto se justifica precisamente por la naturaleza de la conversación coloquial prototípica, género discursivo que necesita de los siguientes rasgos situacionales o coloquializadores (Briz, 2001, p. 41): relación de igualdad entre los interlocutores, relación vivencial de proximidad, marco discursivo familiar y temática no especializada; así como la aparición de rasgos primarios como la ausencia de planificación, la finalidad interpersonal y el tono informal. Dados estos rasgos favorecedores y descriptivos, la recolección del material no puede ser realizada por investigadores desde el equipo central, además de por motivos económicos como hemos señalado arriba, porque su inclusión en el marco interactivo de este género discursivo rompería con la caracterización de la conversación coloquial: es imposible que investigadores externos respeten y cumplan con la relación vivencial de proximidad y que se sitúen dentro de un marco discursivo

familiar. De ahí que sea necesaria la colaboración de personal que sí pueda cumplir con los requisitos mencionados.

Estos equipos son autores de cada uno de los subcorpus, pero se comprometen a seguir la metodología de trabajo propuesta desde el equipo central y a ceder sus corpus para la inclusión en el macrocorpus Ameresco, si bien pueden añadir más sujetos a la muestra y otras especificidades acordes a sus objetivos de investigación, siempre tomando como mínimos los parámetros establecidos desde la dirección del proyecto a través del protocolo de trabajo (Briz *et al*., 2019).

En este momento los equipos colaboradores son los siguientes:

ARGENTINA	**Buenos Aires,** coordinado por Claudia Borzi (Universidad de Buenos Aires). **Tucumán**, coordinado por Silvina Douglas de Sirgo (Universidad Nacional de Tucumán).
CHILE	**Santiago de Chile**, coordinado por Silvana Guerrero y Javier Riffo (Universidad de Chile). **Temuco**, coordinado por Lissette Andrea Mondaca Becerra (Universitat de València).
COLOMBIA	**Barranquilla**, coordinado por Yolanda Rodríguez Cadenas (Universidad del Atlántico Norte). **Medellín**, coordinado por Tetsuji Miyahara y Ana Isabel García Tesoro (Universidad de Antioquia).
CUBA	**La Habana**, coordinado por Ana María Mafud (Universidad de La Habana).
ECUADOR	**Loja**, recogido con la colaboración de Karen Gómez Vicente y coordinado por David Giménez Folqués (Universitat de València).
ESPAÑA	**Las Palmas de Gran Canaria**, coordinado por Marta Samper Hernández (Universidad de Las Palmas de Gran Canaria).
HONDURAS	**Tegucigalpa**, coordinado por Danny Fernando Murillo Lanza (Universitat de València).
MÉXICO	**Ciudad de México**, coordinado por Ricardo Maldonado (Universidad Nacional Autónoma de México). **Monterrey**, coordinado por María Eugenia Flores Treviño (Universidad Autónoma de Nuevo León). **Querétaro**, coordinado por Juliana de la Mora (Universidad de Querétaro).
PANAMÁ	**Ciudad de Panamá**, coordinado por Fulvia Morales de Castillo (Universidad de Panamá).

Tabla 7. Equipos del corpus Ameresco

Además, contamos con los subcorpus cedidos de Iquique (Renata Enghels y Kris Helincks) y Ciudad de México (Katharina Pater), que contienen conversaciones coloquiales prototípicas y periféricas. Además, se han recogido muestras de Bolivia por medio de una colaboración puntual.

En la fecha de redacción de este trabajo, el corpus dispone de un total de 230 conversaciones que contienen 75 horas de grabación, distribuidas como se refleja en la Tabla 8[42].

CIUDAD	**N.º CONVERSACIONES**	**HORAS DE GRABACIÓN**
Barranquilla	8	2h 40 min
Buenos Aires	19	4h 58 min
Ciudad de México	47	22 h
Ciudad de Panamá	12	4 h 55 min
La Habana	36	9 h 6 min
Las Palmas de Gran Canaria	2	1 h
Loja (Ecuador)	12	2h 17 min
Medellín	2	1 h
Monterrey	15	7 h 54 min
Querétaro	12	4 h 12 min
Santiago de Chile	10	3 h 50 min
Tegucigalpa	10	3 h 26 min
Temuco	13	6 h 2 min
Tucumán	5	2 h 6 min
TOTAL	**203**	**75h 26 min**

Tabla 8. Resumen muestras de conversaciones corpus Ameresco

Tras presentar los datos generales sobre el corpus, se muestran las particularidades de su diseño y construcción. En las secciones siguientes se presentarán las diferentes fases de trabajo (1, 2 y 3) y, por último, se ejemplificará con casos reales las dificultades que han

[42] El cómputo se ha realizado sobre los archivos recibidos, si bien, algunas grabaciones aún están en fase de procesado y no están incorporadas a la página web. No se han contabilizado aquí las grabaciones cedidas ni las obtenidas por colaboraciones puntuales.

surgido en cada una de las fases y sus propuestas de solución, así como los problemas observados tanto desde el equipo central como desde los distintos equipos locales.

Téngase en cuenta que, a pesar de que, como hemos visto en la sección 4.2.1., la construcción de un corpus cuenta con unos planteamientos previos a la recogida, momento en el que se toman decisiones sobre todas las áreas que afectan a la construcción del corpus (esto es, objetivo, género discursivo, criterios de representatividad, sistema de transcripción, etc.), la pertinencia de estas decisiones no siempre se mantiene. Es decir, si bien de manera previa se intenta tener una visión de conjunto de las dificultades y obstáculos que pueden aparecer en cada una de las fases, estos no se concretan realmente hasta que no se ejecuta el modelo de diseño elegido y, por tanto, es prácticamente imposible anticipar los problemas venideros. Además, la elección de un marco metodológico previo, en este caso derivado del corpus PRESEEA y Val.Es.Co., conlleva asumir las decisiones adoptadas por dichos equipos siendo conscientes de que pueden existir limitaciones en algunos aspectos. Como señala Wynne (2004, p. 88), si bien es importante lograr una tasa de errores lo más baja posible, existe el peligro de un perfeccionismo excesivo, que puede llevar a una situación en la que el corpus nunca esté terminado, impidiendo su uso y reutilización.

5.2. Diseño del corpus Ameresco

Presentamos a continuación las fases de trabajo fundamentales para la puesta en marcha del corpus Ameresco: la fase 1 de concepción y recogida de los datos (§ 5.2.1.), la fase 2 de tratamiento de los datos obtenidos (§ 5.2.2.) y la fase 3 de archivo, distribución y acceso al corpus por parte de la comunidad de usuarios y usuarias (§ 5.2.3.).

5.2.1. Fase 1. Concepción del corpus y recogida de los datos

En este apartado se muestra cómo se ha configurado la metodología de trabajo del corpus Ameresco. Para comenzar, detallaremos cómo se ha realizado la selección de hablantes (§ 5.2.1.1.); seguidamente, veremos qué requisitos legales se han debido solventar para garantizar una recogida y un tratamiento de datos de acuerdo con el marco ético y legal aplicable (§ 5.2.1.2.). A continuación, presentaremos las características de las grabaciones (§ 5.2.1.3.) así como de la ficha técnica que le corresponde a cada una (§ 5.2.1.4.) y que incluye los metadatos extratextuales de la grabación y los participantes.

Como se ha avanzado más arriba, en la concepción inicial del corpus se plantea que el corpus Ameresco recoja conversaciones coloquiales espontáneas grabadas secretamente

dado que se había observado un déficit de este género discursivo en los corpus existentes (Moreno, 2005c, Briz y Albelda, 2009, Briz, 2012, Enghels, Vanderschueren y Bouzouita, 2015, Rojo 2016a, Solís, 2018, Parodi y Burdiles, 2019, Briz y Carcelén, 2019 y Llisterri, 2021). Además, se concibe como un macrocorpus de carácter panhispánico ya que presenta la voluntad de obtener muestras de las principales variedades del español. Las decisiones concernientes al resto de parámetros se incluirán en la sección correspondiente.

5.2.1.1. Selección de hablantes y tamaño de la muestra

Para la selección de los hablantes se han tomado en consideración los criterios establecidos por los trabajos previos de PRESEEA (Moreno Fernández, 2005a, 2016, 2021a, 2021b) y Val.Es.Co. (Briz y Grupo Val.Es.Co. 1995, 2002) a partir de la estratificación laboviana (Labov, 1966), que sugiere que la variación lingüística está correlacionada con factores sociodemográficos, como la clase social, la educación, la edad y el sexo.

Se ha establecido, por consiguiente, que estos sean o bien nativos de la variedad que se va a recoger, o bien que sean residentes de duración prolongada, siguiendo los parámetros establecidos por Briz y Grupo Val.Es.Co. (2002, p. 14). Así mismo, el tamaño de la muestra se establece según cuotas con afijación uniforme y proporcional (Moreno Fernández, 2021a, p. 13, Briz y Grupo Val.Es.Co., 2002, p. 14), a saber, se establecen estratos atendiendo a criterios sociales (sexo, edad y nivel sociocultural) y se les asigna un número igual de informantes a cada uno. Como señala Moreno Fernández (2021a),

> una razón que nos lleva a preferir este sistema y no una muestra aleatoria o probabilística es que la muestra por cuotas permite una más fácil comparación estadística entre las cuotas internas de la misma muestra y entre muestras diferentes. Por otro lado, las cuotas garantizan la representación de hablantes de los perfiles sociales considerados como fundamentales. (Moreno Fernández, 2021a, p.13)

El establecimiento de estos parámetros responde a la necesidad de homogeneidad para la intercomparabilidad de estudios; sin embargo, dada la sensibilidad del contenido que puede aparecer en una conversación coloquial, los datos referidos a los hablantes deben ser mínimos con respecto a su adscripción a los diferentes estratos establecidos para garantizar su privacidad.

La estratificación final del corpus Ameresco se muestra en la Tabla 9 que se presenta a continuación:

VARIANTE	VARIABLE
Sexo	Mujer Varón
Grupo etario	18-25 26-55 >55
Nivel sociocultural	Bajo: estudios primarios y sin estudios Medio: estudios secundarios y formación profesional Alto: estudios superiores

Tabla 9. Criterios de selección de la muestra (Briz *et al.,* 2019)

Con respecto a la variante *sexo*, se estratifica atendiendo a si es hombre o mujer biológicamente. Por ahora no se ha contemplado la posibilidad de añadir un parámetro que refleje las variables en cuanto a la adscripción de género, si bien es un dato que podría añadirse en la ficha técnica de manera complementaria. Hemos observado que el corpus PRESEEA desde su última actualización de 2021 trabaja con la denominación de sexo/género y contempla la identidad de género no binaria en sus metadatos. Sin embargo, no existe un lugar propio para esta identidad ya que en su metodología indican que en caso de autoadscribirse a esta realidad, el propio hablante deberá decidir en qué categoría (hombre/mujer) inscribirse (Moreno Fernández, 2021a, p. 14). Si atendemos a que género no binario quiere decir que la persona no se identifica con ninguna de estas dos categorías (Ellis y Bartolomé, 2020, p. 21), quizás se podría considerar que se requeriría de una solución alternativa, puesto que esa opción no refleja realmente la casuística. La posibilidad de adscribirse a uno u otro como se ha establecido en este corpus correspondería más bien con la opción de género fluido (Verdejo, 2020, p. 45). Dado que la inclusión de la variable *género* aún no la hemos encontrado, por lo general, en los corpus de español, por el momento no se ha planteado su inclusión hasta que contemos con respaldo metodológico. Si bien, como se ha señalado, esta condición puede recogerse en la ficha técnica.

En lo referente al parámetro *edad*, se toman como referencia los grupos etarios establecidos por el corpus Val.Es.Co. (Briz y Grupo Val.Es.Co. 2002, p. 15): grupo de edad 1 (18-25 años), grupo de edad 2 (26-55 años), grupo de edad 3 (>55 años).

Sin embargo, la estratificación de PRESEEA con respecto a esta variable (20-34, 35-54, 55 en adelante) se basa en estudios más recientes que reflejan de manera más exacta la división etaria actual. Por ello, se decidió modificar la ficha técnica para la recopilación de

los metadatos y, aunque las franjas de edad corresponden al sistema Val.Es.Co., se añade la casilla para incluir la edad concreta, en número, de los participantes (§ 5.2.1.4.). De este modo, no solo se logra que este parámetro pudiera reajustarse en fases futuras, sino que permite a los investigadores que deseen aplicar otros cortes o franjas etarias poder seleccionar los grupos según sus intereses.

La última variante corresponde al *nivel sociocultural*. Para este parámetro, se recoge únicamente el grado de estudios de los participantes y no el salario medio anual, como sí se contempla en la metodología, por ejemplo, de PRESEEA. Se diferencian tres categorías:

- Bajo: hablantes que no tienen estudios (analfabetos) o que tienen estudios primarios.
- Medio: hablantes que tienen estudios secundarios y formación profesional grado medio.
- Alto: hablantes que han terminado estudios superiores y formación profesional grado superior.

Según la metodología del corpus PRESEEA (Moreno Fernández, 2021a, p. 15), el nivel sociocultural puede establecerse además según los años de escolarización cursados, así para el primer nivel corresponderían unos 5 años; para el segundo nivel unos 10-12 años; y para el tercero 15 años de escolarización aproximadamente. En el corpus Ameresco no se ha tenido en cuenta esta segunda capa para el filtrado del nivel sociocultural, sin embargo, esta podría resultar más objetiva y adaptable a las circunstancias de cada país que conforma el corpus, como desarrollaremos en la sección 5.3.1.

En definitiva, la selección de hablantes es homogénea, según principios sociolingüísticos, en base a una muestra tipo que engloba los perfiles necesarios como puede verse en la Tabla 10.

Edad	**Nivel sociocultural**			**Total**
	Alto	**Medio**	**Bajo**	
18-25	V 4 M 4	V 4 M 4	V 4 M 4	V 12 M 12
26-55	V 4 M 4	V 4 M 4	V 4 M 4	V 12 M 12
>55	V 4 M 4	V 4 M 4	V 4 M 4	V 12 M 12
TOTAL	24 (12 V, 12 M)	24 (12 V, 12 M)	24 (12 V, 12 M)	72 (36 V, 36 M)

V (Varón) M (Mujer)

Tabla 10. Resumen de la muestra extraída

Se establece, además, que, por cada celda, que como hemos visto incluye a 8 hablantes, se recojan 3 conversaciones. Esto resulta en una muestra de 27 conversaciones que contendrían un total de 72 hablantes, 8 participantes (4 mujeres y 4 varones) por cada estrato sociolingüístico.

Esta muestra conforma una tabla de mínimos necesarios, sin embargo, como se analizará en la sección 5.3.1., dadas las características de este género discursivo y la imposibilidad de planificar detalladamente las situaciones de grabación, se ha demostrado que 27 conversaciones que recojan a los 72 participantes es un ideal imposible de alcanzar ya que en todos los subcorpus ha sido necesario ampliar esta cifra para contar con la muestra completa de hablantes. El contexto de grabación, esto es, que se realice en situación de familiaridad, espontaneidad y que además la grabación sea secreta, debe ser flexible y amoldarse a estas exigencias.

Las limitaciones en este sentido no se han podido resolver de otra manera en esta fase del proyecto ya que, como se describirá en la sección 5.3., si bien se estableció un protocolo de recogida respaldado metodológicamente, la detección y corrección de estos errores a estas alturas supondría la reelaboración completa del corpus.

Por tanto, somos conscientes de las implicaciones y las limitaciones que pueden surgir en la fase de análisis de los datos, por lo que se ofrecerán propuestas de mejora partiendo de la toma de conciencia de las limitaciones para aportar posibles soluciones de cara al futuro.

5.2.1.2. Requisitos legales: el consentimiento informado

Como hemos visto en el apartado 4.2.1.2., la legislación actual no permite realizar grabaciones si no se cuenta con la autorización previa de los participantes, pero la base de un estudio de la variedad coloquial del español debería garantizar que no se alteren los rasgos primarios puesto que, si se diera un cambio en la familiaridad del marco interactivo del hablante al saberse grabado, el grado de informalidad se vería reducido y, por consiguiente, el grado de prototipicidad de la conversación sería menor (Briz, 1995). Por ello, para el desarrollo del corpus Ameresco se ha diseñado e implantado un modelo de consentimiento informado en tres pasos que garantiza una recogida de datos conforme a la legalidad y que pretende no alterar, o modificar en el menor grado posible, el comportamiento lingüístico de los hablantes.

Este consentimiento informado en el corpus Ameresco instaura un sistema en tres pasos que consiste en (1) una primera autorización previa a la grabación en la que el sujeto otorga su consentimiento a ser grabado en algún momento dentro del espacio de uno o dos meses; (2) una autorización posterior en la que el sujeto confirma que se le ha ofrecido escuchar la grabación y se le da a elegir entre retirarla o cederla con fines de investigación, y (3) una firma de consentimiento para tratar sus datos personales.

En el primer apartado, autorización previa a la grabación, además de rellenar sus datos personales (nombre completo y número de identidad – DNI, cédula o el correspondiente según el país de recogida–), se introduce una de las principales innovaciones del protocolo de recogida de Ameresco: los hablantes son informados de que van a ser grabados en un futuro próximo, normalmente en la horquilla de cuatro a ocho semanas, aproximadamente, pero sin especificar en qué momento exacto; así mismo, se les informa de que, una vez realizada la grabación, podrán escuchar su contenido y retirar el consentimiento inicial en caso de no estar de acuerdo con la cesión del material.

En el segundo apartado, autorización posterior a la grabación, tras finalizar la grabación los hablantes corroboran su voluntad de cederla, afirmando que han sido informados de que acaban de ser grabados, que se les ha ofrecido la posibilidad de escuchar el audio y que pueden retirar su autorización si no están de acuerdo, en cuyo caso se destruiría de inmediato la grabación. Así mismo, autorizan el uso de la grabación con fines estrictamente de investigación, con la particularidad de que el archivo será previamente anonimizado tanto en audio como en el texto transcrito.

Por último, en el tercer apartado, información sobre el tratamiento de los datos personales, los hablantes deben leer y firmar un último apartado sobre cómo se va a proceder al tratamiento de los datos personales. Esta práctica hoy día, como hemos visto en la sección 4.2.1.2., es habitual en cualquier ámbito de la sociedad, especialmente desde la entrada en vigor de la Ley de protección de datos personales y garantía de los derechos digitales (Ley Orgánica 3/2018, de 5 de diciembre) que obliga a informar sobre el uso que se le va a dar esos datos, quién es el responsable de su tratamiento y cómo se puede ejercer el derecho de oposición.

En primer lugar, se ofrece la información pertinente sobre la entidad, empresa o servicio que va a custodiar los datos, a saber, el equipo de investigación y el proyecto concreto que va a trabajar con estas grabaciones. En este caso se trata de la Universitat de València a

través del Proyecto MINECO Es.VaG.Atenuación, dirigido por las doctoras Marta Albelda y Maria Estellés, quienes trabajan en la construcción del corpus oral del español coloquial Ameresco con el objetivo de analizar muestras de habla obtenidas en situaciones de familiaridad y cotidianidad para su posterior análisis lingüístico.

En segundo lugar, ya que hay un registro de audio, aparece un apartado en el que se garantiza que el uso y difusión de este será exclusivamente para fines de investigación y nunca con ánimo de lucro.

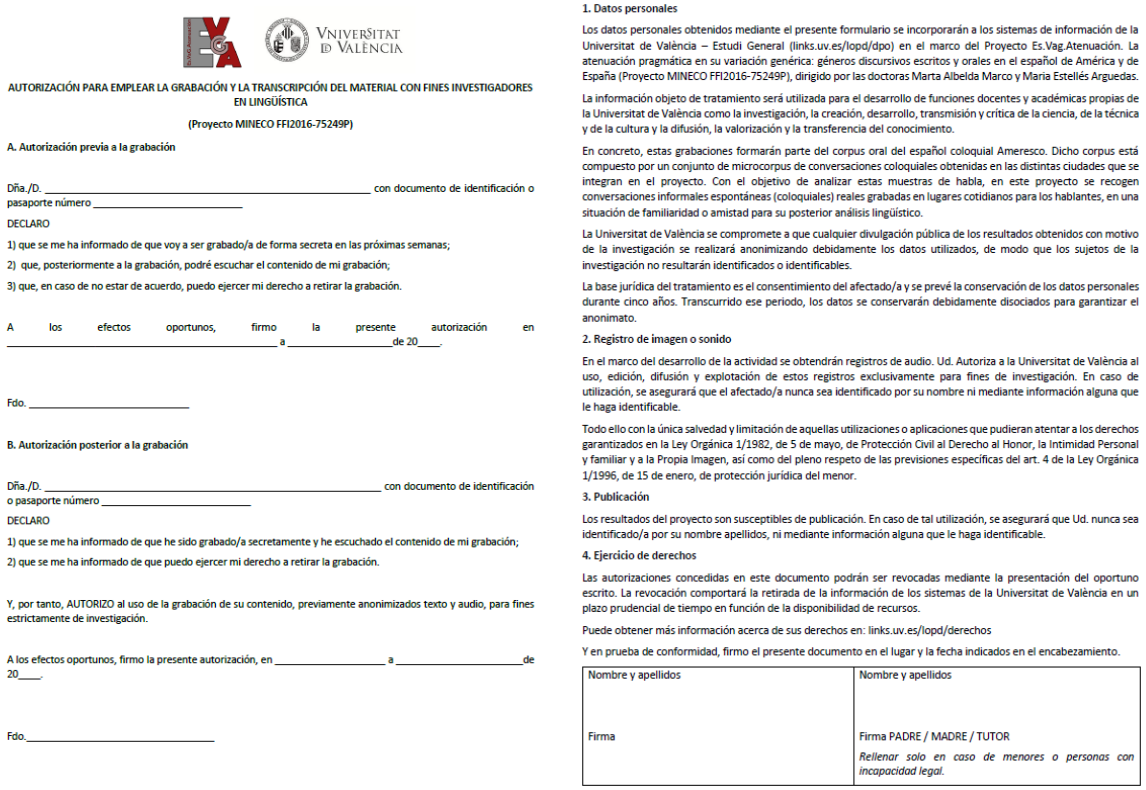
En tercer lugar, se ofrece información sobre las posibilidades de realizar publicaciones de carácter científico en las que se utilicen estos datos, debidamente anonimizados: el proyecto se compromete a anonimizar cualquier dato sensible utilizado en su investigación (publicaciones, congresos, etc.) garantizando que los informantes no podrán ser identificados ni identificables. El sistema de anonimización utilizado en este corpus se detallará en la sección 5.2.2.3.

Por último, se ofrecen los datos necesarios para que el hablante pueda retirar su consentimiento si en algún momento decide que esta grabación deje de formar parte del corpus. Se solicita la firma del participante, así como la firma de la madre, padre o tutor/a en el caso de que en la grabación aparezcan menores de edad o personas con incapacidad legal.

Con este modelo de consentimiento informado el corpus Ameresco cumple por partida triple con la normativa comentada en la sección 4.2.1.2.: obtención del consentimiento previo, del consentimiento posterior a la grabación y del permiso para el tratamiento de datos. La innovación más importante aplicable a la recolección de corpus orales espontáneos la constituye la petición del consentimiento previo no inmediato, que se constituye como una medida esencial para, por un lado, cumplir con los preceptos legales comentados anteriormente y, por otro, para garantizar la naturalidad del comportamiento lingüístico de los participantes. Estos son informados de que van a ser grabados, aunque no sabrán en qué momento exacto. Este lapso de tiempo permite que el hablante olvide que va a ser grabado y, por tanto, su comportamiento lingüístico no debe verse afectado. Se salva así el obstáculo de la paradoja del observador, además de garantizar un comportamiento ético en la recogida de datos.

Este protocolo se instauró tras la entrada en vigor de la nueva política de protección de datos (Ley 3/2018) y fue aprobado por el Comité de Ética y por el gabinete de Servicios

Jurídicos de la Universitat de València tras comprobar que cumplía con la legislación vigente. Se lleva utilizando en la recogida de grabaciones del corpus Ameresco desde ese momento (Briz *et al.*, 2019, Carcelén y Uclés, 2019, Carcelén, en prensa).[43]



AUTORIZACIÓN PARA EMPLEAR LA GRABACIÓN Y LA TRANSCRIPCIÓN DEL MATERIAL CON FINES INVESTIGADORES EN LINGÜÍSTICA

(Proyecto MINECO FFI2016-75249P)

A. Autorización previa a la grabación

Dña./D. _____ con documento de identificación o pasaporte número _____

DECLARO

1) que se me ha informado de que voy a ser grabado/a de forma secreta en las próximas semanas;

2) que, posteriormente a la grabación, podré escuchar el contenido de mi grabación;

3) que, en caso de no estar de acuerdo, puedo ejercer mi derecho a retirar la grabación.

A los efectos oportunos, firmo la presente autorización en _____ a _____de 20____.

Fdo. _____

B. Autorización posterior a la grabación

Dña./D. _____ con documento de identificación o pasaporte número _____

DECLARO

1) que se me ha informado de que he sido grabado/a secretamente y he escuchado el contenido de mi grabación;

2) que se me ha informado de que puedo ejercer mi derecho a retirar la grabación.

Y, por tanto, AUTORIZO al uso de la grabación de su contenido, previamente anonimizados texto y audio, para fines estrictamente de investigación.

A los efectos oportunos, firmo la presente autorización, en _____ a _____de 20____.

Fdo._____

1. Datos personales

Los datos personales obtenidos mediante el presente formulario se incorporarán a los sistemas de información de la Universitat de València – Estudi General (links.uv.es/lopd/dpo) en el marco del Proyecto Es.Vag.Atenuación. La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de América y de España (Proyecto MINECO FFI2016-75249P), dirigido por las doctoras Marta Albelda Marco y Maria Estellés Arguedas.

La información objeto de tratamiento será utilizada para el desarrollo de funciones docentes y académicas propias de la Universitat de València como la investigación, la creación, desarrollo, transmisión y crítica de la ciencia, de la técnica y de la cultura y la difusión, la valorización y la transferencia del conocimiento.

En concreto, estas grabaciones formarán parte del corpus oral del español coloquial Ameresco. Dicho corpus está compuesto por un conjunto de microcorpus de conversaciones coloquiales obtenidas en las distintas ciudades que se integran en el proyecto. Con el objetivo de analizar estas muestras de habla, en este proyecto se recogen conversaciones informales espontáneas (coloquiales) reales grabadas en lugares cotidianos para los hablantes, en una situación de familiaridad o amistad para su posterior análisis lingüístico.

La Universitat de València se compromete a que cualquier divulgación pública de los resultados obtenidos con motivo de la investigación se realizará anonimizando debidamente los datos utilizados, de modo que los sujetos de la investigación no resultarán identificados o identificables.

La base jurídica del tratamiento es el consentimiento del afectado/a y se prevé la conservación de los datos personales durante cinco años. Transcurrido ese periodo, los datos se conservarán debidamente disociados para garantizar el anonimato.

2. Registro de imagen o sonido

En el marco del desarrollo de la actividad se obtendrán registros de audio. Ud. Autoriza a la Universitat de València al uso, edición, difusión y explotación de estos registros exclusivamente para fines de investigación. En caso de utilización, se asegurará que el afectado/a nunca sea identificado por su nombre ni mediante información alguna que le haga identificable.

Todo ello con la única salvedad y limitación de aquellas utilizaciones o aplicaciones que pudieran atentar a los derechos garantizados en la Ley Orgánica 1/1982, de 5 de mayo, de Protección Civil al Derecho al Honor, la Intimidad Personal y familiar y a la Propia Imagen, así como del pleno respeto de las previsiones específicas del art. 4 de la Ley Orgánica 1/1996, de 15 de enero, de protección jurídica del menor.

3. Publicación

Los resultados del proyecto son susceptibles de publicación. En caso de tal utilización, se asegurará que Ud. nunca sea identificado/a por su nombre apellidos, ni mediante información alguna que le haga identificable.

4. Ejercicio de derechos

Las autorizaciones concedidas en este documento podrán ser revocadas mediante la presentación del oportuno escrito. La revocación comportará la retirada de la información de los sistemas de la Universitat de València en un plazo prudencial de tiempo en función de la disponibilidad de recursos.

Puede obtener más información acerca de sus derechos en: links.uv.es/lopd/derechos

Y en prueba de conformidad, firmo el presente documento en el lugar y la fecha indicados en el encabezamiento.

Nombre y apellidos	Nombre y apellidos
Firma	Firma PADRE / MADRE / TUTOR *Rellenar solo en caso de menores o personas con incapacidad legal.*

Figura 23. Modelo de consentimiento informado del corpus Ameresco

Previamente a la promulgación de la ley, en la recogida de los primeros materiales del corpus Ameresco ya se solicitaba la autorización previa y posterior en los términos que se han comentado arriba, esto es, se pedía el permiso previo para grabar sin especificar el momento exacto en el que se realizaría la grabación. Tras la publicación de la Ley 3/2018 se incorpora, además, el apartado específico sobre la política de protección de datos.

Es preciso señalar en este apartado que estos consentimientos se guardan de manera aislada y encriptada, nunca junto al resto de materiales que se recogen en el corpus (grabaciones, transcripciones y fichas técnicas) y que las gestiones relacionadas a este respecto son llevadas a cabo por una persona responsable dentro del equipo central. Así mismo, hay equipos locales que han preferido mantener la salvaguarda de los consentimientos, si bien, han enviado al equipo central un certificado en el que se asegura

[43] Ver Anexo 2.

que se han recogido las autorizaciones necesarias siguiendo el modelo en tres pasos proporcionado por la dirección del proyecto.

5.2.1.3. Grabaciones

Trataremos en esta sección sobre las características que deben reunir las grabaciones. En este sentido, es necesario distinguir dos circunstancias principales que deben considerarse. En primer lugar, detallaremos aquellas que tienen que ver con el papel de la persona encargada de realizar la grabación (§ 5.2.1.3.1.). En segundo lugar, con las características relacionadas con el contexto de grabación que pueden afectar a la calidad del audio, es decir, los requisitos técnicos (§ 5.2.1.3.2.).

5.2.1.3.1. El papel de la persona encargada de recoger la grabación

Dado el carácter secreto que exige este género discursivo, las grabaciones se realizan por medio de dispositivos de grabación situados en lugares estratégicos para garantizar una calidad de audio óptima. Mayoritariamente se han utilizado *smartphones* para la recogida, ya que su presencia pasa completamente desapercibida en el contexto de grabación y, además, se ha comprobado que realizan grabaciones de buena calidad. También se han utilizado otros medios como bolígrafos espías o grabadoras pequeñas camufladas, por ejemplo, por medio de broches que llevaba la persona encargada de realizar la grabación.

Antes de realizar la recogida de las conversaciones, las personas responsables de hacer la grabación reciben unas instrucciones muy precisas. La primera instrucción tiene que ver con el hecho de que el resto de los participantes no deben saber el momento exacto en el que están siendo grabados. Si existe la sospecha de que se ha revelado que están siendo grabados en ese momento, el material recogido no se incorporará al corpus.

Por otro lado, la persona encargada de grabar debe tener claro el grado de implicación y participación que debe adoptar. Sería ideal que el contexto de recogida permitiera que quien se encargue de grabar dejara el dispositivo en una ubicación cercana a los sujetos de grabación y que no participara de la conversación recogida. En este caso, su papel en la interacción sería el de investigador/a no participante. Sin embargo, esto no siempre es posible, ya que en algunos casos el hecho de desaparecer de la escena podría levantar sospechas sobre la posibilidad de estar siendo grabados. Como se detalla en el apartado

anterior, los hablantes han sido informados de que van a ser grabados en algún momento futuro, pero no de cuándo será concretamente.

En los casos en que sea inviable la retirada de escena de la persona encargada de la grabación, es recomendable que esta se encuentre presente durante el transcurso de la conversación, pero que su participación se restrinja al mínimo posible para que no haga sospechar al resto de participantes. Por tanto, su papel en la interacción sería el de investigador/a participante en calidad de hablante pasivo. Una intervención constante y continuada a lo largo de la grabación por parte del investigador/a daría lugar a la desvirtualización de la espontaneidad de la conversación y a la alteración del grado de prototipicidad de la conversación coloquial (Briz, 1998 [2001]).

Siguiendo a Briz y Grupo Val.Es.Co. (2002),

> la grabación secreta, con observación participante o sin ella, ha sido la técnica más empleada en la recogida de datos, ya que constituye la forma más eficaz de obtención del español coloquial y permite soslayar inconvenientes teóricos como la llamada paradoja del observador (Labov, 1983, Hidalgo y Pons, 1991). (Briz y Grupo Val.Es.Co., 2002, p. 17)

5.2.1.3.2. Requisitos técnicos

En cuanto a los requisitos técnicos, las grabaciones recogidas deben cumplir con una calidad de audio suficientemente adecuada, tanto para que posteriormente sea posible la realización de estudios prosódicos, como para garantizar que el proceso de transcripción sea posible con el mayor grado de fidelidad deseable. Por ello se recomienda seguir ciertas indicaciones que garanticen que las características acústicas sean adecuadas, recomendaciones que se han ido ampliando y perfilando a medida que han surgido las dudas y dificultades planteadas por los equipos de grabación, como se muestra en § 5.3.1.

Por un lado, la grabación debe realizarse preferiblemente en lugares interiores particulares, evitándose las grabaciones en lugares abiertos al aire libre o interiores con mucha afluencia de personas, como bares y cafeterías. Las conversaciones recogidas en estos contextos poseen un alto nivel de contaminación acústica y se han tenido que desechar, ya que la transcripción bajo estas circunstancias se convierte en una tarea muy difícil y el grado de fidelidad con lo dicho se ve comprometido, como detallaremos en la sección 5.3.1.

Por otro lado, el siguiente requisito técnico contemplado tiene que ver con la cantidad de participantes que idealmente deben aparecer en la grabación. Así, se ha fijado un mínimo de dos hablantes (excluyendo del cómputo a la persona encargada de recoger la conversación) y un máximo de cuatro o cinco. El máximo tiene que ver con la tarea posterior de transcripción, es decir, se ha comprobado que un número mayor de participantes acaba dando lugar a conversaciones paralelas y a una gran presencia de intervenciones simultáneas, lo que dificulta, cuando no imposibilita, la transcripción. En cuanto al mínimo, una grabación de dos personas implica que el investigador (participante en modo pasivo o no participante) esté presente de una de estas dos maneras, como veremos más adelante.

5.2.1.3.3. Duración de las grabaciones

Como tiempos orientativos, la mayoría de las grabaciones recogen entre 20 y 60 minutos. Debido a las características propias de la conversación coloquial, no puede preverse un tiempo fijo, sino que hay que partir siempre de que no se debe forzar la conversación para obtener mayor tiempo de grabación y que, si se está recogiendo una conversación, esta no debe cortarse por haber superado una duración predeterminada. Como en todos los parámetros, por la propia naturaleza del corpus, se elige siempre la naturalidad por encima del cumplimiento estricto de los requisitos técnicos ideales.

5.2.1.4. Recogida de la ficha técnica (metadatos)

En el corpus Ameresco cada grabación va necesariamente acompañada de una ficha técnica que recoge información general sobre la conversación y sobre los participantes, así como las particularidades de la situación comunicativa. Esta ficha se ha tomado de Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002, pp. 36-37).

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 Grabación secreta
  Observador no participante Grabación ordinaria

- Conversación semidirigida (grabación ordinaria):

f) Descripción de los participantes:

- Número de participantes: Clave:
  activos: pasivos:
- Tipo de relación que los une:
- Sexo:
  varón:
  mujer:
- Edad:
  18-25
  26-55
  >55
- Nivel de estudios:
  analfabetos:
  primarios:
  secundarios:
  medios:
  superiores:
- Profesiones:
- Residencia o domicilio habitual:
- Nivel sociocultural (*tendremos en cuenta únicamente el nivel de instrucción):*
  Nivel alto: estudios superiores
  Nivel medio: estudios de secundaria y formación profesional
  Nivel bajo: estudios primarios o sin estudios

- Lengua habitual:
  monoling. cast.:
  biling.:

g) Grado de prototipicidad coloquial:

conversación coloquial prototípica:
conversación coloquial periférica:

Tabla 11. Modelo de ficha técnica de Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002)

En primer lugar, aparecen los *datos identificadores de la conversación*; por un lado, quién es el investigador responsable (apartado a) y, por otro, los *datos identificadores de la grabación* (apartado b), es decir, en qué momento se recogió, cuál es la duración de la grabación y en qué lugar se grabó. Hay que matizar que este lugar de grabación hace referencia a la ubicación genérica, nunca a la localización exacta, ya que si se incluyera esta se violaría el compromiso de anonimización y protección de datos personales. Por tanto, puede indicarse que la grabación se realizó en una casa particular o en un parque, por ejemplo, pero no se daría la información concreta de dónde se ubica esa localización.

En el siguiente bloque encontramos los ítems referidos a la *situación comunicativa*. En el apartado c) se recogerían los temas tratados en la conversación, por ejemplo, tomando el caso del archivo TGU_016_02_19 de Ameresco-Tegucigalpa incorporado en la página del corpus, corresponderían a *cocina, familiares en el extranjero, problemas familiares, acoso sexual, religión, relaciones maritales, estudio y viajes*. En cuanto al *propósito o tenor funcional* predominante se ha de escoger entre interpersonal o transaccional, clasificación tomada de Briz (1998 [2001]) y que diferencia aquellas conversaciones que responden a la finalidad de hablar por hablar o aquellas que persiguen un fin concreto. El *tono* hace referencia a si es formal o informal; y *el modo o canal* a si es oral o escrito. En el apartado d) se da cuenta del *tipo de discurso*, que debería ser indefectiblemente conversación coloquial espontánea, si bien el modelo de ficha podría ser reutilizable para otro género discursivo[44]. En cuanto a la *técnica de grabación*, se recogen varios parámetros, aunque no corresponden únicamente con los esperables para el género discursivo de la conversación coloquial ya que, como hemos mencionado, la ficha puede ser extrapolable a otros objetivos de investigación y, además, recoge el papel que la persona encargada de recoger la grabación puede adoptar, siendo participante pasivo o directamente no participar de la grabación, según hemos descrito en la sección 5.2.1.3.1. Para el corpus Ameresco, la modalidad de grabación secreta es obligatoria, si bien la participación del observador puede variar.

En el caso de Ameresco, para que la grabación pueda incorporarse al corpus, el material oral debe tener un propósito o tenor interpersonal, un tono informal, un modo o canal oral espontáneo, el tipo de discurso debe ser conversación coloquial espontánea y la técnica de grabación conversación libre, con observador participante o no, y grabación secreta. En el

[44] Inicialmente estaba prevista la posibilidad de que el corpus Ameresco recogiera otros géneros discursivos además de la conversación coloquial. Sin embargo, esta tarea se realizará en fases futuras.

caso de no cumplir con alguna de estas características, la grabación pasaría al repositorio de archivos.

Por último, encontramos el bloque de *información acerca de los participantes*. Se indica en esta parte el número total de participantes y se les asigna una letra identificativa siguiendo el patrón A, B, C, D, etc. Nunca se utilizará la inicial del nombre real por cuestiones de protección de datos. Se identifica quiénes participan de manera activa y quiénes de manera pasiva (los hablantes pasivos no computan en la recogida de la muestra). Se ofrecen a continuación los datos sociolingüísticos de cada participante, esto es, su sexo, edad y nivel sociocultural; así mismo se detalla información relativa a su profesión, su lugar de residencia y su lengua habitual.[45]

En último lugar se señala si la grabación cumple con todas las características esperables de la conversación coloquial prototípica, según la clasificación de Briz (1998 [2001]).

El formato de esta ficha se reconfiguró en un momento dado ya que en el grupo Val.Es.Co. −para sus datos− y con los materiales recogidos por Ameresco hasta entonces, se detectaron errores en su cumplimentación por parte de las personas encargadas de realizar la grabación. Esto puso de manifiesto que quizá la información no estaba presentada con suficiente claridad o que no se estaban ofreciendo las instrucciones necesarias para su recolección. Por tanto, se adoptó un nuevo formato que, si bien recoge los mismos ítems, los presenta en forma de tabla e incluye algunas especificaciones para que el grado de error sea el mínimo posible (§ 5.3.1.). Además, en el protocolo de trabajo se ofrece un modelo ya cumplimentado que sirve como ejemplo sobre cómo debe completarse. La actual ficha técnica se presenta en el siguiente formato, como puede verse en la Figura 24 (y en el Anexo 3):

[45] El corpus Val.Es.Co. recoge conversación en español en una ciudad bilingüe, de ahí que aparezca este ítem en la ficha.

FICHA TÉCNICA

a) Investigador:

b) Datos identificadores de la grabación:

- Fecha de la grabación:

- Tiempo de la grabación

Duración del audio	
Momento de inicio de la transcripción (minuto:segundo)	
Momento de finalización de la transcripción (minuto:segundo)	

- Lugar de grabación:

Municipio	
Espacio concreto (casa, bar, aula…)	

c) Situación comunicativa:

- Temas:

- Propósito o tenor funcional predominante (marcar con una X una de las dos opciones):

Transaccional	
Interpersonal	

- Tono:

- Modo o canal:

d) Tipo de discurso (conversación, debate…):

e) Técnica de grabación (seleccionar una de las dos opciones de cada fila):

Conversación libre / semidirigida	
Investigador participante / no participante	
Grabación secreta / no secreta	

f) Descripción de los participantes:

- Número de participantes:

- Tipo de relación que los une (familia, amigos, hermanos, compañeros de piso…):

- Cuadro de rasgos sociolingüísticos:

Clave hablante	Sexo (V/M)	Edad (especificar edad exacta dentro de una de las tres franjas etarias)			Nivel de instrucción			Activo/Pasivo	Monolingüe cast./Bilingüe	Profesión	Residencia habitual (municipio)
		18-25	26-55	≥56	Bajo	Medio	Alto				
A											
B											
C											
D											
E											
F											
G											

g) Grado de prototipicidad coloquial (marcar con una X una de las dos opciones):

Conversación coloquial prototípica	
Conversación coloquial periférica	

Figura 24. Modelo actual de ficha técnica del Grupo Val.Es.Co. adoptado para Ameresco

5.2.2. Fase 2. Tratamiento de los datos

A continuación, describiremos el protocolo para el tratamiento de los datos adoptado por el corpus Ameresco. En primer lugar, veremos las especificades de los dos modos de trabajo existentes (§ 5.2.2.1.); a continuación, se tratarán los métodos de revisión y validación del etiquetado de los archivos (§ 5.2.2.2.). Posteriormente, se detallará el sistema de anonimización, tanto de audio como de texto (§ 5.2.2.3.). Por último, se explicará la nomenclatura utilizada para la identificación de archivos (§ 5.2.2.4.).

Para la transcripción y codificación de las conversaciones que componen el corpus Ameresco se ha utilizado, por un lado, el sistema de convenciones establecido por el grupo Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002), en combinación, por otro lado, con el sistema de etiquetado XML propuesto por el proyecto PRESEEA (Moreno Fernández, 2021a, 2021b) debido a que, como señalan Carcelén y Uclés (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. (Carcelén y Uclés, 2019, p. 25)

5.2.2.1. Modos de trabajo

Como se desarrollará en el apartado 5.3, inicialmente, no todos los equipos locales han contado con la posibilidad o disponibilidad necesaria para trabajar la alineación audio y texto a través de ELAN. Por ello, en el caso de alguno de los subcorpus, la ejecución de esta fase ha resultado ser más compleja y extendida en el tiempo

En concreto nos enfrentamos a cuatro circunstancias externas que condicionan la adopción de la metodología de trabajo Ameresco:

1. Equipos que trabajan en ELAN alineando audio y transcripción.

2. Equipos que pudiendo trabajar en ELAN no lo hicieron por falta de personal o por insuficiencia de este.

3. Equipos que no podían trabajar directamente en ELAN por motivos técnicos. Un ejemplo de esta situación sería el subcorpus Ameresco-La Habana, donde los requisitos

técnicos necesarios para trabajar directamente en ELAN no se podían alcanzar por limitaciones derivadas de las circunstancias del país: el acceso a internet durante el periodo de recolección del corpus era limitado y no se contó con medios técnicos suficientes. Por lo tanto, hubo que adaptar la fase de transcripción y codificación a estos condicionantes. Tanto en el caso 2 como en el 3, el equipo local realizaba la transcripción por medio de un procesador de textos.

4. Existe una cuarta circunstancia en la que equipo local no ha podido realizar la transcripción, ni con procesador de textos ni en ELAN, por falta de personal, pero sí ha podido recoger las grabaciones; no obstante, a efectos de los modos de trabajo que comentaremos a continuación, la metodología es similar a la de las circunstancias referidas en 2 y 3.

De estas circunstancias se deriva la existencia de dos metodologías de trabajo paralelas dentro del corpus Ameresco, como se detalla a continuación. Actualmente, se procura que los nuevos equipos de trabajo que se incorporan al proyecto trabajen directamente en ELAN, lo que implica la impartición de sesiones de formación síncrona en línea por parte del equipo central cada vez que se incorporan nuevos grupos o cambian los integrantes del equipo de trabajo (§ 5.3). Esta carga adicional se compensa porque, de este modo se agilizan las tareas de codificación y de incorporación de materiales al motor de búsqueda en línea.

Ambas metodologías combinan una parte común, la que aplica a la recogida de las grabaciones, la obtención de las autorizaciones y la ficha técnica. Sin embargo, difieren en lo concerniente al sistema de transcripción y codificación, como veremos en los apartados siguientes. Hay que tener en cuenta que las transcripciones realizadas con un procesador de texto necesitan de una segunda fase de trabajo, como puede verse en la Figura 25. Esta fase se desarrolla en el equipo central y en ella el material recogido se procesa con ELAN partiendo de la transcripción base realizada por los investigadores locales.

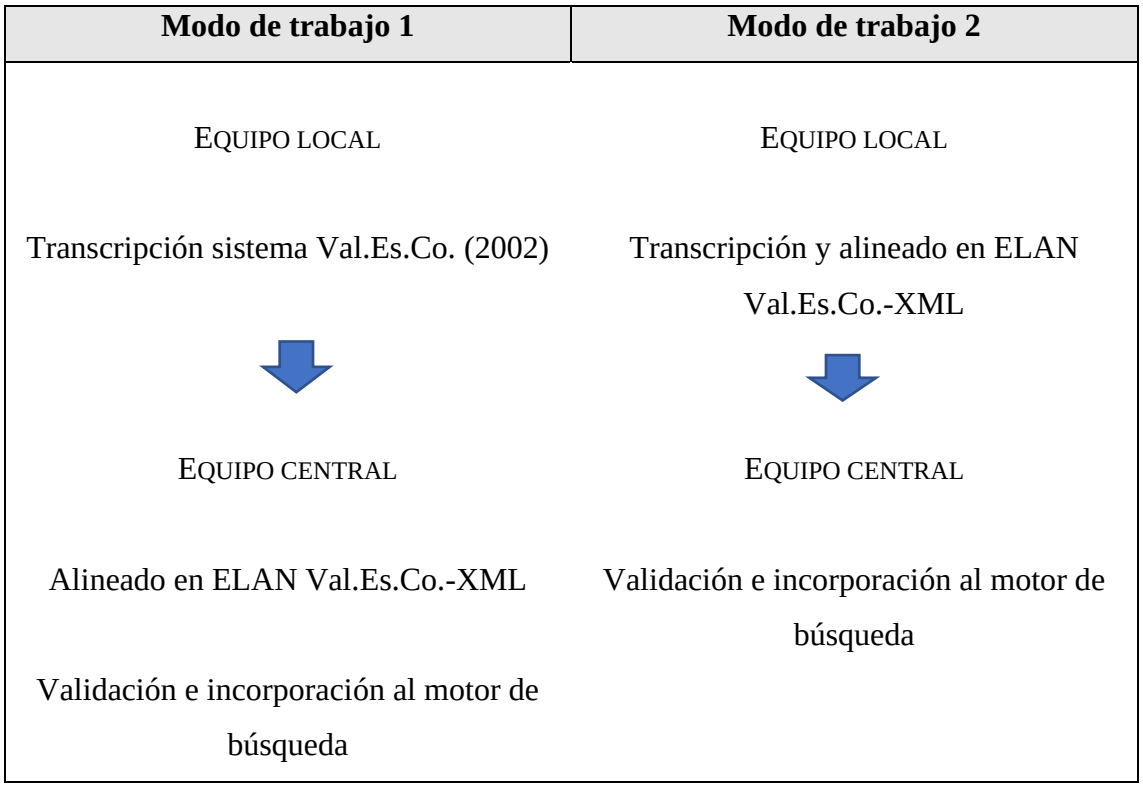


Figura 25. Posibles modos de trabajo del corpus Ameresco

5.2.2.1.1. Protocolo de transcripción y codificación del Modo de trabajo 1

Una vez obtenidas las grabaciones, con sus correspondientes autorizaciones y fichas técnicas, desde el equipo local se procede a la realización de las transcripciones. El sistema de transcripción pautado para el corpus Ameresco se adapta del sistema de transcripción del corpus Val.Es.Co. (2002). Este sistema combina una transcripción que refleja fenómenos propios de la oralidad tales como contracciones, pérdida de la *-d-* intervocálica, aspiraciones o solapamientos, entre otros, con un sistema de signos y símbolos que resulta fácil de interpretar por los usuarios del corpus. En este modo de trabajo, la transcripción se realiza utilizando un procesador de texto, por tanto, no es una transcripción alineada temporalmente.

Se ha optado por una simplificación de las convenciones[46] y como resultado, se utilizarán solamente las siguientes marcas para una transcripción ancha:

[ ] **Simultáneos**

[46] El sistema de transcripción Val.Es.Co. completo puede consultarse en Briz y Grupo Val.Es.Co. (2002, pp. 29-30).

Los corchetes marcan el inicio y el final del habla simultánea, es decir, de aquellos fragmentos que se hayan pronunciado al mismo tiempo por dos o más hablantes. Estos se colocan en las intervenciones de afectadas de todos los hablantes que participan del solapamiento y se visualiza de manera escalar de la siguiente manera:

Ejemplo 1

B: entonces eran [mis favoritos]

A: [aa Rogers]

(SCL_002_03_18)

-
Reinicios y autointerrupciones

El guion se utiliza para marcar reinicios y autointerrupciones realizadas por el hablante durante su discurso. Se coloca justo tras el segmento afectado.

Ejemplo 2

C: andini- andinista lle- lleva ocho días desaparecido/ el radar de siete tazas

(SCL_002_03_18)

/
Pausas

Independientemente de su duración, la barra / se emplea para señalar que hay una pausa. Se coloca entre a continuación de la última palabra emitida.

Ejemplo 3

A: no/ eso fue antes de ayer

(SCL_006_05_18)

(2”)
Silencio

Cuando se produce un silencio prolongado se anota la duración de este por medio de esta convención. Entre paréntesis se añadiría los segundos exactos que dura el silencio.

Ejemplo 4

A: mami ahora ya no llueve (6’) apenas caen unas chispitas

(SCL_006_05_18)

((transcripción))

Se representan entre paréntesis doble aquellos fragmentos en los que el transcriptor no está completamente seguro de la reconstrucción.

Ejemplo 5

E: ¿a qué ((dijiste))?

(SCL_006_05_18)

((…))
Interrupción en la grabación

Indica que ha habido una interrupción en la grabación, por ejemplo, un corte.

Ejemplo 6

A: si quieren ir a la mañana y a la tarde que vayan ((…))

(TUC_001_04_13)

(( ))
Fragmento indescifrable

Los dobles paréntesis sin contenido en su interior se utilizan para indicar que lo dicho en ese momento no se ha podido transcribir por dificultades de comprensión por parte del transcriptor. Es decir, no se ha podido identificar ninguna palabra claramente, bien por una mala calidad de audio, bien por mala vocalización del hablante. En este caso, ni siquiera ha sido posible una reconstrucción aproximada o parcial.

