RESLA
Revista Española de Lingüística Aplicada

https://matrix.aesla.org.es/RESLA
ISSN 0213-2028 | e-ISSN 2254-6774

¿Es posible elaborar corpus orales espontáneos y cumplir la legislación? El modelo en tres fases del corpus Ameresco

Andrea Carcelén Guerrero

Universidad Católica de Valencia San Vicente Mártir, España

Resumen: Hoy día, la construcción de corpus orales presenta un marco propicio para su desarrollo, ya que cuenta con herramientas informáticas que permiten recoger y procesar muestras de habla fácilmente. Sin embargo, la recogida de corpus orales de conversación espontánea plantea dos retos: por un lado, el cumplimiento de las diferentes legislaciones que velan por el derecho a la intimidad y la privacidad de las personas, así como de las políticas de protección de datos; estas implican la obligatoriedad de obtener el consentimiento informado previo de los hablantes y la aplicación de un sistema de anonimización de datos personales. Por otro, evitar la paradoja del observador laboviana para obtener conversaciones en las que los hablantes no alteren su comportamiento lingüístico. Con el objetivo de establecer un modelo de buenas prácticas en materia de protección y tratamiento de los datos, en este trabajo repasaremos las prácticas de tratamiento de datos de corpus orales españoles (PRESEEA, COSER y ESLORA); analizaremos la legislación que aplica a la construcción de corpus y presentaremos el modelo de consentimiento en tres pasos empleado en el corpus Ameresco. Este garantiza, por un lado, una recogida legal, puesto que se informa al hablante de que va a ser grabado próximamente sin especificar cuándo y, por otro, la naturalidad de la recogida, ya que el hablante será grabado en un momento posterior.

Palabras clave: *lingüística de corpus, corpus orales, protección de datos, anonimización, español coloquial.*

Is it possible to produce spontaneous oral corpora and comply with legislation? The three-phase model of the Ameresco corpus

Abstract: Nowadays, the construction of oral corpora presents a favorable framework for its development, since there are computers tools that allow to collect and process speech samples easily. However, the collection of oral corpora of spontaneous conversation poses two challenges: on the one hand, compliance with the different legislations that ensure the right to privacy and privacy of individuals, as well as data protection policies, these imply the obligation to obtain the prior informed consent of the speakers and the implementation of a system of anonymization of personal data. On the other hand, to avoid the Labovian observer paradox in order to obtain conversations in which the speakers do not alter their linguistic behavior. With the aim of establishing a model of good practices in data protection and data processing, in this paper we will review the data processing practices of Spanish oral corpora (PRESEEA, COSER and ESLORA). We will analyze the legislation that applies to corpus construction and we will present the three-step consent model used in the Ameresco corpus. This guarantees a legal collection, since the speaker is informed that he/she is going to be recorded soon without specifying when, as well as the naturalness of the collection, since the speaker will be recorded at a later time.

Keywords: *Corpus linguistics, oral corpora, data protection, anonymization, colloquial Spanish.*

How to cite: Carcelén Guerrero, A. (2026). Is it possible to produce spontaneous oral corpora and comply with legislation? The three-phase model of the Ameresco corpus. *Revista Española de Lingüística Aplicada, 39*(2), 42-67. https://doi.org/10.58859/resla.1535

*Corresponding author: carguean@gmail.com

AESLA
Asociación Española de Lingüística Aplicada

Available under the
CC BY 4.0 license. ©

1. Introducción

En el presente trabajo abordaremos las obligaciones que se desprenden de las normativas vigentes en cuestión de protección de datos personales y privacidad a la hora de recoger corpus orales, concretamente, para el caso de corpus de conversación coloquial espontánea. Se pretende, por tanto, 1) ofrecer la base legal que afecta a la construcción de corpus orales, y 2) establecer un modelo de buenas prácticas en materia de protección y tratamiento de los datos extrapolable a la construcción de otros corpus de estas características, a raíz de la revisión de la metodología de otros corpus orales del español y de lo establecido para el corpus Ameresco.

La particularidad de los corpus de conversación coloquial espontánea reside en que, para poder realizar una investigación lingüística lo más rigurosa y objetiva posible, los hablantes no deberían saber que están siendo grabados. De ser así, se produciría la *paradoja del observador* laboviana (Labov, 1983, p. 66), según la cual el investigador debe encontrar la manera de ver cómo habla una persona cuando no está siendo observada de manera sistemática. Sin embargo, para acceder a estos datos la observación sistemática es necesaria. Lo ideal sería poder recoger estos materiales en un contexto de grabación que diera lugar a muestras de habla naturales sin que los hablantes fueran conscientes de que están siendo grabados; sin embargo, desde un punto de vista ético plantearía problemas legales, puesto que parecería entrar en conflicto con el obligado cumplimiento de la legislación sobre el derecho a la intimidad y a la propia imagen, así como a la protección de los datos personales y los derechos digitales.[1]

La conversación coloquial se caracteriza, justamente, por la presencia de rasgos situacionales como la relación de igualdad entre los interlocutores, su desarrollo dentro de un marco discursivo familiar en el que hay un saber y unas experiencias compartidas entre los hablantes, y por una temática no es especializada (Briz, 1995, 2001). Estas características, además, se manifiestan en una serie de rasgos primarios como son la ausencia de planificación de la conversación, es decir, la toma de turnos no está preestablecida, en la presencia de un fin comunicativo socializador y en el tono informal predominante (Briz, 1995, 2001).

La base de un estudio de la variedad coloquial del español debería garantizar que no se alteren estos rasgos 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, 2001). Por ello, antes de plantear cualquier investigación que implique el uso de grabaciones de conversación coloquial espontánea debe instaurarse una metodología que establezca

[1] La legislación que opera, en este sentido, sería la siguiente: 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 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.

cómo van a ser las tareas de recopilación y cómo va a realizarse la anonimización de los datos para evitar la identificación de los hablantes. Así mismo, debe plantearse un modelo de consentimiento informado que estos deberán firmar.

Para lograr un grado de naturalidad óptimo existen dos posibilidades. Una de ellas, 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.

No obstante, no hay estudios sistemáticos al respecto que justifiquen esa impresión de manera objetiva: se cree que los hablantes acaban olvidando la presencia de mecanismos de grabación, pero no se ha comprobado con corpus de grabación secreta para ratificar si, en efecto, es así, y si lo es en todas las etapas de la grabación, solo en las partes centrales, etc.

La segunda posibilidad, sugerida en este artículo, consiste en utilizar un modelo de consentimiento informado que proponga una manera de lograr la naturalidad de la muestra lingüística respetando a su vez la normativa legal. En este sentido, se presenta la metodología que se ha utilizado en el corpus Ameresco, donde se emplea un modelo de consentimiento en tres fases en el que los informantes (1) consienten ser grabados previamente a la recogida de la muestra, pero sin especificar en qué momento exacto se realizará esta; (2) ratifican su consentimiento en el momento posterior a la grabación, y, (3) junto a estas dos autorizaciones, recoge, a su vez, el consentimiento para el tratamiento de los datos personales.

La estructura de este trabajo es la siguiente: en primer lugar, haremos un repaso a las particularidades de la recogida de corpus orales (sección 2); en segundo lugar, describiremos los principios éticos y legales que operan en la recopilación de corpus orales grabados secretamente (sección 3); a continuación, examinaremos el modo de proceder de diferentes corpus orales del español a este respecto (sección 4), y, por último, detallaremos la metodología de trabajo seguida en la construcción del corpus Ameresco (sección 5) en cuanto a la recogida del