Ejemplo 7

B: ahí cacho/ un- un- y pone sistema cuánto DP no sé cuánto/ (( ))]

(SCL_011_03_18)

(en)tonces
Reconstrucción

Los paréntesis simples se emplean para reconstruir una unidad léxica que se ha pronunciado de forma incompleta.

Ejemplo 8

C: pa(ra) la embarra(da) oo no puedo mover denante no podía mover (SCL_002_03_18)

pa'l **Fenómenos de fonética sintáctica**

La apostrofación refleja fenómenos de fonética sintáctica entre palabras.

Ejemplo 9

C: (es)tai clarito (RISAS)/ pa'cá al la(d)o o pa'acá al frente

(SCL_002_03_18)

h **Aspiración**

A través de este carácter se refleja la aspiración de la *s* implosiva.

Ejemplo 10

B: [a loh máh grande]

(TUC_001_04_13)

° ( ) ° **Susurro**

Esta marca se emplea para indicar que un fragmento se ha pronunciado con intensidad baja, próxima al susurro. Se utilizan los signos combinados de grado ° y el paréntesis. Dentro del paréntesis se colocaría la transcripción que se ve afectada por esta circunstancia.

Ejemplo 11

D: °(pero él pa' arriba no sube)°

(LOJ_001_04_20)

PESADO **Pronunciación enfática**

Mediante el uso de la mayúscula se representan aquellos elementos que se han pronunciado de una manera marcada o enfática.

Ejemplo 12

A: entonces Barbarita ¡QUÉÉÉ CHIQUILLAA!

(HAV_069_03_17)

pe sa do **Silabeo**

Se utiliza la división en sílabas para señalar que esa palabra o fragmento se ha pronunciado separadamente.

Ejemplo 13

A: wave/ mi cro/ [wa ve]

(HAV_002_04_12)

(RISAS) (TOSES) (GRITOS)

Estas tres unidades indican que aparecen risas, toses o gritos en la conversación. La marca exige la escritura en mayúscula de la circunstancia correspondiente.

Ejemplo 14

A: intento levantarme a las 4 de la mañana y cuando miro son como las nueve de la noche (RISAS)

(HAV_069_03_17)

Ejemplo 15

B: [(TOSES)]

A: [veni- venían] Los Tolcos en un camión entonce(s)

(TGU_07_03_09)

aa nn **Alargamientos vocálicos y consonánticos**

Para reflejar que se ha producido un alargamiento, ya sea vocálico o consonántico, este sistema utiliza la duplicación del carácter correspondiente. Independientemente de la duración del alargamiento, solo se utilizarán dos caracteres para su representación.

Ejemplo 16

C: otra cosa que veo mal es/ que que / le venden demasiadoo / cosaa a la gente

(HAV_069_03_17)

Cursiva **Estilo directo**

Se emplea esta tipografía para identificar la reproducción e imitación de emisiones en estilo directo.

Ejemplo 17

B: y le decía *no pa(ra) allá no y me decía noo si piola si no pasa nada*

(SCL_002_03_18)

¿? ¡! ¿¡!?

Los fragmentos interrogativos, exclamativos o interrogativos-exclamativos se delimitan con sus correspondientes signos de puntuación. Recordemos que son los únicos signos de puntuación que se utilizan en este sistema de transcripción.

Ejemplo 18

C: ¡qué bueno mija!/ ay pero mija las colas en las tienda estáan /vaya

B: ¿muchas colas?

C: ¡demasiadoo!

(HAV_069_03_17)

Notas a pie de página

Se utiliza este recurso para añadir aclaraciones u observaciones relevantes sobre algún aspecto de la grabación o la transcripción, como, por ejemplo, anotaciones pragmáticas u otras informaciones necesarias para la correcta interpretación de determinadas palabras.

Ejemplo 19

B: tú no tienes caja *

* Hablante A mirando el televisor

(HAV_069_03_17)

Además, como hemos adelantado arriba, no se utilizan signos de puntuación, a excepción de las interrogaciones y las exclamaciones. En consecuencia, no se usan las mayúsculas por razones de puntuación, salvo para los casos de nombres propios y siglas que aparezcan, así como para marcar la pronunciación enfática. Las cifras y símbolos se transcriben con letra.

En el caso de los nombres propios o de lugares que deban ser anonimizados, la instrucción es la de su sustitución por otro ficticio (§ 5.2.2.3.).

Según este sistema,

> los signos ortográficos convencionales son de gran utilidad por dos razones: en primer lugar, porque el usuario del sistema de transcripción ya está familiarizado con estos signos y, por lo tanto, su descodificación no presenta ningún problema […]. En segundo lugar, estos signos se integran en el propio proceso de lectura y representación de la muestra de habla. (Briz y Grupo Val.Es.Co., 2002)

5.2.2.1.2. Protocolo de transcripción y codificación del Modo de trabajo 2

Para la ejecución de este modo de trabajo partimos de dos contextos previos: el primero, consistente en la recepción de la transcripción ancha en un formato obtenido a través de un procesador de texto; el segundo, omisión de la primera transcripción ancha utilizando el sistema Val.Es.Co. (2002), y realización de la transcripción directamente con el *software* de anotación ELAN.

En el primer caso, será el equipo central el encargado de alinear audio y texto por medio del programa ELAN, siguiendo el sistema de codificación combinando símbolos del sistema Val.Es.Co. con un sistema de etiquetado XML de base TEI, adaptado de la metodología PRESEEA (Moreno Fernández, 2021b). En el segundo caso, el equipo local realizaría la transcripción directamente en ELAN siguiendo las mismas indicaciones.

Una de las ventajas del uso de ELAN, además de que permite trabajar con audio y texto a la vez desde la misma interfaz de trabajo, es que admite la creación de líneas de hablantes (*tiers*) en las que se incluyen las intervenciones de cada participante, además de permitir la creación de otras líneas dedicadas a la inclusión de observaciones, así como para automatizar, como se detallará en 5.2.2.3., las tareas de anonimización.

Recuperando lo dicho en la sección 4.2.2.4., TEI (*Text Encoding Initiative*) constituye un conjunto de directrices y estándares para la codificación de textos electrónicos originado con el fin de desarrollar un sistema común para la creación y el intercambio de textos digitales. Esta iniciativa proporciona una serie de reglas y pautas para la codificación de textos en diferentes formatos y para diferentes propósitos, incluyendo la investigación académica, la preservación de textos antiguos y la creación de ediciones críticas. La codificación de textos

según las pautas de TEI permite una mayor flexibilidad y capacidad de reutilización de los datos, y facilita la creación de herramientas para el análisis y la visualización de los textos. Por otro lado, XML (*eXtensible Markup Language*) es un lenguaje de marcado que se utiliza para almacenar y transportar datos de manera estructurada. Consiste en un conjunto de reglas para etiquetar documentos de texto con información sobre su estructura y contenido. Este lenguaje es el más utilizado en el procesamiento de lenguaje natural y en consecuencia en la construcción de corpus lingüísticos.

Las etiquetas XML se utilizan para definir elementos que, a su vez, pueden contener atributos y valores para anidar unos elementos dentro de otros, por tanto, son muy útiles a la hora de codificar rasgos propios de la oralidad ya que permite diferenciar entre caso y tipo, entre forma y lema. Es decir, se pueden crear etiquetas personalizadas para adaptarlo a las necesidades específicas de una aplicación, en nuestro caso, la codificación de un corpus oral. Se puede utilizar en diferentes sistemas operativos y aplicaciones, y permite el procesamiento informático del corpus para la posterior recuperación de la información por medio de un motor de búsqueda (Carcelén y Uclés, 2019, p. 27).

En este sistema de transcripción se diferencian dos clases de etiquetas: las etiquetas simples y las etiquetas dobles.

- Las etiquetas simples estarían compuestas por un solo elemento: <ininteligible/>, <obs t=“ ”/>. Se colocan entre espacios, con excepción de la etiqueta <alargamiento/> que se escribe pegada al carácter al que afecte.
- Las etiquetas dobles se componen de dos elementos, uno de apertura y otro de cierre: <cita></cita>, <énfasis t=“ ”></énfasis>. Se escriben pegadas al fragmento de texto al que afecta.

Ambas permiten, además, la introducción de atributos, esto es información adicional que no forma parte de la transcripción y que se incluye en la sección entrecomillada de la etiqueta de inicio: <énfasis t=“pronunciación_marcada”></énfasis>, <obs t=“habla con la boca llena”/>.

Además, la posibilidad de añadir información en el atributo es una característica muy útil a la hora de codificar lengua hablada respetando las características propias de la producción oral. Permite reflejar distintas realizaciones fónicas bajo la misma forma, por tanto, tras su

incorporación al motor de búsqueda agiliza y simplifica las tareas de búsqueda por parte del usuario. Con este propósito, se ha desarrollado la etiqueta doble de fonética sintáctica reducida que recoge, por un lado, la representación normativa y, por otro, la realización oral específica <fsr t=“ ”></fsr>.

Esta diferenciación afecta especialmente a los fenómenos discursivos en los que la realización oral y su representación normativa es variable, no solo en cuestiones relacionadas con la fonética sintáctica (<fsr t=“ ”> </fsr>), sino también con la aparición de extranjerismos o fragmentos en otra lengua (<extranjero t=“ ”> </extranjero>), así como de siglas (<siglas t=“ ”> </siglas>). En esta doble representación se ha elegido 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 ejemplo en el que se dice *voy pa casa* se representaría de la siguiente manera: voy <fsr t=“pa”>para</fsr> casa (Carcelén y Uclés, 2019, p. 27).

Tras este repaso a las características generales de este modelo de transcripción, desarrollamos a continuación las particularidades de este sistema.

(a) Etiquetas simples

<alargamiento/>

Esta etiqueta se utiliza para señalar que el hablante ha realizado algún alargamiento tanto vocálico como consonántico, con independencia de la duración que tenga. Se escribe pegado al carácter al que afecte, sea cual sea su posición dentro de la palabra. Sustituye a la convención del sistema Val.Es.Co. que consistía en duplicar la vocal o la consonante afectada en la transcripción (*eeh, mááss*)

Ejemplo 20

C: mi tío dijo que me desheredaba de<alargamiento/>l la comida

(BUE_001_03_19)

<ininteligible/>

Se utiliza para indicar que ese fragmento no se ha podido transcribir. A diferencia de la convención establecida para la transcripción dudosa, en ese caso, el transcriptor no puede reconstruir el enunciado emitido por el hablante ni siquiera de manera aproximada. Se escribe entre espacios. Sustituye a la convención del sistema Val.Es.Co. (( ))

Ejemplo 21

A: [mi<alargamiento/>ra] si fuera <ininteligible/>/ si él fuera inteligente/

(HAV_015_03_12)

<obs t=“ ”/>

La función de esta etiqueta es la de incluir comentarios u observaciones del transcriptor sobre el texto transcrito o sobre la grabación. Sustituye a los comentarios en nota a pie de página del sistema Val.Es.Co. y puede aparecer tanto en la línea del hablante como en la línea de observaciones, dependiendo de si su contenido afecta solamente a la intervención de un participante o si la información afecta al conjunto de la conversación. Se escribe entre espacios.

Ejemplo 22

A: ¿quién lo ha dibujado?/ ¿vos? <obs t="señala un papel"/> ¡me encanta!

(TUC_005_02_15)

<risas/>

Marca que alguno de los participantes se está riendo. A diferencia de la etiqueta doble <entre_risas> </entre_risas>, en este caso señala que se está riendo sin que haya más intercambio comunicativo. En el sistema Val.Es.Co. se señalaba por medio de la convención (RISAS)

Ejemplo 23

B: <risas/> mira cómo se ríe con<alargamiento/> tu mamá

(TCO_012_03_20)

<tos/>

Se emplea para indicar que alguno de los participantes tose. En el sistema Val.Es.Co. se reflejaba mediante la convención (TOS)

Ejemplo 24

C: [<ininteligible/> venga venga venga]

A: [<tos/>] (TCO_012_03_20)

<gritos/>

Indica que se producen gritos en la comunicación, sin que correspondan con ninguna enunciación. En el sistema Val.Es.Co. se señalaba por medio de la convención (GRITOS)

Ejemplo 25

B: cuando no es con tu celular [es otra] cosa <anónimo>Carlos</anónimo>

A: [<gritos/>]

(MEX_002_05_19)

Para otros ruidos que se consideren relevantes reflejar se utilizará la etiqueta simple de observación.

Ejemplo 26

Desconocido: <obs t="ruido aspiración nariz"/>

(LPA_002_04_18)

(b) Etiquetas dobles

<énfasis t=“ ”></énfasis>

Esta etiqueta delimita un fragmento que se ha pronunciado bien de manera marcada o bien silabeado. Esta diferenciación se incluye en el atributo de la siguiente forma: <énfasis t=“pronunciación_marcada”></énfasis> y <énfasis t=“silabeo”></énfasis>

Ejemplo 27

C: <énfasis t="pronunciación_marcada">¿hace años?</énfasis>

(BAQ_003_03_15)

Ejemplo 28

C: <énfasis t="silabeo">odio</énfasis> eso

B: ¿odio por qué?/ [no sabes lo que te pierdes]

(PTY_001_04_17)

La etiqueta de apertura se escribe antes del primer elemento del enunciado afectado, pegado a él, sin espacio, y la etiqueta de cierre irá después del último elemento del enunciado afectado, pegado a él con espacio posterior.

Sustituye a la convención del sistema Val.Es.Co. que reflejaba cualquier pronunciación enfática por medio de la transcripción en mayúsculas del segmento afectado o separada por sílabas para indicar silabeo.

<susurro></susurro>

Se utiliza para reflejar pronunciaciones en voz baja, cercanas al susurro. La etiqueta de apertura se escribe antes del primer elemento del enunciado afectado, pegado a él, y la etiqueta de cierre se sitúa inmediatamente después del último elemento afectado. Reemplaza la convención del sistema Val.Es.Co. °( )°

Ejemplo 29

A: [pero]/ fuimos incluso <susurro>est- esto a vos te encantaría</susurro>

(BUE_002_03_20)

<cita></cita>

Indica que el fragmento delimitado reproduce un segmento en estilo directo. La etiqueta de apertura se escribe antes del primer elemento del segmento afectado y la etiqueta de cierra después del último elemento, sin dejar espacio entre etiqueta y fragmento. En el sistema Val.Es.Co. el estilo directo se marcaba mediante el uso de la cursiva.

Ejemplo 30

B: no si eso le dije yo/ <cita>es por la niña <anónimo>Vivián</anónimo></cita>

(TCO_003_04_20)

<fsr t=“ ”></fsr>

Esta etiqueta 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 recoge cómo se ha pronunciado y entre la etiqueta de inicio y la de cierre la forma ortográfica normativa.

Ejemplo 31

A: no sé algún <fsr t="lao">lado</fsr>

(BAQ_001_03_16)

Ejemplo 32

A: <fsr t="enton">entonces</fsr> no

(MDE_002_02_12)

Ejemplo 33

A: <fsr t="pus">pues</fsr>

(MEX_001_03_19)

En el sistema Val.Es.Co. estos fenómenos (contracciones, pérdida de sonidos, etc.) se marcaban de diferentes formas, bien mediante el uso de paréntesis simples para la reconstrucción de sonidos omitidos, bien mediante el apóstrofo para reflejar contracciones.

<obs t=“ ”></obs>

De manera complementaria a la etiqueta simple de observación comentada anteriormente, la etiqueta doble se utiliza cuando la información que se quiere añadir afecta a un segmento de habla y no es una información puntual que ofrece el transcriptor, en cuyo caso se utilizaría la etiqueta simple. Señala que todo el fragmento delimitado por la etiqueta de inicio y de cierre se ve afectado por la información que se aporta en el atributo.

Ejemplo 34

D: [por eso mejor no] ((no haré))// <obs t="hace sonido con los labios">prrrrrr</obs>

(MEX_009_06_20)

<entre_risas></entre_risas>

A diferencia de la etiqueta simple <risas/>, esta etiqueta delimita fragmentos en los que el hablante se está riendo mientras habla, por tanto, incluye fragmentos de transcripción entre la etiqueta de inicio y de cierre. En el sistema Val.Es.Co. este hecho se señalaba mediante nota al pie de página.

Ejemplo 35

B: [<risas/>]/// (1.3) <entre_risas>ya</entre_risas>

(SCL_006_05_18)

<extranjero t=””></extranjero>

Esta convención se utiliza para señalar palabras o intervenciones en otro idioma; en este caso, dentro del atributo se incluye la pronunciación y entre las etiquetas de apertura y cierre, la transcripción ortográfica normativa. Esta etiqueta no aplica en el caso de nombres propios extranjeros. Esta decisión viene dada por el interés de marcar el cambio de código en la conversación, y los nombres propios no responden a este patrón.

Ejemplo 36

B: una música e<alargamiento/>h <extranjero t="contri">country</extranjero>

(SCL_002_03_18)

Ejemplo 37

A: [tu título]/ <risas/> tipo Margaret Atwoo<alargamiento/>d ahí

(BUE_002_03_20)

<anónimo></anónimo>

Siguiendo las indicaciones establecidas para la anonimización de identificadores directos e indirectos, en aquellos momentos en los que se ha visto necesaria la ocultación de estos datos, la sustitución del original se ha marcado con esta etiqueta y se ha sustituido por un nombre ficticio.

Ejemplo 38

A: [((salta))] el coronavirus <anónimo>Rebeca</anónimo>

(MEX_006_06_20)

(c) Signos de transcripción Val.Es.Co. que se han mantenido en ELAN

[ ] Solapamientos **- Reinicios y autointerrupciones** **((transcripción))**

Se han mantenido, no obstante, algunos signos de transcripción del sistema Val.Es.Co. que no entran en conflicto con el lenguaje XML y que, además, facilitan la comprensión a

usuarios no familiarizados con el entorno ELAN. Estos signos se utilizan de la misma manera descrita más arriba. Hablamos de las marcas para los solapamientos, los reinicios y autointerrupciones y los fragmentos de transcripción dudosa.

Así mismo, este modo de trabajo coincide con el anterior en que no se utilizan signos de puntuación, a excepción de las interrogaciones y las exclamaciones, ni mayúsculas por razones de puntuación excepto para nombres propios y siglas. Las cifras y símbolos se transcriben con letra.

(d) Convenciones del sistema Val.Es.Co. que han desaparecido en ELAN

Otras convenciones del sistema Val.Es.Co. han desaparecido ya que con la alineación temporal audio-transcripción y la visualización que permite el programa ELAN, son prescindibles. Es el caso de las pausas, señaladas en la transcripción por medio del símbolo / y los silencios, ya que pueden visualizarse en el oscilograma que aparece en la interfaz del programa, así como los tonemas, que se marcaban por medio de flechas que indicaban si el tonema era ascendente ↑, descendente ↓ o suspendido →, y que si bien, no se obtienen de primera mano con el uso de ELAN, pueden generarse utilizando en combinación ELAN y el programa para el análisis del habla Praat.

No obstante, algunas de estas convenciones se recuperan tras el procesamiento informático y se incluyen en la transcripción que se encuentra disponible en línea en formato texto plano con el objetivo de favorecer la legibilidad del texto.

5.2.2.2. Revisión y validación

Las conversaciones, debidamente codificadas y alineadas, junto con los documentos asociados (ficha técnica y autorización) pasan en este punto al equipo central. Es este el encargado de realizar la revisión y validación final antes de incorporar los archivos al motor de búsqueda disponible en línea. Este último paso se organiza en dos fases; la fase primera consiste en la revisión manual del conjunto de archivos entregados. Se revisa que los archivos sean los correctos y que estén debidamente ejecutados; junto a esta se hará también la revisión de la división en grupos entonativos, el alineado audio-texto, y se comprueba que todos los identificadores directos e indirectos se hayan anonimizado. La segunda fase tiene

como objetivo comprobar que las etiquetas XML que se emplean en la codificación del corpus se hayan compuesto correctamente, es decir, que no haya errores de escritura o falte algún elemento de la etiqueta. Estos descuidos pueden provocar problemas a la hora de su procesamiento para la incorporación al motor de búsqueda.

Este proceso se realiza de manera automática mediante el uso del *software* Oxygen XML, un editor de XML de acceso libre que permite crear, editar y validar documentos XML, así como también otros lenguajes relacionados. Es una herramienta completa que incluye opciones como el resaltado de la sintaxis del documento, el autocompletado de código, la validación del esquema, la transformación y la publicación de documentos, y una variedad de herramientas para la gestión y organización de proyectos XML.

Para ello, se ha creado un documento base DTD (*document type definition*) que reconoce la estructura correcta del etiquetado XML y que detecta cualquier error de composición y sintaxis, permitiendo el reemplazo automático por la etiqueta correcta. Es decir, la DTD define la estructura y las reglas de validación de un documento XML y describe los elementos y atributos que se permiten, especificando las relaciones jerárquicas entre ellos. Además, establece las reglas para la validación del documento, como la obligatoriedad de ciertos elementos o atributos, el tipo de datos que se pueden utilizar para los atributos y la secuencia adecuada de los elementos. Por ejemplo, hemos señalado anteriormente que la etiqueta <pronunciación_marcada t=“ ”> </pronunciación_marcada> solamente permitía añadir en el atributo las opciones “silabeo” o “énfasis", por tanto, en el caso de que durante la transcripción se haya añadido otro valor o en las opciones permitidas haya habido un baile de letras, el programa detectará el error. En la Figura 26 puede verse un ejemplo, en este caso, el atributo de la etiqueta <énfasis t=“ ”><énfasis> exige la presencia de una barra baja (_) entre *pronunciación* y *marcada* que no aparece, por tanto, Oxygen lo señala como un error de sintaxis.

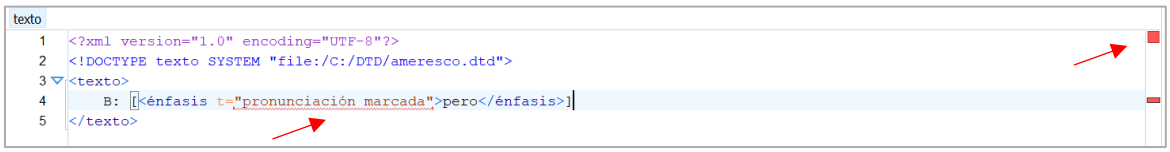


Figura 26. Detección de error de codificación en Oxygen XML

En el caso del corpus Ameresco, la DTD para la validación es la siguiente:

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT ameresco (texto)>
   <!ELEMENT texto ANY>
   <!-- etiquetas -->
   <!ELEMENT alargamiento EMPTY>
   <!ELEMENT ininteligible EMPTY>
   <!ELEMENT risas  EMPTY>
   <!ELEMENT tos  EMPTY>
   <!ELEMENT gritos  EMPTY>
   <!ELEMENT sic (#PCDATA | alargamiento | ininteligible | risas | tos | gritos | entre_risas |
susurro | anónimo | cita | extranjero | siglas | énfasis | fsr | obs | obse)*>
   <!ELEMENT entre_risas (#PCDATA | alargamiento | ininteligible | risas | tos | gritos | sic |
susurro | anónimo | cita | extranjero | siglas | énfasis | fsr | obs | obse)*>
   <!ELEMENT susurro (#PCDATA | alargamiento | ininteligible | risas | tos | gritos |
entre_risas | sic | anónimo | cita | extranjero | siglas | énfasis | fsr | obs | obse)*>
   <!ELEMENT anónimo (#PCDATA | alargamiento | ininteligible | risas | tos | gritos |
entre_risas | susurro | sic | cita | extranjero | siglas | énfasis | fsr | obs | obse)*>
   <!ELEMENT cita (#PCDATA | alargamiento | ininteligible | risas | tos | gritos | entre_risas
| susurro | anónimo | sic | extranjero | siglas | énfasis | fsr | obs | obse)*>
   <!ELEMENT extranjero (#PCDATA | alargamiento | ininteligible | risas | tos | gritos |
entre_risas | susurro | anónimo | cita | sic | siglas | énfasis | fsr | obs | obse)*>
      <!ATTLIST extranjero t CDATA #REQUIRED>
   <!ELEMENT siglas (#PCDATA | alargamiento | ininteligible | risas | tos | gritos |
entre_risas | susurro | anónimo | cita | extranjero | sic | énfasis | fsr | obs | obse)*>
      <!ATTLIST siglas t CDATA #REQUIRED>
   <!ELEMENT énfasis (#PCDATA | alargamiento | ininteligible | risas | tos | gritos |
entre_risas | susurro | anónimo | cita | extranjero | siglas | sic | fsr | obs | obse)*>
      <!ATTLIST énfasis t (silabeo|pronunciación_marcada) #REQUIRED>
   <!ELEMENT fsr (#PCDATA | alargamiento | ininteligible | risas | tos | gritos | entre_risas |
susurro | anónimo | cita | extranjero | siglas | énfasis | sic | obs | obse)*>
      <!ATTLIST  fsr t CDATA #REQUIRED>
   <!ELEMENT obs (#PCDATA | alargamiento | ininteligible | risas | tos | gritos | entre_risas |
susurro | anónimo | cita | extranjero | siglas | énfasis | fsr | sic | obse)*>
      <!ATTLIST obs t CDATA #REQUIRED>
   <!ELEMENT obse EMPTY>
      <!ATTLIST obse t CDATA #REQUIRED>
```

Figura 27. DTD del corpus Ameresco

Una vez revisadas y validadas, las conversaciones están preparadas para ser incorporadas al motor de búsqueda.

5.2.2.3. Anonimización

Dada la necesidad y obligatoriedad de mantener la privacidad de los participantes en la conversación, cumpliendo con los preceptos legales señalados en la sección 4.2.1.2. para el corpus Ameresco, se ha desarrollado un proceso de anonimización en dos capas por las que se eliminan los identificadores directos. Estos se sustituyen por nombres ficticios y se valora la eliminación de los identificadores indirectos si se estima necesario según el contexto.

La primera capa corresponde al nivel textual, esto es, a las transcripciones. En este caso se ha optado por el reemplazo del identificador directo por otro ficticio[47]. En la medida de lo posible se ha intentado que la sustitución respete el patrón fónico del término original. Otro factor que se debe tener en cuenta es el relativo a cuestiones sociales y culturales, es decir, el término sustituto debe respetar y adecuarse al contexto social, dialectal o histórico del término sustituido. Este reemplazo no ha estado exento de problemas como se detalla en la sección 5.3.2.

Los identificadores directos que han sido anonimizados se marcan con la etiqueta XML <anónimo></anónimo>. Para los identificadores indirectos que, tras su valoración, se ha decidido anonimizar, se ha optado por el empleo de esta etiqueta, aunque en otros momentos ha sido necesaria la eliminación del fragmento con la indicación en línea de observaciones de que se ha producido un corte en la grabación. Por ejemplo, en grabaciones en las que un hablante narraba un suceso extraordinario que pudiera llevar a su identificación.

En esta capa de anonimización, el proceso es completamente manual, es decir, la persona encargada de realizar el borrado de audio y texto de estos identificadores debe señalarlos uno por uno en el archivo para después proceder al silenciado.

La segunda capa afecta al nivel oral (audio). En esta fase, se ha aplicado un sistema semiautomático de anonimización del audio. Para borrar los segmentos de audio que debían ser anonimizados se han utilizado los programas ELAN y Audacity. En el caso del corpus

[47] Para la realización de esta tarea, ha sido útil el empleo de páginas web como https://www.nombres.top/ y https://www.apellidos.top/, que ofrecen sustituciones por ítems como nombres que empiezan por la misma letra, que terminan igual o que son parecidos.

Ameresco se ha optado por silenciar el audio, mecanismo menos molesto que la aparición de un indicador sonoro como un pitido.

El programa ELAN permite acotar los fragmentos que posteriormente se van a silenciar. En este caso, se crea una línea o *tier* llamada *anónimo* donde se marcarán los fragmentos de tiempo susceptibles de ser anonimizados, como refleja la Figura 28. Una vez marcados en todo el audio, este programa permite exportar únicamente la línea como texto tabulado, seleccionando las opciones *excluir los nombres de las líneas del output y excluir los nombres de los participantes en el output*.

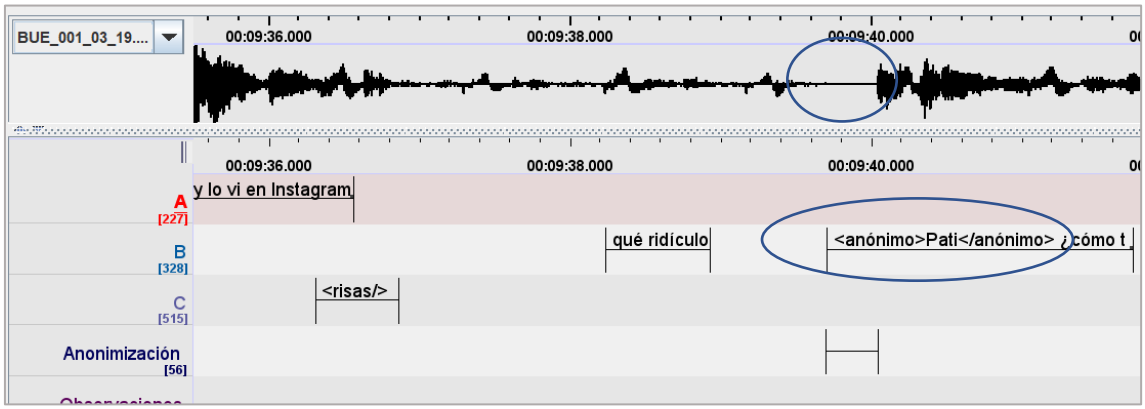


Figura 28. Captura de pantalla de ELAN en fase de anonimización

Una vez realizada la exportación, el archivo resultante se cargará en el programa de edición de audio Audacity. Al realizar la importación se seleccionará la opción *etiquetas*. Con esta configuración, el programa reconoce los segmentos marcados en la fase anterior como un audio etiquetado y puede silenciar en una única operación todos los segmentos seleccionados. Bastará con vincular este nuevo audio anonimizado de nuevo al programa ELAN para poder trabajar con texto y audio alineados y completamente anonimizados (en la capa 1 de texto y en la capa 2 de audio).

En la imagen anterior podemos ver el borrado del segmento en el oscilograma de sonido (capa de audio) y la sustitución que se ha hecho del nombre original por un nombre ficticio, en este caso Pati, marcado con la etiqueta correspondiente (capa de texto). Una vez finalizada la anonimización del audio, la grabación original se elimina definitivamente.

Con este proceso se ha intentado garantizar al máximo la no reidentificación de los participantes. No obstante, es muy difícil lograr un compromiso entre la comodidad del proceso de anonimización semiautomático y la necesidad de una vigilancia externa ya que,

como se mencionaba más arriba, un mismo nombre, en distintos contextos, puede ser susceptible de proporcionar datos personales. Por tanto, la mano humana sigue siendo imprescindible.

5.2.2.4. Identificación de archivos

Para una correcta identificación de los archivos se ha establecido un sistema alfanumérico que adjudica a cada conversación un código particular. Este código tiene la particularidad de que ofrece información sobre los metadatos de la conversación, es decir, sin necesidad de abrir los archivos, con esta ID conocemos la ciudad en la que se ha realizado la grabación, el número de hablantes que participa y el año de la grabación. Así, a cada archivo se le asigna un código ID formado por cuatro bloques, separados entre sí por barra baja (_) como se detalla a continuación.

En primer lugar, la identificación del archivo recibe la abreviatura de la ciudad en la que se ha recogido la grabación. Primeramente, se optó por adoptar la nomenclatura IATA (*International Air Transport Association*) que consta de tres letras; sin embargo, al percibir que el código podía ser opaco para el usuario del corpus, es decir, no se reconocía fácilmente la ciudad a la que hacía referencia, se decidió elegir una abreviatura que permitiera identificar el origen del archivo de una manera clara. Por ejemplo, en el caso de las conversaciones procedentes de Buenos Aires, si se toma como referencia el código IATA, obtendríamos el identificador AEP. Esta combinación de caracteres no permite reconocer la ciudad, así que se optó por adoptar el código identificador BUE. En el caso de las conversaciones obtenidas en Monterrey, su código IATA coincide con la abreviatura reconocible del nombre de la ciudad, MTY, por tanto, sí que se ha mantenido. El listado completo de códigos puede verse en la Tabla 12.

En segundo lugar, aparece una secuencia numérica que corresponde al código de grabación interno y que se adjudica por orden correlativo según se van incorporando conversaciones al corpus. La primera grabación obtenida llevará el código 001, la siguiente el código 002, y así sucesivamente.

El tercer bloque ofrece información sobre el número de hablantes activos que aparecen en la conversación, como mínimo dos (por tanto 02) sin que haya un máximo. Si bien, como

hemos visto en el 3.2.1.3.2., no se recomienda que haya más de 4-5 participantes debido a las dificultades de transcripción que implicaría.

Por último, encontramos dos dígitos que se corresponden con el año de grabación. Si la grabación se obtuvo en 2016, se selecciona 16; si fue obtenida en 2021, se tomaría 21.

El resultado final puede verse en los ejemplos siguientes:

BUE_001_03_19: indica que estamos la primera conversación de Buenos Aires en la que participan tres hablantes y que fue grabada en el año 2019.

MTY_022_03_15: significa que este es el vigesimosegundo archivo de Monterrey, en él aparecen cuatro hablantes y se grabó en 2015.

Ciudad	**Código Ciudad**
Buenos Aires	BUE
Tucumán	TUC
Iquique	IQQ
Santiago de Chile	SCL
Temuco	TCO
Barranquilla	BAQ
Medellín	MED
La Habana	HAV
Loja	LOJ
Las Palmas de Gran Canaria	LPA
Tegucigalpa	TGU
Ciudad de México	MEX
Monterrey	MTY
Querétaro	QRO
Ciudad de Panamá	PTY

Tabla 12. Códigos de las ciudades del corpus Ameresco

5.2.3. Fase 3. Archivo, distribución y acceso al corpus por parte de los usuarios

En esta sección, describiremos el proceso por el cual los materiales del corpus Ameresco se ponen a disposición de la comunidad científica y las posibilidades de descarga para el análisis que pueden realizarse sobre estos materiales. De esta manera, dado que los datos del

corpus se constituyen en contenido de libre acceso, se ha optado por incluirlos en dos sistemas de comunicación informática distinta. El primero, que se describe en la sección 5.2.3.2., explicita la página web base del proyecto; mientras que el segundo, descrito en la sección 5.2.3.3., se articula como una herramienta de transformación y consulta del corpus mediante un entorno de libre acceso y distribución denominado Orastats Aroca (Cabedo y Carcelén, 2022).

5.2.3.1. Aspectos generales

El uso de páginas electrónicas como lugares de centralización de la información y de comunicación con el exterior es una práctica recurrente en todos los proyectos de investigación en los que, además, se pretende ofrecer algún tipo de herramienta o resultado de investigación específica. Son, por tanto, muchos los proyectos de investigación que recurren a la creación de una página web como se ha visto en el Capítulo 2.

En la línea de lo que hemos apuntado en la sección 4.2.3. sobre el almacenamiento y el mantenimiento de los datos, contamos con la ventaja de que en los últimos años se han desarrollado sistemas gratuitos de libre acceso y fácil administración que permiten a los grupos de investigación e investigadores particulares, sin necesidad de poseer especiales conocimientos informáticos, crear, mantener, administrar y ampliar recursos en la web. Los sistemas con un uso más general dentro de los conocidos como CMS (*Content Management System*) son Wordpress, Drupal o Joomla.

Estos sistemas comparten unas mismas características: tienen un panel de administración interno visible para la gestión y diseño de la interfaz, habitualmente a través de plantillas[48] ya generadas y que solo deben descargarse y aplicarse al entorno general de la página, si bien pueden ser adaptadas a las necesidades o gustos de quien las gestiona.

Por otro lado, más allá del diseño, los sistemas CMS permiten crear páginas de consulta que aparecen automáticamente en los menús de la página una vez han sido creados. Así, es relativamente sencillo, como veremos en la sección 5.2.3.2.1. en la que se ubica el *sitemap* (mapa del sitio) de la web de Ameresco, crear diferentes apartados dentro de la página principal en las que incluir información sobre el grupo, sobre el proyecto, etc.

[48] En este sentido, las plantillas más utilizadas proceden de Bootstrap (https://getbootstrap.com/).

Así pues, en este capítulo, nos centraremos en cómo se diseñó y construyó la parte de divulgación web tanto de la información presente en el proyecto (constitución, motivación, equipo de investigadores) como del material recogido en el corpus (audios, archivos de transcripción, etc.). En este apartado, destacaremos cómo el uso de una estructura de base de datos relacional (con MYSQL, PostgreSQL y RSQLite) nos ha permitido adquirir, almacenar y organizar grandes cantidades de datos de manera eficiente para, a su vez, ser puestas a disposición pública. En las siguientes secciones, examinaremos la estructura informática de consulta del corpus y también explicaremos cómo esta arquitectura de base de datos relacional permite una flexibilidad casi ilimitada en la anotación de los textos, al mismo tiempo que ofrece un rendimiento sólido incluso para corpus más extensos. Posteriormente, exploraremos cómo la estructura del corpus y su interfaz se combinan para brindar a los usuarios acceso a diversos tipos de consultas que son difíciles de realizar con otras arquitecturas informáticas; por ejemplo, las páginas web de acceso estático[49] no facilitan el acceso inmediato a la información, aunque permitan descargar los datos de modo directo (Du Bois *et al*., 2000).

Cabe subrayar que toda la metodología usada y diseñada específicamente para la transformación y consulta del corpus, como Oralstats Aroca (§ 5.2.3.3.), es gratuita y de libre acceso para todas aquellas personas que la quieran utilizar, bien para exponer sus propios datos de modo público, bien para usar el entorno de consulta únicamente en un entorno local, es decir, en su propio ordenador y sin opción de exposición en línea.

5.2.3.2. Web de consulta

Antes de describir el estado actual de la web del corpus Ameresco y del entorno Oralstats Aroca, cabe recordar que el corpus Ameresco, que empieza a gestarse en 2010, adquiere su primera página web en el marco del proyecto Es.Var.Atenuación, en 2014, cuando se registra el dominio http://esvaratenuacion.es por primera vez a través de la página de gestión de dominios y alojamientos web Dinahosting S.L. La web inicial se creó en primera instancia con el sistema Wordpress; con la entrada en vigor del proyecto Es.VaG.Atenuación en 2016, la web se adaptó al actual sistema Drupal, que comentaremos en los próximos apartados.

[49] Una página estática es aquella que no varía con frecuencia y que mantiene el mismo contenido hasta que se actualiza; mientras que una página dinámica es aquella generada desde un servidor en el momento en que un cliente hace una petición desde su navegador, es decir, se puede obtener un contenido actualizado cada vez que un usuario realiza una petición (Camazón, 2010, p. 19).

En la Figura 29 puede verse la interfaz de la página de 2014, que únicamente servía como repositorio[50] del material del corpus (archivos de audio, fichas técnicas y transcripción) y que no dispuso de buscador del corpus hasta 2016.

Figura 29. Interfaz de la página Es.VaG.Atenuación

La web del corpus actual puede consultarse a través de las URL http://esvaratenuacion.es y http://www.corpusameresco.com, página marco que sirve como entorno de acceso a los datos que se suben y actualizan en la primera. La web principal, por su parte, tiene el nombre de *esvaratenuacion* porque corresponde al primer proyecto vigente cuando se gestó y recogió el primer material del corpus (Proyecto Es.Var.Atenuación, 2013-2016); dos proyectos posteriores (Es.Vag.Atenuación y Esprint) han mantenido y financiado la infraestructura web. El diseño y entorno general puede observarse inicialmente en la Figura 30, que muestra el primer tercio de la página inicial de la web:

[50] No obstante, este repositorio se ha mantenido en la versión actual de la página y permite la descarga directa y completa de audios, transcripciones y fichas técnicas.

Figura 30. Interfaz actual de la página de inicio del corpus Ameresco

Como puede observarse, en la página inicial se dispone la información general de los distintos proyectos que han permitido la creación y el mantenimiento del entorno web. En la parte inferior, se hacen referencias a cómo buscar en los datos y qué opciones están disponibles, así como una mención al total de conversaciones, ciudades, hablantes y palabras que constituyen el corpus en el momento actual.

En la Figura 31 se muestra el resto de esa página principal de entrada al corpus. En ella, se ha considerado interesante presentar la cantidad de datos que integran el corpus; de este modo, la frecuencia de palabras se distribuye a lo largo de un mapa y, también, de dos gráficos circulares en los que se amplía la distribución de las palabras tanto por país (al igual que en el mapa), como por ciudad. Cuando el usuario desplaza el cursor por encima de las zonas verdes del mapa o por encima de los triángulos de los gráficos circulares, una notificación emergente o *popup* se despliega con el número preciso de palabras.

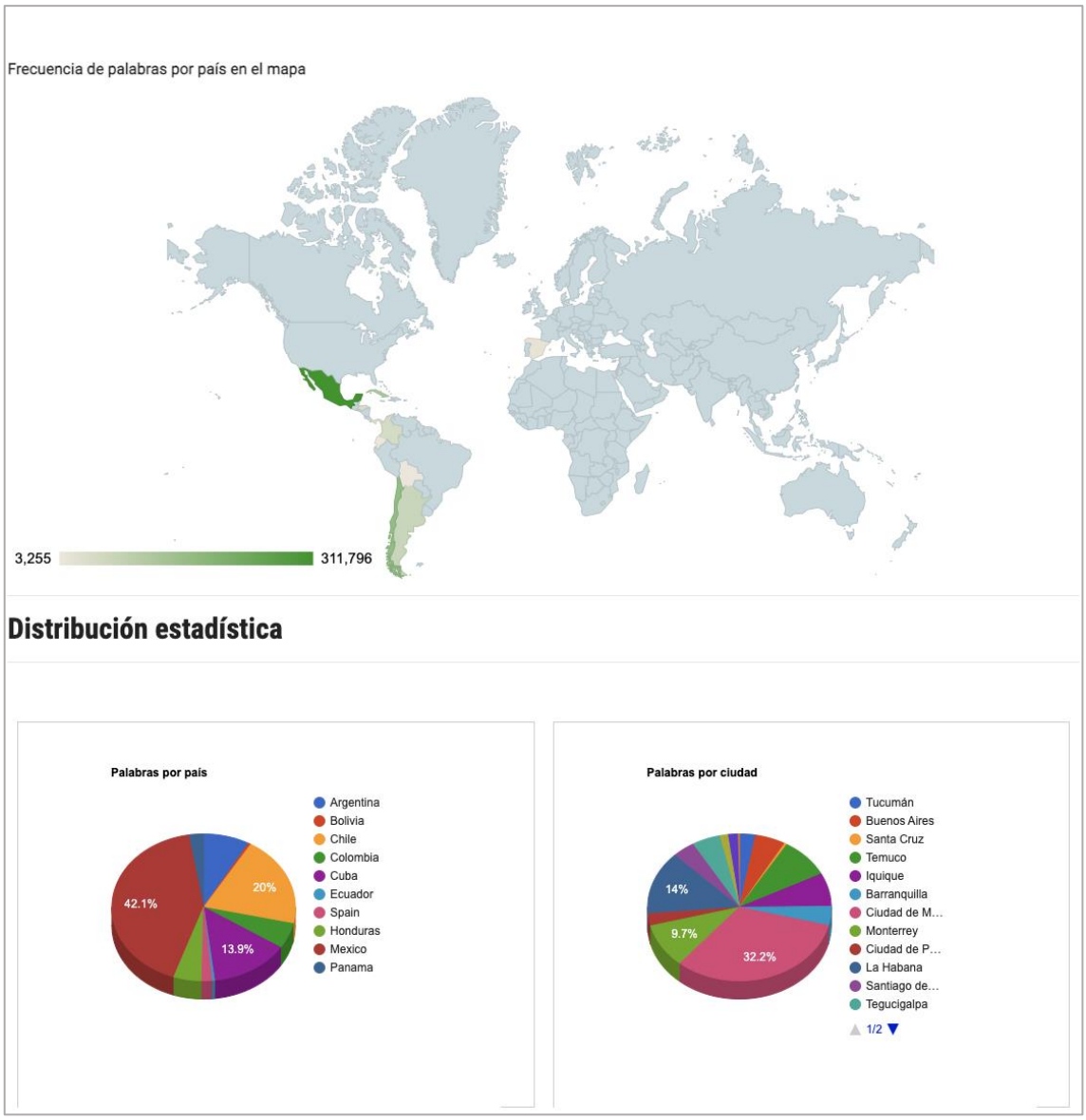


Figura 31. Interfaz actual de la página de inicio del corpus Ameresco

La frecuencia de palabras que se muestra en estos gráficos es estática y se actualiza manualmente por el administrador de la página, no directamente desde la base de datos. Así, el *script* concreto para poder proyectar el mapa, desarrollado con código Javascript a partir de las indicaciones que aparecen en la web de gráficos de Google[51], se ubica en el HTML de la página inicial y es como se muestra en la Figura 32:

[51] https://developers.google.com/chart/interactive/docs/gallery/map?hl=es-419

```
<<p><strong>Frecuencia de palabras por país en el mapa</strong></p>
<script type="text/javascript"
src="https://www.gstatic.com/charts/loader.js"></script><script
type="text/javascript">
      google.charts.load('current', {
        'packages':['geochart','corechart'],
      });
      google.charts.setOnLoadCallback(drawRegionsMap);
google.charts.setOnLoadCallback(drawcountryChart);
google.charts.setOnLoadCallback(drawcityChart);

      function drawRegionsMap() {
        var data = google.visualization.arrayToDataTable([
          ['País', 'Palabras'],
['Argentina', 60450],
['Bolivia', 3255],
['Chile', 147759],
['Colombia',  41825],
['Cuba',  102890],
['Ecuador', 3587],
['Spain', 12528],
['Honduras',  36722],
['Mexico',  311796],
['Panama',  19538]
        ]);

        var options = {datalessRegionColor: '#C5D9DE'};

        var chart = new
google.visualization.GeoChart(document.getElementById('regions_div'));

        chart.draw(data, options);
      }

    </script>

<div id="regions_div" style="width: 100%; height: 500px;"> </div>

<hr />
<p>
<style type="text/css"><!--td {border: 1px solid #ccc;}br {mso-data-
placement:same-cell;}-->
</style>
</p>
```

Figura 32. *Script* de la página de inicio del corpus Ameresco

Por su parte, los gráficos circulares de distribución de palabras por país y ciudad se han generado también manualmente a partir de una hoja de cálculo de Google. Teniendo en cuenta lo comentado tanto para el mapa como para los gráficos circulares, hay que recalcar que son gráficos no vinculados con la base de datos original, por lo que, cuando esta base de datos se amplía o modifica, el gestor del corpus debe cerciorarse de que se han cambiado también los números finales en la página inicial del corpus. Este hecho, no obstante, no ocurre con las funcionalidades de otros recursos disponibles, como veremos en la exposición de Oralstats Aroca (§ 5.2.3.3.).

Al mismo tiempo, en todas las páginas de la web se incluye un pie de página con datos concretos sobre el proyecto vigente que financia parcialmente el corpus y sobre la decisión de los autores y gestores acerca de la legalidad y eticidad en la consulta de los datos.

Figura 33. Datos sobre la financiación del corpus Ameresco

Así pues, la web del proyecto Ameresco explicita que para todo el material disponible (páginas, corpus, archivos, métodos…) se aplica una licencia *Creative Commons Reconocimiento No Comercial Compartir Igual 4.0 Internacional License*, que básicamente acepta el uso libre por parte de usuarios externos, siempre que haya una mención o citación a los autores o editores del entorno web o, en su caso, del corpus.

En un futuro, siempre que la capacidad de financiación lo permita, la web de corpusameresco.com se convertirá en un dominio con un entorno de *hosting* propio y no en un dominio que, como sucede actualmente, sirve de marco para apuntar al dominio esvaratenuacion.es con su propio entorno de alojamiento local.

En todo caso, el dominio esvaratenuacion.es, aunque por nombre se refiere a un proyecto ya finalizado, se mantendrá, ya que las referencias en línea que los investigadores hayan podido incluir en documentos de investigación previos deben tener una referencia localizable para mejorar su propia legibilidad.

5.2.3.2.1. Sitemap

Un *sitemap*, también conocido como mapa del sitio, es un archivo XML que recoge la jerarquía de una página web, es decir, señala las páginas de consulta que un usuario puede ir abriendo mientras navega por la interfaz de la página. La principal funcionalidad es la de mejorar la localización de los entornos y rastreadores de búsqueda, con lo que las posibilidades de indexación de la web son mayores. Dicho de otro modo, la estructura de la web facilita a los motores de búsqueda que encuentren la página y puedan apuntar a las páginas más frecuentes. Al mismo tiempo, aporta una visión general de la disposición o estructura de la web y, así, es más fácil para un gestor o administrador plantearse si deben

incluirse nuevas páginas o si, por el contrario, algunas de las páginas deben ser renombradas, reubicadas o, también, eliminadas.

La estructura actual de la página principal del corpus Ameresco es la siguiente:

1. **El proyecto**
2. **Corpus**
   - Consulta
   - Oralstats-aroca
     - Acceso
     - Sobre etiquetado
   - Cómo citar
   - Archivos
3. **Estadísticas**
4. **Equipo Ameresco**
5. **Materiales**
   - Documentos de trabajo
   - Enlaces
6. **Histórico**
   - Es.Vag.Atenuación
     - Equipo Es.Vag.Atenuación
     - CIAL 2016
     - Comunicaciones
     - Paneles
7. **Bibliografía**
   - Artículos de miembros del equipo
   - Prosodia
   - ELE
   - Atenuación
   - Evidencialidad
   - Intensificación

Figura 34. Mapa del sitio web del corpus Ameresco

En este índice general hay varios aspectos que deben comentarse. El primer enlace, denominado *El proyecto*, dirige a la propia página de entrada, es decir, la primera página que se proyecta cuando se entra en la web por primera vez. En ella, como se ha dicho previamente, se indican cuestiones generales sobre la formación del proyecto, la fecha de inicio y, también, los datos generales del corpus, esto es la frecuencia de palabras proyectadas en un mapa y en dos gráficos circulares.

El segundo enlace, *Corpus*, es el más importante de la web, ya que dirige a la consulta de las conversaciones del corpus Ameresco. El primer subenlace que puede marcarse es

Consulta, y desde ahí se puede acceder a las intervenciones del corpus según las convenciones que se desarrollan en la sección 5.2.3.2.5.1. de este capítulo. Esta primera consulta se realiza directamente por medio de recursos disponibles en el propio entorno de Drupal. Se trata de un acercamiento sencillo de búsqueda y no dispone de las funcionalidades que sí desarrolla Oralstats Aroca, que se articula precisamente como el siguiente subenlace en el menú general, con el epígrafe de Oralstats-aroca. Por último, hay un breve texto bajo el enlace *Sobre etiquetado,* en el que se mencionan las particularidades del etiquetado automático morfosintáctico del corpus, realizado con la librería UDPipe (Wijffels, 2023), aunque también ha sido anotado originariamente con otras herramientas automáticas como Freeling o Treetagger.

Tras los enlaces a las posibilidades de consulta, tanto mediante Drupal como mediante Oralstats Aroca, se incluye el epígrafe *Cómo citar*, ya que es importante mencionar que se atribuye la autoría a los editores/coordinadores de los corpus de cada ciudad y, al mismo tiempo, a las editoras del corpus general. Por ejemplo, en el caso de la Ciudad de México, la manera de citar el corpus sería como sigue:

> Maldonado Soto, Ricardo (en línea): "Corpus de conversaciones Ameresco-Ciudad de México", en Albelda y Estellés (coords.): Corpus Ameresco, www.corpusameresco.com, Universitat de València, ISSN: 2659-8337.

El último enlace de menú anidado en la sección Corpus es *Archivos*[52]; como ampliaremos en la sección 5.2.3.2.5.2., en esta página se ubican todos los archivos relacionados con los datos del corpus (fichas técnicas, audios, archivos de ELAN, Textgrids y transcripción en texto plano TXT).

El siguiente enlace del menú, denominado *Estadísticas*, recoge las estadísticas generales del corpus, según su agrupación por ciudad, sexo, nivel, edad y usuario en cualquiera de sus combinaciones. Esta información la ampliaremos en la sección 5.2.3.2.5.3.

En el enlace *Equipo Ameresco* se incluyen los nombres más importantes relacionados con el proyecto y el corpus, como el director académico, Antonio Briz; las coordinadoras académicas, Marta Albelda y Maria Estellés; los investigadores que gestionan las grabaciones y las transcripciones, como Andrea Carcelén, Gloria Uclés o Lissette

[52] Corresponde a lo que hemos venido denominando *Repositorio*.

Mondaca;[53] las personas que coordinan los diversos corpus de cada área geográfica, como Ricardo Maldonado, Marta Samper, Silvana Guerrero, María Eugenia Flores, entre otros; y, finalmente, los muchos de los colaboradores en la recogida de los corpus particulares de cada ciudad.

En el caso del corpus Ameresco, la colaboración de los equipos es gratuita y desinteresada, como hemos comentado a lo largo de este trabajo; por tanto, es fundamental su reconocimiento a través de la página: cada grupo colaborador aparece aquí mencionado y se ofrece, además, información como su universidad de origen, su cargo y su perfil académico. Por ejemplo, en la web aparecen citados, con fotografía y filiación, los coordinadores de los distintos equipos de recolección de corpus, así como los grupos de trabajo propios de cada universidad, como se ha señalado en la sección 5.1.3. y como muestra la Figura 35 en la que se puede observar a parte de los coordinadores y coordinadoras de los distintos subcorpus.

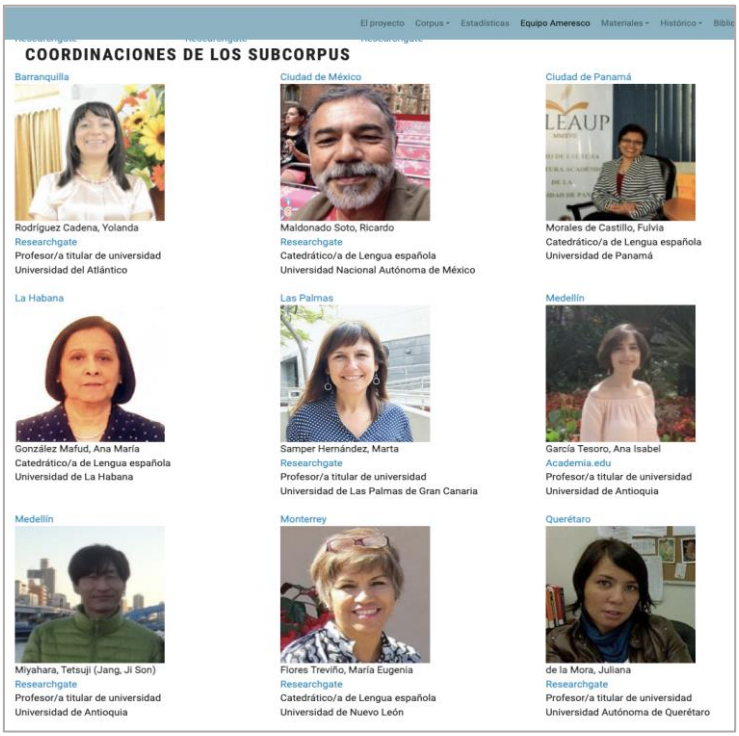


Figura 35. Extracto de la sección *Equipos* en la página del corpus Ameresco

[53] En etapas anteriores han participado en la gestión de grabaciones y transcripciones del corpus Ameresco, así como en el mantenimiento de la web, Jorge Martí, Elena López-Navarro, M.ª Amparo Soler, Dorota Kotwica y Amparo García Ramón.

La siguiente sección, con el nombre de *Materiales,* incluye el enlace *Documentos de trabajo*, que presenta los documentos que, desde el equipo central, se facilitan a los investigadores de los distintos equipos previamente a la recogida de material y que cualquier persona interesada en conocer de primera mano el protocolo de trabajo aplicado al corpus puede consultar. A continuación, incluimos una referencia a esos documentos y a algunas referencias bibliográficas que se consideran más significativas (la presentación y el orden es el mismo que aparece en la web). Desde el equipo responsable de Ameresco se ha contemplado en todo momento la accesibilidad en abierto a la metodología de trabajo a toda la comunidad científica, explicitando de la manera más exhaustiva posible las decisiones adoptadas en cada fase de la recogida del corpus.

Figura 36. Sección *Materiales* de la página del corpus Ameresco

El apartado que sigue es, precisamente, el nombrado como *Enlaces*. Se trata de enlaces que los responsables del proyecto creen que pueden ser interesantes para todas aquellas personas que se acerquen a la web y quieran acceder a planteamientos cercanos relacionados con el ámbito de la pragmática, así como grupos de investigación afines. También se incluye una referencia a otros corpus textuales, orales y escrito, del español. Los enlaces que aparecen actualmente en la web son los siguientes:

Figura 37. Sección *Materiales* de la página del corpus Ameresco

En el enlace de *Histórico* se recogen actividades anteriores que, vinculadas con los proyectos de investigación originales y el propio corpus Ameresco, se realizaron en un momento pasado. En ese sentido, por ejemplo, cabe destacar la referencia al Congreso Internacional de Atenuación Lingüística (CIAL), que se realizó en Valencia del 15 al 18 de junio de 2016.

Finalmente, en el último enlace del menú se encuentran referencias bibliográficas que se consideran importantes para los distintos proyectos vinculados a Ameresco. En estas referencias se incluyen tanto publicaciones de los propios miembros del proyecto, como referencias bibliográficas de autores ajenos que han estado directamente relacionadas con los planteamientos asociados a este grupo de investigación.

El *sitemap* que acabamos de exponer sobre el corpus Ameresco, en definitiva, despliega su anatomía como una página web; este menú permite una navegación más eficiente para los usuarios y mejora la visibilidad para los motores de búsqueda. Al explorar la estructura de la página principal del corpus Ameresco, se aprecia la claridad y organización en la presentación de información, lo que facilita a los visitantes encontrar los recursos necesarios. También permite a los administradores de los recursos de la web actualizar la información

constantemente en las páginas ya creadas e, incluso, habilitar nuevas páginas en cuanto a futuras necesidades. Por ejemplo, sería posible crear un conjunto de páginas específicas anidadas para dar cuenta del desarrollo de seminarios o congresos futuros.

5.2.3.2.2. Tecnología

Como hemos adelantado, la página web del corpus se crea en base a Drupal. Algunos autores de manuales de uso de Drupal recalcan la facilidad de creación web por medio de este sistema y, para el caso de los proyectos de investigación, en los que suele necesitarse la contratación de personal externo, señalan que plataformas como esta son fácilmente manejables (Beighley, 2009, pt. I).

Un sistema como Drupal, alojado además en un servicio de *hosting* web facilitado en algunas universidades, pero que puede contratarse de modo externo, permite superar la traba de confiar todo el aparato comunicativo y funcional de un proyecto a una única persona especializada. Así mismo, incluso cuando se requiera contratar a personal del ámbito informático, puede tomarse Drupal como punto de partida, para no tener que reinventar sistemas o procesos que, durante los últimos veinte años, han sido ya desarrollados, explorados y mejorados con éxito.

En cualquier caso, el aspecto que constituye probablemente el mayor aliciente para instalar un sistema como Drupal es su compatibilidad con las bases de datos gratuitas con mayor uso a nivel internacional, como MYSQL (versiones superiores a 5.7.8), MariaDB (versión superior a 10.3.7), PostgreSQL (versión superior a 10) e, incluso, SQLite (versión superior a 3.26) (Sipos, 2023, Capítulo 1). Todas estas bases de datos pueden incluir fácilmente los elementos que, posteriormente, se proyectan en la parte visual de la web. Al mismo tiempo, tanto la opción de realizar copias de seguridad como la posibilidad de migrar entre versiones es un proceso sencillo, que puede realizarse desde el propio sistema de Drupal en su panel de administración a partir de la herramienta de gestión de la base de datos (*Phpmyadmin* para MySQL, por ejemplo).

Sin embargo, más allá de las cuestiones técnicas de partida, lo que verdaderamente lo distingue es su capacidad de adaptación, ya que la modularidad es uno de sus principios fundamentales, es decir, las herramientas que ofrece facilitan la creación de contenido versátil y estructurado en torno a diferentes módulos, factor esencial para las experiencias

web dinámicas. No obstante, esta plataforma es una elección más en un mercado de CMS (*Content Management System*) que también incluye otras opciones muy similares y que también son de fácil uso, como Wordpress o, en menor medida, Joomla, como señalamos al comenzar el apartado.

Drupal, por tanto, permite modificar de manera fácil e intuitiva el diseño de la página, añadir o suprimir páginas o secciones, ocultar contenidos y secciones a determinados usuarios, por ejemplo, para contenidos relacionados únicamente con la gestión del entorno o que aún no estén preparados completamente para su visualización (Beighley ,2009, cap. 1).

Con el sistema Drupal, se utiliza PHP (*Hypertext Preprocessor*) un lenguaje de código abierto para procesar solicitudes del usuario (denominado *cliente*), interactuar con bases de datos y generar contenido dinámico, entre otras opciones. Quizá la posibilidad de utilizar bases de datos relacionales con Drupal es, como decíamos previamente, el mayor aliciente para utilizarlo, ya que es una práctica habitual en el contexto de corpus lingüísticos (véase, por ejemplo, Davies, 2005, 2009, 2021). En primer lugar, el sistema permite crear *entidades*, es decir, unidades de uso que luego pueden ubicarse en las páginas de la manera que el investigador considere más oportuna. Un ejemplo de entidades para el proyecto del corpus Ameresco sería la opción de crear entidades como *conversación, hablante, grupo entonativo* o *intervención.* Para cada una de estas entidades pueden crearse etiquetas o categorías; de este modo, una conversación tendrá como categorías ciudad, país, duración, prototipicidad, etc., y también permitirá crear campos de archivo, en los que podrán subirse el archivo de audio, la ficha técnica en PDF, el documento de TextGrid de Praat o el documento .eaf de ELAN. Posteriormente, cada una de esas entidades pueden hacer uso de campos de vínculo que manifiestan las relaciones entre las entidades ya creadas.

En el caso del corpus Ameresco, los campos de vínculo suelen ser la identificación de la conversación (BAQ_001_03_16, por ejemplo), del hablante (BAQ_001_03_16_A, BAQ_001_03_16_B) y, también, de las unidades lingüísticas como el campo de identificación de la intervención y el grupo entonativo, ya que, como es fácil de entender, una palabra pertenece a un grupo entonativo que a su vez forma parte de una intervención que ha realizado a un hablante y así sucesivamente.

Todas las entidades que se crean en el sistema se muestran en la interfaz como fichas de análisis que, en el momento en que se modifican, quedan almacenadas en la base de datos

con la última modificación y, por tanto, la disponibilidad en la parte visual de la web es inmediata.

Durante los primeros años de andadura de Ameresco y desde la implementación de un sistema de consulta en 2015, esta era la manera en la que se gestionaba el corpus. Los datos procedentes de las transcripciones en ELAN o Praat se transformaban, se etiquetaban morfosintácticamente y se actualizaban en el entorno web. No obstante, desde 2019, y con la posibilidad de transformación de Oralstats Aroca (§ 5.2.3.3.), se ha optado por una manera distinta de proceder.

Drupal permite importar tablas en su sistema para ser utilizadas como si hubieran sido creadas dentro del propio entorno. En el caso de Ameresco, la tabla de intervenciones creada por Oralstats Aroca se importa para, posteriormente, poder ser explorada mediante los módulos de consulta del sistema nativo de Drupal y, también, para poder realizar gráficos sencillos.

Básicamente, Oralstats Aroca genera, entre sus muchas formas de salida o *output* para los datos transformados, una tabla con el nombre de *intervenciones*. En ella se incluyen referencias a la conversación, al hablante y sus características y, también, una columna con todas las intervenciones del corpus. Como se ha comentado anteriormente, la experiencia de usuario en análisis de la conversación con la que estamos más familiarizados es aquella por la cual un investigador busca una palabra o conjunto de palabras, sin etiquetado morfosintáctico, y tiene como interés recuperar las intervenciones (y contextos) en los que esa palabra o grupo aparecen.

En la Figura 38 se incluye una captura de pantalla de la tabla TBL_interv2, que se ha creado específicamente para importar la tabla de intervenciones generada por Aroca.

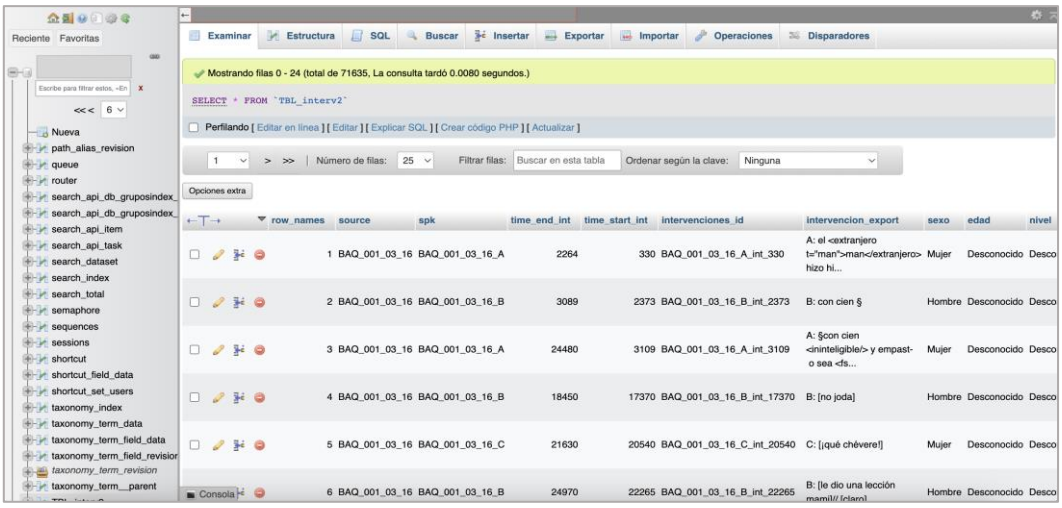


Figura 38. TBL_interv2

Como se ve, dentro del entorno de gestión PHPmyadmin para la base de datos MYSQL que permite el funcionamiento interno de la web, la disposición es exactamente igual a una hoja de cálculo, en la que básicamente hay columnas, filas y celdas. Esta información, convenientemente transformada para su disposición visual por Drupal, será consultada por el usuario en la consulta básica de la web, es decir, dentro del menú, en el epígrafe *Corpus* y, luego, *Consulta*.

La voluntad actual del corpus Ameresco únicamente pretende ofrecer los datos a la comunidad científica en el modo que se considera más general para la realización de estudios en el marco de análisis del discurso en español. Aun así, los datos del corpus están recogidos en multitud de formatos y se ponen a disposición de los investigadores de modo gratuito para que cada cual pueda adaptarlos a sus propias necesidades de investigación.

5.2.3.2.3. Administración interna

Al sistema de administración interna de Drupal puede accederse con un módulo de usuario y contraseña. En el sistema, además de configurar el aspecto de la web y los elementos o entidades que van a ser utilizados, pueden determinarse los usuarios que pueden participar en la administración de la web y, también, crear roles con mayores o menores privilegios de uso. En la Figura 39 se puede observar la interfaz correspondiente a un usuario con perfil de administrador.

Por ejemplo, aunque en Ameresco esta administración suele ser efectuada habitualmente por una única persona, podría darse el caso de que diferentes investigadores en varias partes del mundo contribuyeran en la gestión de una parte controlada de los datos, materiales u otros aspectos de la web. Por ejemplo, cada equipo local podría subir sus propios materiales y alimentar de este modo de manera inmediata la base de datos interna del sistema si así se decidiera desde el equipo central.

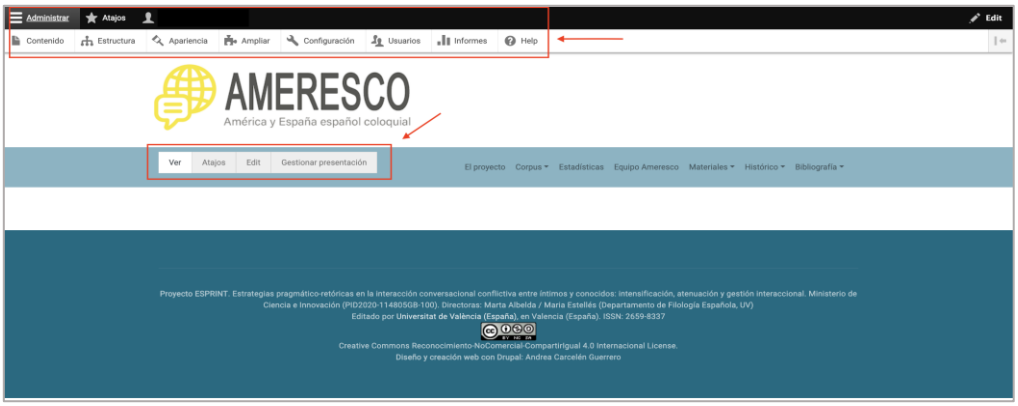


Figura 39. Interfaz de usuario administrador

Las cajas delimitadas en rojo y marcadas con una flecha en la figura anterior indican la parte de administración que el sistema Drupal ofrece a los gestores del sitio. De los ítems que aparecen en el menú superior, más allá de las secciones de *Usuarios*, *Informes* y *Help*, las más importantes, con una pequeña descripción de sus funciones, serían las siguientes:

1. **Contenido**. En Drupal, el contenido es la información que se muestra en el sitio web. Puede incluir texto, imágenes, vídeos, noticias, blogs, productos, páginas estáticas y cualquier otro tipo de información que se desee presentar a los visitantes. La información incluida se gestiona mediante tipos de contenido, que son definidos por el administrador del sitio. En la página de Ameresco son contenido tanto las páginas que consulta el usuario, como los perfiles de los miembros de los equipos de investigación, e, incluso, las unidades que forman parte del corpus lingüístico, como los hablantes, los grupos entonativos, o las propias conversaciones. Para cada uno de estos *contenidos*, los usuarios pueden crear fichas individuales que, inmediatamente, se convierten en filas dentro de tablas de la base de datos MYSQL de Drupal.

2. **Estructura**. Se refiere a cómo se organiza y se presenta el contenido en el sitio web. Drupal ofrece varias herramientas para crear una estructura lógica y jerárquica para

tus contenidos, lo que facilita la navegación y la búsqueda de información. En la estructura pueden configurarse los bloques, las visualizaciones, los menús y, también, pueden crearse los diferentes tipos de contenido.

3. **Apariencia**. Se relaciona con el aspecto visual del sitio web. Drupal utiliza temas para controlar la apariencia del sitio. Un tema es un conjunto de archivos que determina cómo se ven las páginas web, incluyendo los colores, la tipografía, la disposición de los elementos, etc. Puede seleccionarse un tema preinstalado en Drupal o crear uno personalizado.

4. **Ampliar**. En esta sección, Drupal permite actualizar las funcionalidades básicas del sistema nativo con módulos especializados. Por ejemplo, la versión actual del corpus Ameresco utiliza unos cuarenta módulos específicos; de entre ellos, posiblemente uno de los más utilizados y con más repercusión sería *Better Exposed Filters.* Se trata de un módulo que aumenta la capacidad de filtrado de los datos del corpus, es decir, de cómo aparece dispuesto el campo de Búsqueda (y la obligatoriedad de incluir una cadena de texto en él para poder buscar) en la sección Consulta.

5.2.3.2.4. Diseño visual

En Drupal puede seleccionarse un amplio número de plantillas que proporcionan una estructura y una gama de colores directamente aplicables sobre la interfaz de la web. Para el estado actual hemos seleccionado un subtema llamado Barrio[54] que consiste en una ampliación de la librería *Bootstrap*[55] para la paleta de colores, la tipografía y la configuración de las cajas de búsqueda o los bloques, hemos utilizado el tema conocido como Yeti[56]. Así pues, el subtema Barrio sirve de arquitectura general para la disposición de botones, cajas de formulario, enlaces y otros elementos HTML, que tienen o muestran un tema predefinido en la librería (en este caso, Yeti).

[54] https://www.drupal.org/project/bootstrap_barrio
[55] https://bootswatch.com/
[56] https://bootswatch.com/yeti/

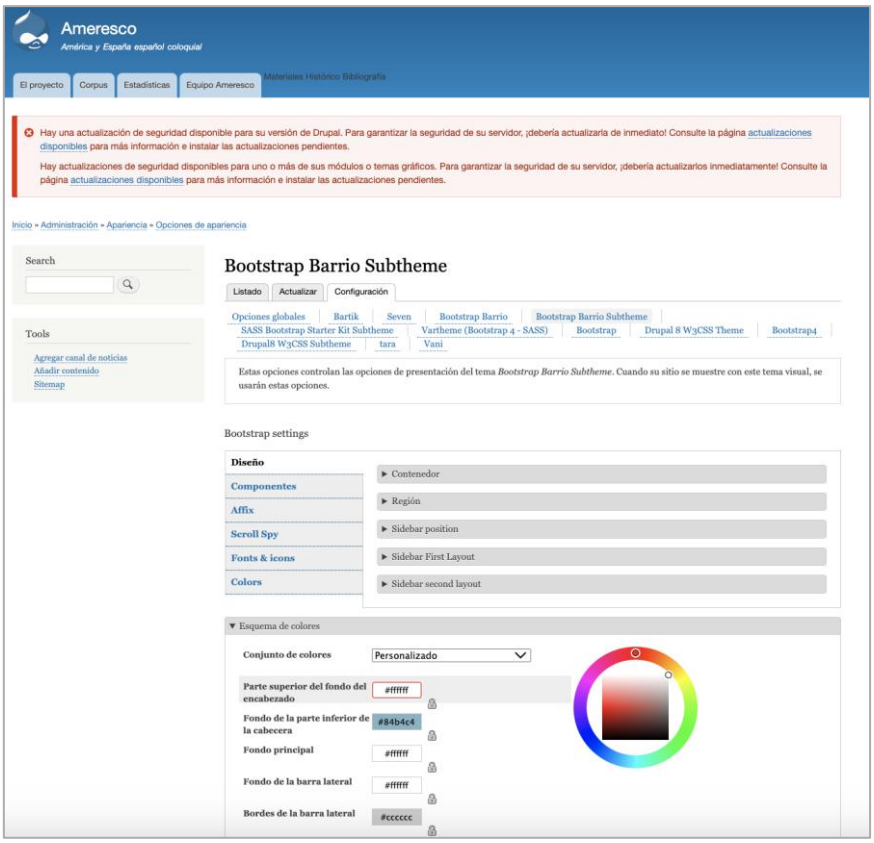


Figura 40. Interfaz configuración de la apariencia

Se observa cómo, dentro de la interfaz general de Drupal, puede cambiarse el color del fondo y, también, de otros elementos de la web que, por la extensión de esta página de modificación, no caben en la Figura 40.

En cualquier caso, la intención principal para la parte visual de la web es que se trate de una web accesible, limpia, sin demasiados colores y que, por otro lado, no esté sobrecargada de elementos que puedan dificultar el acceso a la información o la consulta de los datos del corpus. Por eso se han desestimado la inclusión de imágenes transitorias en forma de carrusel en la página principal y la presencia de *pop-ups*.

Se ha intentado, en la medida de lo posible, emplear un tema básico y, en definitiva, otorgar el protagonismo a los datos y al proyecto en sí. La interfaz o la configuración visual dota de importancia únicamente en tanto que contribuye a mejorar la experiencia del usuario.

5.2.3.2.5. Principales funcionalidades de uso

Como se ha mencionado en los capítulos anteriores (3 y 4), en un corpus lingüístico, una sección de búsqueda es esencial para permitir a los investigadores encontrar rápidamente datos relevantes, facilitando así la investigación y el análisis lingüístico (§ 5.2.3.2.5.1.). La capacidad de descargar archivos del corpus es crucial para trabajar *offline*, realizar análisis avanzados y garantizar la preservación de datos (§ 5.2.3.2.5.2.). Además, las estadísticas del corpus ofrecen una visión general de la composición de los datos, guían la investigación y ayudan en la validación y comparación de corpus § 5.2.3.2.5.3.), lo que es fundamental para estudios lingüísticos más informados y precisos. Estos tres aspectos combinados facilitan la accesibilidad, comprensión y análisis efectivo de los datos lingüísticos en diversas investigaciones lingüísticas.

En las siguientes secciones expondremos cada uno de esos aspectos, ya que consideramos que son básicos en la explotación de los corpus lingüísticos, en concreto, detallando las especificidades del corpus Ameresco.

5.2.3.2.5.1. Consulta básica por intervención

En esta sección introduciremos el apartado de la web que permite consultar el corpus Ameresco. Como se observa en la Figura 41, esta consulta inicial es una consulta sencilla, en la que hay pocos campos de búsqueda, pero uno de ellos es obligatorio: la búsqueda por secuencia de caracteres es el único campo que acepta escritura por parte del usuario. El resto de los campos incluye una lista desplegable de la que puede seleccionarse un único elemento cada vez que se realice una consulta. Finalmente, dos botones, *Buscar* y *Reiniciar*, permiten iniciar la búsqueda, en el caso del primer botón, y dejar la búsqueda en blanco, en cuanto al segundo botón.

Figura 41. Consulta básica por intervención en el corpus Ameresco

Esta figura muestra, por tanto, los cinco campos de búsqueda generales de la plataforma en el corpus Ameresco que utiliza el sistema originario de Drupal:

1. Campo obligatorio en el que hay que incluir una secuencia de búsqueda.
2. Nivel de instrucción del hablante.
3. Sexo del hablante.
4. Edad del hablante.
5. Ciudad de procedencia del hablante.

El primer campo, *Texto,* permite expresiones regulares y consultas básicas textuales. Por eso mismo, un primer mensaje advierte al usuario de que debe utilizar un espacio en blanco antes y después de la secuencia de búsqueda que se quiera realizar. De esta manera, si el usuario busca la palabra *casa*, el buscador ofrecerá como resultados: *casas, casamiento*… En el corpus actual, la búsqueda de *casa* ofrece 926 resultados. Sin embargo, si buscamos con un espacio en blanco antes y después (“ casa ”), los resultados descienden a 510.

Por sintetizar, exponemos algunos ejemplos de búsquedas que pueden realizarse a través de esta página básica de consulta:

- **Búsqueda de cadena de texto**. Si se busca *que* los resultados incluirán la palabra *que*, pero también otros resultados como *queriendo, queso, dique, etc.*
- **Búsqueda de palabra simple**. Si se busca exactamente la palabra *que*, puede buscarse añadiendo un espacio en blanco antes y después de la palabra.
- **Búsqueda de palabras al inicio del grupo entonativo**: .*: que
- **Búsqueda de palabras al final del grupo entonativo**: que$
- **Búsqueda de palabras que constituyan un único grupo entonativo**: .*: sí$

En la Figura 42 mostrada abajo se representa la búsqueda de la secuencia *yo digo que*. En la página de resultados se ofrece un gráfico sencillo de barras horizontales con los datos absolutos; se trata de una primera aproximación a la consulta realizada y, en este caso, no está relativizada sobre un millón de datos, como suele ser práctica habitual en el uso de corpus lingüísticos (McEnery y Wilson, 2001, p. 82).

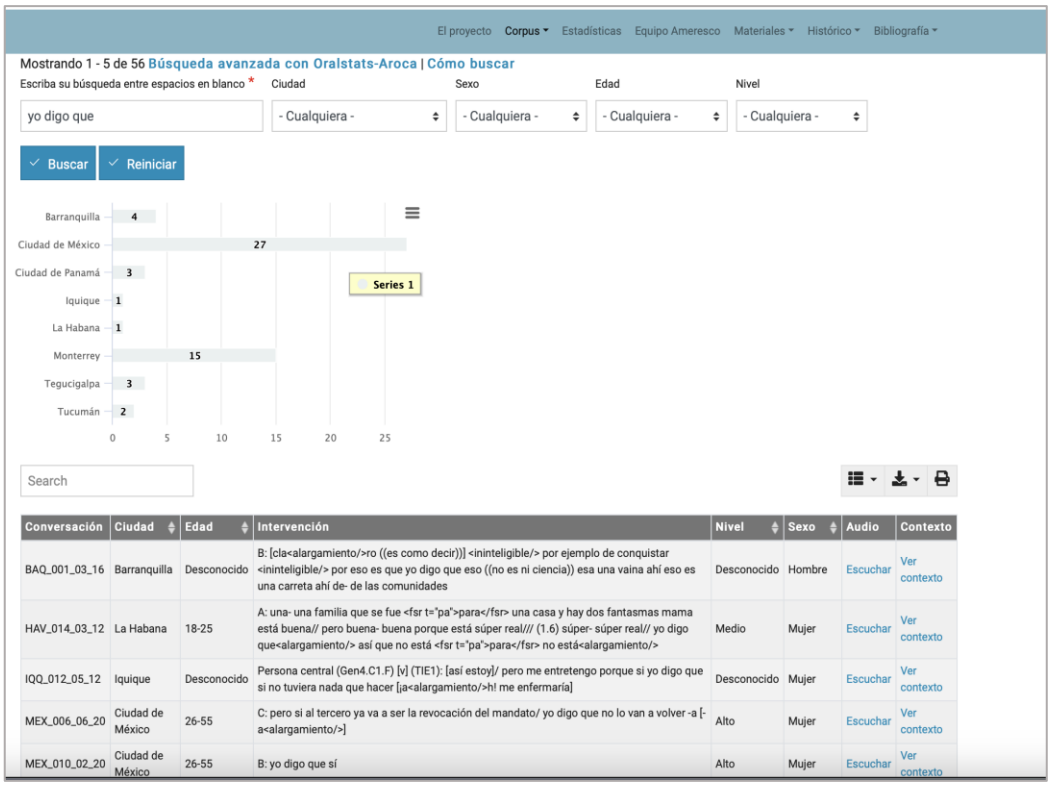


Figura 42. Búsqueda y resultados para *yo digo que* en el corpus Ameresco

Así pues, la búsqueda mediante la interfaz básica de Drupal y los resultados obtenidos tienen como objetivo ofrecer una panorámica general sobre los datos; el mayor valor no se concede al gráfico de barras, que es meramente orientativo, sino a las concordancias que aparecen dispuestas en la parte inferior por grupos de cinco. Estos grupos pueden descargarse en documentos Excel, TXT, XML, etc. Al mismo tiempo, las concordancias pueden reproducirse utilizando el botón *Escuchar* y, también, puede verse un contexto de la intervención en la que aparece la concordancia. Este contexto, como se observa en la Figura 43, incluye cinco intervenciones antes y cinco intervenciones después de la intervención encontrada.

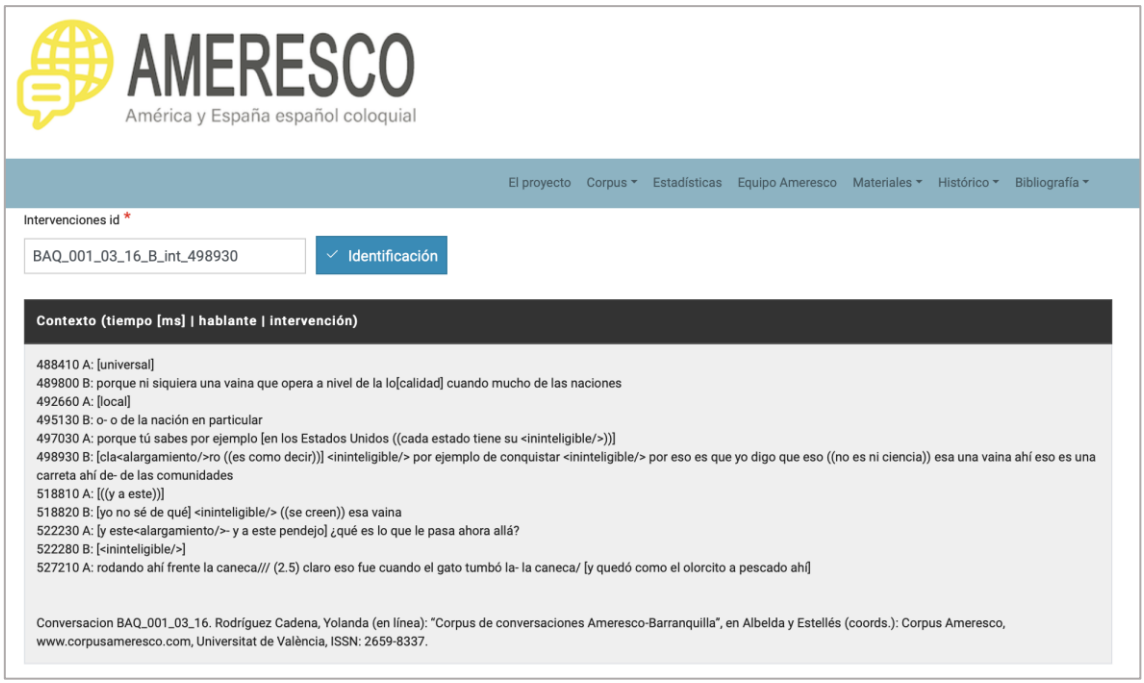


Figura 43. Resultados búsqueda por intervención en el corpus Ameresco

En el contexto del ejemplo de la Figura 43 aparece una referencia numérica a los milisegundos concretos en los que se ubican las intervenciones; de esta manera, en caso de que se requiera aún más contexto, el investigador puede acceder a la conversación completa (BAQ_001_03_16) y buscar ese tiempo exacto. Uno de los aspectos más funcionales es que los contextos incluyen la manera de citar para que el investigador tenga la posibilidad de incluir esta información directamente en sus trabajos y pueda realizar la referencia de manera directa.

En caso de que el investigador requiera de alguna búsqueda más compleja, puede descargar el corpus entero en su dispositivo clicando en la sección *Archivos*. Allí podrá encontrar los archivos del corpus en formato de ELAN, Textgrid y una transcripción conversacional en formato de texto plano; el programa ELAN ya facilita búsquedas complejas, pero también puede transformar los datos según los propios intereses particulares de cada investigador, dado que el formato de este programa es un XML. Otra opción que puede realizar el investigador es descargar las transcripciones en el formato que desee y utilizar otros programas accesibles en el mercado como Wordsmith, Antconc o Atlas.ti.

5.2.3.2.5.2. Descarga de archivos

Algunos de los corpus lingüísticos actuales ofrecen sus materiales para que el usuario pueda descargarlos, como se ha visto en la sección 3.4.1.; en unos casos, ese material se ofrece mediante el pago de una cantidad económica, como, por ejemplo, sucede con algunas de las bases de datos de los corpus de Mark Davies (Davies 2009, 2012, 2021; Davies y Kim, 2019). En otras ocasiones, el material se facilita de modo gratuito a quienes quieran hacer uso del corpus recogido; en el marco de los corpus orales, sería el caso, como ejemplo, del *Santa Barbara Corpus of Spoken American English* (Du Bois *et al*., 2000). Ante esta situación, como hemos señalado con anterioridad, una de las apuestas de la construcción del corpus Ameresco, ha sido, desde sus orígenes, la de facilitar el intercambio y la reusabilidad de los datos de manera que cualquier persona pueda servirse tanto de la metodología de trabajo, como de los datos recopilados.

En la Figura 44 que se muestra a continuación se observa la sección *Archivos* de la web de Ameresco; en ella, se accede a un buscador al inicio en el que puede se puede filtrar por factores como la ciudad, el género discursivo[57], el país o el grado de prototipicidad de la conversación. Esta sección conforma el *repositorio de archivos* del corpus desde la cual puede accederse y descargarse todo el material en bruto.

[57] Inicialmente, como ya hemos comentado, se consideró la opción de recoger otros formatos orales como entrevistas, aunque de momento no se ha llevado a cabo esta tarea.

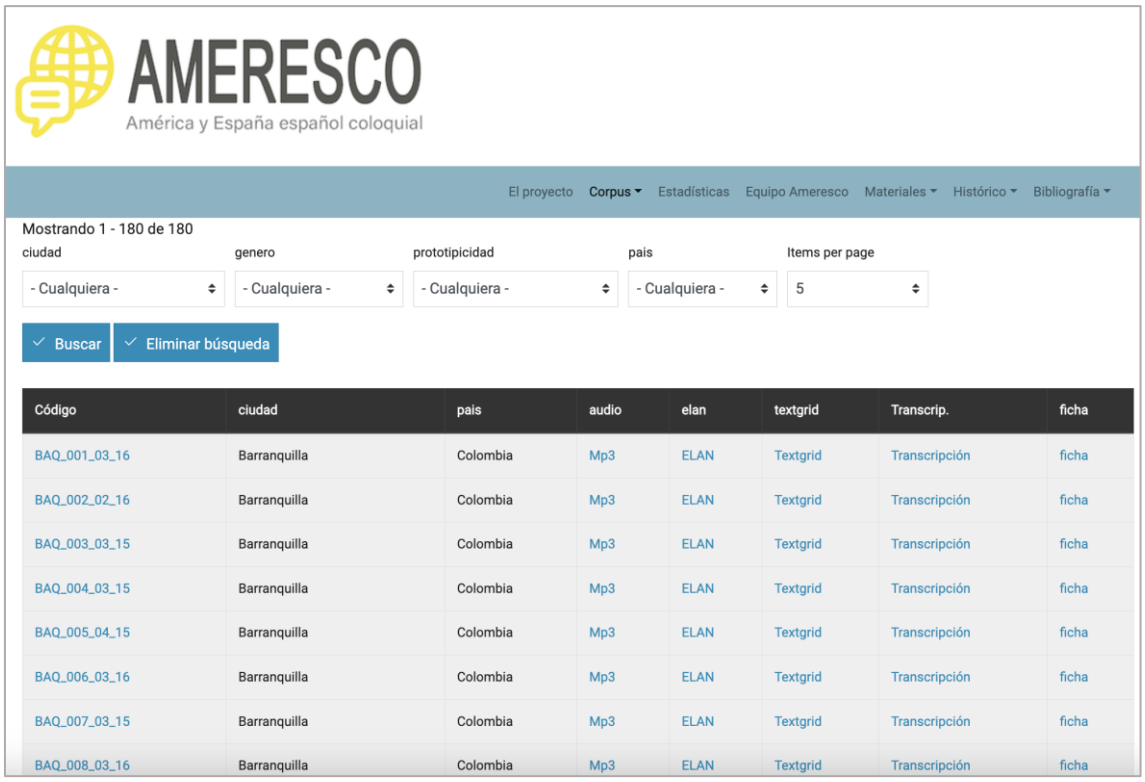


Figura 44. Sección Archivos

La Figura 44 pone de manifiesto la variedad de formatos en los que se ofrece la información del corpus. Según el interés de estudio, se pueden descargar los archivos de texto o transcripción, aptos para estudios de carácter conversacional o discursivo (en los que no sea tan importante la consulta de los audios), o acceder tanto a los audios, en formato MP3, como a los archivos de transcripción en formato .eaf de ELAN y en formato TextGrid para estudios relacionados con el ámbito fonético.

Los audios que conforman el corpus Ameresco se han transcrito y alineado con el texto mediante ELAN (§ 5.2.2.1.2.). Este sistema permite exportar posteriormente a archivos de Praat (formato TextGrid); por su parte, la transcripción en formato de texto plano y con estructura de transcripción clásica conversacional se genera mediante una de las secciones de Oralstats Aroca, como veremos en la correspondiente sección. Por su parte, las fichas técnicas se recogen en un documento Word y se incluyen en la web como documento PDF.

En la línea de accesibilidad que guía toda la construcción del corpus, la puesta a disposición de manera gratuita de todo el material recogido, procesado, codificado y almacenado en Ameresco se extiende incluso a la base de datos en formato MYSQL, ya que

en caso de que se realizara alguna investigación que lo requiriera, esta podría ser solicitada a los gestores del proyecto y, en tal caso, sería facilitada sin problemas.

5.2.3.2.5.3. Estadísticas generales del corpus

Las estadísticas del corpus, como se ha dicho, se incluyen de manera estática, en formato de tabla y sin gráficos, en la sección *Estadísticas de la web*. Estas estadísticas incluyen los siguientes elementos:

1. **Ciudad.** El nombre de la ciudad o región a la que se refieren los datos.
2. **Hablantes.** El número de personas que intervienen en esa ciudad o región.
3. **Conversaciones.** El número de conversaciones que se han registrado o tienen lugar en esa ciudad o región.
4. **Grupos.** La cantidad de grupos entonativos.
5. **Intervenciones.** El número de veces que se ha intervenido o participado en una conversación o grupo en esa ciudad o región.
6. **Palabras.** La cantidad total de palabras utilizadas en las conversaciones de esa ciudad o región.

Cada fila de datos representa una ciudad o región específica y proporciona información sobre estas métricas en ese lugar en particular. Estos datos podrían ser útiles para comprender la actividad de habla y la dinámica social en diferentes áreas geográficas.

En el caso de las estadísticas que aparecen en la web, estas responden a un criterio de combinación entre factores, es decir, primero se aportan las estadísticas sobre frecuencias absolutas por ciudad, luego por ciudad y sexo, después por ciudad, sexo y edad, y así sucesivamente, hasta terminar en la combinatoria máxima que incluye un hablante.

Dado que las estadísticas son muy amplias, incluimos a modo de ejemplo únicamente las totales de todo el corpus, por ciudad y las primeras diez referidas a los hablantes. Estas estadísticas se han generado desde la base de datos original recogida en palabras que utiliza Oralstats Aroca para su funcionamiento interno.

Estadísticas totales

países	ciudades	hablantes	conversaciones	grupos	intervenciones	palabras
11	16	675	180	141203	81259	742170

Tabla 13. Estadísticas totales del corpus Ameresco

Estadísticas por ciudad

ciudad	hablantes	conversaciones	grupos	intervenciones	palabras
Barranquilla	29	8	5504	2358	31008
Buenos Aires	28	8	7558	4276	39540
Ciudad de México	185	54	47040	27756	237270
Ciudad de Panamá	20	5	3704	1924	19538
Iquique	99	11	11183	8108	53662
La Habana	118	36	17406	9260	102890
Las Palmas	9	2	2040	1358	12528
Loja	4	1	866	574	3587
Medellín	4	2	1495	1131	10817
Monterrey	45	15	13882	7234	71782
Querétaro	3	1	508	138	2744
Santa Cruz	2	1	714	283	3255
Santiago de Chile	29	8	5542	3221	28963
Tegucigalpa	37	10	6815	4063	36722
Temuco	48	13	13012	7519	65134
Tucumán	15	5	3665	1787	20910

Tabla 14. Estadísticas por ciudad del corpus Ameresco-Tucumán

Estadísticas por ciudad, sexo, edad, nivel, edad, hablante

ciudad	sexo	nivel	edad	hablante	grupos	intervenciones	palabras
Barranquilla	H	Alto	18-25	BAQ_003_03_15_C	215	129	1208
Barranquilla	H	Alto	26-55	BAQ_005_04_15_C	387	107	1747
Barranquilla	H	Alto	26-55	BAQ_005_04_15_D	50	31	202
Barranquilla	H	Alto	Más de 56	BAQ_006_03_16_C	217	85	804
Barranquilla	H	Desconocido	Desconocido	BAQ_001_03_16_B	83	61	997
Barranquilla	H	Medio	Más de 56	BAQ_008_03_16_B	45	39	208
Barranquilla	H	Medio	Más de 56	BAQ_008_03_16_C	309	126	1623
Barranquilla	M	Alto	18-25	BAQ_003_03_15_A	220	111	1561
Barranquilla	M	Alto	18-25	BAQ_003_03_15_B	284	123	1795

Tabla 15. Estadísticas por ciudad, sexo, edad, nivel, edad, de un hablante del corpus Ameresco-Barraquilla (resultado parcial)

Aunque podría ser interesante mostrar esta información de manera dinámica directamente desde la base de datos, cabe recordar que la base de datos que se incluye en Drupal es únicamente la de las intervenciones. En el caso de Oralstats Aroca, como veremos a continuación, sí que dispone de una opción de visualización mediante gráficos de barra porque hay un acceso a todas las tablas que conforman la base de datos del corpus: palabras, grupos entonativos, intervenciones, hablantes y conversaciones.

5.2.3.3. Oralstats Aroca

En esta sección presentamos de qué manera una herramienta de *software*, desarrollada por los propios lingüistas del proyecto Ameresco, sirve tanto para gestionar y procesar los datos como para analizar el corpus. En tal sentido, Oralstats Aroca (Cabedo y Carcelén, 2022) se articula como un entorno informático que permite transformar y analizar datos de cualquier interacción verbal y que, focaliza, especialmente, en las conversaciones recogidas en conjunción con un grupo de metadatos de carácter sociolingüístico (edad, nivel de instrucción y sexo).

El nuevo enfoque supone cruzar toda la información recogida y unirla en un mismo análisis, con la posibilidad de que este sea tanto exploratorio como inferencial. Para este propósito, los datos pueden relacionarse en un ecosistema de variabilidad común, con un conjunto de variables numéricas y categóricas generadas automáticamente; estas últimas pueden seleccionarse de un repertorio cerrado (sexo, edad, hablante...) o incluso diseñarse y personalizarse de acuerdo con las necesidades específicas de investigación (estudio de la atenuación, (des)cortesía, humor, ironía...).

Este sistema computacional gratuito desarrollado con R (R Core Team, 2021) llamado *Oralstats* (Cabedo Nebot, 2021), permite analizar transcripciones alineadas con el audio mediante códigos de tiempo, teniendo en cuenta factores comunes en el análisis de corpus, como la frecuencia general de palabras, las categorías de partes del discurso y los bigramas o trigramas, pero también variables menos conocidas, como el tono, la intensidad o la duración. Por lo tanto, los datos procedentes de las transcripciones se enriquecen de modo automático con anotaciones procedentes de otros niveles, como la morfosintaxis o la prosodia.

Es en el marco de la aplicación de nuevas tecnologías y del uso de lenguajes de programación en el que Oralstats se inserta. La finalidad última es ofrecer una herramienta de uso que facilite el acceso al corpus y que, con código de libre acceso, permita a expertos con un conocimiento informático más avanzado el poder customizar el código base y poder, en tal sentido, añadir nuevas capas de información a las ya existentes.

Dadas las necesidades específicas de un corpus conversacional como Ameresco, se propuso la modificación del código base de Oralstats[58]; el objetivo originario del autor del

[58] Este código base puede encontrarse en https://github.com/acabedo/oralstats

programa fue realizar el estudio fonético de las transcripciones procedentes de programas como ELAN o Praat, con la finalidad última de establecer patrones de comportamiento verbal y paraverbal asociados a estratos sociolingüísticos (edad, sexo, nivel de instrucción) o incluso a hablantes particulares.[59]

Por ello, en el caso de un corpus conversacional, con las expectativas además de los usuarios del análisis del discurso, se requirieron unas sustanciales modificaciones y simplificaciones del código inicial del programa, sobre todo por lo que respecta al módulo de visualización y consulta de la base de datos interna. En tal sentido, surgió la propuesta de Oralstats Aroca[60] y cuyo código es de libre acceso[61]. Esta modificación tuvo como voluntad enfatizar más en la experiencia de usuario y menos en la transformación concreta y en la realización de pruebas de estadísticas inferencial (mapas de calor, árboles de decisiones, etc.), que sí permite realizar la versión original de Oralstats.

En esta modificación, Cabedo (2021), autor del programa inicial, se ocupó de simplificar y minimizar la parte del código correspondiente a la transformación de los datos, mientras que la gestora actual del corpus (la autora de esta investigación) se preocupó de diseñar, confeccionar y disponer la parte de visualización por parte del usuario, mediante el enfoque en la sección *ui* (*user interface*), del entorno de ejecución Shiny (Chang *et al*., 2021). Shiny es el módulo de R que utiliza toda versión de procesos realizados con R (transformaciones, estadísticas, visualizaciones…) que quieran disponerse *online*. Los programas que se escriben con Shiny tienen una estructura como la que sigue:

```
library(shiny)
ui <- fluidPage()
server <- function(input, output){}
shinyApp(ui = ui, server = server)
```

Figura 45. Código R Shiny

La parte de *ui* del código es la que configura *cómo* va a ver la información el usuario; por su parte, la sección *server* es en la que se realizan todas las tareas de programación como transformaciones o modificaciones internas de los datos. Esta sección correspondería al *qué* se está haciendo con los datos, pero no es una zona del código que el usuario final vaya a ver. Un ejemplo creado *ad hoc* sería el que sigue:

[59] Algunas de estas investigaciones pueden consultarse en Cabedo (2022a), Cabedo (2022b) y Cabedo e Hidalgo (2023).
[60] Actualmente, se encuentra alojado en https://adrin-cabedo.shinyapps.io/aroca_viewer/
[61] Disponible en https://github.com/acabedo/aroca

```
library(shiny)
library(dplyr)

datosinventados <- data.frame(
                  intervenciones = c('hola', 'adiós',
'hola','antes','sí', 'no', 'antes'))

ui <- fluidPage(

  textInput(inputId = "busqueda",label = "Búsqueda"),
  dataTableOutput("resultado")

)
server <- function(input, output){

  resultado <- reactive({

    datosinventados <- datosinventados%>%filter(intervenciones ==
input$busqueda)
  })

  output$resultado <- renderDataTable(resultado())

}
shinyApp(ui = ui, server = server)
```

Figura 46. Ejemplo de código

En el ejemplo anterior, en la sección *ui* hay solo dos elementos: un textInput, que creará un campo de texto, y un dataTableOutput, que creará una tabla. La parte del código que dice al programa qué debe aparecer en la tabla está en la sección *server*; en ella, básicamente se indica que se va a filtrar la variable intervenciones con todo aquel texto que coincida exactamente con los que se escriba en el campo de texto. El resultado puede verse en la Figura 47:

Figura 47. Pantalla de visualización tras ejecutar el código de la Figura 46

Así pues, Oralstats Aroca, en cuanto a lenguaje de programación, hace uso de la estructura Shiny anteriormente mencionada. En general, permite realizar cualquier operación que R

puede llevar a cabo en un entorno local mediante R Studio (Posit team, 2023), pero con la novedad de que puede hacerse desde un entorno en línea mediante una página web.

Una imagen de la interfaz principal del programa, con el uso de la base de datos correspondiente al corpus Ameresco, se encuentra en la Figura 48:

Figura 48. Interfaz página de inicio Oralstats Aroca

En resumen, Oralstats Aroca es una herramienta informática dinámica para explorar datos de habla, un procesador de minería de datos con varias opciones disponibles. Puede mostrar concordancias frecuentes y medidas de variables léxicas, morfológicas, prosódicas y posicionales con una variable independiente como entrada (hablante, sexo, edad, ciudad). En las siguientes secciones (§ 5.2.3.3.1. y § 5.2.3.3.2.) desglosaremos las funcionalidades del *script* del programa, tanto en su parte de transformación como en la parte de consulta y visualización de los datos; para ello, intentaremos ejemplificar el uso mediante casos prácticos que pongan de manifiesto la posible utilidad de la herramienta.

5.2.3.3.1. Módulo de transformación de los datos: *script* Oralstats.creación

En Oralstats Aroca se han modificado parte de las funcionalidades del *script* original de Oralstats; este permitía realizar transformaciones de datos directamente en el navegador vía aplicación Shiny. Con la voluntad de facilitar la adaptación a la gestión de corpus interaccionales, en los que las modificaciones de los datos no requieren de una especificidad

en palabras o fonemas alineados al tiempo, se ha generado una modificación local del *script* original que solo funciona en entorno local y no es accesible desde conexión en línea, se ha llamado Oralstats.creación. Los datos transformados que se generan en esta sección de código de Oralstats Aroca son los que posteriormente nutren al módulo de visualización, que se expone en la sección 5.2.3.3.2. Si bien todo el código de la parte Oralstats.creación es muy extenso, vamos a detenernos en este apartado en comentar los fragmentos más importantes.

El principio del código está destinado a realizar una serie de tareas relacionadas con la creación y verificación de directorios en el sistema de archivos. Aquí se explica lo que hace cada parte del código:

1. **Librerías**. Al principio se cargan varias librerías necesarias para el funcionamiento del código. Estas librerías proporcionan funciones y herramientas adicionales para trabajar con datos y crear aplicaciones Shiny. Entre las librerías más relevantes destacaremos DBI (Wickham, and Müller, 2022), que permite conectar nuestro programa con una base de datos externa (PostgreSQL en este caso) y tidyverse (Wickham *et al.,* 2019), que integra una serie de paquetes que permiten transformar, filtrar y visualizar los datos.

2. **Establecer el directorio de trabajo**. Utiliza la función **setwd** para establecer el directorio de trabajo. Ese directorio es donde se buscarán y crearán aquellos otros en los que se guardarán los archivos que se generen (por ejemplo, las transcripciones en formato TXT y compiladas en intervenciones que son posteriormente las que se ubican en la sección *Archivos de la web*). El directorio de trabajo se establece utilizando la ubicación del archivo oralstats.creación, ejecutado desde RStudio. Establecerlo en el entorno local de un ordenador es muy importante porque la referencia de los archivos que se vayan creando o modificando harán referencia a esta ruta en nuestro ordenador.

3. **Crear directorios**. A continuación, el código verifica si varios directorios específicos ya existen en el sistema de archivos. Si no existen, los crea; de lo contrario, imprime "ya existe". Cada uno de estos directorios están relacionados con diferentes aspectos de la aplicación Shiny, como el almacenamiento de archivos de audio, texto, metadatos, resultados, etc. Estos directorios se utilizan para organizar y gestionar los datos utilizados por la aplicación. Por ejemplo, los primeros tres directorios

(**eaf_sample**, **txt_sample**, **textgrid_sample**) están destinados a recoger los archivos de ELAN, texto tabulado o Textgrids de Praat que contengan las transcripciones realizadas por los investigadores; evidentemente, lo habitual es que sea uno solo de los tres tipos de archivos los que sean utilizados. En los directorios **outputs** y sus subdirectorios (**conversaciones**, **rds**, **txts**) es donde se almacenan los resultados generados por la aplicación. Por ejemplo, el directorio *conversaciones* almacenará los archivos de texto que correspondan a las transcripciones en formato de intervención, después de haberse efectuado las modificaciones pertinentes desde el formato original (convencionalmente, los grupos entonativos que se hayan transcrito y alineado al audio mediante ELAN o Praat).

4. **Almacenamiento de datos transformados.** Los datos transformados pueden guardarse en diferentes formatos, si bien, la opción más conveniente es utilizar una base de datos. En cuanto a estas, dos de las más extendidas son PostgreSQL y MySQL, aunque también puede usarse la opción de SQLite. En un entorno local, la base de datos RSQLite es más que suficiente, mientras que, si la aplicación se aloja en un Shiny Server externo o en http://Shinyapps.io, es recomendable, de cara a la celeridad de la carga de los datos, que estos se alojen en un servidor externo con buena disponibilidad.

Quizá la parte más relevante en la transformación de los datos es aquella en la que los grupos entonativos extraídos de las transcripciones se anotan con valores fónicos procedentes de los archivos de intensidad y tono originarios de Praat. Al mismo tiempo, estas anotaciones, que habitualmente corresponden a las medias que se encuentran de cada variable por grupo entonativo, se complementan con información gramatical, cantidad de categorías morfológicas por grupo (sustantivos, verbos, adjetivos…) y cantidad de palabras catalogadas como positivas o negativas, según la coincidencia con una lista de palabras etiquetadas como positivas o negativas *a priori* y que pueden modificarse según los intereses particulares de investigación.

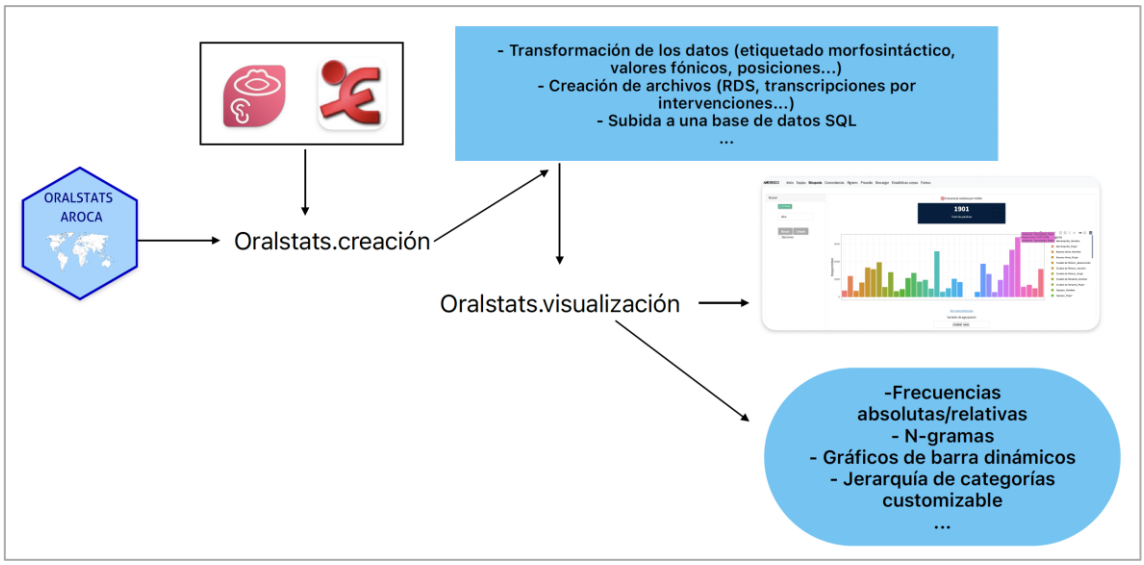


Figura 49. Esquema del procesado del corpus en Oralstats Aroca

Así pues, el formato de entrada general puede ser un archivo .eaf de ELAN, un TextGrid de Praat o un archivo tabulado TXT, que puede, a su vez, proceder de la exportación también desde ELAN. En las siguientes líneas se incluyen ejemplos del mismo fragmento de transcripción en formatos .eaf, TextGrid y TXT tabulado:

```
    <TIER DEFAULT_LOCALE="es" LINGUISTIC_TYPE_REF="default-lt"
      PARTICIPANT="B" TIER_ID="B">
      <ANNOTATION>
          <ALIGNABLE_ANNOTATION ANNOTATION_ID="a198"
              TIME_SLOT_REF1="ts81" TIME_SLOT_REF2="ts99">
              <ANNOTATION_VALUE>con cien</ANNOTATION_VALUE>
          </ALIGNABLE_ANNOTATION>
      </ANNOTATION>
      <ANNOTATION>
          <ALIGNABLE_ANNOTATION ANNOTATION_ID="a70"
              TIME_SLOT_REF1="ts503" TIME_SLOT_REF2="ts543">
              <ANNOTATION_VALUE>no joda</ANNOTATION_VALUE>
          </ALIGNABLE_ANNOTATION>
      </ANNOTATION>
      <ANNOTATION>
          <ALIGNABLE_ANNOTATION ANNOTATION_ID="a71"
              TIME_SLOT_REF1="ts655" TIME_SLOT_REF2="ts735">
              <ANNOTATION_VALUE>le dio una lección
mami</ANNOTATION_VALUE>
          </ALIGNABLE_ANNOTATION>
      </ANNOTATION>
      <ANNOTATION>
          <ALIGNABLE_ANNOTATION ANNOTATION_ID="a72"
              TIME_SLOT_REF1="ts759" TIME_SLOT_REF2="ts787">
              <ANNOTATION_VALUE>claro</ANNOTATION_VALUE>
          </ALIGNABLE_ANNOTATION>
      </ANNOTATION>
      <ANNOTATION>
          <ALIGNABLE_ANNOTATION ANNOTATION_ID="a73"
              TIME_SLOT_REF1="ts1133" TIME_SLOT_REF2="ts1231">
              <ANNOTATION_VALUE>¡ah! pero eso fue en cuarto yo pensé
que era ahora</ANNOTATION_VALUE>
          </ALIGNABLE_ANNOTATION>
      </ANNOTATION>
```

Figura 50. Estructura interna de un archivo .eaf de ELAN

```
    item [2]:
      class = "IntervalTier"
      name = "B"
      xmin = 0
      xmax = 609.985
      intervals: size = 167
      intervals [1]:
          xmin = 0
          xmax = 2.373
          text = ""
      intervals [2]:
          xmin = 2.373
          xmax = 3.089
          text = "con cien"
      intervals [3]:
          xmin = 3.089
          xmax = 17.37
          text = ""
      intervals [4]:
          xmin = 17.37
          xmax = 18.45
          text = "no joda"
      intervals [5]:
          xmin = 18.45
          xmax = 22.265
          text = ""
      intervals [6]:
          xmin = 22.265
          xmax = 23.665
          text = "le dio una lección mami"
```

Figura 51. Estructura interna de un archivo TextGrid de Praat

```
tier  spk tmin  tmax  dur annotation  filename
B   B   2373    3089    716 con cien    BAQ_001_03_16.eaf
B   B   17370   18450   1080    no joda BAQ_001_03_16.eaf
B   B   22265   23665   1400    le dio una lección mami BAQ_001_03_16.eaf
B   B   24640   24970   330 claro   BAQ_001_03_16.eaf
```

Figura 52. Estructura interna de un archivo tabulado

Con la transformación efectuada por Oralstats.creación, se pueden obtener secuencias como la siguiente:

```
1   00:00:00.33 A: el <extranjero t="man">man</extranjero> hizo hizo treinta y ella salió con
2   00:00:02.37 B: con cien §
3   00:00:03.10 A: §con cien <ininteligible/> y empast-  o sea <fsr t="enmarcao">enmarcado</fsr> <i
ninteligible/> no se algún <fsr t="lao">lado</fsr> en todo caso lo llevó<alargamiento/>// tremendo de
trabajo <fsr t="ice">dice</fsr> <cita>me voy a sacar un cinco</cita> dice <anonimo>Wendy</anoni
mo> esa fue la lección de mejor dicho/ lo que el profesor le puso <obs t="ruido"/> fue un tres <fsr t="
pelao">pelado</fsr> [<risas/>] <entre_risas>un tres <fsr t="pelao">pelado</fsr> le puso</entre_risas
> [<ininteligible/>] eso fue una lección [una lección] que le dio <obs t="ruido"/>
4   00:00:17.37 B: [no joda]
5   00:00:20.54 C: [¡qué chévere!]
6   00:00:22.26 B: [le dio una lección mami]// [claro]
```

Figura 53. Visualización final de la transcripción

En la Figura 53, se observa cómo los grupos entonativos transcritos con ELAN se disponen en formato conversacional clásico, según las convenciones de transcripción establecidas en el seno del grupo Val.Es.Co. (Briz y Grupo Val.Es.Co., 2002); así, símbolos como las barras /, // o ///, que marcan pausas de menos a mayor duración, o §, que indica intervenciones inmediatas temporalmente[62], o incluso la creación de unidades que empiezan con un hablante seguido de dos puntos, se realiza directamente desde Oralstats.creación.

Al mismo tiempo, los grupos entonativos también se pueden enriquecer con información prosódica como la que se observa en la Figura 54. Cabe mencionar que solo podemos incluir algunas de las variables, pero se exponen todas ellas más abajo. Para crear estas configuraciones prosódicas, el sistema necesita información de tono e intensidad en las carpetas habilitadas para ello, concretamente *pitch* e *intensity.* De estos dos últimos archivos incluimos ejemplos en la Tabla 16 y en la Tabla 17:

	tier	spk	time_start	time_end	dur	pitch_mean	intensity_mean	inflexion_st	range_st	palabras	velocidad	dif_pitch	dif_range	dif_inten	dif_inflexion	dif_dur	dif_velocidad	sexo	edad	nivel	content
1	A	BAQ_001_03_16_A	330	2264	1934	272.5	56.3	−8.9	13.8	9	4.65	2.76	3.34	−2.07	−8.78	−890.71	1.66	NA	NA	NA	el <extranjero t="man"
2	A	BAQ_001_03_16_A	3109	9160	6051	229.0	59.2	0.3	13.9	18	2.97	−0.25	3.44	0.83	0.42	3226.29	−0.02	NA	NA	NA	con cien <ininteligible/
3	A	BAQ_001_03_16_A	9510	14440	4930	230.0	56.9	−2.0	8.9	19	3.85	−0.17	−1.56	−1.47	−1.88	2105.29	0.86	NA	NA	NA	tremendo de trabajo <f
4	A	BAQ_001_03_16_A	14900	24480	9580	231.2	60.0	1.8	20.0	26	2.71	−0.08	9.54	1.63	1.92	6755.29	−0.28	NA	NA	NA	lo que el profesor le pu
5	B	BAQ_001_03_16_B	2373	3089	716	235.2	59.5	16.2	29.0	2	2.79	8.99	14.35	1.24	15.44	−3175.73	−0.12	NA	NA	NA	con cien
6	B	BAQ_001_03_16_B	17370	18450	1080	170.2	59.1	9.0	9.2	2	1.85	3.39	−5.45	0.84	8.24	−2811.73	−1.06	NA	NA	NA	[no joda]
7	B	BAQ_001_03_16_B	22265	23665	1400	153.3	60.6	−6.5	15.9	5	3.57	1.58	1.25	2.34	−7.26	−2491.73	0.66	NA	NA	NA	[le dio una lección mam
8	B	BAQ_001_03_16_B	24640	24970	330	96.7	60.9	−0.7	2.8	1	3.03	−6.40	−11.85	2.64	−1.46	−3561.73	0.12	NA	NA	NA	[claro]
9	C	BAQ_001_03_16_C	20540	21630	1090	160.7	56.7	−4.2	22.8	2	1.83	−0.04	9.67	−0.37	−3.27	$0.00	−1.49	NA	NA	NA	[¿qué chévere!]

Figura 54. Ejemplo de información prosódica de los grupos entonativos

time	pitch
0.62056689342406113	397.74717921332342
0.63056689342406114	398.63901966808919
0.64056689342406115	399.08493989547208
0.65056689342406115	399.53086012285498
0.66056689342406116	399.97678035023768
0.67056689342406117	400.42270057762056
0.68056689342406118	400.86862080500345

Tabla 16. Ejemplo de sección de archivo tabulado con información de tiempo y tono

[62] Cabe recordar que en la fase de transcripción y codificación estos símbolos se habían eliminado (ver § 5.2.2.1.) priorizando un sistema de transcripción ancho. Sin embargo, gracias a Oralstats Aroca se pueden recuperar de manera automática.

time	intensity
0.62056689342406113	49.04498890536659
0.63056689342406114	52.10858066497191
0.64056689342406115	54.47310765656532
0.65056689342406115	59.53086012285498
0.66056689342406116	61.97678035023768
0.67056689342406117	63.42270057762056
0.68056689342406118	58.86862080500345

Tabla 17. Ejemplo de sección de archivo tabulado con información de tiempo e intensidad

A partir de lo expuesto anteriormente, se percibe que Oralstats Aroca realiza un amplio número de transformaciones; tanto a nivel particular para cada unidad de análisis como para la combinación entre ellas. Por ejemplo, las palabras se etiquetan morfosintácticamente con el etiquetador UDPipe (Wijffels, 2023), pero los grupos entonativos se nutren de esa información, ya que adquieren la cantidad de sustantivos, de verbos, de conjunciones, etc., de las palabras que los componen. Así, los datos permiten al investigador, si es objeto de su interés científico, saber qué grupos entonativos del corpus tienen mayor cantidad de conjunciones o de verbos y, al mismo tiempo, conocer otras características fónicas de ellos, como la media de tono, de intensidad o el rango tonal. La parte más importante de esa transformación es la siguiente:

```
prosodia_resumen <-
prosodia_q%>%group_by(id)%>%summarise(pitch_mean=mean(pitch,na.rm =
TRUE), intensity_mean= mean(intensity, na.rm = TRUE),inflexion_st =
12*log2(mean(pitch[quartil=="q4"],trim=0.1,na.rm=TRUE)/mean(pitch[quar
til=="q1"],trim=0.1,na.rm=TRUE)),range_st =
12*log2(max(pitch,na.rm=TRUE)/min(pitch,na.rm=TRUE)))%>%mutate_if(is.n
umeric,round,digits=1)
prosodiaq4 <-
prosodia_q%>%filter(quartil=="q4")%>%group_by(id)%>%summarise(pitch_me
an=mean(pitch,na.rm=TRUE),trim=0.1)
prosodiaq1 <-
prosodia_q%>%filter(quartil=="q1")%>%group_by(id)%>%summarise(pitch_me
an=mean(pitch,na.rm=TRUE),trim=0.1)

(...)

grupos_ampliados <- grupos%>%left_join(prosodia_resumen,
by="id")%>%rename(grupo_id = id)
grupos_ampliados <-
grupos_ampliados%>%left_join(intervencionesdb%>%select(intervenciones_
id,grupo_id),by="grupo_id")
grupos_pos <- tokenized_tagged%>%group_by(grupo_id = id)%>%summarise(

  qnoun = sum(upos=="NOUN",na.rm = TRUE),
  qverb = sum(upos=="VERB",na.rm = TRUE),
  qadj = sum(upos=="ADJ",na.rm = TRUE),
  qadv = sum(upos=="ADV",na.rm = TRUE),
  qadp = sum(upos=="ADP",na.rm = TRUE),
  qcconj = sum(upos=="CCONJ",na.rm = TRUE),
  qdet = sum(upos=="DET",na.rm = TRUE),
  qaux = sum(upos=="AUX",na.rm = TRUE),
  qintj = sum(upos=="INTJ",na.rm = TRUE),
  qnum = sum(upos=="NUM",na.rm = TRUE),
  qpron = sum(upos=="PRON",na.rm = TRUE),
  qpropn = sum(upos=="PROPN",na.rm = TRUE),
  qsconj = sum(upos=="SCONJ",na.rm = TRUE)
)%>%ungroup()
grupos_ampliados <- grupos_ampliados%>%left_join(grupos_pos,
by="grupo_id")

grupos_sent <- tokenized_tagged_sent%>%group_by(grupo_id =
id)%>%summarise(

  qpos = sum(sentimiento=="positivo",na.rm = TRUE),
  qneg = sum(sentimiento=="negativo",na.rm = TRUE)

)%>%ungroup()
```

Figura 55. Fragmento de Código Oralstats Aroca

El código anterior ejemplifica como Oralstats Aroca realiza anotaciones complementarias a los grupos entonativos segmentados por el investigador; se trata, por tanto, de la modificación de la tabla de grupos entonativos. Las variables que se crean en este proceso son las siguientes:

1. **pitch_mean**. Calcula la media de la variable **pitch** para cada grupo.
2. **intensity_mean**. Calcula la media de la variable **intensity** para cada grupo.
3. **inflexion_st**. Calcula una medida basada en la diferencia entre los valores medios de **pitch** en el cuartil “q4” y el cuartil “q1” para cada grupo.
4. **range_st**. Calcula una medida basada en la diferencia entre los valores máximos y mínimos de **pitch** para cada grupo.

5. Varias variables que comienzan con “q” y están relacionadas con la cantidad de categorías morfosintácticas (por ejemplo, **qnoun**, **qverb**, **qadj**, etc.); aluden a la cantidad de nombres, verbos, adjetivos, conjunciones… de cada grupo.

6. Las variables **qpos** y **qneg,** por su parte, se crean para contar la cantidad de sentimientos positivos y negativos en los datos tokenizados.

No obstante, los grupos entonativos no solo se tokenizan y anotan morfológicamente mediante el etiquetador UDPipe (Wijffels, 2023), sino que los grupos entonativos también se utilizan para crear otras entidades, ya que se emplean para elaborar la tabla de intervenciones, unidades monológicas que suponen la compilación de grupos entonativos ubicados entre pausas previamente a que otro hablante participe oralmente en la conversación.

Así pues, en resumen, la idea básica de Oralstats Aroca es completar y ampliar la información que procede de las transcripciones originales. El propio *script* genera tablas en una base de datos PostgreSQL o RSQLite, según la elección del investigador. También exporta las mismas tablas en formato RDS o CSV en las respectivas carpetas del directorio *output.* En general, las tablas creadas son: hablante, conversación, intervención, grupos y palabras. Desde la palabra hasta la conversación, todas las tablas comparten por lo menos una o dos columnas de relación; por ejemplo, una palabra tendrá siempre en la fila de su tabla correspondiente una columna en la que un dato hará referencia al grupo entonativo del que procede; lo mismo el grupo en relación con la tabla intervención y esta con el hablante. De esta manera, todas las tablas se vinculan entre sí.

5.2.3.3.2. Módulo de visualización: *script* Oralstats.Aroca

Una de las principales fortalezas de Oralstats Aroca es su capacidad para crear visualizaciones interactivas de datos. Se pueden crear gráficos, tablas dinámicas y otros tipos de visualizaciones para explorar los datos del corpus de manera efectiva. Los usuarios pueden interactuar con estas visualizaciones para obtener información detallada y descubrir patrones.

Al mismo tiempo, Oraltats Aroca permite crear cuadros de mando interactivos que contienen múltiples visualizaciones y elementos, como filtros y acciones. También admite análisis avanzados mediante la creación de cálculos personalizados, el uso de funciones

estadísticas y la integración absoluta con las posibilidades de explotación gráfica de R y su librería Ggplot2 (Wickham, 2016).

En las siguientes líneas, se comentará la parte más importante dentro de la sección *Server* del entorno Shiny, sobre todo el uso de código SQL para realizar las consultas a la base de datos, y las diferentes partes de la estructura visual del menú.

5.2.3.3.2.1. Consulta SQL

Como se ha indicado en apartados anteriores, una de las funcionalidades de utilizar R como lenguaje de programación es que permite consultar bases de datos y utilizar la información procedente de ellas como material para un visualización o transformación posterior.

Son múltiples las instancias de consulta SQL que se realizan en Oralstats Aroca, por lo que, a modo de ejemplificación, incorporamos el código más simple, es decir, aquel que se ejecute cuando el usuario, en el campo de búsqueda de la sección de visualización, introduce una palabra para ser buscada:

```
texto <- dbGetQuery(conexion1,paste0("WITH cte AS (SELECT lag(token,2)
over(partition by palabras.id order by posicion_grupo ASC) as
prev2,lag(token,1) over(partition by palabras.id order by
posicion_grupo ASC) as prev1,lag(upos,2) over(partition by palabras.id
order by posicion_grupo ASC) as prevpos2,lag(upos,1) over(partition by
palabras.id order by posicion_grupo ASC) as prevpos1,
token,lead(token,1) over(partition by palabras.id order by
posicion_grupo ASC) as post1,lead(token,2) over(partition by
palabras.id order by posicion_grupo ASC) as post2,upos,lead(upos,1)
over(partition by palabras.id order by posicion_grupo ASC) as
postpos1,lead(upos,2) over(partition by palabras.id order by
posicion_grupo ASC) as
postpos2,posicion_grupo_tag,posicion_intervencion,ulemma,alargamiento,
solap, cita,interr, grupos.content as content,
grupos.source,hablantes.sexo,hablantes.nivel,hablantes.edad,conversaci
ones.ciudad,
grupos.time_start,grupo_id,grupos.spk,grupos.intervenciones_id FROM
palabras left join grupos on palabras.id=grupo_id LEFT JOIN hablantes
on hablantes.spk = grupos.spk LEFT JOIN conversaciones on
hablantes.source = conversaciones.source) SELECT * from cte where
token = '", input$filter1,"' OR upos ='",input$filter1,"' OR ulemma
='",input$filter1,"'"))}
```

Figura 56. Código de búsqueda

En general, el código anterior combina las tablas de palabras con grupos entonativos, hablantes e intervenciones y recoge al mismo tiempo los valores de hasta un máximo de dos

palabras antes y dos palabras después de la palabra buscada. De esta manera, si se busca la palabra *parece*, puede luego completarse la búsqueda indicando palabras antes y después; por ejemplo, podría completarse con *no parece que*.

En cualquier caso, en la versión actual de Oralstats Aroca aún quedan por configurar, desde la experiencia de usuario, muchas posibles búsquedas y combinaciones. Por el momento, las posibilidades de consulta son muy amplias, pero todavía no suficientes para cubrir las opciones que disponen otros gestores de búsqueda de corpus, como los configurados por Mark Davies (Davies y Kim, 2019; Davies 2005); los ideados por Guillermo Rojo para la Real Academia Española (Rojo, 2021); o los disponibles en otros corpus de conversaciones y entrevistas de español (Barcala *et al*., 2018).

5.2.3.3.2.2. Estructura del menú

La estructura actual del menú se compone de los siguientes enlaces:

1. **Inicio.** En él aparece la entrada al entorno de Oralstats Aroca y las fotografías de las personas relacionadas con la coordinación o dirección académicas del corpus Ameresco y con su creación y mantenimiento. También se muestra el mismo mapa que aparece en la web general del proyecto en el que se proporcionan las frecuencias de palabras por país.

2. **Equipo.** De nuevo, la misma referencia que aparece en la web del proyecto, en la que aparece una lista con los coordinadores y colaboradores de los distintos equipos de investigación por ciudad.

3. **Búsqueda.** Es la parte más importante y de elaboración más compleja; en ella, hay un campo de búsqueda principal en el que se puede consultar una palabra simple o, también, un conjunto de palabras. Al mismo tiempo, hay una serie de categorías de búsqueda (ciudad, sexo, nivel…) que incluyen las características sociolingüísticas generales del corpus; existe también otro grupo de campos que hacen referencia a más aspectos de las palabras que pueden buscarse en el corpus, como cita, solapamiento, interrogación, alargamiento. En estos campos, se puede determinar si se quiere buscar la palabra o palabras deseadas en cualquier condición o en alguna de estas situaciones: por ejemplo, se pueden buscar palabras con o sin alargamiento,

en el interior o en el exterior de una cita discursiva o con o sin solapamiento; por otro lado, hay campos que marcan la distancia de palabras y categorías gramaticales en con la búsqueda realizada. Finalmente, hay dos campos que permiten buscar la palabra por su posición inicial, intermedia, final o única de grupo entonativo e intervención. Los resultados de la búsqueda se proyectan en un gráfico que puede subdividirse por categorías sociolingüísticas, si bien la configuración por defecto es la ciudad. Así mismo, los resultados pueden contabilizarse por frecuencia relativa (número de apariciones por millón de palabras).

4. **Concordancias**. Permite la consulta de las concordancias de las palabras buscadas en la sección anterior. Cuando se pulsa en una de las filas de resultados, un contexto con la referencia concreta a los coordinadores y editores del corpus aparece en la parte de abajo de la pantalla; también se puede navegar por la página y escuchar la conversación completa.

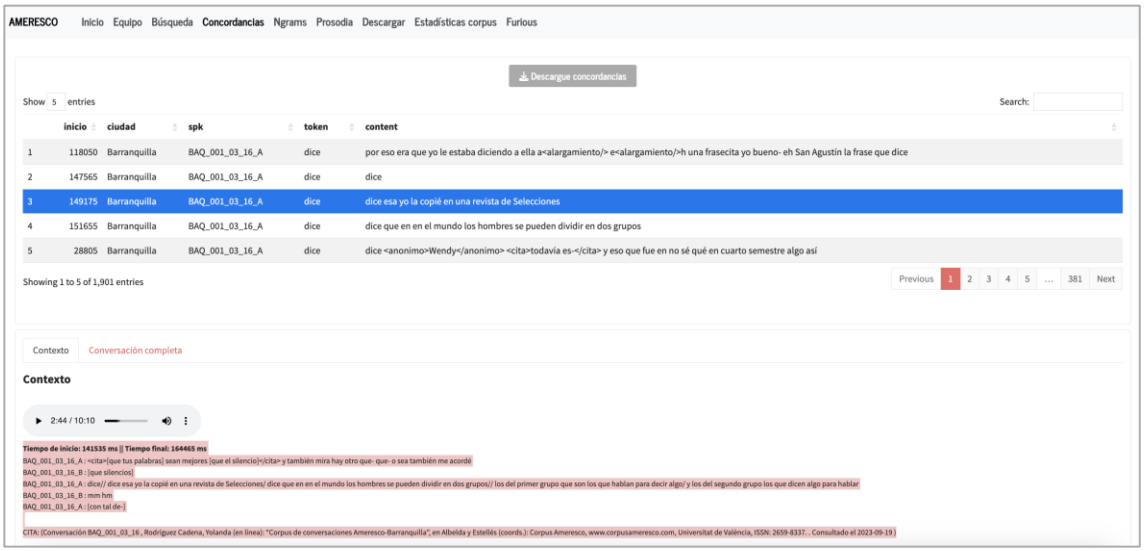


Figura 57. Obtención de resultados por concordancias

5. **N-gramas**. Se puede consultar de manera general los n-gramas (hasta una extensión de cuatro) que aparecen en el corpus. Esta consulta puede configurarse combinando los diferentes estratos sociolingüísticos (ciudad, sexo, nivel…) y los resultados se ofrecen según frecuencia absoluta o relativa.

6. **Prosodia**. Con una búsqueda similar a la anterior, se pueden consultar variables prosódicas como media de intensidad, tono medio, rango tonal, velocidad de habla o

duración a partir de gráficos *lollipop* o diagramas de caja. Por ejemplo, en los resultados de la Figura 60, puede observarse la media de tono por edad y ciudad en un conjunto agrupado de diagramas de caja. Cuando se desliza el cursor por encima de las cajas aparecen los datos concretos que muestra el gráfico en un mensaje emergente o *popup*.

Figura 58. Visualización por diagrama de caja de resultados

7. **Descargar**. En esta sección se ofrece la posibilidad de descargar los archivos de audio o ELAN y de generar, directamente desde la base de datos, un documento PDF, Word o HTML con la conversación que se decida. En todos estos casos se ha cuidado la referencia a los editores y coordinadores del corpus.

8. **Estadísticas corpus**. Como se verá más abajo, se dispone un gráfico de barras que puede explorarse por frecuencia de palabras y frecuencia de hablantes según la combinación de los factores sociolingüísticos del corpus. Así, por ejemplo, puede conocerse la cantidad de palabras emitidas por hombres y mujeres en las distintas ciudades donde se han recogido conversaciones o, también, conocer el número exacto de hablantes hombres y mujeres por ciudad.

La novena sección se corresponde a Furious, aún en fase de desarrollo, herramienta con la que se pretende explorar las conversaciones del corpus Ameresco a partir del establecimiento por parte del investigador de una serie de criterios prosódicos y/o morfológicos que, en combinación, puedan considerarse como demarcativos de estados de

emocionales negativos y que, por tanto, corresponda con la transmisión de un posible conflicto verbal[63].

5.2.3.2.3.3. Ejemplos de uso de Oralstats Aroca

Se presentan a continuación, a modo de ejemplificación, casos concretos de búsquedas por medio de Oralstats Aroca.

Búsqueda por frecuencias y estrato sociolingüístico

En la figura 59 puede observarse la pantalla de resultados por frecuencia para la búsqueda *dice*. En este caso, el diagrama de barras presenta la frecuencia relativa por ciudades, mostrando un total de 1901 palabras.

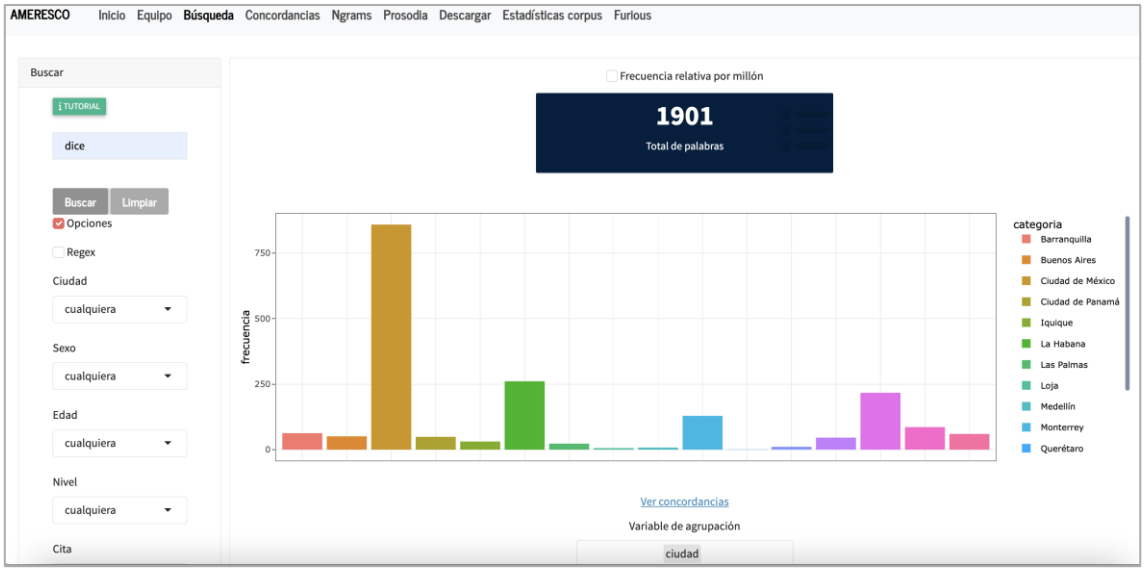


Figura 59. Búsqueda *dice* en el corpus Ameresco: resultados por frecuencia

En el gráfico de barras que aparece en la Figura 60 mostrada abajo se visualiza el resultado de la búsqueda *dice*, pero, en este caso, se arrojan los resultados según el estrato sociolingüístico *sexo*.

[63] De acuerdo con la línea de investigación del corpus ESPRINT, que recopila conversaciones conflictivas y en la que el corpus Ameresco se utiliza como corpus de control.

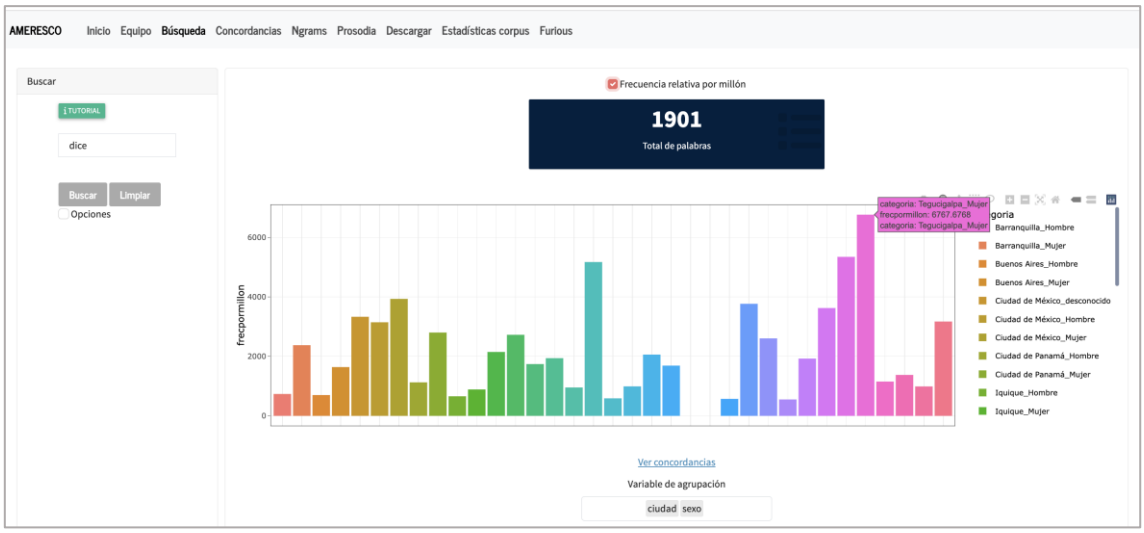


Figura 60. Búsqueda *dice* en el corpus Ameresco: resultados por estrato sociolingüístico

Búsquedas por posiciones y otras etiquetas de anotación

A continuación, la Figura 61 muestras los resultados filtrados según la combinación de patrones de búsqueda *sexo*, *distancia*, en intervenciones que contengan *cita* en cuatro ciudades concretas (Ciudad de México, La Habana, Santiago de Chile y Temuco).

Figura 61. Búsqueda por posición y otras etiquetas en el corpus Ameresco

N-gramas más frecuentes por estratos sociolingüísticos

Con respeto a los n-gramas más frecuentes del corpus Ameresco, los resultados de la búsqueda seleccionando las variables *ciudad* y *sexo* y la combinación de tres unidades (n-

grama3) arroja el resultado de *no sé qué* como n-grama más frecuente como puede verse en la Figura 62.

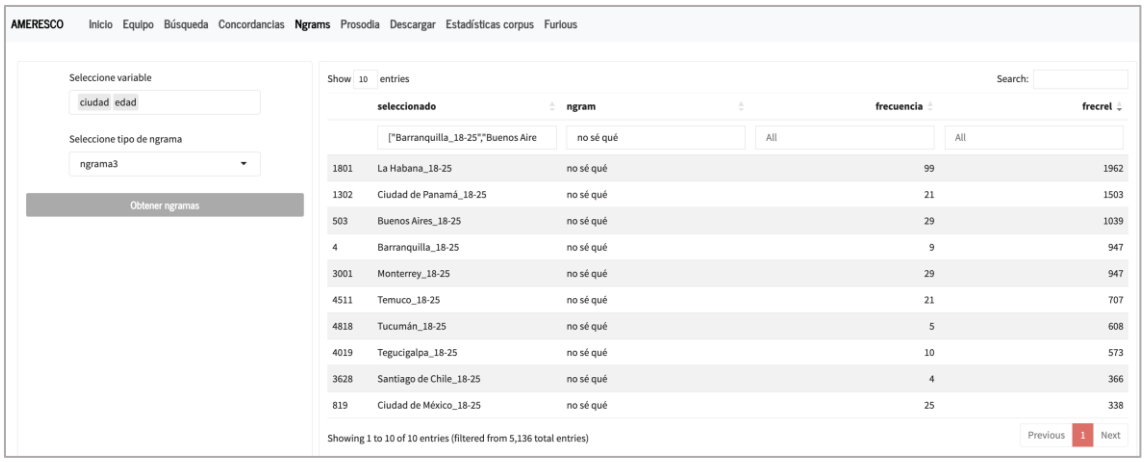


Figura 62. N-gramas más frecuentes del corpus Ameresco

Estadísticas generales como apoyo de gestión

La Figura 63 muestra las estadísticas generales del corpus Ameresco. En este caso, los resultados obtenidos ofrecen la información actualizada de los datos totales del corpus en cuanto al número de ciudades recogidas, conversaciones, grupos entonativos, intervenciones, palabras totales del corpus y número de hablantes.

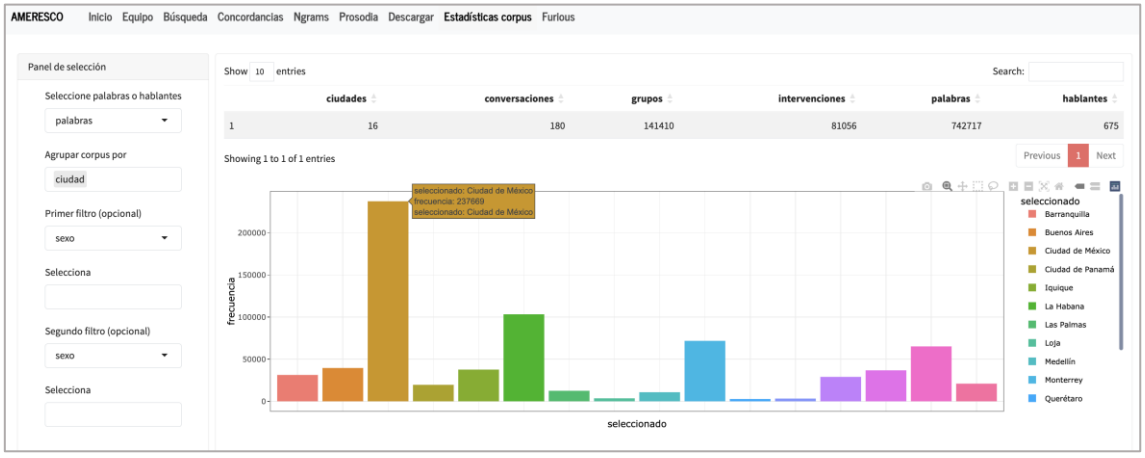


Figura 63. Estadísticas generales del corpus Ameresco

Descargar directamente desde la base de datos

Como hemos mencionado más arriba, desde Oralstats Aroca es posible la descarga de los archivos en diversos formatos (Figura 64):

Figura 64. Posibilidades de descarga del corpus Ameresco

Desde esta sección de la web se accede al audio en formato MP3 y al archivo .eaf de ELAN, pero además puede descargarse la transcripción en formato PDF, HTML o Word. En la Figura 65, a continuación, puede verse un extracto de la transcripción en formato PDF para MTY_037_03_15.

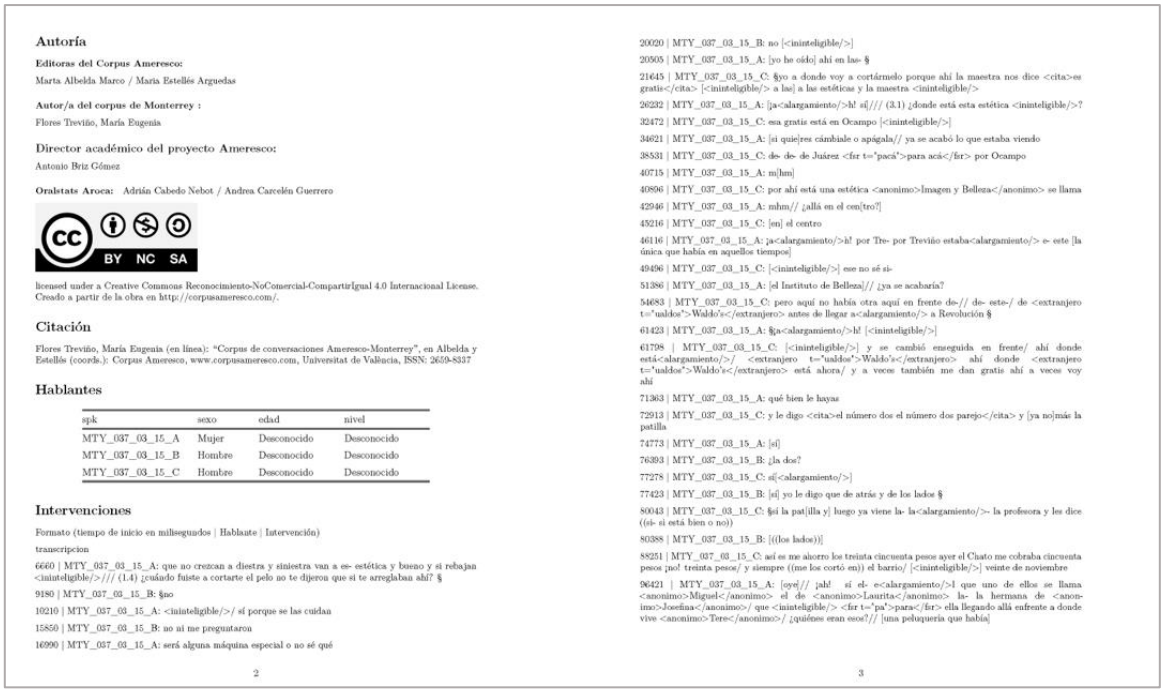


Figura 65. Ejemplo descarga en PDF en el corpus Ameresco

5.3. Dificultades y propuestas de solución del corpus Ameresco en cada una de las fases

Retomando lo señalado en el apartado 5.2. sobre la imposibilidad de anticipar los posibles problemas y dificultades que puedan surgir en cada una de las fases de diseño y construcción de un corpus oral si bien el protocolo de recogida y transcripción ha sido pensado de manera minuciosa y redactado con el mayor grado de detalle posible, hasta que no es usado como herramienta de trabajo y se ponen en práctica las indicaciones pautadas, no se empiezan a descubrir posibles fallos y necesidades de mejora en su ejecución. Una parte importante del trabajo en este sentido la ha constituido la comunicación con los equipos locales, y su formación, además de la propia experiencia desde el equipo central. Esta comunicación no solo es necesaria desde el punto de vista del correcto funcionamiento del proceso y del cumplimiento del protocolo, sino que cada una de las interacciones supone también una oportunidad de mejora del corpus.

Así, por un lado, las consultas revelan la existencia de espacios de mejora en la redacción de los protocolos de recogida y tratamiento de los datos (fases 1 y 2), sea por la necesidad de reformularlos de manera más clara o de incluir nueva casuística que ejemplifique situaciones especialmente problemáticas; por otro, muchas consultas vienen acompañadas de problemas que ponen de manifiesto la necesidad de repensar el sistema o de dar espacio a las peculiaridades de una comunidad lingüística concreta que no quedan reflejadas en las etiquetas, parámetros sociolingüísticos, etc., planteados inicialmente.

De igual modo, para la fase 3, de archivo de los datos y acceso a ellos a través de la página electrónica, se ha visto la necesidad de implementar mejoras o realizar cambios en la interfaz de las diferentes versiones para dar respuesta a las necesidades planteadas por investigadores usuarios del corpus, pero también para satisfacer las necesidades que han ido surgiendo en los propios miembros del grupo de investigación. Además, estas modificaciones han afectado a la manera de almacenar los datos en línea y a su consiguiente visualización y descarga, como veremos más adelante.

5.3.1. Dificultades en la fase 1. Concepción y recogida de los datos

Se contemplan en esta sección las dificultades relacionadas por un lado con la gestión de los equipos de trabajo y, por otro, las concernientes a la recogida de los datos, esto es, obtención de las grabaciones, los permisos y las fichas técnicas.

Con respecto a la **gestión de los equipos de trabajo**, como se mencionó al inicio de este capítulo, la construcción del corpus es posible gracias a la inestimable colaboración de equipos locales ubicados en diferentes ciudades de Hispanoamérica y España. En este sentido, el primer obstáculo es, precisamente, el de la distancia geográfica (y también, en consecuencia, temporal) a la hora de formar a los equipos en el protocolo de recogida. Estas formaciones suponen un gran coste en tiempo ya que, deben ser preparadas anticipadamente y se debe acordar una fecha concreta con cada uno de los equipos, teniendo en cuenta la diferencia horaria. Además, la formación necesariamente se tiene que realizar en varias sesiones ya que contiene los siguientes módulos formativos[64] imposibles de abarcar en una sola reunión:

- Recogida de las grabaciones: entornos ideales de grabación, obtención de muestras que se ajusten a los objetivos del corpus.
- Protocolo de protección de datos y recogida de autorizaciones.
- Transcripción de las muestras siguiendo las convenciones fijadas por el proyecto, bien sea a través del modo de trabajo 1 o del modo de trabajo 2.
- Entrega de materiales al equipo central.

A este hecho debe sumarse la circunstancia del cambio de personal en los equipos, en primer lugar, porque en un alto porcentaje las personas encargadas de recoger las grabaciones son estudiantes universitarios, y estos no pueden colaborar de manera estable durante todo el tiempo que dura la recogida. Por lo tanto, en cada curso académico habitualmente participan colaboradores nuevos con los que se debe repetir la formación. Aquellos grupos que cuentan con personal investigador fijo han podido desarrollar esta tarea con mayor estabilidad y con menos necesidad de repetir la formación.

Además, no se pueden dejar de señalar las condiciones excepcionales sucedidas durante los años 2020 y 2021 relacionadas con la pandemia mundial del coronavirus. Como es bien sabido, en este periodo se sufrieron las consecuencias de esta enfermedad que llevó a que los diferentes gobiernos decretaran, entre otras medidas, la orden de confinamiento en los domicilios durante diferentes periodos de tiempo según cada país; y una vez levantada esta orden, se aplicaron otras restricciones de movilidad y de distancia social que impedían igualmente reanudar las tareas de recogida. Estas circunstancias que afectaron sin duda a

[64] Las formaciones impartidas se acompañan de material de apoyo para los grupos, en concreto, el protocolo de trabajo para los equipos (Briz *et al*., 2019) y un tutorial de ELAN (Uclés, 2020), ambos disponibles en línea en la página web del corpus.

todo proceso de investigación en general fueron especialmente duras para la elaboración del corpus Ameresco, dada la imposibilidad de reunirse presencialmente en contextos vivenciales de proximidad, (sin contar, además, con otras consecuencias que afectaron al estado de salud de las diferentes personas que integraban los equipos en esos momentos).

Por otro lado, como se apuntaba en la sección 5.2.2.1., no todos los equipos contaban con las condiciones ideales para realizar la recogida del corpus en todas las fases pactadas (recogida y transcripción), bien por falta de personal, bien por falta de medios para acceder a estas herramientas. Esta circunstancia llevó a que inicialmente se implantaran los dos modelos de trabajo comentados en el Capítulo 4. Este hecho dio lugar a la aparición de nuevos obstáculos que hubo que superar desde el equipo central como, por ejemplo, los problemas asociados a la transcripción por parte de personal español de una variedad dialectal diferente a la materna y las implicaciones que esto conlleva, como se explicará más adelante.

En determinadas circunstancias, bajo la recepción de subvenciones de la Universitat de València por medio del programa Convenio Marco, o mediante los fondos de los proyectos de investigación MINECO y MICINN, ha sido posible realizar formaciones *in situ* tanto por estancias de la coordinadora técnica en algunas de las universidades colaboradoras como por viajes de los coordinadores locales a València.

Respecto a los cronogramas de recogida y entrega de los materiales por parte de los equipos locales al equipo central, estos no pueden ser inamovibles. Si bien, se pacta un calendario de actuación ajustado a las posibilidades de cada equipo, este no siempre se puede seguir precisamente por la necesidad de cumplir con otras obligaciones laborales y de investigación de cada uno de los equipos. En el protocolo de trabajo (Briz *et al*., 2019) se establece un ejemplo de cronograma en el que se prevé una entrega en dos partes de todo el material, una primera entrega de doce conversaciones transcritas, y una segunda entrega de las quince restantes, no obstante, pueden establecerse otras entregas según las circunstancias de cada equipo hasta completar la muestra establecida (desarrollada en § 5.2.1.1.). En la Figura 66 de abajo, puede verse un ejemplo de la previsión de entrega de materiales en la primera fase del corpus Ameresco-Tegucigalpa en la que se detalla el envío de las diez primeras conversaciones y transcripciones recogidas.

CALENDARIO DE ENTREGA CONVERSACIONES AMERESCO-TEGUCIGALPA

Por: Danny Fernando Murillo Lanza

Código de conversación	Fecha de entrega	Duración total	Tiempo de transcripción
TGU_001_03_19	06 – 12 de abril de 2020	28:05	20:00
TGU_007_03_19	13 - 19 de abril de 2020	19:07	19:07
TGU_009_04_19	20 – 26 de abril de 2020	19:32	19:32
TGU_010_04_19	27 de abril al 03 de mayo de 2020	21:13	21:13
TGU_011_03_19	04 al 10 de mayo de 2020	22:05	22:05
TGU_012_04_19	11 al 17 de mayo de 2020	21:37	21:37
TGU_015_02_19	18 al 24 de mayo de 2020	24:34	20:00
TGU_016_02_19	25 al 31 de mayo de 2020	20:55	20:55
TGU_017_04_19	01 al 7 de junio de 2020	30:53	20:00
TGU_003_03_19	8 al 21 de junio de 2020	50:03	40:00

Figura 66. Cronograma de entrega Ameresco-Tegucigalpa

La siguiente dificultad está relacionada con la selección de la muestra de hablantes. Desde el equipo central (§ 5.2.1.1.) se proporcionan unos criterios para la selección de hablantes, estratificados sociolingüísticamente en *sexo*, *grupo etario* y *nivel sociocultural*. Esta muestra opera sobre los criterios mínimos de recogida de 72 hablantes en 27 conversaciones, cifra con la que hipotéticamente se ocupaban todas las casillas y se obtenía una muestra representativa. Sin embargo, como se adelantó en la sección 5.2.1.1., según se han ido recogiendo los materiales se ha comprobado que se necesitan muchas más conversaciones para obtener ese número de hablantes. Las propias características de la conversación coloquial y la necesidad de grabar de manera secreta condicionan la obtención de las grabaciones que, para este género discursivo, son menos controladas que para un género como por ejemplo la entrevista.

Con respecto a la muestra, se ha registrado otra problemática, aún pendiente de solucionar, también apuntada en la sección 5.2.1.1. Se trata, precisamente, de la adecuación de los criterios sociolingüísticos con respecto a las características sociales de hoy en día y, además, de las circunstancias particulares de cada país, en muchos casos con particularidades que hacen que no encajen bien en esta estratificación.

En primer lugar, con respecto a la variable *sexo*, se contempla la opción masculino o femenino, si bien, en la actualidad existen otros modelos de identidad de género que no quedan aquí recogidos y que cuya integración en esta estratificación cabría plantearse. En segundo lugar, los grupos etarios con los que trabaja Ameresco, proceden de los establecidos originariamente por Briz y el Grupo Val.Es.Co. (2002) quienes delimitan los grupos de 18-

25 años, 26 a 55 y >55 esto es un primer grupo para jóvenes, un segundo grupo para adultos y un tercer grupo para adultos mayores. Sin embargo, se ajustan más a la realidad actual la estratificación marcada por PRESEEA, quienes distinguen tres generaciones: de 20 a 34 años, de 35 a 54 años y de 55 años en adelante, división establecida "a la vista de lo que se ha decidido en otras investigaciones del mundo hispánico y con un deseo de primar la simplicidad sobre la casuística (Moreno Fernández, 2021a, p. 14).

En cuanto al *nivel sociocultural*, establecer estas franjas de acuerdo con las diferentes etapas de escolarización como sucede en Ameresco, es menos objetivo que fijarlo según los años de escolarización como hace PRESEEA (Moreno Fernández, 2021a, p. 15), ya que la correspondencia para con los sistemas educativos de cada uno de los países integrantes del corpus puede no ser equivalente.

Asociado a la recogida de la muestra se hace necesario solventar otro problema como es la sobrerrepresentación de algunos estratos, frente a la escasez o imposibilidad de acceder a otros. Por ejemplo, al menos en las ciudades más desarrolladas, las opciones de acceder a población joven analfabeta o con estudios primarios se reducen bastante, ya que existe un periodo de escolarización obligatorio, aunque este sea diferente en cada país (15 años de escolarización obligatoria en el caso de México, 13 para Chile y once años en Panamá,[65] por ejemplo). Por otro lado, cuando los encargados de recoger las grabaciones son estudiantes, estos accederán fácilmente a los estratos de edad más cercanos a su entorno y experiencia, esto es el joven y el adulto, mientras que los adultos mayores son más difíciles de alcanzar.

Tomando conciencia de estas limitaciones, en la medida de lo posible, se ha tratado de adaptar la ficha técnica para que recoja la mayor cantidad posible de información con valor numérico; por ejemplo, en 2019 desde el grupo Val.Es.Co. se modificó la redacción para solicitar la edad exacta de cada hablante de cara a una posible reagrupación en el futuro. Este nuevo modelo de ficha técnica se adoptó en el corpus Ameresco. Esta posibilidad de reagrupación queda abierta con la naturaleza cuantitativa de los datos recogidos, y así dependerán de la sociología y la sociolingüística la propuesta de nuevos modelos o franjas de estratificación. En cuanto al déficit o exceso de hablantes, por el momento no se ha encontrado una solución a este problema. Se anima a los grupos locales a que recojan el mayor número de hablantes posibles de los pactados en la muestra, y aquellos huecos que

[65] Datos recogidos por la UNESCO, *Sistema de Información de Tendencias Educativas en América Latina* https://siteal.iiep.unesco.org/

no se han podido rellenar quedan a la espera mientras no se den unas circunstancias propicias para acceder a ellos.

A continuación, se desarrollan aquellas dificultades relacionadas con la **recogida de los datos** propiamente dicha, esto es, problemas asociados a aspectos técnicos en la recogida de las grabaciones, de las fichas técnicas y los consentimientos informados, así como a otros problemas encontrados en las grabaciones y que las invalidaba para ser incorporadas al corpus; por último, la entrega del material al equipo central también ha debido solventar algunos escollos.

Los principales problemas técnicos a los que deben enfrentarse las personas, encargadas de recoger las grabaciones, esbozados en la sección 5.2.1.3., tienen que ver con encontrar unas condiciones óptimas para la recogida del audio. Esto es, se necesita una ubicación adecuada, a ser posible en interiores, para evitar la aparición de ruidos de fondo que entorpezcan la inteligibilidad del audio y la posterior tarea de transcripción. Como se apuntaba más arriba, si la grabación se realiza en exteriores, es esperable que aparezcan ruidos no controlables como contaminación acústica procedente del tráfico, sirenas, animales, murmullos de fondo, etc., de modo que se les sugiere, si no evitar la grabación en exteriores (puesto que podría provocar la no aparición de determinados actos de habla −por ejemplo, algunos actos directivos−, léxico común relacionado con el tráfico, la ciudad, etc.), al menos sí restringir la grabación en exteriores a los casos en los que el ruido no interfiera de manera significativa. En este sentido, indudablemente la grabación en espacios interiores permite al investigador un mayor control de los factores ambientales, sin embargo, la retroalimentación de los grupos no ha permitido observar otros problemas relacionados con este aspecto en estas conversaciones.

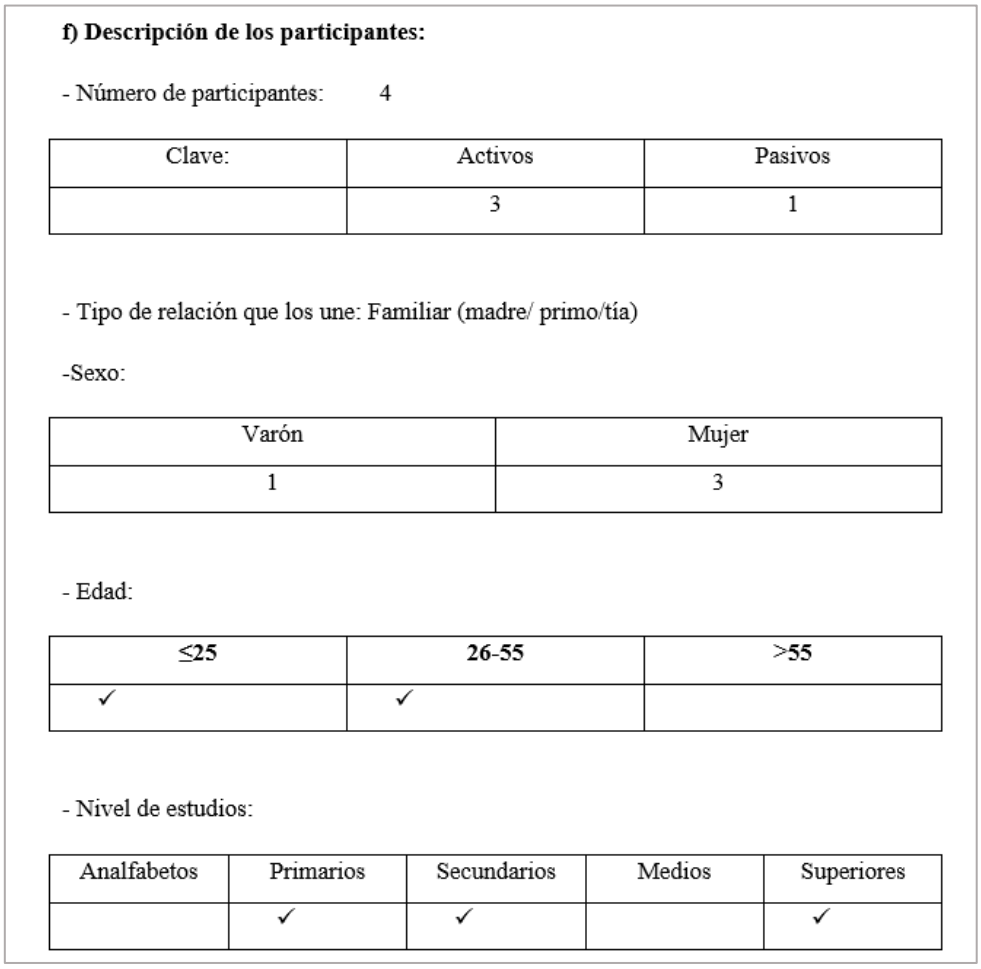
En los espacios interiores privados, dada una de las características de la conversación coloquial prototípica (relación de cercanía y familiaridad), con frecuencia el investigador debe salvar otros obstáculos que entorpecen la calidad del audio: ruidos de platos mientras comen, televisiones de fondo, música, visualización de vídeos en el teléfono móvil, entre otros. La recomendación general en esta circunstancia es que no se traten de evitar estos ruidos, en la medida en que hacerlo podría resultar un comportamiento sospechoso y restar naturalidad o hacer intuir a los sujetos de la grabación en curso; en cambio, sí se recomienda descartar la grabación en aquellos casos en que los ruidos no son puntuales, sino continuos

(ruidos de máquinas de coser, animales domésticos que ladran, pían, etc. de manera constante, música de fondo, etc.). No obstante, si el material lingüístico producido en esos casos es valioso desde un punto de vista no fónico, las conversaciones rechazadas para el corpus por este motivo pueden incluirse como parte del repositorio, puesto que son una fuente interesante de datos para investigadores de campos en los que únicamente se considere la transcripción.

Por otro lado, la elección de espacios interiores públicos como cafeterías, restaurantes, etc., hace que los factores contextuales tampoco puedan ser controlados. Además, a los ya mencionados ruidos, música, etc., se suma la presencia frecuente de hablantes ajenos a la grabación que mantienen conversaciones de fondo y que, en ocasiones, interactúan con las personas grabadas, de modo que se crea una situación en la que se graba a personas que no han firmado el consentimiento informado y que podrían suponer un problema legal. Las intervenciones de estos sujetos se transcriben y anonimizan, pero no se computan sociolingüísticamente ni se consideran sujetos de la investigación.

Respecto a las condiciones técnicas de la grabadora, una buena ubicación de esta es fundamental para este tipo de recogidas, ya que al no ser planeadas y tener que ser grabadas secretamente, no permite que cada hablante cuente con su propio micrófono. Al mismo tiempo, la disponibilidad de medios técnicos es muy diversa dependiendo del equipo local. Algunos equipos han grabado con teléfonos inteligentes, que en sus versiones más recientes proporcionan una excelente calidad de grabación y reproducción de audio y son, por tanto, aparatos muy adecuados para la recogida de conversaciones; no obstante, se podían utilizar otros dispositivos siempre y cuando no estuvieran a la vista de los hablantes y se asegurara una calidad óptima de grabación. El mayor problema derivado del uso de *smartphones,* especialmente en las primeras fases de recogida del corpus en la que estos dispositivos eran menos avanzados, era el hecho de la aparición en la grabación de diferentes notificaciones propias del teléfono, si bien hoy día la mayoría de los modelos silencian las notificaciones por defecto cuando se está usando la grabadora. Por otro lado, dada la impredecibilidad de movimiento de los participantes que no saben que están siendo grabados, en ocasiones el dispositivo de grabación no ha podido recoger bien todas las intervenciones cuando los hablantes se alejaban o ha recogido grabaciones con ruidos muy predominantes cuando estos, por ejemplo, jugaban a algún juego de mesa o abrían envoltorios de comida.

Tras la recogida de las conversaciones, las personas encargadas de la grabación deben recoger los metadatos referentes a los hablantes y el contexto. En las primeras fases de la recogida, el modelo de ficha técnica utilizada era la del corpus Val.Es.Co. (2002), sin embargo, se detectaron errores en su cumplimentado, por un lado, porque notamos que algunos de las informaciones no habían sido explicadas de manera clara y, por tanto, los equipos locales manifestaron que no sabían exactamente qué datos incluir, lo que provocó que algunas de estas fichas tuvieran errores; por otro, se daba también cierta confusión en la manera de rellenarlas, especialmente los datos referidos a los hablantes, como se ejemplifica en la Figura 67. La localización de estos errores, también en los trabajos del corpus Val.Es.Co. 2.0 y 3.0, llevó a la modificación de la ficha técnica a una apariencia más amable y, por tanto, menos propensa a errores.



f) Descripción de los participantes:

- Número de participantes: 4

Clave:	Activos	Pasivos
	3	1

- Tipo de relación que los une: Familiar (madre/ primo/tía)

-Sexo:

Varón	Mujer
1	3

- Edad:

≤25	**26-55**	**>55**
✓	✓	

- Nivel de estudios:

Analfabetos	Primarios	Secundarios	Medios	Superiores
	✓	✓		✓

Figura 67. Ficha técnica antigua para la recogida del corpus con errores

Si nos fijamos en la figura de arriba, se puede comprobar que no se han identificado correctamente a los hablantes. En el protocolo se indica que deben registrarse como A, B, C,

D… porque deben estar anonimizados. Sin esta identificación codificada es imposible asociar las intervenciones con el hablante que las emitió para posteriormente poder filtrar las búsquedas en la página de consulta. Con el cambio en la estructura de la ficha que se observa en la Figura 68 a continuación, se minimizó el riesgo de error, y, además, se aprovechó para registrar la edad exacta de cada participante de cara a una futura reestructuración de la estratificación como hemos comentado arriba.

Clave hablante	Sexo (V/M)	Edad (especificar edad exacta dentro de una de las tres franjas etarias)			Nivel de instrucción			Activo/Pasivo	Monolingüe cast./Bilingüe	Profesión	Residencia habitual (municipio)
		18-25	26-55	≥56	Bajo	Medio	Alto				
A	M		32				x	Activo	Monolingüe	Contadora	Quilmes
B	V		27				x	Activo	Monolingüe	Docente	Quilmes
C	V		33				x	Activo	Monolingüe	Bancario	Quilmes
D	M		30				x	Activo	Monolingüe	Docente	Quilmes

Figura 68. Ficha técnica modificada para la recogida del corpus correctamente cumplimentada

Ha habido casos en los que no ha sido posible la subsanación de estos errores ya que, entre el momento en que el equipo local entrega el material y la revisión del equipo central pueden pasar varios meses. Además, como se señaló en § 4.2.1.2., debido a cuestiones legales, el tratamiento de los datos impide que el investigador pueda invertir el proceso de obtención de los materiales y recontactar con los sujetos, ya que se estaría violando la ley de protección de datos, por tanto, es fundamental que se recojan de manera correcta.

Así mismo, en el momento de terminar la grabación, junto con la recogida de la ficha técnica, debe completarse la firma del consentimiento informado, tanto en la confirmación del permiso como en el apartado de tratamiento de los datos. Volviendo sobre lo dicho en la 5.2.1.2., recuérdese que esta autorización consta de tres fases y que las personas que van a realizar la grabación han debido recopilar en un momento anterior en el tiempo, la primera firma de los hablantes en las que se les informa de que serán grabados en un futuro sin que ellos sepan el momento exacto.

En cuanto a la gestión por parte de los equipos locales en este sentido no ha presentado ninguna dificultad; sin embargo, por parte del equipo central fue necesaria la modificación de los términos del consentimiento para ajustarlo a las nuevas políticas de protección de

datos surgidas con la publicación de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales, como se comentó en la sección 5.2.1.2. Hasta esa fecha, se solicitaba el permiso de los participantes únicamente de manera posterior a la grabación y no se detallaba la manera en que se iban a tratar sus datos y ni ante qué organismo podían ejercer sus derechos sobre los mismos. Con esta nueva legislación, por tanto, hubo de redactarse un nuevo modelo de consentimiento informado que pasó por la supervisión del departamento jurídico y de protección de datos, así como del Comité de Ética de la Universitat de València.

Con respecto al envío de los materiales al equipo central, las dificultades han estado relacionadas principalmente con el método de envío. En el protocolo inicial de trabajo se pautó esta entrega en base a los materiales que deben enviarse y los formatos en que deben estar, esto es, debía enviarse el audio (MP3 o WAV, si bien se aceptan en otros formatos, puesto que desde el equipo central se realiza la conversión), la transcripción (en formato .doc o .eaf, según se haya trabajado con el modo de trabajo 1 o 2), la ficha técnica (en .doc o .pdf) y las autorizaciones (aunque algunos equipos han sido los responsables de custodiar ellos mismos los consentimientos como se señaló en el apartado 5.2.1.3.2.). Se daba por hecho que estos materiales se agruparían y se enviarían por bloques, pero cada equipo ha ido enviándolos de una manera no homogénea ni sistemática. Por este motivo, se decidió optar por una manera unificada de envío a través de plataformas de almacenamiento en la nube de acceso restringido. En el caso particular de La Habana (Cuba), dados los impedimentos técnicos para poder hacer la entrega de la manera pautada con anterioridad, se aprovecharon varios encuentros cara a cara para hacer el intercambio; un primer momento aprovechando la celebración del Congreso Internacional *Enseñanza de la Lengua Española-Gramática, escritura y oralidad* celebrado en esta ciudad los días 10, 11 y 12 de enero de 2017, y un segundo momento con la estancia de investigación realizada por la coordinadora técnica del equipo Ameresco-La Habana, en la Universitat de València.

Por último, en esta fase 1 de recogida de materiales, debe señalarse que no todos los materiales recogidos por los equipos locales han cumplido con las condiciones necesarias para su incorporación al corpus. La entrega de materiales al equipo central se pautó de forma que hubiera una primera valoración sobre la validez de los archivos, así, el equipo local envía al equipo central una cala de grabaciones con el objetivo de confirmar su idoneidad para el corpus, esto es, que cumplía con las características necesarias para su posterior uso (que habían sido grabadas de manera secreta, que la calidad del audio era lo suficientemente

óptima, que el género discursivo era correcto, etc.). Después de pasar por este primer filtro y primeras valoraciones, el equipo local procedía a recoger la muestra completa de conversaciones atendiendo a las observaciones que se le habían realizado desde el equipo central. No obstante, a pesar de esta primera validación, se han recibido grabaciones que no cumplían con los parámetros establecidos.

Uno de los motivos principales del descarte de grabaciones ha sido la confusión con el género discursivo. En el protocolo de trabajo se explica que el corpus va a recoger conversación coloquial prototípica (Briz, 1998 [2001]), sin embargo, ha habido casos en los que la grabación contenía otros géneros como entrevista o conversación semidirigida que, si bien podrían considerarse conversación coloquial periférica, no son objeto de estudio *a priori*; también se han encontrado casos de conversaciones que no se grabaron de forma secreta.

Además, se han observado problemas relacionados con una mala organización de la situación comunicativa para la conversación y su grabación por parte de la persona encargada de recogerla. Como se vio en el apartado 5.2.1.3.1. ha habido casos en los que aparecen dos hablantes y uno de ellos es la persona encargada de grabar. Este contexto de grabación afecta de varias maneras e invalida el material recogido. Por un lado, porque el grado de espontaneidad y secreto está comprometido al 50 %, por tanto, se podría poner en duda el cumplimiento de los requisitos de este género. Por otro, estas situaciones derivaban en una distorsión del género discursivo ya que, la persona encargada de la grabación tiende a asumir el rol de entrevistador y fuerza la interacción comunicativa para conseguir la participación del otro hablante; por tanto, en lugar de conversación espontánea lo que se está recogiendo es conversación semidirigida, cuasi entrevista. De manera paralela, sucedía que la persona encargada de la grabación, por no participar y condicionar la muestra, optaba por no participar apenas; por lo tanto, lo que se espera que sea conversación espontánea acababa convertida en monólogo del hablante que no sabe que está siendo grabado. En consecuencia, no hay interacción ni se cumple con requisitos del género esperado.

Ejemplo 39. Grabación descartada por ser una entrevista

https://nuvol.uv.es/owncloud/index.php/s/dVBJakY02ExQ4ay

En este ejemplo, la persona encargada de recoger la conversación está tratando de grabar a otra persona. Sin embargo, la hablante que no sabe que está siendo grabada no se presta a conversar y finalmente es el hablante que inicia la grabación el que toma el turno la mayor parte del tiempo y opta por guiar la conversación. Por tanto, no hay interacción y la grabación no es válida para el corpus.

Ejemplo 40. Grabación descartada por ser un monólogo

https://nuvol.uv.es/owncloud/index.php/s/1LAhEEyhhpRZjex

En este otro ejemplo, la persona encargada de grabar asume el rol de entrevistador y la persona que no sabe que está siendo grabada acapara el turno, por lo que la conversación se convierte en monólogo.

Ambos casos, Ejemplo 39 y 40, describen los problemas descritos en cuanto al número de participantes. En estos ejemplos, solo aparecen dos personas, una de ellas la encargada de grabar, y como se ha podido escuchar, no se da conversación coloquial prototípica. Por este motivo, en las formaciones y el protocolo de trabajo, se ha hecho especial hincapié en las debilidades encontradas en este tipo de grabaciones con dos personas y se ha propuesto que, como se vio en 5.2.1.3.1., el investigador deje la grabadora y salga de escena o permanezca como hablante pasivo y limite todo lo posible su participación, descartándose aquellas grabaciones en las que solo haya dos hablantes y uno de ellos es la persona encargada de grabar.

Como se puede ver a continuación en la Figura 69, muchas de las grabaciones descartadas en esta fase de validación se eliminaron por otros motivos. En unos casos, si bien las grabaciones parecían cumplir con las características del género, la grabación no ha sido realizada de forma secreta, ya que los propios hablantes participantes hacen alusión al hecho

de estar siendo grabados. A este respecto, se ha seguido insistiendo encarecidamente a los equipos locales en que esto no suceda ya que conllevaría la distorsión de un análisis posterior.

Envío archivos prueba julio 22

ID Equipo local	Duración	Aceptada	Observaciones
Grabación 1	31' 49"	No	No es secreta, 1' 25" nos dices cuándo
Grabación 2	34' 25"	No	No es secreta pq hace alusión al tiempo en minuto 5.50-6.00, Pájaro de fondo continuamente, hay poca interacción
Grabación 3	29' 17"	No	No es secreta 27' 22" "*hay q borrar eso* ". Música de fondo hasta el minuto 9, a partir del 13 vuelve, ruido de tráfico
Grabación 4	28' 46"	Sí	A partir de 15' empiezan a ver vídeos de Tik Tok, habría que cortar desde ahí y transcribir solo los 15' primeros
Grabación 5	27' 44"	No	No parece secreta. Al principio, una graba y el otro no sabe, es conversación dirigida
Grabación 6	29' 02"	No	Son niñas jugando
Grabación 7	29' 45"	No	Es más un monólogo/entrevista semidirigida, dos hablantes, el que graba y la señora
Grabación 9	30' 48"	No	Calidad de audio regular, los hablantes están lejos y hay un ruido de fondo como de lavavajillas/lavadora

Figura 69. Captura de un registro de validez de los archivos enviados al equipo central

Ejemplo 41. Grabación descartada por no ser secreta

https://nuvol.uv.es/owncloud/index.php/s/WpYQriIkbHqFj5c

En el ejemplo anterior, en un principio parece que estamos ante una conversación prototípica, sin embargo, tras un minuto de grabación los hablantes preguntan cuándo a empezar la grabación. Ante esta pregunta, la persona encargada de grabar les explica el procedimiento de la grabación. Por tanto, tuvo que descartarse la grabación.

Otro motivo para desechar una grabación, aunque cumpla con los requisitos del género discursivo, es la mala calidad del audio, bien por motivos técnicos a la hora de realizar la grabación (posición de la grabadora lejos de los hablantes, sujetar la grabadora con la mano mientras está encendida, uso de un dispositivo de baja calidad), bien por factores contextuales a la grabación: ruido excesivo de fondo, música, televisión encendida, que se estén realizando tareas domésticas o comiendo, lo que implica ruido de platos y cubiertos; movimiento de muebles, grabación obtenida en lugares públicos, ruido de clientela, ruido de tráfico, etc. Estas circunstancias, como muestra el ejemplo siguiente, dificultan enormemente la realización de una transcripción con un grado de fidelidad alto.

Ejemplo 42. Grabación descartada por ruido de máquina de coser

https://nuvol.uv.es/owncloud/index.php/s/2H4NilvVPTBXH10

En el Ejemplo 42 se ha grabado la conversación mientras una de las participantes cose con la máquina de coser. El ruido constante y fuerte de esta máquina hace imposible una transcripción fiable.

Ejemplo 43. Grabación descartada por ruido de TV y ausencia de conversación

https://nuvol.uv.es/owncloud/index.php/s/VO6ULBZZLmNuVHj

En este último caso, el Ejemplo 43, la hablante que inicia es la persona encargada de realizar la grabación. El otro hablante está en casa viendo la televisión que, además de ser un ruido de fondo constante, impide el desarrollo de la conversación. Por tanto, nos encontramos con un doble motivo para descartar esta grabación: el ruido y la ausencia de conversación.

Estas grabaciones descartadas si −aun no siendo válidas para el corpus− eran enviadas con todo el material asociado (transcripción, ficha técnica y autorización), son aprovechadas para crear un repositorio de conversaciones no prototípicas que pueden constituir un material de estudio y análisis válido para otras finalidades.

5.3.2. Dificultades en la fase 2. Tratamiento de los datos

En este apartado se atiende a la fase 2, de tratamiento de los datos. Por un lado, se detallan las dificultades relativas a las convenciones de transcripción y codificación, obtenidas a través de las dudas planteadas tanto por los equipos locales, como por el personal externo

asociado[66] al equipo central; por otro, a los obstáculos surgidos en la revisión de las transcripciones por parte del equipo central. Estas dudas han sido planteadas al equipo central por dos medios principales: a través de la comunicación directa con la coordinadora técnica (correo electrónico, mensajería instantánea, talleres y reuniones presenciales o virtuales) y, en algunos casos, por medio de un archivo de dudas compartido entre el equipo local y central en el que se iban volcando las consultas para que fueran resueltas desde el equipo central y quedaran archivadas en la nube.

Con respecto a las dudas surgidas sobre la transcripción y codificación, podrían categorizarse en torno a los siguientes ejes: (a) transcripción de fenómenos de diversa naturaleza; (b) empleo de marcas y etiquetas en la transcripción; (c) anonimización.

(a) Transcripción de fenómenos de diversa naturaleza

Se recogen aquí las dudas y dificultades surgidas en torno a la transcripción general de fenómenos de diversa naturaleza, esto es, consultas sobre la manera de representar particularidades de cada variedad dialectal (seseo, ceceo, aspiración generalizada, etc.), la reproducción del estilo directo, palabras o fragmentos en otra lengua, la representación de elementos funcionales, etc. y qué grado de exhaustividad es necesario a la hora de reflejar otras cuestiones como, por ejemplo, los ruidos.

Respecto de las **variedades dialectales**, esta consulta tiene su origen sobre todo en el personal contratado por el equipo central en Valencia. A la hora de enfrentarse a la transcripción de variedades dialectales diferentes a la propia, no conocían de qué manera debían hacerlo; así, han sido recurrentes las consultas sobre si debían reflejarse y en qué forma fenómenos como el seseo, el ceceo, la aspiración de la /s/, los cambios en el paradigma verbal que suceden en variedades como la argentina (cantas>cantás) o la chilena (cantas>cantái).

Con respecto a los **fenómenos propios del habla**, como el seseo, el ceceo o la aspiración, se ha optado por tomar la decisión operativa de no reflejarlos cuando estos son una característica propia de la variedad dialectal que se está transcribiendo ya que, de hacerlo, conllevaría una mayor inversión de tiempo y, lo que es más importante, generaría conflicto

[66] Con *personal externo asociado al equipo central* nos referimos a los transcriptores contratados con cargo al proyecto para agilizar las tareas de procesamiento de los datos (transcripción si era necesaria y alineado de texto y audio de los materiales procedentes de equipos que optaron por el modo de trabajo 1).

en el motor de búsqueda que no solo deberá buscar en base a la forma normativa (§ 5.2.2.1.2.), sino que también, debería incluirse una marca o etiqueta para todas las realizaciones encontradas. Por tanto, el esfuerzo de procesado informático sería mayor y ralentizaría la extracción de los resultados. Este hecho es, en este caso, inviable ya que va en contra de la eficiencia que se quiere lograr con las búsquedas. Sí que se ha reflejado en momentos concretos en los que el hablante ha realizado alguno de estos fenómenos con una intención expresa, como, por ejemplo, imitando dentro de una emisión de estilo directo a otra persona. En cuanto a los cambios en el paradigma verbal, en el caso de Argentina, la recomendación ha sido no marcarlos y transcribirlos según su realización, decisión apoyada porque en el propio *DLE* se registra esta variación en la conjugación. En el caso chileno, no recogido de manera sistemática por el *DLE*, se ha marcado como puede verse en el ejemplo más abajo, etiqueta que explicaremos en el bloque (b). En general, cuando ha habido dudas sobre la representación de diversas variedades léxicas, la recomendación general ha sido la de transcribir la forma recogida por el *DLE* o por el *Diccionario de Americanismos*. Tal ha sido el caso, por ejemplo, cuando se nos planteó la consulta sobre el término *wey* en el español de México; se recomendó el uso de *güey*, puesto que es la forma que recoge el *Diccionario de Americanismos*.

En cuanto al **concepto de estilo directo o cita**, las principales dudas han girado en torno a qué se considera estilo directo, es decir, si solo se aplicaba a los momentos en los que un hablante reproduce las palabras textuales de alguien o si también debían incluirse aquellas reproducciones de las palabras textuales emitidas por uno mismo, a veces introducidas por verbos de pensamiento. En el ejemplo siguiente puede verse un caso en el que se distingue claramente el estilo directo, frente al Ejemplo 45, donde se recoge el segundo caso, que hemos considerado marcar también como estilo directo:

Ejemplo 44

A: ah no yo no/ no me-// no me chocó tanto

B: ya<alargamiento/> pero igual brígido que<alargamiento/>

A: sí

B: <fsr t="digái">digas</fsr> <cita>mi mamá me [quiso matar</cita>]

(SCL_005_05_18)

Ejemplo 45

A. sí porque la gente se piensa <cita>no ahora le entra al servicio y se arregla</cita>

(HAV_058_02_17)

En referencia a **nombres propios extranjeros y otros extranjerismos** o fragmentos pronunciados en otra lengua que han sido registrados en las conversaciones, las indicaciones se han dado en dos sentidos: por un lado, los nombres propios extranjeros se han reproducido manteniendo su ortografía original sin ningún tipo de marca, retomando el ejemplo incluido en 5.2.2.1.2. para explicar el uso de la etiqueta *extranjero* encontramos el caso:

Ejemplo 46

A: [tu título]/ <risas/> tipo Margaret Atwoo<alargamiento/>d ahí

(BUE_002_03_20)

Mientras que las intervenciones emitidas en otra lengua sí que han sido marcadas con la etiqueta <extranjero t=“ ”></extranjero> (ver 5.2.2.1.2.):

Ejemplo 47

A: el tractor se las lleva <fsr t="pa">para</fsr> allá pero cargarlas es el <extranjero t="chou">show</extranjero>

(TCO_012_03_20)

Ejemplo 48

A: estos están<alargamiento/> mega <extranjero t="lait">light</extranjero>//

(MEX_043_03_21)

En estos ejemplos puede observarse que se ha transcrito la emisión respetando la ortografía original, aunque en el atributo se incluye cómo se ha pronunciado dicha emisión.

El último aspecto problemático, que quizás es el que más dudas ha generado, ha sido el de la **representación de elementos funcionales, interjecciones, onomatopeyas y ruido**s, tanto producidos por los propios hablantes, como sucedidos de manera externa.

Respecto a los elementos funcionales como *mm, eemm, uhum*, como se señaló en la sección 4.3.2., no existe un consenso generalizado sobre cómo deben transcribirse, si bien parece haber algunos de estos elementos que están más convencionalizados, y lo mismo ocurre con las interjecciones y las onomatopeyas. En la medida de lo posible se ha recomendado la utilización de las convenciones establecidas por la RAE y ASALE, bien en sus obras, bien a través de Fundéu. Aun así, queda pendiente la extracción de un listado estandarizado que se pueda aplicar de manera sistemática y que dé respuesta a este problema. El motivo por el cual no se ha hecho mayor hincapié en este aspecto se debe a que la interfaz de búsqueda permite recuperar el fragmento de audio en el que aparecen estos elementos y, por tanto, puede escucharse de primera mano cómo ha sido pronunciado. En los casos en que se hacía necesario incluirlos con un mayor grado de detalle, se ha hecho utilizando la etiqueta de *observación* y en ella se recogía la información contextual que pudiera ser relevante, como se ve en el ejemplo que aparece a continuación:

Ejemplo 49

A: [ah sí es cierto]// y yo <obs t="imita sonido de tallar">pff</obs> [tallar tallar] <entre_risas>tallar esa es mi vida</entre_risas>

(MEX_032_02_21)

En cuanto a qué nivel de detalle se espera con respecto a la transcripción de los ruidos, se estableció que solo se iba a reflejar en la transcripción aquellos momentos en que estos supongan una interferencia significativa con la conversación, como, por ejemplo, si el ruido de fondo hace que no se entienda lo que se ha dicho o si ha provocado un corte en la conversación porque los ha interrumpido. Para otros casos, queda a criterio de la persona que realiza la transcripción valorar la pertinencia de su inclusión y, de ser necesario, estas anotaciones se harían en la línea dedicada a las observaciones en ELAN con la etiqueta <obs t="hay ladridos de fondo"/>, o en nota al pie si se trabaja con procesador de texto.

Además, desde algunos equipos ha habido una preocupación por reflejar el significado pragmático de algunos de los ruidos, elementos funcionales e interjecciones mencionados anteriormente. Estos equipos han planteado la posibilidad de anotar estos significados, por ejemplo, cuando se produce un *¡ay!* que refleja emoción, o una aspiración imposible de

transcribir que indica sorpresa. Desde el protocolo de trabajo no se ha pedido este grado de detalle, pero, como hemos señalado, los equipos pueden añadir aquellas capas de información que necesiten según sus intereses de investigación. Por tanto, para dar salida a esta consulta se estableció que se hiciera a través de la etiqueta de observación.

Ejemplo 50

B: <obs t="sonido de asco">ah</obs>

(MEX_040_02_21)

Ejemplo 51

B: <obs t="dos chasquidos con los labios que significa atención, incredulidad o asombro">mts</obs>

(MEX_042_03_21)

(b) Empleo de marcas y etiquetas en la transcripción

Se presentan ahora los problemas en torno al uso de las marcas y etiquetas del sistema de transcripción de Ameresco planteadas por el personal transcriptor, como, por ejemplo, cómo se combinan varios signos o cuál es el significado y aplicación exactos de algunas de ellas.

Una de las consultas más frecuentes ha sido la relativa a las marcas <sic> </sic> y <fsr t=“ ”> </fsr>, que en algunos momentos han sido confundidas. Debe matizarse, como se señaló en 5.2.2.1.2., que la primera se utiliza para reflejar algún error en la pronunciación por parte del hablante, como sería el caso de <sic>pusistes</sic>, cuyo uso señala que no es un error o descuido del transcriptor; mientras que la segunda recoge fenómenos de fonética sintáctica y en general aquellos casos en los que la ortografía y la pronunciación de una palabra no coinciden. Un ejemplo representativo del uso de esta etiqueta lo constituye la transcripción de la conjunción *pues*, cuya realización en el medio oral puede producirse de múltiples maneras: *pues, ps, pss, pus, pos, s*. Dada esta casuística y pensando en las posibilidades de búsqueda futuras desde la plataforma en línea, debía privilegiarse la búsqueda por la forma normativa, pero intentando dar cuenta de su representación oral; de ahí que se optara por transcribir de la siguiente manera:

<fsr t="pos">pues</fsr>

El buscador recupera la forma normativa *pues*

El usuario/a, sin necesidad de acceder al audio, puede saber cómo se ha pronunciado

La siguiente dificultad en lo relativo a las marcas y etiquetas en la transcripción está relacionada con la representación del habla simultánea. Recordemos que para esta circunstancia se contempla el uso de los corchetes, tanto en el modo de trabajo 1 como en el modo de trabajo 2. La particularidad de esta marca tiene que ver con el hecho de que debe utilizarse en todas las intervenciones, independientemente de a cuántos hablantes afecte, que se produzcan solapadas, por tanto, se representación exige un esfuerzo considerable por parte de los transcriptores. En el caso del modo de trabajo 1, en procesador de texto, se señalan siguiendo el modelo Val.Es.Co. (2002), es decir, los fragmentos solapados se encierran entre corchetes y, además, se alinean visualmente como se puede ver en el siguiente ejemplo:

Ejemplo 52

A: o por último hacete cargo si ya fue tuyo [ya recupéralo si total]

B: [sií po]

(SCL_011_03_18)

Para el modo de trabajo 2, ya que el programa ELAN admite la creación de líneas para cada hablante, no es necesaria tal alineación ya que la interfaz del programa ofrece de por sí esa visualización. Como se puede observar en la Figura 70, los hablantes B y C emiten varias intervenciones solapadas, simultáneos enmarcados con la forma en rojo:

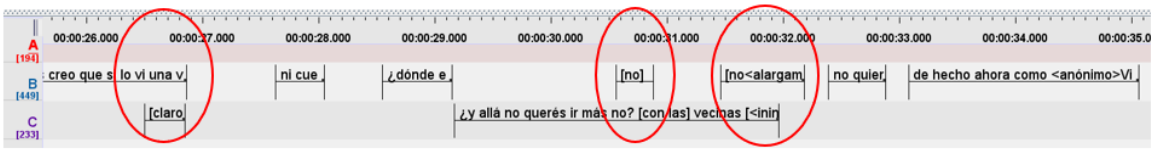


Figura 70. Visualización de un solapamiento de habla en ELAN

La última consulta en este aspecto está relacionada con la manera de combinar el uso de las diferentes marcas y etiquetas. En este sentido, se planteó la cuestión de en qué orden

debían aparecer cuando era necesario incluir varias marcas a la vez, por ejemplo, que hubiera un extranjerismo dentro de una reproducción en estilo directo, afectada, además, por una observación. En estos casos funciona como las fórmulas matemáticas, es decir, se empieza poniendo la más general en la parte exterior y dentro de ella se van marcando el resto hasta llegar a la más específica. En un ejemplo como el que sigue, la etiqueta más general sería la de *cita*, mientras que la más concreta corresponde a *énfasis*:

Ejemplo 53

A: <cita>[y ya de ahí si-] si quieren a lo súper sencillo <énfasis t="pronunciación_marcada">psicología</énfasis></cita>

(MEX_042_03_21)

(c) Anonimización[67]

Este último apartado recoge las indicaciones dadas a los transcriptores y transcriptoras en cuanto a qué se debe anonimizar y cómo deben anonimizarse los segmentos necesarios. Surgieron dudas sobre cómo anonimizar nombres propios en diversos casos, con especial incidencia en nombres y apellidos de personas, lugares y otras informaciones aparecidas como nombres de empresas y lugares relacionadas con los hablantes.

Con respecto a los nombres propios la recomendación general ha sido sustituir por otro con el mismo patrón fonético y adecuado a las características socioculturales y diatópicas, como se señaló en la sección 5.2.2.1.2. Además, debe atenderse, en primer lugar, al contenido completo de la grabación, ya que si un nombre de pila se usa en unos momentos completo y en otros acortado (por ejemplo, Alberto > Alber, Carolina > Caro), el término ficticio debería poder ser adaptado a esta característica (si el nombre real es Cristina y en ocasiones se acorta como Cris, el reemplazo por Carmen, por ejemplo, no sería adecuado). En segundo lugar, si nos encontramos ante una conversación procedente de Colombia, y el término a sustituir es un nombre propio de una mujer de menos de 30 años, no cabría la opción reemplazarlo por otro característico de una zona catalanoparlante. Para agilizar esta tarea, se recomendó el uso de páginas electrónicas que incluyen registros de nombres, apellidos en las que aparece

[67] Los ejemplos presentados en este apartado son ficticios ya que por protección de datos no pueden usarse los ejemplos reales.

información como la media de edad de las personas con esos nombres, como puede verse en la Figura 71.

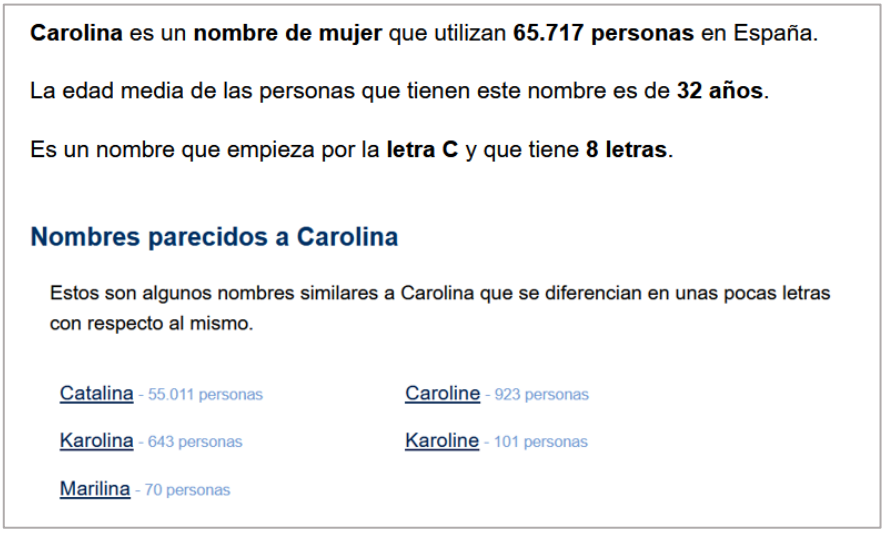


Figura 71. Captura de la búsqueda *Carolina* en la web Nombres Top

Como hemos señalado arriba, son susceptibles de anonimización los nombres de empresas y lugares relacionados con los hablantes. Para este caso, si bien la instrucción general consiste en buscar un nombre ficticio similar, se han dado casos en los que el ficticio no se adecuaba al contexto discursivo y, por tanto, su uso implicaba la pérdida de la coherencia del relato. Si bien se intenta mantener el correlato situacional, en los casos en que se observó que cualquier sustitución iba a alterar la comprensión del relato, se decidió por usar una observación o un nombre genérico, del tipo <anónimo><obs t="ciudad 1"/></anónimo>.

Otro caso controvertido con respecto a la anonimización tiene que ver con grabaciones con contenido sensible o delicado en los que, generalmente previa indicación del equipo local y/o de los propios participantes al finalizar la grabación, se ha procedido al borrado completo del segmento delicado y añadido una observación en la transcripción indicando el corte en la grabación: <obs t="corte en la grabación"/>. Esto ocurrió, por ejemplo, en una grabación en la que una familia relataba que habían sufrido un grave crimen recientemente.

Por otro lado, se observaron casos en los que hubo una sobreanonimización, esto es, ante la instrucción de anonimizar aquellos elementos que pudieran identificar al hablante, hubo transcriptores que entendieron que se debía anonimizar todos los nombres propios y de

lugares que aparecían, sin atender previamente al contexto interactivo. Así, por ejemplo, encontramos anonimizaciones para nombres de dirigentes políticos y personalidades públicas, así como de lugares y localizaciones que no tenían que ver con el hablante en particular, error que se corrigió en la fase de revisión.

De momento, no se ha logrado automatizar las tareas de anonimización, ya que podría haber problemas de reconocimiento de entidades, por ejemplo, Puerta puede ser apellido, pero también un sustantivo común; el uso de herramientas automáticas implicaría inevitablemente la revisión y desambiguación manual por parte de un humano puesto que, si bien hay opciones como la búsqueda automatizada que distinguiera mayúsculas y minúsculas para distinguir en qué casos aparece como apellido y en cuáles como sustantivo común, en el caso de los identificadores indirectos aún se hace más complejo y la opción automatizada no es viable. No obstante, el protocolo actual contempla una manera semiautomática de hacerlo, como se ha comentado en 5.2.2.3.

Con respecto de la fase de **revisión de los materiales**, como hemos adelantado anteriormente, no siempre los equipos locales contaban con la infraestructura suficiente, técnica y de personal, para abordar todas las fases de la recolección del corpus. En algunos casos, solo han podido encargarse de recoger las grabaciones, con sus fichas técnicas y permisos, pero no de ejecutar la fase de transcripción; en otros casos, se han enviado transcripciones anchas en formato texto, pero no transcripciones estrechas ni transcripciones alineadas. Aun sabiendo que este hecho ralentizaría el trabajo posterior de procesamiento del corpus, se decidió contar con este material "incompleto" ya que la parte más valiosa es precisamente la recogida de la grabación y desde el equipo central no se podía asumir su ejecución. Mientras que, la fase de transcripción sí que podía ser completada desde el equipo central de Valencia.

El hecho de asumir esta tarea en el equipo central conllevó el solventado de dos inconvenientes: por un lado, ha sido necesario buscar −y remunerar− a personal externo para la agilización de la transcripción y validación de los materiales; por otro, estos transcriptores, hablantes nativos de la variedad del español peninsular, han debido enfrentarse a la transcripción o revisión de otras variedades dialectales ajenas a su propia realidad, hecho que no ha estado exento de problemas, como comentaremos a continuación.

Las personas encargadas de transcribir las grabaciones enviadas desde los diferentes equipos carecen del contexto necesario para resolver ambigüedades surgidas en la grabación y, a su vez, no poseen conocimientos dialectales suficientes para comprender ni plasmar la realidad de lo dicho en la grabación, de manera que la pérdida de información es incremental e inevitable. Una de las dudas más recurrentes de los revisores es la de la necesidad o no de completar o enmendar información cuando la persona que ha hecho la trascripción ha obviado intervenciones o ha transcrito cosas que, a juicio del revisor, el hablante realmente no ha dicho, puesto que la conciencia de la pérdida de información hace que la sujeción al texto proporcionado sea máxima. En ocasiones, los revisores no actúan corrigiendo directamente, pero sí manifiestan sus dudas sobre si una forma en concreto es una variante dialectal o si, dentro de esta, se considera una forma no estándar y, por tanto, debe etiquetarse como fenómeno de fonética sintáctica (§ 5.2.2.1.2.). Lógicamente, esta situación se da de manera especial entre los revisores cuyo dialecto difiere de la variedad recogida, que corren el peligro de corregir expresiones desconocidas, pero válidas (e incluso generales) en la variedad. Ante esta situación, la recomendación por parte de la coordinación técnica ha sido doble.

En primer lugar, para las personas que revisan transcripciones hechas en procesador de textos, lo más conveniente es comenzar realizando la división en grupos fónicos y, una vez dividida la conversación, volcar el contenido de la transcripción que proporciona el transcriptor local, sobre la cual, en efecto, se puede añadir o corregir todo aquello que el corrector juzgue que falta. Los transcriptores a menudo obvian las repeticiones o los reinicios, omiten fragmentos, solapamientos, etc. No obstante, se insiste en el cuidado especial hacia el léxico empleado, puesto que hay muchos localismos que el revisor no nativo del dialecto puede desconocer y, por tanto, malinterpretar y corregir como un término más familiar en el dialecto propio. Ante esa situación, se recomienda una batería de recursos, dependiendo de la naturaleza del término desconocido, que van desde el diccionario de la RAE hasta el *Diccionario de Americanismos*, y también, según las ocasiones, la consulta directa en buscadores como Google. En el caso recogido a continuación, por ejemplo, la persona española que transcribió no identificó ‘trusa’ y lo corrigió como ‘blusa’, sustitución que fue detectada y corregida en la fase de validación.

Ejemplo 54

B: yo quería otro tipo de trusa/ quería una trusa completa tipo bodi// pe<alargamiento/>r[o<alargamiento/>]

(HAV_082_02_17)

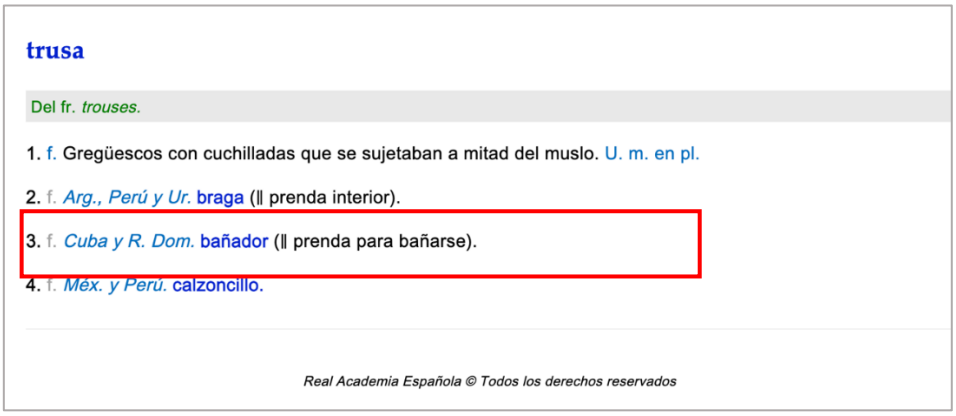


Figura 72. Entrada del *DLE* para el término *trusa*

Otro ejemplo, sería el caso de la interjección *¡nombre!* de la Figura 73, para expresar asombro, sorpresa o desacuerdo en el español de México, en concreto en las conversaciones de Monterrey. Los transcriptores de España sustituyeron este marcador por *¡no, hombre!* y fueron los propios investigadores locales los que dieron cuenta del error.

Ejemplo 55

C: ¡no hombre! ni se gasta

(MTY_022_03_15)

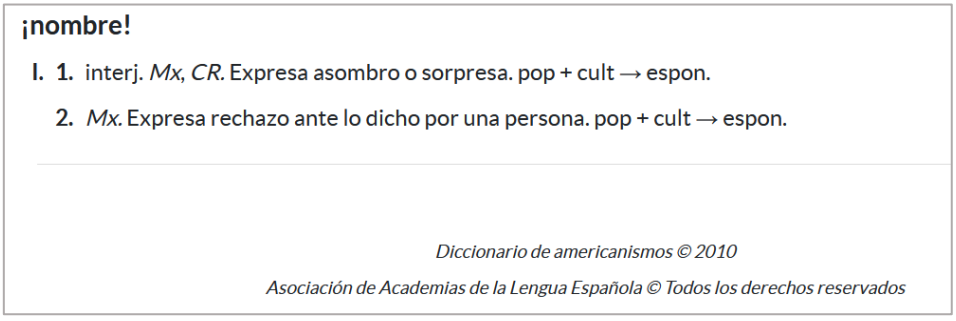


Figura 73. Entrada del *Diccionario Americanismos* para el término *nombre*

En segundo lugar, para las personas que trabajan directamente con archivos en ELAN, la recomendación es la misma, pero teniendo en cuenta que los cambios en las cajas de segmentación son complejos y una mala práctica puede llevar a desajustes en la alineación.

Tras la detección de estos problemas, se ha animado a que los equipos locales trabajen todas las fases de la recogida y transcripción, optando por un único modo de trabajo (el modo 2 que alinea texto y audio), no solo como una medida para evitar estos errores, sino también, para la agilización del procesamiento de los datos y la introducción en el motor de búsqueda. Por tanto, en el protocolo actual, la labor de recogida y transcripción de las muestras de las diferentes ciudades debe pasar de forma obligatoria por las manos de hablantes (miembros) de esa variedad dialectal. Al ser muestras de conversación coloquial grabadas secretamente en entornos vivenciales de proximidad y familiaridad, las grabaciones deben realizarlas sujetos pertenecientes a estas realidades, no investigadores externos, ajenos a la realidad familiar como ya vimos en el apartado 5.2. Así mismo, los equipos locales deben encargarse de la transcripción al ser conocedores de primera mano de la variedad dialectal sujeta a grabación.

5.3.3. Dificultades en la fase 3. Archivo, distribución y acceso al corpus por parte de los usuarios

Uno de los problemas que surgen al ofrecer los archivos a los usuarios tiene que ver con la responsabilidad, por parte de los editores o gestores del corpus, de mantener la infraestructura de alojamiento y, también, de preocuparse por ofrecer formatos de archivo que sean utilizables y perdurables en el tiempo y que no deban ser constantemente adaptados a nuevos formatos más modernos.

> The last challenge for the future of spoken corpora is their continued availability and accessibility. While an increasing number of corpus compilers are eager to make their spoken corpora available to the research community, technological and ethical difficulties have to be met as discussed below. For corpus data stored in nondigital form such as analogue tapes (there is still a lot of historical data that has not been digitised yet) every access means loss of quality. Moreover, many older data formats will not be accessible anymore in the near future. The archiving and dissemination of spoken corpora, even in digital form, thus implies the constant pressure of keeping up with technological advances. For instance, raw video data encoded in

> one format such as MPEG1 will have to be regularly updated to new encoding schemes. (Gut, 2020, p. 250)

Además, ha sido uno de los objetivos del corpus Ameresco dar respuesta al mayor número de necesidades del usuario, contemplando, como se ha detallado en la sección 5.2.3. opciones de acceso y uso más sencillas, en las que el usuario puede descargar los archivos en diversos formatos adaptados a sus necesidades e incluso hacer uso de un motor de búsqueda simple; mientras que se ofrece también la posibilidad a través de la aplicación Oralstats Aroca de acceder a un motor de búsqueda avanzada y de descargar los archivos en formatos como ELAN, que permite una mayor explotación sobre el texto plano. Así mismo, se da respuesta a las propias necesidades del grupo de investigación, como son los estudios prosódicos y el análisis de la conversación pragmática.

La página electrónica ha ido desarrollando diferentes versiones y ampliando las posibilidades de búsqueda y obtención de resultados según se han ido detectando posibilidades de mejora en el sentido señalado en el párrafo anterior, como se especificó en la sección 5.2.3.1.

Por último, se contemplan también las necesidades de almacenamiento, ya que ha sido necesario mantener un equilibrio entre ofrecer las grabaciones en una calidad óptima y no ralentizar el procesamiento de las consultas. Así, si bien ELAN necesita que los archivos estén en formato WAV, para su inclusión en la web se han convertido a formato MP3, menos pesado y más eficiente. Otras limitaciones tienen que ver con el hecho de la visualización de todos los formatos disponibles, esto es, podemos encontrar los archivos de ELAN y de Praat en la plataforma, pero su visualización depende de que cada usuario instale en su ordenador dichos programas y adquiera unas nociones mínimas sobre el empleo de estas herramientas.

5.4. Síntesis del capítulo

En este capítulo se han abordado los ejes de actuación con respecto al diseño y construcción del corpus Ameresco partiendo de la reflexión de debilidades y fortalezas en la concepción y materialización de metodologías de construcción de otros corpus orales del panorama español realizada en los capítulos 3 y 4.

En primer lugar, se han ofrecido las características generales del corpus Ameresco, esto es, cómo surge este proyecto y qué objetivo de investigación persigue, así como el reconocimiento de la extensa red de equipos locales que participan de su construcción junto al equipo central. En segundo lugar, se ha explicitado las decisiones operativas que desde la dirección del corpus se han tomado en todas las fases de su construcción, esto es, se presenta ante la comunidad científica, a modo de metodología práctica que aspira a poder ser replicable en la construcción de nuevos corpus orales. En este sentido, se ha ofrecido la explicación del protocolo de recogida de los materiales que conforman el corpus (grabaciones, consentimientos informados y fichas técnicas), detallando qué y cómo se ha recogido de manera justificada. En segundo lugar, se han incluido los diferentes modos de trabajo atendiendo a diversas circunstancias de trabajo que dependen de los equipos locales que recogen el corpus; así, tras un análisis de las circunstancias que favorecen o desfavorecen la recogida de los datos, se ha propuesto dos modos de trabajo que intentan dar cabida a todas las circunstancias posibles de los equipos locales. Estos modos de trabajo difieren en particular en el sistema de transcripción y codificación, protocolo cuyas características se han detallado de manera minuciosa y exhaustiva, ofreciendo ejemplos concretos, y que, además, contempla las pautas establecidas para la revisión, validación y anonimización de los archivos. En tercer lugar, para la fase de archivo, distribución y acceso al corpus, se detallan las funcionalidades disponibles a través de la página web del corpus a través de sus dos modos de consulta: la web general y la aplicación Oralstats Aroca para consultas avanzadas.

Tras la descripción de las tres fases de diseño del corpus Ameresco, se han recogido las dificultades y propuestas de solución que se han ofrecido para cada una de las fases basándonos en la experiencia concreta de gestión del corpus desde sus inicios y en la comunicación directa con los equipos locales.

Como ha quedado patente, la explicación detallada de cada uno de los factores que se desarrollan en cada fase viene a dar respuesta a la falta de especificación que existe de manera general en los análisis previos del panorama de corpus orales de español.

Capítulo 6

Consideraciones finales

6.1. Principales hallazgos

En la estructura general de este trabajo hemos visto, en el Capítulo 1, la motivación y justificación de este trabajo de acuerdo con los objetivos de investigación relacionados con la metodología para el diseño y construcción de un corpus oral multidialectal de conversación coloquial.

El Capítulo 2 se ha centrado en la aproximación a la lingüística de corpus. Se ha mostrado cómo un corpus lingüístico se define como una colección de textos naturales, escritos u orales, que se almacenan y procesan digitalmente con el propósito de servir como material empírico para el estudio del lenguaje. Hay tres aspectos fundamentales que comparten las definiciones de corpus: los textos deben ser producidos en situaciones reales; la recopilación de los textos debe estar guiada por parámetros explícitos para garantizar su constitución y replicabilidad y, por último, los corpus deben estar disponibles en formato electrónico para su análisis mediante programas computacionales. Se ha mostrado, así mismo, la diferencia entre *archivos de textos, bases de datos* y *corpus*, que se fundamenta en que los primeros son depósitos de textos recogidos oportunamente, mientras que los corpus se crean con criterios específicos de representatividad y equilibrio de muestra. Las bases de datos, por otro lado, pueden ser colecciones de textos que no necesariamente siguen los principios de un corpus.

Un aspecto fundamental planteado en este capítulo es la propia naturaleza de la lingüística de corpus como disciplina, ya que se considera una aproximación teórica y metodológica alejada de la visión generativista del lenguaje basada en la introspección, y existe un vivo debate sobre si la lingüística de corpus es una disciplina por derecho propio o es únicamente una metodología aplicable a otras disciplinas lingüísticas para obtener resultados de diversa índole, en la línea de los autores que la ven como un sustento a la investigación lingüística al proporcionar tanto corpus lingüísticos como la tecnología computacional para procesarlos.

Junto a este debate, la lingüística de corpus se enfrentó a duras críticas iniciales por carecer, supuestamente, de rigor y fiabilidad. Estas críticas se han desvanecido con el tiempo, y la lingüística de corpus se ha establecido como un enfoque científico válido y valioso, puesto que utilizar ejemplos reales en los estudios lingüísticos contribuye a evitar la subjetividad de la introspección. En ese sentido, los corpus ofrecen una base empírica para

la observación, la inducción, la deducción, la comprobación y la evaluación de los datos lingüísticos.

Los materiales para los corpus pueden obtenerse tanto a partir de muestras de habla natural como mediante técnicas de elicitación, de manera que el resultado varía entre datos muy naturales y datos altamente monitorizados. Así mismo, la perspectiva de explotación de los datos de corpus puede ser *corpus-based* y *corpus-driven* según si, respectivamente, se parte de hipótesis previas y se utiliza el corpus como apoyo para respaldar afirmaciones lingüísticas o, en el segundo caso, se utiliza el corpus como punto de partida y se descubren patrones lingüísticos sin preconcepciones teóricas. Para esta identificación de patrones lingüísticos, así como para maximizar la objetividad en la identificación de fenómenos y la posibilidad de análisis cuantitativos y estadísticos, la lingüística de corpus resulta una herramienta imprescindible, puesto que permite el uso de *software* específico que facilita la búsqueda, extracción y clasificación de datos en los corpus, lo que enriquece los estudios lingüísticos.

El desarrollo de la lingüística de corpus se divide en tres generaciones o etapas desde sus inicios hasta la actualidad, marcadas por la aparición de nuevos avances tecnológicos y por un cambio sustancial en la percepción de la metodología. En la primera generación, hasta mediados del siglo XX, la lingüística de corpus aún no se denominaba como tal y se basaba en enfoques empíricos (esto es, en el uso de datos reales) de la lengua, y fue criticada especialmente por Chomsky, quien favorecía la intuición como fuente de conocimiento en lingüística. Los primeros corpus se recopilaban en papel y no se prestaba atención a la representatividad, ya que el análisis se realizaba manualmente**.** En las décadas de los años 60 y 70 surge la segunda generación. En esta etapa, se produjo un avance tecnológico con la mayor disponibilidad de computadoras, lo que permitió la creación de los primeros corpus informatizados, como el *Survey of English Usage Corpus* (SEU), el *Brown University Corpus of American English* (Brown Corpus), el *Lancaster-Oslo/Bergen Corpus* (LOB) y el *London-Lund Corpus of Spoken English* (LLC). Sin embargo, en esta generación predominaron los corpus de textos escritos debido a las dificultades técnicas y de transcripción asociadas con los datos orales. Finalmente, en la tercera generación, a partir de los años 80, la lingüística de corpus experimentó un renacimiento gracias a la influencia de autores como Leech, quienes abogaron por el uso de corpus en combinación con la intuición lingüística. El desarrollo de la lingüística computacional y la mayor disponibilidad de

tecnología avanzada permitieron la creación de macrocorpus, como el *Bank of English* (COBUILD Corpus) y el *British National Corpus* (BNC), en inglés, ylos corpus de referencia académicos como CREA, CORDE y CORPES XXI, en español. Al auge de la lingüística de corpus ha contribuido, además, la aparición de asociaciones y centros de investigación especializados y de revistas científicas.

En cuanto a la caracterización de los corpus, se emplean varios criterios para su clasificación. Así, se atiende al medio, según el cual existen corpus escritos, orales y multimodales y, en cuanto al periodo temporal recogido, pueden ser diacrónicos o sincrónicos; según el número de lenguas tratadas, se clasifican también en corpus monolingües, bilingües, multilingües o paralelos, y según la especificidad de los textos, n generales o de referencia y especializados o técnicos, entre los cuales cabe añadir los corpus de entrenamiento; si es el tamaño lo que se considera, existen macrocorpus o microcorpus, y son abiertos si se actualizan con nuevos materiales, o cerrados si tienen una muestra fija; por último, de acuerdo con la información adicional que proporcionan, algunos están anotados para el posterior análisis automatizado, mientras que otros son simples o puros, sin anotaciones. Todos estos criterios de clasificación se pueden combinar para caracterizar los corpus de manera más precisa.

El objeto de estudio de este trabajo lo constituyen, en concreto, los corpus que se centran en los materiales orales. Los corpus orales han revolucionado la lingüística al permitir el estudio de la variación del lenguaje en contextos naturales, en el hábitat primero del lenguaje, de manera que resultan óptimos para el estudio de fenómenos relacionados con la pragmática y la conversación. Así, la pragmática de corpus surge como una disciplina emergente que combina la lingüística de corpus con la pragmática para estudiar fenómenos que requieren un enfoque tanto cualitativo como cuantitativo

El Capítulo 3 se enfoca en el desarrollo de los corpus orales en el ámbito concreto del español. Como sucede en otras tradiciones, estos corpus hunden sus raíces en la dialectología y los atlas lingüísticos, que favorecen la observación empírica de datos sobre la intuición y que mostraron la insuficiencia de la encuesta como único método de recogida. La lingüística de corpus en español comenzó con cierto retraso en comparación con el mundo anglosajón, pero tuvo un rápido desarrollo cuyo primer hito fue el *Programa Interamericano de Lingüística y Enseñanza de Idiomas* (PILEI), iniciado en la década de 1960, que buscaba estudiar el habla en grandes concentraciones urbanas de América; se crearon corpus de

diferentes localizaciones en Hispanoamérica y España, aunque estos no están disponibles al público en su totalidad. Uno de los primeros corpus que proporcionó acceso al audio en formato CD-ROM fue el *Macrocorpus de la Norma Lingüística Culta de las principales ciudades de España y América* (MC-NC), que incluye entrevistas individuales revisadas y homogeneizadas, y que posteriormente se incorporó al *Corpus de Referencia del Español Actual* (CREA) de la Real Academia Española y al *Corpus del Español* de Mark Davies (CE). Junto al PILEI, otro de los proyectos pioneros más relevantes es el macrocorpus para *el Estudio Gramatical del Español Hablado en América* (EGREHA), que contiene materiales del PILEI y el MC-NC, con una mayoría de sus materiales todavía inéditos. A finales de los años 90 nace el que hoy en día es el proyecto más importante de estudio sociolingüístico del español de España y América, PRESEEA, bajo la dirección de Francisco Moreno Fernández, un corpus que recoge entrevistas semidirigidas en diferentes ciudades de habla hispana, en formato de transcripción ortográfica enriquecida y etiquetado TEI en XML.

Los primeros corpus de habla en español fueron iniciativas privadas, pero surgieron más tarde corpus académicos como el *Corpus de Referencia del Español Actual* (CREA) y el *Corpus del Español del Siglo XXI* (CORPES XXI). CREA se publicó en 1998 con material tanto escrito como oral, aunque sin acceso a los audios de estos últimos. En contraste, CORPES XXI, que se publicó en 2013, marca un avance en la disponibilidad de material oral en español para la investigación lingüística al mejorar la representatividad geolingüística y permitir el acceso al registro sonoro, puesto que sus materiales orales provienen de acuerdos con medios de comunicación, YouTube, etc., y de otros corpus que han cedido sus materiales.

Aunque resulta imposible recoger todos los corpus del español sin que falte alguno y, además, todo recopilatorio está condenado a su propia obsolescencia casi inmediata dada la velocidad de los avances informáticos, se recogen en este capítulo algunas de las principales obras que ofrecen el panorama de corpus orales del español. El artículo de Moreno Fernández (2005c) se enfoca en la recopilación, destacando la importancia de considerar tanto las variedades geolingüísticas del español como la representatividad de los materiales recogidos y distingue los corpus para el desarrollo de tecnologías del habla de aquellos para el estudio lingüístico general de la lengua oral, estos últimos divididos entre corpus orales generales y especializados para estudios lingüísticos. Además, en la línea de su interés por

la sociolingüística, señala la importancia de considerar criterios adecuados para la selección de hablantes y la necesidad de ajustarse a una tipología establecida previamente. El trabajo de Briz y Albelda (2009), una iniciativa del Instituto Cervantes, tiene como objetivo proporcionar una visión general de los corpus de lengua hablada y escrita en español elaborados o en proceso durante la primera década del siglo XXI, e incluye tanto corpus orales como corpus escritos, clasificados (además de según su medio, oral o escrito) teniendo en cuenta el propósito (general o específico), el tamaño (macrocorpus o microcorpus), el formato (textual o motor electrónico de búsqueda) y el modo de acceso (completo al texto o a través de concordancias). También incluyen otros tipos de corpus, como aquellos destinados al desarrollo de tecnologías del habla, lenguajes técnicos y la adquisición y desarrollo del lenguaje. Diez años después, Briz y Carcelén (2019) actualizan la recopilación de corpus orales del español, y los clasifican en corpus panhispánicos, de variedades geográficas concretas y otros repositorios en línea sin acceso a los audios. Junto a estos recopilatorios, se encuentran otros más parciales o menos exhaustivos, como Enghels, Vanderschueren y Bouzouita (2015) Rojo (2016), Solís (2018), Parodi y Burdiles (2019) o Llisterri (2021). De acuerdo con el grado de consenso encontrado en estos trabajos recopilatorios, se puede observar cómo los corpus más relevantes en el ámbito hispánico parecen ser los corpus que cuentan con fines lingüísticos generales.

Este tipo de análisis se revela útil no solo para entender el panorama actual de los corpus orales del español, sino, principalmente y en la línea de nuestro objeto de estudio, para identificar posibles áreas de mejora en la recopilación de estos recursos, algo que se ha aprovechado para el diseño del corpus Ameresco. Los trabajos revisados llenan un vacío importante a la hora de obtener una panorámica del estado de la investigación en corpus, pero no pueden incluir todos los corpus existentes y solo reflejan una fotografía en el momento de su publicación, debido en parte a la ausencia de acceso a corpus más pequeños o creados por investigadores sin el respaldo de grandes instituciones. Los trabajos de Briz y Albelda (2009) y Llisterri (2021) pueden considerarse los más completos en dos sentidos distintos: el primero, por su exhaustividad y el segundo, por la minuciosidad y la información técnica sobre transcripción y anotación, que a menudo falta en otros trabajos.

Uno de los problemas identificados en los recopilatorios y que pone de manifiesto unas carencias generales en la disciplina es la ambigüedad en la terminología y la falta de definiciones claras para términos como *microcorpus* y *macrocorpus*, o como *corpus oral*

frente a *corpus de lengua hablada*, a menudo provocada porque los trabajos se centran en aspectos distintos de los corpus orales del español, lo que a veces resulta en diferencias terminológicas y de clasificación. En consecuencia, se plantea la necesidad de aclarar y definir criterios concretos para clasificar los corpus orales. Por ello, en respuesta a los desafíos y ambigüedades identificados en los trabajos previos, y teniendo en cuenta la ya mencionada imposibilidad de recoger todos los corpus actuales, se ha presentado una revisión actualizada del panorama de corpus oral del español y se han definido con más detalle un conjunto de corpus. Por orden alfabético, se ofrece información detallada sobre los corpus CE, CEMC, CEMC II, CET, COJEM, COLA, COLEH, COLEM, CORDIAL, CORLEC, CORPES XXI, CORPEUU, el *Corpus lingüístico del habla de Almería*, COSER, CREA, *El español hablado en Bogotá*, ESLORA, MESA, PRESEEA, Val.Es.Co. 3.0 y VHW. Esta selección se ha efectuado de acuerdo con la aplicación de varios criterios: el medio, puesto que se seleccionan únicamente los corpus orales, independientemente de si incluyen o no transcripción fonética; su carácter sincrónico; su acceso libre y no comercial, y la no restricción en cuanto a género discursivo, estado (completado o no) y variedad geolectal del español recogida. Así mismo, se trata únicamente de corpus que responden a fines lingüísticos generales (se excluyen, por tanto, los de tecnologías del habla o aprendizaje del español, por ejemplo) y se filtran únicamente los que permiten el acceso en línea a los datos.

El Capítulo 4 parte de los planteamientos previos a considerar ante el propósito de la construcción de un corpus oral, contando con la desventaja de la falta de una protocolización metodológica estandarizada en la literatura. No obstante, existe cierto consenso sobre los pasos que conlleva la recolección de un corpus oral: diseño, obtención de permisos, obtención del material lingüístico, planificación y preparación del método de almacenamiento y procesamiento informatizado del corpus. De los principios básicos que delimitan la caracterización de un corpus, su finalidad, esto es, el objetivo de investigación que hay detrás de su construcción, parece ser el más relevante, si bien deben señalarse otros principios que tienen que ver con la representatividad y el tamaño de la muestra, las cuestiones ético-legales que aplican sobre el material recogido y la elección de un sistema de codificación pertinente al objetivo de investigación primero, entre otros.

Una vez establecidos estos principios básicos en la fase de concepción y delimitación del corpus, el proceso de construcción se concreta en el desarrollo de tres fases: la recogida de

los datos, el tratamiento de los datos y el archivo, distribución y acceso a los datos a través de herramientas y plataformas para su explotación. En la fase 1 de recogida de los datos, además de las grabaciones, debe plantearse la obtención de los metadatos de los hablantes y la situación comunicativa, en forma de ficha técnica, así como de los permisos pertinentes para cada corpus. En el caso que nos ha ocupado en este trabajo, los corpus de conversación espontánea grabados secretamente, el proceso de obtención del consentimiento informado presenta mayor complejidad legal que el consentimiento necesario para grabaciones que no son secretas, puesto que, según la legislación española vigente, no se puede grabar sin que los hablantes hayan sido informados. Este hecho ha supuesto un impedimento, ya que se ha debido buscar una manera lícita de obtener los materiales cumpliendo con este precepto, pero, a su vez, garantizando la naturalidad de los datos obtenidos, así como la anonimización de los datos que pudieran permitir la identificación de los participantes.

Respecto de la fase 2, la que opera sobre el tratamiento de los datos, de acuerdo con lo expuesto en el Capítulo 4, se ha visto que existe cierta vacilación terminológica a la hora de caracterizar los diferentes métodos de representación de la lengua oral por medio de la escritura. Así, se ha establecido una denominación operativa en cuanto a las características de cada uno de estos métodos (*transliteración, transcripción, codificación* y *anotación*) que permitiera la descripción y el reconocimiento de los principales sistemas existentes. Una vez caracterizados, se ha ofrecido una descripción de cada uno de ellos y se han referenciado las principales iniciativas y programas informáticos que permiten el tratamiento computacional de los datos obtenidos.

En la fase 3, se ha reflexionado sobre las necesidades de almacenamiento y distribución de los materiales que componen el corpus, así como los medios necesarios para ponerlos a disposición de la comunidad científica, acceso que en la actualidad se suele conceder en línea por medio de motores de búsqueda o páginas web en las que se disponen los archivos para su descarga.

En la segunda parte de este capítulo, tras la reflexión teórica que conlleva la construcción de un corpus y que afecta a la determinación de su diseño final, se ha realizado un análisis contrastivo de los sistemas de transcripción y codificación de los corpus orales del español considerados más importantes, según el análisis realizado en el capítulo anterior, a saber, COLA, C-Oral-ROM, CORLEC, PRESEEA y Val.Es.Co. (2002). No obstante, para un análisis más completo se ha prescindido de C-Oral-Rom y de CORLEC, por no estar

disponible en formato abierto o por no contar con plataforma en línea para la recuperación de la información, y, en cambio, se han seleccionados otros dos corpus, COSER y CORPES XXI, que no aparecían en la tabla entre los corpus con más consenso porque se publicaron con posterioridad a la elaboración de algunos de los trabajos recopilatorios revisados.

En este sentido, se han establecido tres bloques de revisión comparada entre factores externos (objetivo del corpus, género discursivo, criterios de representatividad, aspectos legales); factores internos (sistema de transcripción, codificación y anonimización de los datos); y factores relacionados con el modo de acceso a los materiales. Esta visión en conjunto de las decisiones adoptadas por otros corpus orales del español ha servido de base para la configuración del corpus Ameresco, al detectarse aquí fortalezas en el diseño que podían aprovecharse en la construcción del corpus que nos ocupa, así como por la localización de debilidades ante las cuales podían establecerse sugerencias de mejora en su aplicación posterior.

El Capítulo 5 constituye el núcleo central de este trabajo y explica cómo se han logrado los objetivos 1 a 7 planteados en el Capítulo 1; esto es, se exponen las bases teórico-metodológicas de la construcción de un corpus multidialectal de conversación coloquial, el corpus Ameresco, y se expone el proceso mismo de construcción y de puesta a disposición de este.

En primer lugar, se ha ofrecido una caracterización general del corpus que atiende principalmente a los orígenes del proyecto y a su marco de desarrollo, a su objetivo inicial de investigación, esto es, la variación de la atenuación pragmática en el español hablado por medio de la conversación coloquial en diversas variedades geolectales; y a los grupos de trabajo participantes en su construcción, sin los que el desarrollo de este corpus no habría sido posible. Además, se detallan otras características como su tamaño y el tipo de corpus que conforma, según la tipología establecida en el Capítulo 2.

A continuación, se han presentado los principios fundamentales que han marcado el diseño final de Ameresco, atendiendo a las tres fases que se han mencionado en los capítulos anteriores. En cuanto a la primera fase de recogida del material oral que compone el corpus, se muestran los criterios adoptados para la selección de hablantes y el tamaño de la muestra, siguiendo la metodología heredada de Val.Es.Co. (2002) y PRESEEA. Así, para el corpus Ameresco se ha establecido una muestra de 72 hablantes divididos en tres estratos

sociolingüísticos (sexo, grupo etario y nivel sociocultural). Con respecto a la recogida del consentimiento informado, se ha partido de los preceptos establecidos en el capítulo anterior para implantar el uso de un consentimiento en tres pasos que cumple con las exigencias legales de la normativa española, así como de la propia institución que acoge este proyecto, la Universitat de València. Este consentimiento, como se ha visto, cuida de que no se produzca la paradoja del observador y se obtengan muestras naturales, solventando el obstáculo de tener que informar previamente a los hablantes de que van a ser grabados y que cumple, además, con lo exigido por la ley de tratamiento de datos personales, según la cual, los participantes deben poder ejercer sus derechos de retirada del material ante los responsables del proyecto.

Para la recogida de las grabaciones, se detallan las diversas situaciones que condicionan el acto de grabar, esto es, la elección de una localización propicia en la que haya cierto control de elementos externos como ruidos y aparición de hablantes no previstos, así como del comportamiento que debe adoptar la persona que va a recoger las conversaciones, atendiendo a la caracterización de la conversación coloquial. Se ha tratado, además, la recogida de la ficha técnica de metadatos, documento imprescindible para dotar de contexto al material sonoro de cara al posterior análisis pragmático.

En cuanto a la fase 2, se ha explicado la existencia de dos modos de trabajo y las circunstancias que han provocado esta situación, esto es, que haya equipos que no podían realizar el tratamiento de los datos por medio del alineado audio y texto y que han debido trabajar de manera más tradicional, a través de la transcripción por medio de procesadores de textos. No obstante, los dos modos comparten la fase 1 de trabajo, aunque difieren en el sistema de codificación adoptado, en lenguaje XML para el procesamiento informatizado, y la transcripción enriquecida para el trabajo en procesador de texto. En ambos casos, se ha incluido una relación de signos, símbolos, marcas y etiquetas empleados en el corpus Ameresco, con la información de uso y ejemplos extraídos del corpus, para su caracterización. Se ha detallado, además, el protocolo de revisión, validación y anonimización de los materiales.

Con relación a la fase 3, tras la caracterización general de las principales plataformas disponibles para alojar el corpus, se han descrito las diferentes opciones de visualización y análisis de los datos desde la página web del corpus Ameresco. A saber, se han implementado dos versiones, una web de consulta básica desde la que pueden descargarse

los materiales completos y realizarse búsquedas sencillas por intervención, y una segunda opción, más avanzada, denominada Oralstats Aroca, que permite transformar y analizar los datos del corpus de manera relacional, más precisa para realizar análisis de tipo prosódico y de análisis de la conversación que la web de consulta básica.

En la última parte de este capítulo, se ha incluido un análisis de las dificultades que han aparecido en cada una de las fases del corpus Ameresco, bien localizadas desde el equipo central, bien señaladas desde los equipos locales que ejecutan el protocolo de trabajo establecido. Para estas dificultades se han propuesto soluciones en todos los casos posibles, mientras que, en otros, se ha asumido la limitación y se contempla la corrección de estas.

6.2. Relevancia de los resultados

Se detallan a continuación, los principales hallazgos obtenidos en este trabajo, a saber, exponer la relevancia de los resultados obtenidos, concretándolos en la aportación particular del corpus Ameresco al panorama internacional de corpus orales y la defensa de la explicitación de la metodología de construcción de corpus como punto fundamental en el desarrollo de la lingüística de corpus.

6.2.1. La aportación del corpus Ameresco al panorama internacional de corpus orales

A partir de este análisis, desde el corpus Ameresco se ha creado una metodología de trabajo que tiene su base en objetivos de estudio pragmáticos y de interacción en conversación coloquial espontánea y que, además, nace con la voluntad de llenar dos huecos importantes: por un lado, el de los corpus panhispánicos, es decir, nace con la voluntad de recoger muestras de las principales ciudades del ámbito hispánico; por otro, responde al déficit señalado en la bibliografía de corpus de este género discursivo, la conversación coloquial espontánea grabada de forma secreta. De esta manera, Ameresco pretende constituirse en un material de estudio que, si bien, nace como extensión natural de los presupuestos del corpus Val.Es.Co., sea espejo del principal corpus de carácter panhispánico del español, el corpus PRESEEA, de manera que permita la intercomparabilidad de investigaciones en diferentes géneros discursivos (la entrevista en el caso de PRESEEA, la conversación coloquial en el caso de Ameresco). De ahí que se hayan tomado como antecedentes metodológicos los establecidos en este corpus para su aplicación en Ameresco,

criterios que tienen que ver con la representatividad sociolingüística y con el sistema de transcripción y codificación a favor de la intercambiabilidad y reusabilidad de los datos.

Sin embargo, no podemos perder de vista una diferencia esencial entre ambos corpus en cuanto a la combinación de los criterios de representatividad y naturalidad. En este sentido, en el corpus Ameresco primará la naturalidad sobre la representatividad ya que nuestro objetivo final es el estudio pragmático. Si bien una posible falta de representatividad podría implicar un desequilibrio en la muestra, siempre podría repararse incorporando nuevos materiales. Como se señaló en el Capítulo 4, en ocasiones el criterio de la representatividad es discutible; esto es, no cabe duda de que un corpus debe optar a alcanzar esta representatividad y así lo señala la literatura, sin embargo, en corpus de naturaleza multidialectal de esta envergadura solo puede ajustarse en cuanto a la cantidad de hablantes recogidos según la estratificación establecida y la localización. Al mismo tiempo hay que reconocer que es imposible que exista un equilibro en el número de formas registradas en cada caso. En ese sentido, si comparamos el nivel de población de cada una de las ciudades que componen el corpus Ameresco, las proporciones de ciudades como Buenos Aires o Ciudad de México deberían ser mayores que las adjudicadas a Panamá o a Loja (Ecuador).

Además, el corpus Ameresco es un trabajo en curso. Aunque se ha recopilado una parte importante (recordemos que actualmente cuenta con unas 75 horas de grabación de 14 ciudades), aún queda mucho trabajo por delante para completar el panorama panhispánico. No obstante, en consonancia con lo mencionado arriba en cuanto a la representatividad, subcorpus que han completado la recolección de los materiales, han manifestado su voluntad, y así lo están haciendo, de recoger nuevas grabaciones que mejoren la naturalidad de las grabaciones. Por tanto, apostamos por una constante voluntad de mejora tanto desde el equipo central como desde los diferentes equipos locales comprometidos con el proyecto.

En esta línea, se han dispuesto mecanismos de reparación de posibles errores y limitaciones del corpus, como es la detección de conversaciones menos prototípicas que, cuando se cuente con conversaciones más representativas, pasarían al repositorio dando paso a los materiales nuevos. Aunque cabe decir que las grabaciones menos adecuadas no serían eliminadas completamente del corpus. Es por ello por lo que desde la coordinación técnica del equipo central se lleva a cabo un registro pormenorizado de las características de cada conversación que se recibe, como se ha visto en el Capítulo 5.

Además, el corpus Ameresco es un corpus abierto en diversos sentidos. Por un lado, se contempla la posibilidad de recibir retroalimentación por parte de los usuarios y usuarias del corpus que pueden contactar con el equipo central para notificar fallos que hayan encontrado. Por otro, también son bien recibidas las sugerencias y necesidades que surjan en los diferentes equipos de trabajo colaboradores, así como de nuevos equipos que quieran adherirse y reflejar otras áreas urbanas aún no recogidas. Y, por último, estamos abiertos a compartir los datos, tanto los que conforman el corpus en sí (grabaciones, transcripciones, metadatos), como aquellas informaciones concernientes a los protocolos de trabajo establecidos, así como a la realización de actividades conjuntas con otros corpus con el objetivo de estar en constante crecimiento y reflexión que permita enriquecer nuestra contribución a la comunidad científica.

No obstante, a pesar de las fortalezas que presenta este corpus, somos conscientes de que el corpus Ameresco tiene errores y limitaciones en su fase actual. Algunos de estos son subsanables y así, se resolverán en la medida en que contemos con presupuesto y personal para realizar las mejoras, así como respetando la metodología adoptada. Pero no podemos olvidar que no podemos deshacernos de algunos de estos errores, especialmente cuando más que errores, son circunstancias del pasado. En este sentido, el corpus cuenta con limitaciones, como se ha analizado en el Capítulo 5, por ejemplo, en los comienzos de recogida del corpus se adoptó un sistema de estratificación sociolingüístico heredado de Val.Es.Co. que, si bien era válido en su momento, hoy en día necesitaría de una revisión para adaptarlo a las condiciones de vida actuales. Dicho esto, en la medida de lo posible, se han llevado a cabo acciones para intentar subsanar esta limitación, como la adopción de un nuevo modelo de ficha técnica que incorpora el dato de la edad exacta de los hablantes de cara a una futura restructuración de los grupos etarios. Así mismo, hay otras limitaciones que no admiten corrección; es el caso de las fichas técnicas recogidas erróneamente desde el equipo local. En este caso la recuperación de la información correcta es imposible ya que, debido a la ley de privacidad y tratamiento de los datos, no se puede revertir el proceso de recogida y volver a contactar con los participantes de las conversaciones. Por tanto, somos conscientes de este hecho y lo aceptamos como posible limitación de cara a investigaciones concretas. Se podría decir que no existen corpus perfectos en este sentido, especialmente cuando parten de un género discursivo como la conversación coloquial que exige unos contextos de grabación muy específicos y poco controlados.

6.2.2. La explicitación de la metodología como vía de avance de la disciplina

Con todo y las limitaciones que se han señalado, cabe destacar lo que supone la aparición del corpus Ameresco en el panorama científico internacional. Más allá de la validez e interés de los datos que aporta para el estudio de la variedad coloquial en el español y que se han listado en la sección 6.2.1, con la redacción de este trabajo se ha pretendido que el corpus Ameresco pueda paliar en parte las carencias, mencionadas a lo largo de estas páginas, en cuanto a los protocolos de construcción de corpus orales. Sin duda, somos plenamente conscientes de la existencia de estos protocolos (todo corpus construido conlleva la elaboración previa de un sistema de trabajo), pero normalmente, o bien son documentos de uso interno de los grupos de investigación, que rara vez se ofrecen con detalle a los usuarios externos, o se transmiten entre investigadores como un modo de actuar y no llegan a plasmarse por escrito. Por tanto, para el desarrollo de la metodología del corpus que nos ocupa en esta tesis, en muchos casos hemos tenido que extraer dicha información de manera inductiva a partir de la experiencia como usuarios de los diferentes corpus analizados.

La revisión realizada en este aspecto ha servido para, partiendo de estas bases previas, analizar los métodos disponibles, ver cuáles son las fortalezas de estos para replicarlas, así como también las debilidades, esto es, errores o limitaciones que se deben solventar. No obstante, cabe reseñar que no es replicable aquello que no se explica, de ahí la reivindicación desde este trabajo.

6.3. El trabajo de compilación de corpus. Una reivindicación necesaria

Tras este estudio pormenorizado del panorama de corpus orales, del desarrollo de las fases de trabajo y construcción implicadas en corpus orales y concretadas en la ejecución del corpus Ameresco, puede verse que la descripción pormenorizada de las cuestiones previas, la ejecución y el seguimiento ocupan un documento de más de 300 páginas que podría, perfectamente, haberse enriquecido con más ejemplos, preguntas y soluciones adoptadas, quizá de menos relevancia para el lector. Ante este volumen de información, puede surgir la pregunta de cuánto cuesta construir un corpus que tarda tanto en describirse.

A la vista de este trabajo, creemos que queda patente cómo componer un corpus oral implica un gran esfuerzo en términos de tiempo, pero también y como consecuencia, en

términos económicos. Esta última consecuencia es menos evidente puesto que, en general, o al menos en el ámbito hispánico, la labor del compilador de corpus suele ser un trabajo altruista. En ocasiones, puede contarse con la ayuda, remunerada o no, de personal estudiante que, si bien pone todo su esfuerzo en desarrollar las tareas adjudicadas, no se ha profesionalizado en esta labor y además no puede dedicarse de manera continuada a la elaboración del corpus; en otras, los propios miembros del grupo de investigación llevan a cabo labores técnicas de procesado de los datos, con el consiguiente menoscabo de su tiempo dedicado a la investigación. Es por esto por lo que se reivindica desde estas líneas la labor realizada por el personal que está detrás de la confección de un corpus oral a la vista de las dificultades de encontrar financiación por parte de la administración.

Para que quede constancia de lo que implica en tiempo la construcción de un corpus de las características de Ameresco, ofrecemos a continuación una estimación de trabajo del corpus realizado en un año. En concreto, desde la coordinación técnica se han llevado a cabo tareas como las que se mencionan a continuación:

1.- Formación de los transcriptores y del personal que va a recoger las grabaciones. Se realiza unas 10 veces al año y con cada incorporación de transcriptor o equipo local; 10-12 horas de formación en cada grupo. Total: 120 horas/año.

Este personal necesita recibir formación específica en el manejo de los programas, en el sistema de transcripción empleado, en las convenciones de marcado de la anonimización, en el etiquetado, etc., dependiendo de su nivel de implicación en el proyecto. Así, hay un formador que debe encargarse de formar equipos en las siguientes tareas:

- Formación en el manejo de ELAN y *software* de tratamiento acústico
- Formación en convenciones de transcripción
- Formación en sistema de etiquetado

2.- Recogida de los materiales. Se dedican entre 5 y 7 horas por semana: 294 horas/año [42 semanas].

- Coordinación de grupos de trabajo locales, en España y América
- Validación de las muestras enviadas antes de pasar a transcripción (comprobación de nivel de ruido, espontaneidad, ajuste al proceso de política

de protección de datos por parte de los equipos locales, estratificación sociolingüística de los participantes, etc.)

3.- Transcripción y codificación de las grabaciones Se dedican entre 3 y 4 horas al día, unas 24 horas por semana: 1008 horas/año [42 semanas].

- Transcripción y codificación de materiales ya recogidos y/o en proceso de recolección
- Segmentación de grupos entonativos en ELAN
- Transcripción en ELAN
- Etiquetado completo en TEI-XML (modelo PRESEEA) adaptado a Ameresco

4.- Procesamiento de las grabaciones: anonimización, revisión, validación. Se dedican entre 1 y 2 horas al día, unas 10 horas por semana: 420 horas al año [42 semanas].

- Anonimización en el audio de los fragmentos previamente etiquetados como <anónimo> </anónimo> con el *software* Audacity
- Revisión de la transcripción
- Revisión del etiquetado y validación de la sintaxis mediante los programas Oxygen y R
- Transformación del formato .eaf de ELAN en texto tabulado apto para el motor de búsqueda web

5.- Procesamiento informático para subida a la web. Se realiza 2 veces al año, 20 horas cada vez. Total: 40 horas/año.

- Tokenización de los grupos entonativos
- Etiquetado morfosintáctico automático
- Revisión del etiquetado morfosintáctico
- Creación de las intervenciones (turnos) a partir de grupos entonativos
- Tratamiento de metadatos como texto tabulado y subida a la web
- Actualización en la base de datos de la web
- Mantenimiento de la web y corrección de problemas de transformación

- Transformación de los audios en .WAV a formato MP3 y subida de audios la web
- Generación de Textgrids (Praat) para el análisis fónico y subida de estos a la web
- Análisis de datos orales a través del sistema Oralstats Aroca

6.- Almacenamiento informático de los materiales (con sistemas encriptados) y copias de seguridad. Se realiza 2 veces al año; 20 horas cada vez. Total: 40 horas/año.

Total de horas para una persona: 1902 horas al año.

Como ejercicio de reflexión puede trasladarse el cálculo, que es solo anual, al total del trabajo realizado hasta ahora (material recogido en 14 ciudades a cuyos integrantes ha habido que formar y atender las dudas y problemas, 75 horas de grabación repartidas en 203 conversaciones que deben ser transcritas, alineadas, anonimizadas y procesadas informáticamente). Dado este esfuerzo en horas de trabajo y contabilizado también en términos económicos (que dejamos a discreción del lector, quien puede otorgar un valor razonable al salario por hora y multiplicar por los años y las conversaciones procesadas), puede entenderse el valor del trabajo de todas y cada una de las personas que han participado en el proceso.

En última instancia, y como resumen, esta tesis viene a ser una reivindicación del trabajo realizado por lingüistas compiladores de corpus. Por un lado, la exhaustividad y metarreflexión que se realiza para establecer los diferentes protocolos de trabajo no se suele poner en valor**.** De hecho, hemos detectado una falta sistemática de explicitación de dichas metodologías en publicaciones científicas, y hemos tenido que conformarnos, en algunos casos, con informaciones publicadas en las páginas electrónicas de cada corpus, y solo de manera más excepcional en publicaciones científicas. Esto, que es general en todo trabajo de compilación de corpus, aplica especialmente a la metodología de trabajo de corpus orales, ya que hemos comprobado en el estudio bibliográfico realizado que los corpus escritos cuentan con más referencias y mayor reflexión metateórica.

En este sentido, cabe señalar la experiencia personal durante este periodo predoctoral cuando, al presentar investigaciones basadas en metodología de construcción de corpus orales en distintos foros (publicaciones, congresos internacionales, seminarios, etc.), se nos

ha revelado repetidamente el interés de la comunidad investigadora por las distintas consideraciones que deben tenerse en cuenta para construir un corpus; algo que contrasta con el hecho de que, si se observan las contribuciones presentadas en congresos especializados en corpus, puede afirmarse que la mayoría son investigaciones basadas en corpus o con corpus, pero no sobre la propia concepción y ejecución de la lingüística de corpus.

De manera incontestable, si no se conoce cuál ha sido la experiencia previa de otros investigadores en otros trabajos, es inviable replicar los aciertos y reparar los errores, lo cual es el principio básico de la investigación científica en cualquier disciplina. Sin embargo, cabe dejar muy claro que la falta de publicaciones y –al menos de manera completa– de protocolos escritos de corpus orales muy probablemente no obedece ni a una voluntad de ocultación (que no sería lógica en trabajadores, como se decía arriba, mayoritariamente altruistas) ni a la desidia por exponer contenido técnico, sino al poco valor otorgado a la construcción del andamiaje y el mucho valor otorgado a su aprovechamiento. Este desequilibrio es muy evidente en la industria editorial, en el propio *scope* de las revistas y en una parte de quienes, en última instancia, se benefician de los corpus para poder realizar sus investigaciones. Es necesaria, en este sentido, una mayor presencia de la lingüística *de* corpus en los foros académicos y un mayor reconocimiento de este trabajo a la hora de evaluar la producción científica en España.

Chapter 7

Conclusions

The doctoral thesis presented here stands as a culmination of the work undertaken during a predoctoral FPI contract, an intensive journey that spanned nearly half a decade and was built upon the foundation of years of personal interest in corpus linguistics. As part of the obligations and responsibilities associated with this contract, we were involved from an early stage in constructing the Ameresco corpus, a project that had been initiated only a few years earlier. Thus, from the outset, our research aims revolved around the methodological challenges and considerations required when building a linguistic corpus, particularly a corpus like Ameresco, which places a spotlight on conversational oral language that has been secretly recorded.

Based on prior experiences and observations, it became evident that there was a lack of uniformity in the various stages of design and construction of the oral corpora available. This observation fueled one of our primary interests: to devise an efficient, standardized, and replicable construction protocol for the Ameresco corpus. To lay the groundwork for this task, our initial steps involved studying the existing corpus panorama. In separate workblocks, we conducted a detailed analysis both of the existing corpora and of their descriptions in the literature, paying especial attention to those based on oral material. Our purpose was to critically assess the strengths, weaknesses and nuances in each phase of design and construction, aiming to shape a working methodology applicable to the Ameresco corpus. In essence, addressing the technical aspects of constructing a spontaneous oral corpus involves dissecting the challenges that emerge during the design and creation of oral corpora and subsequently proposing innovative solutions. These solutions leveraged both technical and methodological tools to address the challenges identified across three pivotal phases: data collection, data processing, and the archiving, distribution, and accessibility of corpus data for the academic community.

To study these aspects, we conducted exploratory analyses of various Spanish corpora with the primary goal of detecting problems in the data collection and digitization phases. This exploration not only paved the way for the creation of a comprehensive descriptive map of the extant Spanish and international oral corpora, but also, allowed for a critical analysis of this corpus landscape after pinpointing the challenges in each design. Additionally, various computer tools for the computational processing of oral corpora were explored, especially those intended for transcription and alignment of audio and video material. These tools were evaluated weighing its pros and cons in relation to the specific requirements of

the Ameresco corpus. based on their advantages and disadvantages in relation to the specific needs of constructing the Ameresco corpus.

Our successive exploratory analyses of Spanish oral corpora detected the main problems experienced in the transcription and encoding phases of the corpus, explored means of information retrieval using computer tools, and culminated in the proposal of a working methodology for the design and construction of the Ameresco corpus across three phases, as detailed in this thesis.

The value of this work, we believe, well beyond the construction of the corpus itself, lies in its novel approach, which revolves around a methodological metareflection in the field of corpus linguistics. Instead of treading the more conventional path of conducting corpus analysis or delving into linguistic variables using corpus data, this thesis pivots its focus towards the methodology of corpus construction, aiming to establish a robust working model. This particular facet has often been overshadowed in academic literature, possibly stemming from the misconception that the true scientific merit lies more on the analyses conducted using corpus data rather than in the intricate, delicate, altruistic and mostly self-taught work involved in designing and standardizing corpus collection. In this sense, we have aspired to design a model that can serve as a reference for other corpora in terms of best practices in oral corpus creation, grounded in transparency, and ensuring that all materials and protocols are readily available, thereby maximizing replicability.

While there are numerous studies *based on* corpora, the same cannot be said for studies offering methodological reflection on the construction of such corpora. The fact that this research was undertaken during the predoctoral phase, as alluded to before, lends it a unique experiential dimension, drawing directly from hands-on experience, extensive fieldwork, and the intricate management and construction of a tangible corpus, Ameresco, which, thanks to this endeavor, is now accessible to the scientific community. Lastly, this work arises from the assertion of the need to value corpus compilation work, both in its technical aspects and in the often overlooked and underrecognized work of conception and preplanning.

7.1. Achievement of the initial research objectives

The research objectives outlined for this work in the introduction were divided into two main blocks: research and methodological, on one hand, and technical /applied, on the other:

From the perspective of theoretical and methodological reflection, this work aimed to:

1. Provide the fundamental methodological foundations for the design and creation of a corpus of conversational oral language, from its conception to its distribution on public networks; in particular, a corpus that aims to collect various varieties and aspires to be a macrocorpus created by multiple teams. This aim was addressed in Chapters 2, 3, and 5, but especially 4.

2. Critically analyze the advantages and disadvantages, strengths, and weaknesses of previous designs of oral corpora as a basis and reference for the proposal presented here. This aim was specifically addressed in Chapter 3.

3. Identify and address all the problems associated with the various phases of designing an oral corpus. Starting from an initial exploration of the different processes involved in designing an oral corpus, from data collection to the incorporation into search engines on the web, the goal is to extract all the problems associated with the design phases. This aim was specifically addressed in Chapter 4.

4. Examine the various technical possibilities for solving problems in the case of the Ameresco corpus in each of the three design and creation phases. This aim was specifically addressed in Chapter 5.

5. Propose solutions in each of these phases, justified based on criteria of cost-effectiveness for their use and ease of access for corpus users, as well as respect for the linguistic-discursive nature of the materials. Both alternatives from the literature and the authors' own experiences in creating the Ameresco corpus will be leveraged for this purpose. This aim was specifically addressed in Chapter 5.

From the applied perspective, this thesis pursued two additional objectives:

6. Considering the objectives 1 to 5, the development of the Ameresco corpus of colloquial conversations in Spanish of America and Spain.

7. Provide the corpus to the scientific community by making it accessible through the internet and building a search interface that achieves a balance between effectiveness and intuitiveness.

After detailing the research objectives and explaining where in the work they have been addressed, the main findings of this thesis will be presented, along with their relevance within the field. Later, a personal insight into the very task of compiling a corpus will be offered, highlighting the value of the metareflection on the construction process and the need for a better consideration of the field.

7.2. Main Findings

After presenting the motivation and justification for this work in accordance with the research objectives related to the methodology for designing and constructing a multidialectal oral corpus of colloquial conversation in Chapter 1, Chapter 2 has focused on the approach to corpus linguistics. It has been shown how a linguistic corpus is defined as a collection of natural, written, or spoken texts that are stored and processed digitally for the purpose of serving as empirical material for language study. There are three fundamental aspects shared by corpus definitions: the texts must be produced in real situations, the collection of texts must be guided by explicit parameters to ensure its constitution and replicability, and finally, the corpora must be available in electronic format for analysis using computational programs. It has also shown the difference between text files, databases, and corpora, based on the fact that the former are repositories of collected texts, while corpora are created with specific criteria of representativeness and sample balance, and databases can be collections of texts that do not necessarily follow the principles of a corpus.

A fundamental issue raised in this chapter is the nature of corpus linguistics as a discipline, as it is considered a theoretical and methodological approach far removed from the generativist view of language based on introspection, and there is a lively debate about whether corpus linguistics is a discipline in its own right or is only a methodology applicable to other linguistic disciplines to obtain various kinds of results, in line with authors who see it as a support for linguistic research by providing both linguistic corpora and computational technology for processing them. Alongside this debate, corpus linguistics faced initial criticisms for allegedly lacking rigor and reliability. These criticisms have faded over time, and corpus linguistics has established itself as a valid and valuable scientific approach since

using real examples in linguistic studies helps to avoid the subjectivity of introspection. In this sense, corpora provide an empirical basis for observation, induction, deduction, verification, and evaluation of linguistic data.

Materials for corpora can be obtained from both samples of natural speech and elicitation techniques, resulting in a range from very natural to highly monitored data. Likewise, the perspective for exploiting corpus data can be corpus-based and corpus-driven, depending on whether, respectively, one starts with prior hypotheses and uses the corpus to support linguistic claims or, in the second case, one starts with the corpus and discovers linguistic patterns without theoretical preconceptions. For the identification of linguistic patterns, as well as for maximizing objectivity in the identification of phenomena and the possibility of quantitative and statistical analysis, corpus linguistics is an essential tool since it allows the use of specific software that facilitates searching, extraction, and classification of data in corpora, enriching linguistic studies.

The development of corpus linguistics is divided into three generations or stages from its beginnings to the present, marked by the emergence of new technological advances and a substantial change in methodology perception. In the first generation, until the mid-20th century, corpus linguistics was not yet called as such, and was based on empirical approaches (i.e., the use of real data) to language, and it was criticized, especially by Chomsky, who favored intuition as a source of knowledge in linguistics. The early corpora were collected on paper, and little attention was paid to representativeness, as analysis was done manually. In the 1960s and 1970s, the second generation emerged. In this stage, there was a technological advance with the greater availability of computers, which allowed for the creation of the first computerized corpora, such as the *Survey of English Usage Corpus* (SEU), the *Brown University Corpus of American English* (Brown Corpus), *the Lancaster-Oslo/Bergen Corpus* (LOB), and the *London-Lund Corpus of Spoken English* (LLC). However, this generation was dominated by written corpora due to technical and transcription difficulties associated with oral data. Finally, in the third generation, starting from the 1980s, corpus linguistics experienced a renaissance thanks to the influence of authors like Leech, who advocated for the use of corpora in combination with linguistic intuition. The development of computational linguistics and the greater availability of advanced technology allowed the creation of macrocorpora, such as the *Bank of English* (COBUILD Corpus) and the *British National Corpus* (BNC) in English, and academic reference corpora such as CREA, CORDE, and CORPES XXI in Spanish. The rise of corpus

linguistics has also been contributed to by the emergence of specialized research associations and centers and scientific journals.

Regarding the characterization of corpora, several criteria are used for their classification. Thus, attention is paid to the medium, according to which there are written, spoken, and multimodal corpora, and in terms of the collected time period, they can be diachronic or synchronic. Depending on the number of languages treated, they are also classified into monolingual, bilingual, multilingual, or parallel corpora, and depending on the specificity of the texts, they can be general or reference and specialized or technical corpora, including training corpora. If size is considered, there are macrocorpora and microcorpora, and they are open if they are updated with new materials or closed if they have a fixed sample. Finally, according to the additional information they provide, some are annotated for subsequent automated analysis, while others are simple or pure without annotations. All these classification criteria can be combined to characterize corpora more precisely.

The creation and development of corpora have led to a revolution in almost all branches of linguistics, given that corpora have become a tool for the grounded study and analysis of linguistic facts. Despite the effort and cost of collecting oral corpora, compared to processing written corpora, there is no doubt that the effort put in collecting oral samples is worthwhile. Their utility for applied studies in the various manifestations of language variation is undeniable, but they are especially useful for studying real language, the context in which speech is realized, given the nature of the material provided. Having language samples produced in a natural context not only allows to study phonetic, phonological, morphosyntactic and lexical variation with accuracy and realism, but also to address other types of variations that affect language but are external to it, such as geographical, social, and situational variation. In this sense, access to oral corpora is a most useful resource in language teaching and learning, specialized discourse and terminology, and, especially, pragmatic studies, the area we will focus on as it is the specific objective of the construction of the Ameresco corpus.

In particular, the collection of spontaneous colloquial conversation oral corpora has favored research and analysis of conversation and the ethnographic aspects of conversation revolving around it. The most conversational a sample, the better possibilities for pragmatic study, since conversation is the most pragmatic and free mode of communication in terms of interactional behavior.

Furthermore, many linguistic phenomena, such as discourse markers, interjections, vocatives, or hesitation markers, which have challenged grammatical analysis due to their formal and functional properties, are mainly —and even only— found in oral production. Thus, with the emergence and development of oral corpora with context, the study and characterization of these phenomena have expanded, allowing for a deeper understanding of their true functional behavior in a way that was impossible in the past. Likewise, the fact of conversational oral corpora allowing for the analysis of the genre 'conversation' itself was the starting point and focus of Conversation Analysis studies, which led to an exponential development of the various facets involved in interaction, both from a structural and sociological point of view. The study of turn-taking dynamics, conversation structure, discourse units, natural speech phenomena such as overlaps, restarts, interruptions, etc., not only allows us to understand how free interactional management is among conversation participants but also, helps us figure out how other discursive genres are managed, and what kind of interpersonal relationships underlie such conversational management.

It should be noted that other methods have been used for the study of pragmatic phenomena, in addition to natural spoken language corpora, such as social habit tests, role-plays, surveys, and discourse completion task questionnaires. The validity and adequacy of these methods depends on factors like the research aims and or the presence of limitations. Thus, in studies aiming to know the speakers' opinions or intuitions on a particular matter, the use of surveys or questionnaires might be a better, faster and more adequate alternative to the use or compilation of corpora.

In short, since the materials collected in corpora were not originally produced for research purposes, they provide authentic, purely empirical material, fully suitable for pragmatic analysis. These corpora often offer sociolinguistic information about the participants, providing researchers with firsthand knowledge about the situational parameters for the appropriate interpretation of the meanings implicit in speech, the coordinates that allow us to detect the frequent mismatch between linguistic forms and communicative functions, the role of the interlocutors in communication, the social values that underly social actions like speech acts, etc.

In this sense, it is worth mentioning the recent development of corpus pragmatics, a new subdiscipline that combines the objectives of corpus linguistics (of a more quantitative nature) and pragmatics (initially, more qualitative), which were considered mutually

exclusive for a long time. This relatively young line of research allows for a thorough and reliable study of pragmatic phenomena.

Chapter 3 has focused on the development of oral corpora in the specific context of Spanish. As in other traditions, these corpora have their roots in dialectology and linguistic atlases, which favor empirical observation of data over intuition and showed the inadequacy of surveys as the sole data collection method. Corpus linguistics in Spanish started somewhat later compared to the Anglo-Saxon world, but it underwent rapid development, culminating in the first milestone of the *Programa Interamericano de Lingüística y Enseñanza de Idiomas* (PILEI), initiated in the 1960s, which sought to study speech in large urban concentrations in the Americas. Corpora from various locations in Latin America and Spain were created, although many are not published, or they are not publicly available in their entirety.

One of the first corpora to provide access to audio in CD-ROM format was the *Macrocorpus de la Norma Culta de las Principales Ciudades de América y España* (MC-NC), which includes individual interviews that have been reviewed and standardized and was later incorporated into the *Corpus del Español Contemporáneo* (Corpus of Contemporary Spanish - CREA) of the Real Academia Española and Mark Davies' *Corpus del Español* (Corpus of Spanish- CE). Alongside PILEI, another of the most relevant pioneering projects is the *Macrocorpus para el Estudio Gramatical del Español Hablado en América* (Macrocorpus for the Grammatical Study of Spoken Spanish in America-EGREHA), which contains materials from PILEI and MC-NC, with most of its materials still unpublished, except for North America and part of the Caribbean area. In the late 1990s, the most important sociolinguistic study project of Spanish in Spain and America, PRESEEA, was born under the direction of Francisco Moreno Fernández. The main outcome of PRESEEA is the creation of a macrocorpus that collects semi-directed interviews in different Spanish-speaking cities from Europe and America, in the format of enriched orthographic transcription and TEI XML tagging.

The first spoken corpora in Spanish were all private initiatives, but later, academic corpora such as the CREA and the *Corpus del Español del Siglo* XXI (Corpus of 21st Century Spanish - CORPES XXI) emerged. CREA was published in 1998 with both written and oral material, although without access to audio for the latter. In contrast, CORPES XXI, published in 2013, takes a step further in what regards the availability of oral material in Spanish for linguistic research by improving geolinguistic representativeness and allowing

access to audio recordings, as its oral materials come from agreements with media, YouTube, etc., and other corpora that have contributed their materials.

Bearing in mind the unavoidable limitation regarding the fact that it is impossible to collect all corpora in Spanish without missing some, and moreover, considering that any compilation is doomed to its own obsolescence almost immediately due to the speed of computer advances, this chapter includes some of the main works that offer an overview of oral corpora in Spanish. Moreno Fernández (2005) focuses on the collection process of corpora, highlighting the importance of considering both geolinguistic varieties of Spanish and the representativeness of the collected materials, and distinguishes between corpora for speech technology development and those for general linguistic study of oral language, the latter being divided between general oral corpora and specialized corpora for linguistic studies. Additionally, in line with his original interest in sociolinguistics, he emphasizes the importance of considering appropriate criteria for the selection of speakers and the need to adhere to a pre-established typology. A few years later, the work of Briz and Albelda (2009), conceived as part of an initiative of the Cervantes Institute, aims to provide an overview of spoken and written language corpora in Spanish developed or in progress during the first decade of the 21st century. It includes both spoken and written corpora, classified considering their purpose (general or specific), their size (macrocorpus or microcorpus), format (textual or electronic search engine), and mode of access (access to the complete text or only through concordances). The work also includes other types of corpora, such as those intended for speech technology development, technical languages, and language acquisition and development. Ten years later, Briz and Carcelén (2019) updated the collection of Spanish oral corpora, also for the Cervantes Institute, classifying them into pan-hispanic corpora, corpora of specific geographical varieties, and other online repositories without access to audio. Alongside these collections, there are others that are more partial or less exhaustive, such as Enghels, Vanderschueren, and Bouzouita (2015), Rojo (2016), Solís (2018), Parodi and Burdiles (2019), or Llisterri (2021), this latter especially exhaustive in what regards the phonic approach to the material.

Comparing the aforementioned compilations of oral corpora allows for a study on the degree of consensus found in collection works. This comparison illustrates how, unsurprisingly, the most well-known and relevant corpora in the Hispanic context seem to be corpora with general linguistic purposes and with a broad spectrum of varieties, but there is one unexpected exception, the Val.Es.Co. corpus, that appears in the greatest number of

compilations despite its limitations in dialect (only Spanish spoken in Valencia, Spain) and genre (colloquial conversations). The importance of this corpus probably lies on its originality and on the adequacy of the oral spontaneous material collected for the purposes of pragmatic research, as we mentioned above.

The analysis of several compilations has proven useful not only for understanding the current landscape of Spanish oral corpora and for providing researchers with a good number of linguistic resources, but also, primarily, and in line with the research objectives of this thesis, for identifying possible areas for improvement in the collection of these resources, something that has been used to our benefit in the design of the Ameresco corpus. The compilatory works reviewed play a crucial role, in providing an up-to-date overview of the state of corpus research, but they cannot include all existing corpora and only reflect a snapshot at the time of their publication, partly due to the lack of access to smaller corpora or to those created by researchers without the backing of large institutions. The works of Briz and Albelda (2009) and Llisterri (2021) can be considered the most comprehensive in two different senses: the first for its thoroughness and the second for the meticulousness and technical information about transcription, annotation, and further technical aspects that are often disregarded or only barely mentioned in other works.

One of the problems identified in the collections, which highlights general deficiencies in the discipline, is the ambiguity in terminology and the lack of clear definitions for terms such as microcorpus and macrocorpus or spoken corpus vs. spoken language corpus. This ambiguity is often caused by the fact that the works focus on different aspects of Spanish oral corpora, resulting in terminological and classification differences. Consequently, there is a need to clarify and define specific criteria for classifying oral corpora. Therefore, in response to the challenges and ambiguities identified in previous works and taking into account the aforementioned impossibility of collecting all current corpora, an updated review of the Spanish oral corpus landscape has been presented, and a selection of corpora has been defined in more detail which was made according to the application of various criteria: (a) the medium, as only oral corpora are selected, regardless of whether they include phonetic transcription or not; (b) their synchronic nature; free and non-commercial access; and (c) the absence of restrictions regarding discourse genre, completion status, or geolinguistic variety of collected Spanish. Furthermore, only corpora that allow online access to the data are considered.

Thus, alphabetically, detailed information is provided about the following corpora *America y España Español Coloquial* - Ameresco (that, being the basis of our study, is developed in Chapter 5), *Corpus del Español* (Corpus of Spanish- CE); *Corpus del Español Mexicano Contemporáneo* (Corpus of Contemporary Mexican Spanish - CEMC and CEMC II), *Corpus del español en Texas*- CET, *Corpus Oral Juvenil del Español de Mallorca* (Youth Oral Corpus of Spanish in Mallorca – COJEM), *Corpus Oral de Lenguaje Adolescente* (Oral Corpus of Teenage Language - COLA), including Madrid (COLAM), Santiago de Chile (COLAS) and Buenos Aires (COLABA), *Corpus Oral de la Lengua Hablada en Honduras* (Oral Corpus of the Spoken Language in Honduras – COLEH), *Corpus Oral de la Lengua Española en Montreal* (Oral Corpus of the Spanish Language in Montreal- COLEM), *Corpus Oral Didáctico Anotado Lingüísticamente* (Didactic Oral Corpus Annotated Linguistically C-Or-DIAL), *Corpus Oral de Referencia de la Lengua Española Contemporánea* (Reference Oral Corpus of Contemporary Spanish Language -CORLEC), CORPES XXI, *Corpus del Español en los Estados Unidos* (Corpus of Spanish in the United States - CORPEUU), the *Corpus del Habla de Almería* (Speech Corpus of Almería), *Corpus Oral y Sonoro del Español Rural* (Oral and Sound Corpus of Rural Spanish- COSER), CREA, *El español hablado en Bogotá* (Spoken Spanish in Bogotá), ESLORA, *Macrosintaxis del Español Actual* (Macrosyntax of Current Spanish – MESA), PRESEEA, *Valencia Español Coloquial* (Valencia Colloquial Spanish- Val.Es.Co)., and *Voices of Hispanic World*- VHW.

Chapter 4 delved into the very process of construction of oral language corpora. Such task requires a meticulous process of planning and a prior reflection to ensure that the collected data are representative and useful for analysis. Although the existing literature addresses the general principles that any good corpus design, whether written or oral, should follow, there is no protocol that dictates step by step how to create a corpus. In general, five fundamental stages have been commonly established in corpus construction: under different names, the tasks involved in these stages are related to recording, transcription, representation or markup, encoding or annotation, and application.

When considering the design of a corpus, the main principle relates to the the research aim. The purpose of its creation depends on the type of linguistic analysis intended and the research questions guiding the corpus collection. Secondly, it is necessary to establish the corpus representativeness, considering the selection criteria to be followed in order to obtain a sample that reflects diversity in terms of diatopic, diastratic, or diafasic variation. Closely

related to this latter aspect is the final size of the corpus. For example, corpora for specific research, such as a doctoral thesis, with limited or no funding at all and involving a single person must necessarily end up being smaller a corpus promoted by a research group or institution, with more technical, human, and financial resources to undertake its construction.

Another fundamental aspect relates to the materials to be collected and the legal implications associated with those materials. In the case of written corpora, issues related to the copyright of the files that will constitute the sample must be considered. In the case of oral corpora, this point becomes more complex, as it involves knowledge of and compliance with privacy and data protection legislation.

Lastly, considerations related to data storage and preservation must be addressed, as well as decisions on how the users will access the corpus and under what conditions; whether the collected materials can be openly shared or if access should be restricted, etc. This latter aspect is particularly relevant if the corpus contains recordings made in private or confidential settings, such as a medical or healthcare environment. Such recordings, even when anonymized, cannot legally be accessible to the general public but are restricted to the research group. Again, this final aspect depends on the context in which each corpus arises, since generally, more resources available also lead to better conditions for corpus access and distribution.

Focusing exclusively on oral corpora, an additional step must be considered. While the inclusion of written materials in a corpus is more straightforward—for example, collecting digital news articles, encoding them for computational processing, and depositing this material in an online search engine—, in the case of oral corpora, before materials can be incorporated into the electronic platform, the materials must be transcribed and labelled. Therefore, questions arise about how the particular phenomena of spoken language will be represented in writing, how comprehensive the transcription and encoding will be, or how metadata and contextual information about the recording circumstances and participants will be collected, including permissions and informed consents from individuals participating in the recordings.

In the present day, thanks to the rapid advances in technology and artificial intelligence, numerous tools have emerged that facilitate automatic transcription. These tools are highly effective in specific cases of recordings with high situational control, as well as in speech recognition and synthesis, in formal monological discourse, or in cases with optimal

recording quality, such as television, radio, or professional content creators. However, these transcription tools still have limitations when compared to human ability. Furthermore, transcription of spoken language, which is inherently lengthy and costly, becomes more challenging as the level of transcription detail increases and, generally, large corpora continue to be transcribed manually or are at least completed manually.

The aforementioned division of the corpus construction into five fundamental stages has been reconsidered here as a three-stage process: Phase 1, of conception and collection, where the corpus is designed, permissions are obtained, and physical data captured; Phase 2, of processing the collected materials, which includes planning and preparation of the transcription and encoding system, the anonymization of materials, corpus processing, and annotation options; and Phase 3 (which may or may not occur, depending on the human, contextual and funding resources) involving the dissemination of corpus data through platforms for exploitation.

In the first phase of data collection, one must reflect on the characteristics that the corpus should have according to the research aims it responds to. Aspects such as size, representativeness, or sample balance determine the selection of the materials to be collected. Bearing this last idea in mind, there is an inherent reflection in the design and conception of a corpus consisting of studying and calculating in advance the costs of time, personnel, and financing, as well as in foreseeing the availability of means, instruments, and technical knowledge to carry out this work. This anticipated reflection must be carried out to assess the degree of feasibility of the data collection, the planning and timing of the work, as well as the preparation of the necessary material for its execution, and to make sure that such material does not exist already, making it unnecessary to collect the corpus.

The fundamental and general criterion that guides any corpus design is, as we have seen, its purpose of use in research, and from this, other questions arise that the researcher must ask. The research purpose implies determining the phenomenon to be studied and evaluating its various dimensions in relation to the type of materials that will be most suitable for its research. These dimensions include the linguistic nature of the data (phonic, prosodic, morphosyntactic, lexical, discursive, etc.); The variationist nature of the data, be they diastratic or sociolectal (age, sex, level of instruction), diaphasic or situational (relationship between speakers, degree of familiarity of the physical recording environment, speech topics, interpersonal or transactional purpose), and diatopic or geolectal (circumscription to particular dialectal areas, jargons, specific language vs. general language); The discursive

genre, be it monologal (talk, class, lecture, humor monologue, etc.), dialogal (conversation, interview, debate, work meeting, commercial transaction, tutoring, medical consultation, press conference, etc.) or computer-mediated discourse (video conference, spontaneous messaging audios, etc.). Further aspects must be considered like the need for discursive context, the impact of interaction and type of recipient or a forecast of the frequency index of the phenomenon studied, the required degree of naturalness of the data or the need for image/video.

Thus, for a hypothetical study on the functions of a given interjection, it would be necessary to determine whether the prosodic, grammatical, and pragmatic study of this interjection is to be carried out, for instance, among young Chileans or in general standard Spanish; it should also be considered the convenience of a conversational corpus, given the dialogic character of interjections; the researchers should assess to what extent contextual data is needed since, considering the nature of interjections, these can be formulated in response to various external stimuli, meaning that the discursive genre for the data would have to be well selected, as there would be genres in which the interjection would be practically absent.

After evaluating the previous characteristics, the design requires advancing to further preliminary decisions. These involve questions about the size of the corpus (number of speakers, duration of recordings, amount of speech segments needed for each speaker), the representativeness or number of samples needed based on the source population, the balance or variety of samples required, and the characteristics of the recordings. This includes considerations about the collection time span, technical aspects of the recording, and the treatment of contextual information (metadata). Finally, special attention must be paid to ethical and legal issues concerning the copyright of the materials, consent from the speakers, and the various laws in place in different countries.

Using oral corpora introduces a challenge to the planning process: the Labovian observer's paradox suggests that researchers should find ways to observe natural speech without the speaker's knowledge to obtain genuine and reliable data. This implies that subjects should be unaware they are being observed. However, systematic observation is essential to gather this data, leading to an ethical dilemma. Another legal paradox emerges when considering that researchers should capture natural speech samples without the speaker's awareness. Not informing subjects that they are being recorded conflicts with laws

protecting privacy, personal image, and data protection rights, resulting in the impossibility of recording.

While valid anonymization methods exist in many research fields, a standard doesn't appear to be established in linguistics. Early oral corpora were produced without prioritizing participants' privacy. For example, in the PRESEEA corpus, participants were aware they were being recorded. In contrast, for the Val.Es.Co. corpus, which captured spontaneous conversations, authorization was obtained post-recording. Although ethical measures were implemented in early corpus collections, they lacked profound reflection on their underlying reasons. Fortunately, many previous mistakes in corpus linguistics regarding ethical concerns have been addressed. Currently, there's a consensus that oral corpus collection should involve obtaining speakers' authorization and informed consent. Data protection is imperative, and standards should vary based on the data type. Focusing on Spain's regulations, Organic Law 1/1982 protects fundamental rights to honor, personal and family privacy, and individual image. Any unauthorized breach of these rights without explicit legal approval or clear consent from the impacted individual is punishable. Recording someone unknowingly can be a significant privacy infringement. While permissible in certain situations, in most cases, it's a severe privacy violation.

Moreover, recording spontaneous informal conversations often captures personal data. The European Regulation 2016/679 and Organic Law 3/2018 outline the treatment of personal data, emphasizing that data protection principles apply to any information linked to an identifiable individual. Anonymized data, where all identifying elements are removed, aren't deemed personal data. However, for oral corpora, full anonymization is challenging, especially if the corpus is meant for public access.

Two primary strategies address this dilemma. The first is to inform participants they're being recorded, even if it compromises spontaneity. The second strategy employs an informed consent model, allowing natural linguistic sampling while complying with legal regulations, like the informed consent crafted in 2015 for the Ameresco project, updated as per changing needs and laws.

Regarding informed consent, it is vital that participants give their authorization willingly, having been adequately informed, and without ambiguity. The regulation stresses transparency, ensuring informants understand the specific purposes for which their data will be used. They should also recognize their right to retract consent at any moment.

Regarding phase 2, which deals with data processing, this chapter discusses the terminological hesitation when characterizing different methods of representing oral language in writing. An operational denomination has been established for each of these methods (transliteration, transcription, coding, and annotation) to facilitate the description and recognition of the main existing systems. After characterizing them, a description of each method is provided, along with references to the main initiatives and software programs that enable the computational processing of the collected data.

The literature review in this research highlights the use of various terms to describe the different stages involved in processing collected data, which initially exist in audio format but are transformed into written form for computer processing. The distinction between *transliteration, transcription, encoding, labeling, marking*, and *annotation* is not always clear. Transliteration involves converting oral material into written form following grammatical and normative orthographic conventions, including punctuation. Transcription, however, captures oral characteristics in writing, typically without punctuation to avoid introducing subjectivity. Encoding assigns labels, marks, or codes to elements in a corpus to facilitate computer-based search and analysis. Annotation adds additional linguistic information to enrich the data, often related to grammar, semantics, or lexicon.

In relation to the decision of which transcription system to adopt to convert oral data into written form, several factors must be considered. Firstly, it is necessary to determine which aspects of orality one wishes to reflect and in what level of detail. This depends on the purpose of the corpus and may vary depending on the analytical events one intends to represent. In the literature, some requirements for an adequate transcription system have been established, such as defining clear categories, using symbols that are accessible and easy to learn, employing widely available characters, being economical in representation, and being adaptable to allow users to introduce their own transcription categories.

Transcribing orality is a complex process that involves converting oral grammar into written form, which can result in information loss. Therefore, a balance must be struck between fidelity to orality, system economy, and data processing ease. Several models of transcription for oral materials have been proposed, including phonetic and phonological transcription, transliteration, and enriched transcription.

Phonetic transcription is used to accurately capture the pronunciation of speech sounds, using systems like the International Phonetic Alphabet (IPA) or SAMPA, or like ToBI,

SAMPROSA, and INTSINT, used to represent intonation. In transliteration, oral information is converted into writing without including special marks or codes. It adheres to the spelling and grammatical rules of the language and includes punctuation. This approach focuses on lexical analysis and is not as concerned with phonetic or pragmatic phenomena. It has the advantage of being easy to read but does not fully reflect the characteristics of orality. In enriched transcription, semi-orthographic marks are added to represent aspects of orality, such as intonation, overlaps, and elongations. This approach allows for a more accurate representation of orality but may be less suitable for computer processing.

The encoding of oral corpora involves assigning labels or codes to specific elements in the corpus to facilitate their search and analysis. This can be extratextual (information outside the text) or intratextual (internal text structure). Extratextual encoding includes bibliographic information, speaker data, and technical details of the material. Intratextual encoding refers to the structure of the transcription and can include labels to mark different linguistic phenomena or conventions.

International standards for the encoding of oral corpora have been proposed, such as the Text Encoding Initiative (TEI), which uses XML tags to represent transcriptions of spoken interactions. Other standards include the Network of European Reference Corpora (NERC) and the Expert Advisory Group on Language Engineering Standards (EAGLES). The encoding of oral corpora makes data more accessible and reusable, facilitating their analysis and computer processing. Programs like ELAN Annotation and EXMARaLDA are used to carry out the encoding and transcription of oral corpora in a synchronized manner.

Annotating an oral corpus, i.e., providing additional linguistic information beyond transcription and encoding, can be done in several ways, depending on research objectives and specific needs. These annotations vary in the level of linguistic analysis they involve. Common types of annotation include tokenization, lemmatization, part of speech tagging (POS), syntactic parsing, and semantic analysis, as well as other typologies such as pragmatics, stylistics, and error annotation in learner corpora. Tokenization involves identifying and characterizing each grammatical element in the analyzed sequence, resulting in a list of all the words in the transcription. Lemmatization allows researchers to extract and examine all variants of a lemma (base form) without listing all possible inflected forms, enhancing text analysis and topic identification. Part of speech tagging (POS) assigns grammatical categories to words, often automatically, with potential manual post-editing. It aids in various applications, from disambiguating homographs to studying word class

occurrences in a corpus. Syntactic parsing adds labels to indicate the syntactic structure of a segment, facilitating natural language processing tasks, such as clause type analysis. Finally, semantic annotation assigns codes to words based on their semantic function, marking either semantic relations within constituents of a sentence or the semantic characteristics of words in a text.

These annotation layers are typically generated using computer programs, although manual review and disambiguation, especially for morphological, syntactic, and semantic annotation, are common and recommended, especially for oral, informal material. Tools for these tasks are available in programming languages like Python, R, or C++, and online platforms like Linguakit (for Spanish) and Xiada (for Galician) perform these processes.

Phase 3, the final phase to consider in designing an oral corpus, involves archiving, distribution, and user access to the corpus. This includes decisions on how to store the materials to ensure long-term preservation, how to make the corpus known to the scientific community, such as through networks or initiatives that promote its dissemination, and how researchers and users will access the collected materials for linguistic analysis.

The most significant factor influencing decisions in this phase is often economic, as hosting and archiving an online corpus that provides access to materials require investment and technical maintenance by specialized personnel. The extent of these implications depends on whether only transcriptions are uploaded online or if audio materials are also included, which significantly increases storage space requirements. Additionally, if both audio and transcriptions are to be provided, a search engine needs to be created to filter and process the previously encoded and/or annotated material. Each added feature incurs space and economic costs, and the more options provided to users, the greater the economic and personnel needs. Free resources and platforms are not always viable due to privacy commitments related to the data collected, which depends on the type of corpus constructed. Researchers and creators of the corpus should be capable of ensuring its long-term computational maintenance to reduce dependence on external hiring, considering potential changes in personnel and the evolving nature of technology.

When contemplating open public access to the corpus, several factors must be considered. Firstly, the format in which materials will be provided, ideally using non-commercial formats like XML or plain text to ensure compatibility with international standards. However, practical storage needs may require alternative formats. Secondly, compliance

with data protection agreements and copyright rights is essential to secure necessary permissions from participants or rights holders before publicly sharing the data collected from sources like radio, television, or social media.

In the second part of this chapter, after the theoretical reflection on corpus construction and its impact on the final design, a comparative analysis of transcription and coding systems of the most important Spanish oral corpora is conducted. However, for a more comprehensive analysis, C-Oral-Rom and CORLEC have been omitted due to their unavailability in open format or lack of an online platform for data retrieval. Instead, two other corpora, the *Corpus Oral y Sonoro del Español Rural* (COSER) and the *Corpus del Español del siglo XXI* (CORPES XXI), which were not included in the table of corpora with the most consensus, were selected as they were published after the compilation of previous corpora.

In this context, three blocks of comparative review are established: external factors (corpus aims, discursive genre, representativeness criteria, legal aspects); internal factors (transcription system, data coding, and anonymization); and factors related to data access. This comprehensive view of decisions made by other Spanish oral corpora serves as the basis for the configuration of the Ameresco corpus, identifying strengths in design that could be utilized and weaknesses that could be improved in its subsequent application.

In what regards the external factors, the decisions are detailed of the studied corpora regarding aspects related to the data collection phase, the research aims of each corpus, the discourse genres they collect, whether they follow any criteria for sample representativeness, and the legal aspects they need to consider and how they have addressed them. Looking at the data collection objectives of the corpora considered for this analysis, CORPES XXI serves as a reference corpus, and its objective is to describe the state of the Spanish language from a pan-Hispanic perspective. PRESEEA also aims to be a pan-Hispanic sociolinguistically representative corpus. The COLA and COSER corpora are multidialectal, with the former primarily studying adolescent language and the latter focusing on rural speech to document grammar-related phenomena and morphosyntactic variation. Finally, the Val.Es.Co. (2002) corpus is a monodialectal corpus for the study of Spanish colloquial conversation.

Depending on the discourse genre of the corpus, CORPES XXI, being a reference corpus, includes a variety of genres, whereas PRESEEA and COSER focus exclusively on semi-

directed interviews, and COLA and Val.Es.Co. (2002) opt for collecting informal spontaneous colloquial conversation, with Val.Es.Co. doing so secretly.

The representativeness criteria of the academic corpus, CORPES XXI, is based on the origin and medium of the materials and does not aim to be representative in a sociolinguistic sense. However, PRESEEA and Val.Es.Co. were built with considerations for stratification by gender, age groups, and education level. In the case of the COLA corpus, which exclusively collects speech samples from adolescents, the representativeness criteria are based on gender and age, collecting recordings from the age group of 13 to 19. As for the COSER corpus, only one speaker per selected enclave, who is typically elderly, and native to the area.

The legal aspects of CORPES XXI involve signing agreements with media or safeguarding copyright, but no consent from the speakers is required, since the oral material collected in this corpus is incorporated based on the original corpus's data collection premises, and permissions were obtained at the time by the group that managed it initially. In the case of COLA, participants have signed informed consent, although for underage informants, these were signed by their parents or guardians. For the COSER, PRESEEA, and Val.Es.Co. (2002) corpora, the informants have given their consent at some point before or after recording, although it is not explicitly mentioned in some cases.

In what regards internal factors, related to the data processing phase, we include the description of the methodology adopted by the selected corpora regarding their transcription and coding systems, the use of conventional or non-conventional spelling, phonetic marks, marks reflecting noise, laughter, and functional elements, marks of orality, lexical marks, transcription marks, and anonymization.

The international TEI standard forms the basis for the transcription and coding systems of the COLA, CORPES XXI, and PRESEEA corpora, while COSER and Val.Es.Co. develop their proprietary systems. Regarding the use of spelling criteria in transcription, as a general trend, except for COSER, the transcription tends to break the rules of writing.

The COLA corpus is transcribed orthographically, but punctuation marks are not used, as they are considered an interpretation of the text. CORPES XXI behaves similarly, although in this corpus, interrogative and exclamatory structures are indicated. PRESEEA states that it uses conventional normative spelling, including accentuation, without punctuation marks except for questions and exclamations. Like CORPES XXI, it does not use capital letters

except for proper names and acronyms. This decision makes sense since the PRESEEA corpus has provided some of its materials to the academic corpus and, therefore, seeks compatibility. Val.Es.Co. (2002) opts for conventional spelling with exceptions, without punctuation marks except for questions and exclamations, and COSER is oriented towards the study of morphosyntactic variation, so the transcription of materials is carried out following conventional spelling conventions.

Phonetic Marks, such as pauses and silences, hesitations, self-interruptions, and marked pronunciations such as emphasis or whispering are recorded in COLA. CORPES XXI, on the other hand, has defined a series of phonetic marks that include pauses, silences, cut-off words, and hesitations, whereas COSER uses a minimal set of marks to reflect pauses, silences, and types of pronunciation. PRESEEA uses several phonetic marks, including pauses, silences, cut-off words, hesitations, elongations, and emphasis, and Val.Es.Co. (2002) marks phonetic features such as pauses, silences, intonation, and emphasis.

In COLA, there are two types of marks to reflect noise and laughter in the transcribed materials. CORPES XXI uses three types of marks to indicate noise, laughter, and functional elements in the transcribed materials, just like COSER, PRESEEA and Val.Es.Co. COLA does not specify the use of marks of orality on its website, so it is unclear whether they use any specific symbols or marks to indicate features of spoken language.

CORPES XXI, COSER, PRESEEA, and Val.Es.Co. (2002) use four types of marks to indicate features of spoken language that are characteristic of orality, like fillers, repetitions, non-speech events, and substitutions.

COLA, CORPES XXI, COSER, and PRESEEA do not specify the use of lexical marks on their respective websites, so it is unclear whether they use any specific symbols or marks to indicate lexical features. Val.Es.Co. (2002) uses the symbol [NVL] to indicate neologisms or words of non-standard or regional usage.

Prosodic marks, such as stress, pitch, and rhythm are not reflected in COLA, CORPES XXI, COSER and PRESEEA, and Val.Es.Co. (2002) uses only tonemes.

In addition to the marks discussed above, transcription corpora may use other symbols or marks to indicate various linguistic features or phenomena. COLA, COSER, and PRESEEA do not specify the use of other marks on their respective websites, so it is unclear whether they use any additional symbols or marks to indicate linguistic features or phenomena. CORPES XXI uses several additional symbols or marks in its transcriptions, including

annotations, code-switching, disfluencies, paraphrases, etc., whilst Val.Es.Co. (2002) also uses several additional symbols or marks, including code-switching, metalinguistic comments, punctuation, repetitions, and words of foreign origin.

Since the specific symbols and marks used in transcription corpora can vary, it is essential to refer to the documentation provided by each corpus for precise details on their transcription conventions.

Finally, the third stage considers the ways to access data from each corpus. Except for the Val.Es.Co. (2002) corpus, which was published in print and can only be accessed by reading the transcriptions, the rest of the analyzed corpora have an open online search engine available to the public. In the case of COLA, users can access transcriptions along with corresponding audio fragments online and features a search engine that retrieves information through concordances, including access to audio fragments. CORPES XXI has a search engine that filters results through concordances, and in some cases, users can access linked audio for their searches. Due to copyright reasons, these materials cannot be downloaded beyond search results. The COSER corpus also retrieves information through concordances and offers the option to listen to complete audio recordings, although downloading the audio is not possible. PRESEEA retrieves searches through concordances via its search platform and allows for the download of audio and complete transcriptions, including metadata, in their entirety.

Chapter 5 constitutes the core of this work and explains how objectives 1 to 7, outlined in Chapter 1, have been achieved. It presents the theoretical and methodological foundations for the construction of a multidialectal corpus of informal conversation, the Ameresco corpus, and outlines the construction process and its availability.

First, a general characterization of the corpus is provided, focusing on the project's origins and development framework, initial research objective (pragmatic attenuation variation in colloquial Spanish conversation in different dialects), and the working groups involved in its construction. The Ameresco corpus, as referenced by Albelda and Estellés online, was conceived to gather authentic speech samples for the study and characterization of colloquial Spanish across its various dialectal varieties, as highlighted by Briz in his works from 1995, 2001, and 2016. This corpus specifically focuses on prototypical colloquial conversations that were secretly recorded in major cities across Latin America and Spain.

As highlighted in research objective number 6 (referenced in Chapter 1), the outcome of this thesis indicates that the corpus is in an advanced stage of construction. As of October 2023, it represents 9 countries, covering 14 cities with a total of 742,170 words included in the search engine. However, there are more recordings that have yet to be processed. The collection of materials is ongoing, with plans to incorporate new research teams from countries not yet represented, thus envisioning Ameresco to be a pan-Hispanic corpus.

Next, the fundamental principles that have shaped the final design of Ameresco are presented, considering the three phases mentioned in previous chapters. Regarding the first phase of collecting oral material for the corpus, the criteria for selecting speakers and the sample size are outlined, following the methodology inherited from Val.Es.Co. (2002) and PRESEEA. For the Ameresco corpus, a sample of 72 speakers divided into three sociolinguistic strata (gender, age group, and sociocultural level) has been established. In terms of informed consent collection, the approach is based on the precepts established in the previous chapter to implement a three-step consent process that complies with the legal requirements of Spanish regulations and the institution hosting the project, the Universitat de València. This consent process ensures that the observer's paradox is avoided, and natural samples are obtained while also meeting the requirements of personal data processing laws, which allow participants to exercise their rights to withdraw their data from the project.

For the recording collection, various factors that affect the recording process are discussed, including the selection of a suitable location with some control over external elements such as noise and unexpected speakers and the behavior of the person collecting conversations, considering the characteristics of colloquial conversation. The collection of metadata technical sheets, essential for providing context to the audio material for subsequent pragmatic analysis, is also addressed.

In phase 2, the existence of two working modes and the circumstances leading to this situation are explained, where some teams could not perform data processing using audio-text alignment and had to work more traditionally through transcription using word processors. However, both modes share the phase 1 of work but differ in the coding system adopted: XML for computerized processing and enriched transcription for word processor-based work. In both cases, a list of signs, symbols, marks, and tags used in the Ameresco corpus is included, along with usage information and examples extracted from the corpus for characterization. The protocol for material review, validation, and anonymization is also detailed.

Regarding phase 3, after the general characterization of the main platforms available for hosting the corpus, different options for viewing and analyzing data from the Ameresco corpus via the website are described. Two versions have been implemented: a basic consultation web interface for downloading complete materials and conducting simple searches by intervention, and a more advanced option called Oralstats Aroca, which allows for transforming and analyzing corpus data relationally. This advanced option is more precise for prosodic analysis and conversation analysis than the basic consultation web interface.

In this section, a software tool is presented, Oralstats Aroca, developed by the linguists of the Ameresco project (Cabedo and Carcelén, 2022) to manage, process and analyze the corpus. In this regard, the tool is articulated as a computer environment that allows transforming and analyzing data from any verbal interaction, focusing especially on conversations collected in conjunction with a set of sociolinguistic metadata (age, level of education, and gender).

Oralstats Aroca is a dynamic computer tool for exploring speech data, a data mining processor with several options available. It can display frequent concordances and measures of lexical, morphological, prosodic, and positional variables with an independent variable as input (speaker, gender, age, city). This new approach involves cross-referencing all the collected information and combining it into a single analysis, with the possibility of it being both exploratory and inferential. For this purpose, data can be related in a common variability ecosystem, with a set of automatically generated numerical and categorical variables; the latter can be selected from a closed repertoire (gender, age, speaker...) or even designed and customized according to specific research needs (study of mitigation, (im)politeness, humor, irony...).The free computational system developed with R (Cabedo Nebot, 2021) from which Oralstats Aroca stems, Oralstats, allows analyzing transcriptions aligned with audio through time codes, considering common factors in corpus analysis, such as general word frequency, parts of speech categories, and bigrams or trigrams, but also lesser-known variables in the field, such as tone, intensity, or duration. Therefore, data from transcriptions are automatically enriched with annotations from other levels, such as morphosyntax or prosody.

Oralstats is situated within the framework of the application of new technologies and the use of programming languages. The ultimate goal is to offer a user-friendly tool to access the corpus and, with open-source code, to allow experts with more advanced IT knowledge to customize the code and add new layers of information to the existing ones. This was

precisely the case in Ameresco since, given the specific needs of a conversational corpus, the base code of Oralstats was modified and simplified, especially regarding the visualization and consultation module of the internal database. In modifying the original base-code, Oralstats Aroca emerged to emphasize more the user experience and less the specific transformation and the performance of inferential statistical tests (heat maps, decision trees, etc.), which were the main aims of the original version, Oralstats.

In this modification, the author of the initial program minimized the part of the code corresponding to data transformation, while we were concerned with designing, making, and providing the user visualization part, focusing on the *ui* (user interface) section of the Shiny runtime environment (Chang *et al*. 2021); In general, Oralstats_Aroca allows performing any operation that R can carry out in a local environment through R Studio (Posit team, 2023), but with the novelty that it can be done from an online environment through a website.

In the last part of this chapter, an analysis of the difficulties encountered in each of the phases of the Ameresco corpus is included, whether identified by the central team or reported by the local teams following the established work protocol. Solutions have been proposed for these difficulties wherever possible, while in other cases, limitations have been acknowledged, and plans for correction have been made. The experience faced during the collection and transcription protocol of a linguistic corpus project highlights the importance of communication with local teams and the need for continuous improvement. Although the collection and transcription protocol has been meticulously planned and drafted in as much detail as possible, potential flaws and areas for improvement in its execution are not discovered until it is used as a working tool, and the prescribed instructions are put into practice. Effective communication with local teams has been a crucial part of this work, as well as the experience gained by the central team. This communication is not only essential for the proper functioning of the process and protocol compliance but also provides opportunities to enhance the corpus.

Queries have revealed the need for improvements in the drafting of data collection and processing protocols (phases 1 and 2). This may involve clarifying them or including new scenarios that illustrate particularly challenging situations. Additionally, many inquiries have brought to light the need to rethink the system or accommodate the specific characteristics of a particular linguistic community that are not reflected in the initial labels and sociolinguistic parameters. For phase 3, data archiving and access through the electronic platform, the need for enhancements or changes in the interface of different versions has

emerged to meet the requirements of researchers using the corpus. These modifications have also affected how data is stored online and its subsequent visualization and download.

Challenges in phase 1, the conception and data collection, include training geographically dispersed teams, the turnover of team members, and the impact of the COVID-19 pandemic. Some teams lacked ideal conditions for data collection, leading to the implementation of two different working models. These challenges necessitated adjustments to the recording process and led to issues related to dialectal variations. In exceptional cases, on-site training sessions were conducted thanks to grants from the Universitat de València. Due to the voluntary nature of team collaboration, timelines for data collection and submission had to be flexible, considering the other work and research obligations of team members.

The selection of speakers for the corpus also posed challenges, as the initially proposed criteria did not always align with the sociolinguistic characteristics of specific communities. Issues related to overrepresentation of certain groups and the difficulty in accessing others were encountered. Technical challenges during data collection included ensuring optimal audio recording conditions, especially when recording in outdoor settings with uncontrollable background noise. Smartphones were used for recording in some cases, but they posed challenges related to notifications and interference from ambient sounds. Metadata collection and obtaining informed consent also required attention to detail.

The delivery of materials to the central team presented difficulties, with different teams sending materials in various formats and manners. A unified approach through restricted-access cloud storage platforms was eventually adopted. Finally, not all materials collected by local teams met the necessary criteria for inclusion in the corpus, despite an initial validation process.

One of the main reasons for discarding recordings has been confusion about the discourse genre. The working protocol explains that the corpus will collect typical casual conversation (Briz, 2001). However, there have been cases where the recording contained other genres such as interviews or semi-directed conversations, which, although considered to be peripheral casual conversation, are not the primary focus of the study. Also been cases of staged secret conversations have been found.

In addition, there have been issues related to poor organization of the recording: cases where two speakers appear, and one of them is the person responsible for recording. This recording context affects the collected material in several ways and invalidates it as a valid

sample of colloquial conversation. On one hand, the degree of spontaneity and secrecy is compromised by 50%, thus raising doubts about meeting the requirements of this genre. On the other hand, these situations led to a distortion of the discourse genre because the person responsible for recording tends to assume the role of an interviewer and forces communicative interaction to involve the other speaker. Therefore, instead of spontaneous conversation, what is being recorded is semi-directed conversation, almost like an interview. Simultaneously, it often happened that the person responsible for recording, in order not to participate and influence the sample, chose to participate very little. Therefore, what is expected to be spontaneous conversation turned into a monologue of the speaker who doesn't know they are being recorded. Thus, there is no interaction, and the requirements of the expected genre are not met.

For this reason, during training and in the work protocol, special emphasis has been placed on the weaknesses found in recordings with two people, and it has been proposed that the researcher should leave the recorder and exit the scene or remain as a passive speaker and limit their participation as much as possible.

In some cases, even though the recordings seemed to meet the genre's characteristics, it was later demonstrated that they had not been made in secret, as the participants themselves referred explicitly to the fact that they were being recorded. In this regard, it has been strongly requested that this should not happen, as it would distort subsequent analysis. And, even if the recordings meet the requirements of the discourse genre, sometimes the audio has poor quality. This can be due to technical issues during recording (recorder placed far from the speakers, holding the recorder by hand while it is on, using a low-quality device) or contextual factors like excessive background noise, music or television playing, household chores or eating, which involve the noise of dishes and utensils, as well as those derived of a public environment such as a park or cafe, like traffic noise or customer noise. These circumstances greatly hinder the production of a high-fidelity transcription.

These discarded recordings, though not valid for the corpus, if sent with all associated materials (transcription, technical information, and authorization) are stored to create a repository of non-typical conversations that might serve as valid material for other research purposes.

In Phase 2, data processing, the difficulties are related to transcription and coding conventions, obtained through questions raised by both local teams and external personnel

associated with the central team, but also by the central team. These questions have been raised with the central team primarily through direct communication with the technical coordinator (email, instant messaging, workshops, and in-person or virtual meetings) and, in some cases, through a shared document for questions between the local and central teams, where queries were logged and resolved by the central team and stored in the cloud.

Regarding questions that arose about transcription and coding, they can be categorized around the following axes: transcription; use of marks and labels in transcription; anonymization.

The difficulties related to the general transcription of phenomena are of various natures. This includes inquiries about how to represent characteristics of each dialectal variety (seseo, ceceo, widespread aspiration, etc.), the reproduction of direct speech, words or fragments in another language, and the representation of functional elements, among others. It also addresses the level of detail required to reflect other issues like noises.

Regarding dialectal varieties, this query mainly originated from personnel hired by the central team in Valencia. When facing the transcription of dialectal varieties different from their own, they were uncertain about how to do it. Therefore, there have been frequent questions about whether phenomena such as seseo, ceceo, /s/ aspiration, changes in verbal paradigms in varieties like Argentinean Spanish (cantas > cantás) or Chilean Spanish (cantas > cantái) should be reflected and how.

For phenomena related to speech, such as seseo, ceceo, or aspiration, the operational decision has been not to reflect them when they are characteristic of the dialectal variety being transcribed. Doing so would require more time investment and could lead to conflicts in the search engine, which should not only search based on the normative form but also include a mark or label for all occurrences found. Therefore, it would be computationally more challenging and slow down result extraction. However, in specific cases where a speaker intentionally produces these phenomena, for example, imitating someone else during direct speech, they have been included. Regarding changes in verbal paradigms, in the case of Argentina, the recommendation has been not to mark them and to transcribe them according to their realization, a decision supported because this variation in conjugation is recorded in the *DLE*. In the case of Chilean Spanish, which is not systematically documented in the *DLE*, it has been marked, as shown in the example below, a label we will explain in section (b). In general, when there were doubts about representing various lexical varieties,

the general recommendation has been to transcribe the form recorded in the *DLE* or the *Dictionary of Americanisms*. Regarding the concept of direct speech or quotation, the main questions revolved around what is considered direct speech, i.e., whether it only applied to moments when a speaker reproduces someone else's words verbatim or if it should also include reproductions of one's own words, sometimes introduced by thought verbs. In the example below, you can see a case where direct speech is clearly distinguished from a case where we have also marked it as direct speech.

Regarding foreign proper names and other foreign words or fragments spoken in another language recorded in conversations, instructions have been given in two directions. On one hand, foreign proper names have been reproduced while maintaining their original spelling without any marks. On the other hand, interventions in another language have been marked with the <extranjero t=""></extranjero> tag.

The last item, which has perhaps generated the most doubts, is the representation of functional elements, interjections, onomatopoeias, and noises, both produced by the speakers themselves and occurring externally. Regarding functional elements like *mm*, *eemm*, *uhum*, there is no widespread consensus on how to transcribe them. However, some of these elements seem to be more conventionalized, and the same applies to interjections and onomatopoeias. Whenever possible, the use of conventions established by the RAE and ASALE has been recommended, either in their works or through Fundéu. Nevertheless, the extraction of a standardized list that can be applied systematically and address this issue is pending. The reason greater emphasis has not been placed on this aspect is that the search interface allows users to retrieve the audio segment where these elements appear, so they can listen to how they were pronounced firsthand. In cases where it was necessary to include them in more detail, the observation label was used, and contextual information relevant to the context was recorded. As for the level of detail expected in transcribing noises, it was established that only those moments where noise significantly interferes with the conversation would be reflected in the transcription. For example, if background noise makes it impossible to understand what was said or if it interrupted the conversation. For other cases, it is up to the transcriptionist to assess their relevance, and if necessary, these notes would be made in the observation line in ELAN, with the <obs t="ladridos de fondo"/> tag or in footnotes when working with a word processor.

Furthermore, some teams have been concerned with representing the pragmatic meaning of some of the noises, functional elements, and interjections mentioned earlier. These teams

have raised the possibility of annotating these meanings, for example, when there is an "ay" that expresses emotion or an untranscribable aspiration that indicates surprise. The protocol did not request this level of detail, but as noted, teams can add layers of information as needed for their research interests. Therefore, to address this query, it was established that it should be done through the observation label.

Regarding the use of marks and tags in transcription several issues were raised by the transcribers, such as how various signs are combined or the exact meaning and application of some of them. This happens, for example, with the marks <sic> </sic> and <fsr t=“ ”> </fsr>, which have been confused at times. It should be clarified that the first is used to reflect an error in pronunciation by the speaker, indicating that it is not an error or oversight by the transcriber; while the second captures phenomena of phonetics and syntax and, in general, cases where the spelling and pronunciation of a word do not match. A representative example of the use of this tag is the transcription of the conjunction *pues*, which can be realized in oral speech in multiple ways: *pues, ps, pss, pus, pos, s.* Given this variability and considering future search possibilities from the online platform, the search for the normative form should be prioritized, while attempting to account for its oral representation.

The next most frequently reported difficulty is related to representing simultaneous speech. For this circumstance, brackets are considered for use in both working modes 1 and 2. The peculiarity of this mark has to do with the fact that it should be used in all interventions, regardless of how many speakers are affected, and when they occur simultaneously, their representation requires a significant effort on the part of the transcribers. In the case of mode 1, in a word processor, they are marked following the Val.Es.Co. (2002) model, meaning that overlapping segments are enclosed in brackets and also aligned visually.

For mode 2, since the ELAN program allows the creation of lines for each speaker, such alignment is not necessary, as the program's interface already provides that visualization.

The last query in this aspect is related to how to combine the use of different marks and tags. In this regard, the question was raised about the order in which they should appear when multiple marks need to be included at once, for example, when there is a foreign word within a direct speech reproduction, which is also affected by an observation. In such cases, it works like mathematical formulas, meaning that the most general mark is placed on the outside, and within it, the others are marked until reaching the most specific one.

This final section covers the instructions given to transcribers regarding what should be anonymized and how the necessary segments should be anonymized. Doubts arose about how to anonymize proper names in various cases, with a particular focus on names and surnames of individuals, places, and other information appearing as the names of companies and places related to the speakers.

Regarding proper names, the general recommendation has been to replace them with another name that has the same phonetic pattern and is suitable for the sociocultural and diatopic characteristics. Names of companies and places related to the speakers are also subject to anonymization. While the general instruction is to find a similar fictitious name, there were cases where the fictitious name disrupted the coherence of the narrative. While situational correlation is usually maintained, in cases where any substitution would alter the understanding of the story, an observation or a generic name, such as <anónimo><obs t="city 1"/></anónimo>, was chosen.

There were also cases of over-anonymization, where transcribers misunderstood the instruction to anonymize elements that could identify the speaker, resulting in the anonymization of all proper names and places mentioned, without considering the interactive context. For example, anonymizations had been applied to the names of political leaders and public figures, as well as places and locations unrelated to the specific speaker.

Regarding the phase of reviewing the materials, as mentioned earlier, not all local teams always had the necessary infrastructure, technical resources, and personnel to complete all phases of corpus collection. In some cases, they were only able to collect the recordings along with their technical information and permissions, but not to carry out the transcription phase. In other cases, they sent broad transcriptions in text format, but not narrow transcriptions or aligned transcriptions. Despite knowing that this would slow down the subsequent corpus processing work, it was decided to accept this "incomplete" material because the most valuable part is precisely the collection of the recordings, which the central team could not execute. However, the transcription phase could be completed from Valencia.

Taking on this task in the central team led to solving two inconveniences: firstly, it was necessary to find and compensate external personnel to expedite transcription and validate the materials; secondly, these transcribers, native speakers of Peninsular Spanish variety, had to deal with the transcription or review of other dialectal varieties foreign to their own reality, which was not without its challenges, as discussed below.

The individuals responsible for transcribing the recordings sent from different teams lacked the necessary context to resolve ambiguities that arose in the recording, and they did not possess sufficient dialectal knowledge to understand and accurately represent what was said in the recording. This resulted in an incremental and inevitable loss of information. One of the most common doubts among reviewers is whether or not to complete or amend information when the transcriber has omitted interventions or transcribed things that, in the reviewer's judgment, the speaker did not actually say, given the awareness of information loss, adherence to the provided text is maximal. Sometimes, reviewers do not directly correct, but they express their doubts about whether a specific form is a dialectal variant or if, within that, it is considered a non-standard form and should therefore be labeled as a phenomenon of phonetics and syntax (§ 5.2.2.1.2). Naturally, this situation is especially pronounced among reviewers whose dialect differs from the recorded variety, as they are at risk of correcting unfamiliar yet valid (and even common) expressions in the variety. Faced with this situation, the technical coordination's recommendation has been twofold.

Firstly, for people who review transcriptions made in a word processor, it is most convenient to start by dividing the speech into phonetic groups, and once the conversation is divided, transfer the content of the local transcriber's transcription. This allows the reviewer to add or correct anything that they believe is missing. Transcribers often omit repetitions or restarts, omit fragments, overlaps, etc. However, special care is recommended when it comes to the lexicon used, as there are many localisms that non-native reviewers of the dialect may not be familiar with and, therefore, may misinterpret and correct as more familiar terms in their own dialect. Faced with this situation, a range of resources is recommended, depending on the nature of the unfamiliar term, ranging from the RAE dictionary to *Dictionaries of Americanisms*, and even direct searches on platforms like Google. In one Cuban transcription, the Spanish reviewer did not recognize 'trusa' and corrected it to 'blusa', a substitution that was detected and corrected during the validation phase.

Secondly, for those working directly with files in ELAN, the recommendation is the same, but they should consider that changes in segmentation boxes are complex, and poor practice can lead to misalignment.

After identifying these problems, it has been encouraged for local teams to handle all phases of data collection and transcription, opting for a single working mode (mode 2, which aligns text and audio). This is not only a measure to avoid these errors but also for expediting

data processing and input into the search engine. Therefore, in the current protocol, the collection and transcription of samples from different cities must go through the hands of speakers (members) of that dialectal variety. Since the recordings are of colloquial conversation secretly recorded in close and familiar living environments, the recordings must be made by individuals belonging to these realities, not external researchers who are unfamiliar with the family context. Local teams should also be responsible for transcription since they have firsthand knowledge of the dialectal variety subject to recording.

The challenges in Phase 3, Archiving, Distribution, and User Access to the Corpus, arise when offering files to users is related to the responsibility of corpus editors or managers to maintain the hosting infrastructure and to provide file formats that are usable and enduring over time, without constantly needing adaptation to new, more modern formats. Furthermore, it has been one of the goals of the Ameresco corpus to respond to a wide range of user needs by providing simpler access and usage options where users can download files in various formats tailored to their needs, even using a basic search engine. Additionally, through the Oralstats Aroca application, advanced search capabilities are provided, along with the ability to download files in formats such as ELAN, which allows for greater exploitation of plain text. It also caters to the research group's needs, including prosodic studies and pragmatic conversation analysis.

Lastly, storage needs have also been considered. While ELAN requires files to be in WAV format, they have been converted to MP3 format for inclusion on the website, as it is less resource-intensive and more efficient. Other limitations relate to the visualization of all available formats. While ELAN and PRAAT files are available on the platform, their visualization depends on each user installing these programs on their computer and acquiring basic knowledge of how to use these tools.

7.2. Importance of Ameresco corpus

Based on this analysis, the Ameresco corpus has developed a methodology rooted in pragmatic and interactional study objectives in spontaneous colloquial conversation. It aims to fill two important gaps: firstly, in the realm of pan-Hispanic corpora, it seeks to collect samples from the major cities within the Spanish-speaking world; secondly, it addresses the deficit identified in the corpus literature for this discursive genre, secret recordings of spontaneous colloquial conversation. Ameresco aims to become a study resource that, while

a natural extension of Val.Es.Co. corpus principles, mirrors the main pan-Hispanic Spanish corpus, PRESEEA, enabling the comparability of research across different discursive genres (interviews for PRESEEA and colloquial conversation for Ameresco). Hence, the methodological precedents established in this corpus have been applied in Ameresco, in what regards criteria related to sociolinguistic representativeness and the transcription and coding system in favor of data interchangeability and reusability.

However, we cannot lose sight of an essential difference between both corpora regarding the combination of representativeness and naturalness criteria. In this sense, in the Ameresco corpus, naturalness will take precedence over representativeness since our ultimate goal is pragmatic study. A possible lack of representativeness, while it may lead to an imbalance in the sample, is a fact that can be addressed by the incorporation of new materials. As noted in Chapter 4, sometimes the representativeness criterion is debatable; that is, there is no doubt that a corpus should strive to achieve representativeness, as the literature suggests. However, in a multidialectal corpus of this magnitude, it can only be adjusted in terms of the quantity of speakers collected according to the established stratification and location. It is impossible to achieve a balance in the number of recorded forms in each case. In this sense, if we compare the population level of each of the cities that make up the Ameresco corpus, the proportions of cities like Buenos Aires or Mexico City should be greater than those assigned to Panama or Loja (Ecuador).

Furthermore, the Ameresco corpus is a work in progress. Although a significant part has been compiled (currently with approximately 75 hours of recordings from 14 cities), there is still much work ahead to complete the Pan-Hispanic panorama. However, in line with what was mentioned above regarding representativeness, sub-corpora that have completed the collection of materials have expressed their willingness, and are doing so, to collect new recordings to improve the naturalness of the recordings. Therefore, we are committed to a constant desire for improvement, both from the central team and from the various local teams involved in the project.

In this regard, mechanisms have been established to address possible errors and limitations of the corpus, such as the detection of less prototypical conversations, which, when more representative conversations are available, will be moved to the repository to make way for new materials. It should be noted that the less suitable recordings would not be completely removed from the corpus. This is why the technical coordination of the central

team keeps a detailed record of the characteristics of each conversation received, as seen in Chapter 5.

Additionally, the Ameresco corpus is open in various ways. On one hand, the possibility of receiving feedback from users of the corpus who can contact the central team to report any issues they have found is considered. On the other hand, we are also open to suggestions and needs that arise from different collaborating work teams, as well as from new teams that want to join and reflect other urban areas that have not yet been covered. Lastly, we are open to sharing data, both those that make up the corpus itself (recordings, transcriptions, metadata) and information concerning the established work protocols, as well as engaging in joint activities with other corpora with the aim of continually growing and reflecting to enrich our contribution to the scientific community.

However, despite the strengths of this corpus, we are aware that the Ameresco corpus has errors and limitations in its current phase. Some of these are correctable and will be addressed as we have the budget and personnel to make improvements, while respecting the adopted methodology. However, it should be noted that we cannot completely eliminate some of these errors, especially when they are more like circumstances of the past. In this regard, the corpus has limitations. For example, in the early stages of corpus collection, a sociolinguistic stratification system inherited from Val.Es.Co. was adopted, which, while valid at the time, would now need to be revised to adapt to current living conditions. Having said that, to the extent possible, actions have been taken to try to address this limitation, such as the adoption of a new technical form that includes the exact age of speakers for future age group restructuring. Likewise, there are other limitations that cannot be corrected; this is the case with technical forms that were collected incorrectly by the local team. In this case, retrieving the correct information is impossible because, due to privacy and data handling laws, the data collection process cannot be reversed to contact the conversation participants again. Therefore, we are aware of this fact and accept it as a possible limitation for specific research. It could be said that there are no perfect corpora in this regard, especially when they are based on a discursive genre like casual conversation that requires very specific and uncontrolled recording contexts.

Despite the limitations mentioned, it is worth noting what the appearance of the Ameresco corpus means in the international scientific community, beyond the validity and interest of the data it provides for the study of colloquial Spanish, as listed in 6.2.1. With the writing of this work, the intention has been for the Ameresco corpus to partially address the deficiencies

mentioned throughout these pages regarding the protocols for constructing oral corpora. We are fully aware of the existence of these protocols (every corpus construction involves prior development of a working system), but they are often internal documents of research groups, rarely detailed for external users, or transmitted among researchers as a way of acting and are not put in writing. Therefore, for the development of the corpus methodology discussed in this thesis, in many cases, we had to extract this information inductively from the experience as users of the various analyzed corpora.

The review conducted in this regard has served to analyze the available methods based on these previous foundations, identify their strengths to replicate them, as well as their weaknesses, that is, errors or limitations that need to be addressed. However, it should be noted that what is not explained cannot be replicated, hence the advocacy from this work.

7.3. Final remarks

After this detailed study of the panorama of oral corpora, as well as of the development of the phases of work and construction involved in oral corpora, as manifested in the execution of the Ameresco corpus, it can be seen that the detailed description of the previous issues, execution, and follow-up takes more 300 pages to be described.

In view of this work, we believe it is evident how composing an oral corpus involves a significant effort in terms of time, but also, consequently, in terms of money. This latter consequence is less obvious because, in general, or at least in the Hispanic context, the work of corpus compilers is usually altruistic. At times, one can count on the assistance, paid or unpaid, of student personnel who, while putting in their best efforts to carry out the assigned tasks, have not professionalized in this work and cannot dedicate themselves continuously to corpus development. In other cases, the research group members themselves carry out technical data processing tasks, thereby reducing the time dedicated to research. This is why we advocate the work done by the individuals behind the creation of an oral corpus considering the difficulties in obtaining funding from the administration.

To give an idea of the time required for constructing a corpus with the characteristics of Ameresco, we offer an estimate of the work done in one year. Specifically, the technical coordination has undertaken tasks such as the following:

1. Training of transcribers and personnel responsible for recording the audio. This is done about 10 times a year and with each addition of a transcriber or local team who might need

training; Since the training process takes 10-12 hours (in two or three sessions) for each group, 120 hours are needed, per year, to properly train the people in recording, transcribing, validating, etc. This personnel needs specific training in the use of programs, the transcription system employed, anonymization marking, conventions, labeling, etc., depending on their level of involvement in the project. Thus, there is a trainer responsible for training teams in the following tasks:

- Training in the use of ELAN and acoustic processing software.

- Training in transcription conventions.

- Training in labeling system.

2. Collection of materials. Considering that it takes approximately 5 to 7 hours per week to collect materials, 294 hours per year (calculated based on 42 working weeks) would be devoted to this task. This particular activity involves:

- Coordination of local working groups in Spain and America; the coordinator sets a delivery schedule, sends reminders when deadlines are close or the teams are behind schedule.

- Validation of samples sent before transcription (checking noise level, spontaneity, compliance with data protection policy by local teams, sociolinguistic stratification of participants, etc.); This is the one of the most time-consuming activity and cannot be automatized or shortened without a loss in accuracy.

3. Transcription and encoding of recordings. Approximately 3 to 4 hours per day, about 24 hours per week: 1008 hours/year [42 weeks].

- Segmentation of intonational groups in ELAN to provide a scaffolding for the transcription. Pauses or tonal breaks must be detected and a frontier placed between them so that some factors, like the pause length, can be calculated later.

- Transcription in ELAN, aligning audio and text.

- Full labeling in TEI-XML (PRESEEA model) adapted to AMERESCO, together with the explicit mention to those situational cues needed to understand the conversation (explained in the tier ‘observations’).

4. Processing of recordings: anonymization, review, validation. This phase takes approximately 1 to 2 hours per day, adding up to 10 hours per week and 420 hours per year

[42 weeks]. This particular activity comprises several stages involving software with different functions:

- Anonymization of audio fragments previously labeled as <anonymous> </anonymous> with Audacity software.

- Review of transcriptions to guarantee their accuracy.

- Review of labeling and syntax validation using OXYGEN and R programs, making sure the expressions are correctly formulated, there are no opening signs without the corresponding closing sign, etc.

- Transformation of ELAN's .eaf format into tabulated text suitable for the web search engine.

5. Computer processing for web upload. This is done twice a year and takes circa 20 hours each time. Total: 40 hours per year. This activity is subdivided into:

- Tokenization of intonational groups.

- Automatic morphosyntactic tagging with the POS tagger.

- Review of morphosyntactic tagging, given that informal speech is filled with non-standard tokens that challenge the POS tagger efficacy. Frequently, the automatic attributions must be corrected manually.

- Creation of interventions (turns) from intonational groups.

- Treatment of metadata as tabulated text and upload to the web.

- Updating the web database.

- Web maintenance and resolution of transformation issues.

- Conversion of audio from .wav to .mp3 format and uploading of audio to the web.

- Generation of textgrids (PRAAT) for phonetic analysis and uploading them to the web to make them available for experts.

6. Computer storage of materials (with encrypted systems) and backups. Done twice a year; 20 hours each time. Total: 40 hours/year

If we consider the average annual working hours in Spain is 40 hours per week, it amounts to 1680 hours a year, considering 42 working weeks (excluding holidays and public holidays). If we consider, instead, the total amount of hours devoted to activities in 1 to 6,

the total amount of hours for one person in one year is higher, 1902, meaning that one person, fully devoted to the corpus from 9 to 5, would not be able to deliver the work on time, not to mention the fact that other activities (teaching, writing, attending conferences) would be out of the table. As a final reflection exercise, one can extrapolate this annual calculation to the total work done so far (material collected in 14 cities, whose members had to be trained and assisted with their doubts and problems, 75 hours of recording distributed across 203 conversations that need to be transcribed, aligned, anonymized, and computer-processed). Given this effort in terms of working hours and also accounted for in economic terms (which we leave to the discretion of the reader, who can assign a reasonable value to the hourly wage and multiply it by the years and conversations processed), one can understand the value of the work of each and every person who has participated in the process.

Ultimately, and as a summary, this thesis serves as an advocacy for the work done by linguists who compile corpora. On one hand, the thoroughness and metareflection carried out to establish different work protocols are often undervalued. In fact, we have observed a systematic lack of explicitation of these methodologies in scientific publications, and in some cases, we have had to rely on information published on the electronic pages of each corpus, on personal communications by the authors or compilers, and only more rarely in scientific publications. This, which is common in all corpus compilation work, particularly applies to the methodology of oral corpora work, as we have found in the bibliographic study that written corpora have far more references and greater metatheoretical reflection.

In this regard, it is worth noting the personal experience during this predoctoral period when presenting research based on oral corpus construction methodology at various international conferences. There has been repeated interest from the research community in the various considerations to be taken into account for corpus construction. This interest contrasts with the fact that, when taking a look at the contributions presented at these conferences, most are investigations *based on* corpora or *with* corpora, whereas those works dealing with the conception and execution of corpora are definitely scarcer.

Undoubtedly, if the previous experience of other researchers is not registered in scientific works, it is impossible to replicate successes and to correct errors, which is the basic principle of scientific research in ours or in any other discipline. However, it should be made very clear that the lack of publications and (at least in a comprehensive manner) written protocols for oral corpora most likely does not result from a desire to conceal (which would not be logical in researchers who are predominantly self-taught and altruistic, as it was

mentioned above) or from a reluctance to present technical content, but rather from the little value placed on the construction of the framework vs. the high value placed on its utilization, discouraging scholars to further advance in the field and to make the effort to make complex technical content available for others. This imbalance is very evident in the publishing industry, in the scope of relevant journals, and among those who ultimately benefit from corpora to conduct their research. Therefore, a greater presence of corpus linguistics in academic forums and greater recognition of this work when evaluating scientific production in Spain is necessary.

Bibliografía

Adolphs, S. y Knight, D. (2010). Building a spoken corpus. What are the basics? En A. O'Keeffe y M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*, pp. 38-52, Routledge.

Agencia Española de Protección de Datos. (2016). *Orientaciones y garantías en los procedimientos de anonimización de datos personales*. En línea, https://datos.gob.es/es/documentacion/orientaciones-y-garantias-en-los-procedimientos-de-anonimizacion-de-datos-personales

Agencia de los Derechos Fundamentales de la Unión Europea y Consejo de Europa, (2014). *Manual de legislación europea en materia de la protección de datos,* Oficina de Publicaciones de la Unión Europea.

Albelda, M. (2022). Los corpus del español hablado y los estudios pragmáticos. En G. Parodi, P. Cantos-Gómez y C. Howe (eds.), *The Routledge Handbook of Spanish Corpus Linguistics,* pp. 223-238, Routledge.

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

Albelda, M. y Jansegers M. (2019). From visual perception to evidentiality: A functional empirical approach to *se ve que*. *Lingua: International review of general linguistics*, vol. 220, núm. 2, pp. 76-97.

Ajmer, K. (2018). Corpus Pragmatics: From Form to Function. En A. H., Jucker, K. P. Schneider y W. Bublitz (Eds.), *Methods in Pragmatics*, pp. 555-585. De Gruyter Mouton.

Asociación de Academias de la Lengua (2010). *Diccionario de americanismos*. En línea, https://www.asale.org/damer/

Atkins, S., Clear, J. y Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), pp. 1-16.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh University Press.

Baker, P. Hardie, A. y McEnery, T. (2006), *A Glossary of Corpus Linguistics,* Edinburgh University Press.

Barcala, M., *et al*. (2018). El corpus ESLORA de español oral: diseño, desarrollo y explotación. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 5(2), pp. 217–237.

Beighley, Lynn. (2009). *Drupal*. John Wiley & Sons, Incorporated.

Bejarano, D. *et al*. (2018). *Protocolo de transcripción ortográfica Corpus Lingüísticos del Instituto Caro y Cuervo* (CLICC). En línea https://clicc.caroycuervo.gov.co/

Blas Arroyo, J. L. (2010). Confluencia de normas sociolingüísticas en un hecho de variación sintáctica: factores sociales en la selección de la variante *deber de + infinitivo* (vs. deber) en un corpus oral, *Hispania*, 93(4), pp. 624-649.

Berber Sardinha, T. (2004). *Lingüística de corpus*, Manole.

Biber, D., Conrad, S., y Reppen, R. (1998). *Corpus Linguistics: Investigating Language structure and use*. Cambridge University Press.

Bolaños Cuéllar, S. (2015). La lingüística de corpus: perspectivas para la investigación lingüística contemporánea. *Forma y Función*, 28(1), pp. 31-54.

British National Bank. (En línea). *The British National Corpus Online*. Disponible en http://www.natcorp.ox.ac.uk/corpus/index.xml

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

_____. (1998 [2001]). *El español coloquial en la conversación. Esbozo de pragmagramática*. Ariel.

_____. (2010a). El registro como centro de la variedad situacional. Esbozo de la propuesta del Grupo Val.Es.Co. sobre las variedades diafásicas. En I. Fonte y L. Rodríguez Alfano (Eds.), *Perspectivas dialógicas en estudios del lenguaje*. pp. 21-56, Universidad Autónoma de Nuevo León.

_____. (2010b). Lo coloquial y lo formal, el eje de la variedad lingüística. En R. M. Castañer Martín y V. Lagüéns García (Eds.), *De moneda nunca usada: Estudios dedicados a José M.ª Enguita Utrilla*, pp. 125-133, Institución Fernando el Católico.

_____. (2012a). Los déficits de los corpus orales del español (y de algunos análisis). En Jiménez, T. E. *et al*. (coord.), *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*, pp. 115-137, Universidade de Santiago de Compostela.

_____. (2012b). La constelación comunicativa coloquial. Hacia un modo más dinámico de entender lo coloquial. *Español Actual*, 98, pp. 217-225.

_____. (2016). El proyecto Ameresco: la idea de un corpus de conversaciones coloquiales del español de América. En A. M. Bañón Hernández *et al*. (ed. lit.), *Oralidad y análisis del discurso: homenaje a Luis Cortés Rodríguez,* pp. 81-104, Universidad de Almería.

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

Briz Gómez, A. y Carcelén Guerrero, A. (2019). El futuro iberoamericano del español: la investigación del español oral y en español. En *El español en el mundo 2019, Instituto Cervante*s, 189-217, Bala Perdida.

Briz Gómez, A. *et al*. (2019). *Protocolo de trabajo para los equipos Ameresco*. En línea https://esvaratenuacion.es/protocolo-de-trabajo (versión actualizada enero de 2020).

Briz Gómez, A. y García Ramón, A. (2020). La conversación coloquial como prototipo de lo dialogal. En O. Loureda y A. Schrott (Eds.), *Manual de lingüística del hablar*, pp. 261-286, De Gruyter.

Briz Gómez, A. y Grupo Val.Es.Co. (2002). *Corpus de conversaciones coloquiales*, Anejos Oralia. Arco Libros.

_____. (2014). Las unidades del discurso oral. La propuesta Val.Es.Co. de segmentación de la conversación (coloquial). *Estudios de Lingüística del Español*, 35, pp. 13-73.

Cabedo Nebot, A. (2021). *Oralstats*. https://github.com/acabedo/oralstats

_____. (2022a). Visualizing melody with multiple acoustic and tagging values using the visualization module of the Oralstats tool. *Estudios de Fonética Experimental*, 31, pp.135-148.

_____. (2022b). Análisis melódico del habla como herramienta distintiva para el perfil idiolectal de hablantes. *Revista da ABRALIN*, vol. 21(2), 48-70.

Cabedo Nebot, A. y Carcelén Guerrero, A. (2022). *Oralstats Aroca*. Valencia. https://github.com/acabedo/aroca

Cabedo Nebot, A., e Hidalgo Navarro, A. (2023). Caracterización fónica de la (des)cortesía en el español hablado de Valencia. Aproximación cualitativo-cuantitativa. *Círculo de lingüística aplicada a la comunicación*, 93, pp. 131-149.

Camazón, J. N. (2010). *Aplicaciones web*. Editex.

Carcelén Guerrero, A. (en prensa). ¿Es posible elaborar corpus orales espontáneos y cumplir la legislación? El modelo en tres fases del corpus Ameresco.

Carcelén Guerrero, A. y Uclés Ramada, G. (2019). Diseño y construcción de un corpus oral multidialectal. El corpus Ameresco. *Normas: Revista de Estudios Lingüísticos Hispánicos*, 9 (1), pp. 17-36.

Cestero Mancera, A. M.ª (1994). *Análisis de la conversación: alternancia de turnos en la lengua española*. Universidad de Alcalá.

_____. (1999), *Repertorio básico de signos no verbales*, Arco Libros, Madrid.

Chang, W. *et al*. (2021). *Shiny: Web Application Framework for R*. https://cran.r-project.org/web/packages/shiny/index.html

Childs, B., Van Herk, G. y Thorburn, J. (2011). Safe harbour: Ethics and accesibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* (7-1), pp. 163-180.

COLA (en línea). *Corpus oral de lenguaje adolescente*. Disponible en https://blogg.hiof.no/colam-esp/el-corpus-cola/

Cortés Rodríguez, L. (2002). Las unidades del discurso oral, *Boletín de Filología*, 17, pp. 7-29.

Crystal, D. (1991). *The Cambridge Encyclopedia of Language.* Cambridge University Press.

D'Arcy, A. y Bender, E. (2023). Ethics in Linguistics, *Annual Review of Linguistics,* 9 (1), pp. 49-69.

Davies, M. (2002-) Corpus del Español: Hiistorical/Genres. En línea http://www.corpusdelespanol.org/hist-gen/

———. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International journal of corpus linguistics* 10 (3), pp. 307–34.

———. (2009). The 385+ million word Corpus of Contemporary American English (19902008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14 (2), pp. 159–90.

———. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7 (2), pp.121–57.

———. (2021). The TV and Movies corpora: Design, construction, and use. *International journal of corpus linguistics* 26 (1), pp. 10–37.

Davies, M. y Jong-Bok, K. (2019). The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus. *Linguistic research* 36 (1), pp. 1–34.

Davies, M. y Parodi, G. (2022). *Constitución de corpus crecientes del español, en Lingüística de corpus en español*. En G. Parodi, P. Cantos-Gómez y C. Howe (eds.), *The Routledge Handbook of Spanish Corpus Linguistics*, pp. 13-32, Routledge.

Díaz Sánchez, A. (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)

Du Bois, J. W. (1991), Transcription design principles for spoken discourse research, Pragmatics, pp. 71,-106 International Pragmatics Association

_____. (2015). *Representing discourse*. MS, Linguistics Department, University of California, Santa Barbara.

Du Bois, J. W. *et al*. (2000-2005). *Santa Barbara corpus of spoken American English*, Parts 1-4. Linguistic Data Consortium.

_____. (2015). Outline of discourse transcription. *Talking Data: Transcription and Coding, Discourse Research*, pp. 45-90.

Ellis Montalbán, P. y Bartolomé Peral, E. (2020). La condición de No Binario en la legislación europea: estudio comparativo sobre definiciones y marcos legales y político. *Inguruak: Soziologia eta zientzia politikoaren euskal aldizkaria, Revista vasca de sociología y ciencia política*, 69, pp. 20-38.

Egbert, J., Biber, D. y Gray, B. (2022). *Designing and Evaluating Language Corpora. A Practical Framework for Corpus Representativeness*, Cambridge University Press.

Enghels, R., Vanderschueren, C. y Bouzouita, M. (2015). Panorama de los corpus y textos del español peninsular contemporáneo. En M. Iliescu y E. Roegiest (Ed*.), Manuel des anthologies, corpus et textes romans*, pp. 147-170, De Gruyter.

Escarrá, A. y Díaz, I. (2011). Cardiocor, un corpus para uso de los traductores de la medicina. *Matices en Lenguas Extranjeras*, *5*, pp. 134-142.

ESLORA. Corpus para el estudio del español oral <http://eslora.usc.es>, versión 2.0 de septiembre de 2020, ISSN: 2444-1430.

Espinosa, G. y García Ramón, A. (2019). A Preliminary Typology of Interactional Figures Based on a Tool for Visualizing Conversational Structure. En O. Loureda, I. Recio, L. Nadal y A. Cruz (Eds.), *Empirical Studies of the Construction of Discourse,* pp. 94-130, John Benjamins.

Estellés Arguedas, M., y Albelda Marco, M. (2020). The Boundaries between Perception and Evidentiality. Dialectal and Diachronic variation in 'se ve que'. *Anuari De Filologia. Estudis De Lingüística*, 10, pp. 163–193.

Fernández Ordóñez, Inés (dir.) (2005-). Corpus Oral y Sonoro del Español Rural, http://www.corpusrural.es

Fuentes Rodríguez, C. (2021) (dir.). Corpus MEsA 2.0. [Recurso electrónico]. http://www.grupoapl.es/materiales-corpus/corpus-mesa

Gadet, F. *et al*. (2012). Un grand corpus de français parlé: le CIEL-F. *Revue française de linguistique appliquée*, 17(1), pp. 39-54.

Gallardo Paúls, B. (1996). *Análisis conversacional y pragmática del receptor*. Episteme.

Givon, T. (1979). On Understanding Grammar**.** Academic Press.

Goffman, E. (1983). The interaction order: American Sociological Association, Presidential Address. *American Sociological Review*, 48(1), pp. 1-17.

Gómez Molina, J. R. (2013). Norma y usos de las perífrasis 'deber + infinitivo' / 'deber de + infinitivo'. En J. R. Gómez Molina (coord.), *El español de Valencia. Estudio sociolingüístico*, pp. 71-108, Peter Lang.

Gries, S. (2006). Some Proposals towards a More Rigorous Corpus Linguistics. *Zeitschrift für Anglistik und Amerikanistik*, 54(2), pp. 191-202.

Gries, S. Th. (2009). What is Corpus Linguistics?, *Language and linguistics compass* 3 (5), pp. 1225–41.

Gries, S. Th. (2010). *Useful statistics for corpus linguistics*, pp. 269-272.

Gut, U. (2020). Spoken Corpora. In, M. Paquot y S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics,* pp. 235–56. Springer International Publishing.

Halliday, M. A. K. (1990). *Language, context and text*, pp. 29- 49. Oxford University Press.

Hardie, A. (2012). CQPweb Combining Power, Flexibility and Usability in a Corpus Analysis Tool. *International Journal of Corpus Linguistics* 17 (3), pp. 380–409.

Heritage, J. (2013). Language and Social Institutions: The Conversation Analytic View. *Journal of Foreign Languages,* 36 (4), pp. 2-27.

Hidalgo, A. y Sanmartín, J. (2005). Los sistemas de transcripción de la lengua hablada. *Oralia: Análisis del discurso oral*, 8, pp. 13-36.

Hincapié, D. y Bernal, J. A. (2018). *Lingüística de corpus*, Páramo, Instituto Caro y Cuervo.

Hinrichs, E. y Krauwer S. (2014). The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. En *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA), pp. 1525–1531, Cambridge University Press.

Jørgensen, A. M. y Eguía, E. (2014). Presentación de COLA, un corpus oral de lenguaje adolescente en línea. En Eriksdottir, S. A. (ed.). *Actas/actes/atti, Rom14*, pp.1-17, Vigdís Finnbogadóttir Institute of Foreign Languages.

Jucker, A.H., Schneider, K.P., y Bublitz, W. (2018). *Methods in Pragmatics*, De Gruyter.

Julià Luna, C. (2021), Del atlas lingüístico tradicional al corpus geolingüístico digital: diseño de un proyecto. *Scriptum digital: revista de corpus diacrònics i edició digital en llengües iberoromàniques*, 10, pp. 109-147.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman.

Krug, M., y Schlüter, J. (Eds.). (2013). *Research Methods in Language Variation and Change*. Cambridge University Press.

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics.

_____.(1972). Some Principles of Linguistic Methodology. *Language in Society* 1 (1), pp. 97–120.

_____. (1983). Modelos sociolingüísticos. Cátedra.

Leech, G. (1992). Corpora and theories of linguistic performance. En J. Svartvik (Ed.), *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, pp. 105-122, De Gruyter.

Leech, G., Myers, G., y Thomas, J. (Eds.) (1995). *Spoken English on computer*. Longman.

Ley Orgánica 1/1982, de 5 de mayo, de Protección civil y derecho al honor, la intimidad personal y a la propia imagen.

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales.

Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.

Ley de Propiedad Intelectual 1/1996, de 12 de abril, y sus posteriores modificaciones.

Llisterri, J. (en línea). La lingüística de corpus. Disponible en https://joaquimllisterri.cat/language_resources/lang_res/linguistica_corpus.html#La_noci%C3%B3n_de_corpus

_____. (2021). Corpus para investigar sobre el componente fónico en español como LE/L2. En M. Cruz y J. Muñoz (Eds*.), e-Research y español LE/L2.: investigar en la era digital*, pp. 164-196, Routledge.

Llisterri, J. y Mariño, J. B. (1993). *Spanish adaptation of SAMPA and automatic phonetic transcription*. Report SAM-A/UPC/001/V1. Dispoinible en línea https://joaquimllisterri.cat/publicacions/SAMPA_Spanish_93.PDF

Llisterri, J. *et al*. (2005). Corpus orales para el desarrollo de las tecnologías del habla en español, *Oralia: Análisis del discurso oral*, 8, pp. 289-328.

López Morales, H. (1997). Corpora orales hispánicos. En A. Briz Gómez, J. R. G. Molina y M. J. M. Alcalde (coords.), *Pragmática y Gramática del español hablado*. Actas del II Simposio sobre el español coloquial, pp. 137-145, Pórtico.

Lüdeling, A. & Kytö, M. (2008). *Corpus Linguistics: An International Handbook,* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science). De Gruyter.

Manjón Cabeza, A. (2017). Deber (de) + infinitivo en el corpus PRESEEA de Granada, *Lingüística en la Red*, XV, pp. 1-16.

Martín de Santa Olalla Sánchez, A. (1999). Una propuesta de codificación morfosintáctica para corpus de referencia en lengua española, *Estudios de Lingüística del Español* (ELiEs). En línea, http://elies.rediris.es/elies3/index.htm

Martín Herrero, C. (2009). Aproximación a ciertas perspectivas en Lingüística de Corpus. En P. Cantos y A. Sánchez (Eds.), *A survey of corpus-based research,* pp. 1020-1032. Recurso electrónico https://www.um.es/lacell/aelinco/contenido/index.html

McEnery, T. y Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.

McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-Based Language Studies*, Routledge.

McEnery, T., y Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press.

Méndez Guerrero, S. (2015). Corpus Oral Juvenil del Español de Mallorca (COJEM), *Linred: Lingüística en la Red*, 13, pp. 1-186.

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*, Cambridge University Press.

Moreno Fernández, F. (2005a). Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA. *Oralia: Análisis del discurso oral*, 8, pp. 123-140.

_____. (2005b). Geografía lingüística de Hispanoamérica. En J. M.ª Enguita et al. (coords.), *Jornadas Internacionales en memoria de Manuel Alvar*, pp. 89-108, Institución Fernando el Católico.

_____. (2005c). Corpora of Spoken Spanish Language. The Representativeness Issue. En Y. Kawaguchi et al. (Eds.), *Linguistic Informatics, State of the Art and the Future*, pp. 120-144, John Benjamins.

_____. (2016). En torno a PRESEEA: Notas de investigación y de sociología de la ciencia. *Boletín de Filología Universidad de Chile*, vol. 51(2), pp. 369-376.

_____. (2021a). *Metodología del Proyecto para el estudio sociolingüístico del español de España y de América (PRESEEA).* Editorial Universidad de Alcalá.

_____. (2021b). *Marcas y etiquetas mínimas obligatorias para materiales de PRESEEA*. Editorial Universidad de Alcalá.

Nicolás Martínez, C. (2012), *C-Or-DiAL (Corpus Oral Didáctico Anotado Lingüísticamente)*, Liceus.

Niemants, N. (2018). Des enregistrements aux corpus: transcription et extraction de données d'interprétation en milieu médical. *Meta*, 63(3), pp. 665-694.

Nijs, V. (2023). *Radiant: Business Analytics Using R and Shiny*. https://cran.r-project.org/web/packages/radiant/index.html

Ochs, E. (1979). Transcription as theory. En E. Ochs, y Schieffelin, B. (Eds.), *Developmental Pragmatics*. Pp. 43-72, Academic Press.

Ogrodniczuk, M., *et al*. (2019). From the National Corpus of Polish to the Polish Corpus Infrastructure. *Jazykovedný časopis* 70 (2), pp. 315–23.

O'Keeffe, A. y M. J. McCarthy (Eds.) (2010). *The Routledge Handbook of Corpus Linguistics*, Routledge.

Open Language Archives Community & Linguistic Data Consortium. (2010) *Open Language Archives Community OLAC. United* States. [Web Archive] Retrieved from the Library of Congress, https://www.loc.gov/item/lcwaN0003992/.

Payrató, L. (1995). Transcripción del discurso coloquial. En L. Cortés (Ed.), *El español coloquial. Actas del I Simposio sobre Análisis del Discurso Oral*, pp. 43-70, Universidad de Almería.

Parodi, G. (2006). El Grial: interfaz computacional para anotación e interrogación de corpus en español, *RLA. Revista de Lingüística Teórica y Aplicada*, Concepción (Chile), 44 (2), II Sem., pp. 91-115.

_____. (2008). Lingüística de corpus: una introducción al ámbito. *RLA. Revista de Lingüística Teórica y Aplicada*, Concepción (Chile), 46 (1), I Sem., pp. 93-119.

Parodi, G. y Burdiles, G. (2019). Corpus y bases de datos (Corpora and databases). En J. Muñoz *et al*. (coord*.), The Routledge Handbook of Spanish Language Teaching: metodologías, contextos y recursos para la enseñanza del español,* pp. 596-612, Routledge.

Pascual Aliaga, E. (2019*). Los truncamientos en la conversación coloquial: estudio de las huellas de formulación discursiva desde un modelo de unidades de lo oral*. [Tesis doctoral, Universitat de València]. Repositorio en línea Roderic, https://roderic.uv.es/handle/10550/72956

Paquot, M. y Gries. S. Th. (2021). *A Practical Handbook of Corpus Linguistics*. Springer International Publishing AG.

Pons Bordería, S. (2022). *Creación y análisis de corpus orales: saberes prácticos y reflexiones teóricas*. Peter Lang.

Pons Bordería, S. (dir.). Corpus Val.Es.Co 3.0. <http://www.valesco.es>

Pons, S. y Gurillo, L. (2005). Corpus para el estudio de la conversación coloquial: el corpus Val.Es.Co (Valencia Español Coloquial). *Oralia: Análisis del discurso oral*, 8, pp. 243-264.

Posit team. (2023). *RStudio: Integrated Development Environment for R*. Boston, MA: Posit Software, PBC. https://posit.co/

Poyatos, F. (1994). *La comunicación no verbal,* Itsmo.

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

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Wickham, H. y Müller, K. (2022). DBI: R Database Interface. *R Special Interest Group on Databases (R-SIG-DB)*, https://cran.r-project.org/

Real Academia Española. (en línea) Diccionario de la lengua española (23.ª ed., versión 23.6). https://dle.rae.es.

_____. Banco de datos (CORPES XXI) [en línea]. Corpus del Español del Siglo XXI (CORPES). <http://www.rae.es>

Reglamento (EU) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

Reppen, R. (2010). Building a corpus: what are the key considerations? En A. O'Keeffe, y M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*, pp. 31-37, Routledge.

Rock, F. (2001). Policy and Practice in the Anonymisation of Linguistic Data, *International Journal of Corpus Linguistics* 6(1), pp.1-26.

Rojo, G. (2016a), Corpus textuales del español. En J. Gutiérrez Rexach (coord.), *Enciclopedia de Lingüística Hispánica,* vol. 2, pp. 285-296.

_____. (2016b) Citius, maius, melius. Del CREA al CORPES XXI. En J. Kabatek, C. de Benito Moreno (coords.), *Lingüística de corpus y lingüística histórica iberorrománica*, pp. 197-212, De Gruyter.

_____. (2021), *Introducción a la lingüística de corpus en español*, Routledge.

Rojo, G. *et al*. (en línea), *Descripción del sistema de codificación: textos orales*, CORPES XXI. https://www.rae.es/corpes/assets/rae/files/codOral.PDF

Romero Trillo, J. (2020). Corpus Pragmatics. En D. Koike y C. Félix-Brasdefer, *Handbook of Spanish Pragmatics*, Routledge.

Rufino Morales, M.(2020). Estudio comparativo de métodos de transcripción para corpus orales: el caso del español. *Revista Nebrija de Lingüística aplicada a la enseñanza de Lenguas*, vol. 14, 29, pp. 126-146.

Rühlemann, C. y Ajmer, K. (2015). *Corpus Pragmatics. A Handbook*. Cambridge University Press.

Sacks, H. (1992). *Lectures on Conversation*. Blackwell.

Sacks, H., Schegloff, E. A., y Jefferson, G. (1974). A Simpest Systematics for the Organization of Turn-taking for Conversation. *Language*, 50(4), pp. 696-735.

Samper, J. A., Hernández, C. E. y Troya, M. (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, Servicio de Publicaciones y Difusión Científica de la ULPGC.

Samper Padilla, J. A. (2005). Macrocorpus para el estudio de la norma lingüística culta. *Oralia: Análisis del discurso oral*, 8, pp. 105-122.

Sánchez Sánchez, M. (2005). El corpus de referencia del español actual (CREA): el CREA oral. *Oralia: Análisis del discurso oral*, 8, pp. 37-56.

Schneider, K. P. (2018). Methods and ethics of data collection. En A. H. Jucker, K. P. Schneider y W. Bublitz (ed.), *Methods in Pragmatics*, pp. 37-93, De Gruyter.

Schegloff, E. A. (2007). *Sequence Organization in Interaction. A Primer in Conversation Analysis I*. Cambridge University Press.

Senft, G. (2009). Culture and Language Use. En G. Senft, J. Östman y J. Verschueren (Eds.), *Handbook of Pragmatics Highlights*, 2, pp. 105-109.

Sierra Martínez, G., (2017). *Introducción a los Corpus Lingüísticos*. Universidad Nacional Autónoma de México-Instituto de Ingeniería.

Sinclair, J. (1996). *Preliminary Recommendations on Corpus Typology*. EAGLES (Expert Advisory Group on Language Engineering Standards).

_____. (1997). Corpus Evidence in Language Description. En A. Wichmann *et al*. (Eds.), *Teaching and Language Corpora*, pp. 27-39, Longman.

_____. (2004). Corpus and Text. Basic Principles. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* En línea: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Sipos, D. (2023). *Drupal 10 Module Development: Develop and Deliver Engaging and Intuitive Enterprise-Level Apps*. Packt Publishing, Limited.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Textbooks in Language Sciences. Language Science Press.

Solís García, I. (2018). Corpus españoles dialógicos para el análisis de la conversación. *CHIMERA: Romance Corpora and Linguistic Studies*, 5 (1), pp. 117-129.

Survey of English Usage Corpus. (En línea). Disponible en https://www.ucl.ac.uk/english-usage/about/history.htm

Svartvik, J. (ed.) (1990). *The London-Lund Corpus of Spoken English: Description and Research*. Lund University Press.

Thompson, P. (2004). Spoken language corpora. En M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice.* En línea: http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*, John Benjamin.

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

Torruella, J. y Llisterri J. (1999). Diseño de corpus textuales y orales. En J. M. Blecua *et al*. (Eds.), *Filología e Informática. Nuevas tecnologías en los estudios filológicos*, Universidad Autónoma de Barcelona, pp. 45-77, Milenio.

Uclés Ramada, G. (2019). Transcripción de conversaciones con ELAN. En línea https://esvaratenuacion.es/protocolo-de-trabajo

Ueda, H. (2021). *Cómo usar el sistema LYNEAL. Letras y Números en Análisis Lingüístico*. En línea https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/doc/how-to-es.PDF

Vázquez Rozas, V. (2014). *ESLORA: diseño, codificación y explotación de un corpus oral del español de Galicia*, II Workshop de Procesamiento Automatizado de Texto y Corpus (WOPATEC-2014). Viña de Mar: Pontificia Universidad Católica de Valparaíso, 13-14 de noviembre de 2014 [en línea] https://gramatica.usc.es/~vvazq/PDF_publ/corpus_eslora_pres.PDF

Vázquez, V. y Recalde, M. (2009). Problemas metodológicos en la formación de corpus orales. En P. Cantos y A. Sánchez (Eds.), *A survey of corpus-based research*, pp. 51-64.

Recurso electrónico https://www.um.es/lacell/aelinco/contenido/index.html

Verdejo Muñoz, M. (2020). Conocimientos y actitudes de estudiantes universitarios hacia la diversidad de género y la diversidad sexual en un contexto multicultural. *MODULEMA. Revista científica Sobre Diversidad Cultural*, 4, pp. 42–65.

Villayandre Llamazares, M. (2008). Lingüística con corpus (I). *Estudios humanísticos. Filología*, 30, pp. 329-349.

Walsh, S. (2013). Corpus Linguistics and Conversation Analysis at the Interface: Theoretical Perspectives, Practical Outcomes. En J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics*, pp. 37-50. Springer.

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. https://ggplot2.tidyverse.org/

Wickham, H. *et al.* (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wijffels, J. (2023). Udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. https://cran.r-project.org/web/packages/udpipe/index.html

Wynne, M. (2005), *Developing Linguistic Corpora: a Guide to Good Practice*. En línea http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm (12.10.2005)

Xiao, R. (2008). Well-known and influential corpora. En A. Lüdeling and M. Kytö (2008) (Eds.), *Corpus Linguistics: An International Handbook*, pp. 383-456, De Gruyter.

Índice de tablas y figuras

Figuras

Tablas

Anexos

Índice de anexos

ANEXO 1

Listado de corpus mencionados

ACUAH	Análisis de la Conversación de Alcalá de Henares
ADESSE	Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español
ALCORE	Alicante Corpus Oral del Español
Ameresco	América y España Español Coloquial
ARTHUS	Archivo de Textos Hispánicos de la Universidad de Santiago
Atlas interactivo de la entonación española	
Brown Corpus	Brown University Corpus of American English
CAES	Corpus de Aprendices de Español
CARACAS-77 y CARACAS-78	Estudio sociolingüístico de Caracas 1977 y 1978
Cardiocor	Cardio Corpus
CE	Corpus del Español de Mark Davies
CEAP	Corpus de Encuestas en Asunción de Paraguay
CECBNA	Corpus del Español Conversacional de Barcelona y su Área metropolitana
CEMC	Corpus del Español Mexicano Contemporáneo I-II

CET	Corpus del Español en Texas
CHEM	Corpus Histórico del Español de México
CLHA	Corpus Lingüístico del Habla de Almería
COBUILD Corpus	Bank of English
COEM	Corpus Oral del Español de México
COGILA	Corpus del Grupo de Investigación Lingüística Aplicada
COJEM	Corpus Oral Juvenil del Español de Mallorca
COLA	Corpus Oral de Lenguaje Adolescente
COLEH	Corpus Oral de la Lengua Hablada en Honduras
COLEM	Corpus Oral de la Lengua Española en Montreal
C-ORAL-ROM	Corpus Oral de las Lenguas Romances
CORDE	Corpus Diacrónico del Español
CORDIAL	Corpus Oral Didáctico Anotado Lingüísticamente
COREC	Corpus Oral de Referencia del Español en Contacto
COREMAH	Corpus Español Multimodal de Actos de Habla
CORLEC	Corpus de Referencia de la Lengua Española Contemporánea

CORLEC	Corpus Oral de Referencia de la Lengua Española Contemporánea
CORPES XXI	Corpus del Español del siglo XXI
CORPEUU	Corpus del Español en los Estados Unidos
Corpus CORALES	
Corpus de Lovaina	
Corpus de Puerto Cabello	
Corpus de Valencia (Venezuela)	
Corpus del Español Oral en Bilbao	
Corpus del español Oral en Bilbao y su Área metropolitana	
Corpus Lingüístico de Referencia de la Lengua Española en Argentina	
Corpus Lingüístico de Referencia de la Lengua Española en Chile	
COSER	Corpus Oral y Sonoro del Español Rural
COVJA	Corpus Oral de la Variedad Juvenil Universitaria del Español Hablado en Alicante
CRATER	Corpus Resources and Terminology Extraction
CREA	Corpus de Referencia del Español Actual
CSC	Corpus para el estudio del español hablado en Santiago de Compostela

CSCM	Corpus Sociolingüístico de la Ciudad de México
CSMV	Corpus Sociolingüístico de Mérida, Venezuela
CUMBRE	
EGREHA	Estudio Gramatical del Español Hablado en América
El español hablado en Bogotá	
El habla popular	
ENTREVIS	
ESLORA	Corpus para el Estudio del Español Oral
GRIAL	
LEXESP	Léxico Informatizado del Español
LLC	London-Lund Corpus of Spoken English
LOB	Lancaster-Oslo/Bergen Corpus
MC-NC	Macrocorpus de la Norma Lingüística Culta de las principales ciudades de España y América
MeSA	Macrosintaxis del Español Actual
METAPRES	El discurso metalingüístico en la prensa española
NERC	Network of European Reference Corpus
PaGeS	Corpus paralelo alemán-español
PAROLE	PARallèle Oral en Langue Etrangère

PILEI	Programa Interamericano de Lingüística y Enseñanza de Idiomas
PRESEEA	Proyecto de Estudio Sociolingüístico del Español de España y América
SBCSAE	Santa Barbara Corpus of Spoken American English
SEE	Survey of Spoken English
SEU	Survey of English Usage Corpus
Spokes Corpus	
Val.Es.Co.	Valencia Español Coloquial
VHW	Voices of Hispanic World
VUM	Vernáculo Urbano Malagueño

ANEXO 2

Modelo de consentimiento informado del corpus Ameresco

AUTORIZACIÓN PARA EMPLEAR LA GRABACIÓN Y LA TRANSCRIPCIÓN DEL MATERIAL CON FINES INVESTIGADORES EN LINGÜÍSTICA

(Proyecto MINECO FFI2016-75249P)

A. Autorización previa a la grabación

Dña./D. _____ con documento de identificación o pasaporte número _____

DECLARO

1) que se me ha informado de que voy a ser grabado/a de forma secreta en las próximas semanas;

2) que, posteriormente a la grabación, podré escuchar el contenido de mi grabación;

3) que, en caso de no estar de acuerdo, puedo ejercer mi derecho a retirar la grabación.

A los efectos oportunos, firmo la presente autorización en _____ a _____de 20____.

Fdo. _____

B. Autorización posterior a la grabación

Dña./D. _____ con documento de identificación o pasaporte número _____

DECLARO

1) que se me ha informado de que he sido grabado/a secretamente y he escuchado el contenido de mi grabación;

2) que se me ha informado de que puedo ejercer mi derecho a retirar la grabación.

Y, por tanto, AUTORIZO al uso de la grabación de su contenido, previamente anonimizados texto y audio, para fines estrictamente de investigación.

A los efectos oportunos, firmo la presente autorización, en _____ a _____de 20____.

Fdo._____

1. Datos personales

Los datos personales obtenidos mediante el presente formulario se incorporarán a los sistemas de información de la Universitat de València – Estudi General (links.uv.es/lopd/dpo) en el marco del Proyecto Es.Vag.Atenuación. La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de América y de España (Proyecto MINECO FFI2016-75249P), dirigido por las doctoras Marta Albelda Marco y Maria Estellés Arguedas.

La información objeto de tratamiento será utilizada para el desarrollo de funciones docentes y académicas propias de la Universitat de València como la investigación, la creación, desarrollo, transmisión y crítica de la ciencia, de la técnica y de la cultura y la difusión, la valorización y la transferencia del conocimiento.

En concreto, estas grabaciones formarán parte del corpus oral del español coloquial Ameresco. Dicho corpus está compuesto por un conjunto de microcorpus de conversaciones coloquiales obtenidas en las distintas ciudades que se integran en el proyecto. Con el objetivo de analizar estas muestras de habla, en este proyecto se recogen conversaciones informales espontáneas (coloquiales) reales grabadas en lugares cotidianos para los hablantes, en una situación de familiaridad o amistad para su posterior análisis lingüístico.

La Universitat de València se compromete a que cualquier divulgación pública de los resultados obtenidos con motivo de la investigación se realizará anonimizando debidamente los datos utilizados, de modo que los sujetos de la investigación no resultarán identificados o identificables.

La base jurídica del tratamiento es el consentimiento del afectado/a y se prevé la conservación de los datos personales durante cinco años. Transcurrido ese periodo, los datos se conservarán debidamente disociados para garantizar el anonimato.

2. Registro de imagen o sonido

En el marco del desarrollo de la actividad se obtendrán registros de audio. Ud. Autoriza a la Universitat de València al uso, edición, difusión y explotación de estos registros exclusivamente para fines de investigación. En caso de utilización, se asegurará que el afectado/a nunca sea identificado por su nombre ni mediante información alguna que le haga identificable.

Todo ello con la única salvedad y limitación de aquellas utilizaciones o aplicaciones que pudieran atentar a los derechos garantizados en la Ley Orgánica 1/1982, de 5 de mayo, de Protección Civil al Derecho al Honor, la Intimidad Personal y familiar y a la Propia Imagen, así como del pleno respeto de las previsiones específicas del art. 4 de la Ley Orgánica 1/1996, de 15 de enero, de protección jurídica del menor.

3. Publicación

Los resultados del proyecto son susceptibles de publicación. En caso de tal utilización, se asegurará que Ud. nunca sea identificado/a por su nombre apellidos, ni mediante información alguna que le haga identificable.

4. Ejercicio de derechos

Las autorizaciones concedidas en este documento podrán ser revocadas mediante la presentación del oportuno escrito. La revocación comportará la retirada de la información de los sistemas de la Universitat de València en un plazo prudencial de tiempo en función de la disponibilidad de recursos. Puede obtener más información acerca de sus derechos en: links.uv.es/lopd/derechos

Y en prueba de conformidad, firmo el presente documento en el lugar y la fecha indicados en el encabezamiento.

Nombre y apellidos Firma	Nombre y apellidos Firma PADRE / MADRE / TUTOR *Rellenar solo en caso de menores o personas con incapacidad legal.*

ANEXO 3

Modelo de ficha técnica de los corpus Val.Es.Co. y Ameresco

FICHA TÉCNICA

a) Investigador:

b) Datos identificadores de la grabación:

- Fecha de la grabación:

- Tiempo de la grabación

Duración del audio	
Momento de inicio de la transcripción (minuto:segundo)	
Momento de finalización de la transcripción (minuto:segundo)	

- Lugar de grabación**:**

Municipio	
Espacio concreto (casa, bar, aula…)	

c) Situación comunicativa:

- Temas:

- Propósito o tenor funcional predominante (marcar con una **X** una de las dos opciones):

Transaccional	
Interpersonal	

- Tono:

- Modo o canal:

d) Tipo de discurso (conversación, debate…):

e) Técnica de grabación (seleccionar una de las dos opciones de cada fila):

Conversación libre / semidirigida	
Investigador participante / no participante	
Grabación secreta / no secreta	

f) Descripción de los participantes:

- Número de participantes:

- Tipo de relación que los une (familia, amigos, hermanos, compañeros de piso…):

- Cuadro de rasgos sociolingüísticos:

Clave hablante	**Sexo (V/M)**	**Edad (especificar edad exacta dentro de una de las tres franjas etarias)**			**Nivel de instrucción**			**Activo/Pasivo**	**Monolingüe cast./Bilingüe**	**Profesión**	**Residencia habitual (municipio)**
		18-25	**26-55**	**>55**	**Bajo**	**Medio**	**Alto**				
A											
B											
C											
D											
E											
F											
G											

g) Grado de prototipicidad coloquial (marcar con una **X** una de las dos opciones)**:**

Conversación coloquial prototípica	
Conversación coloquial periférica	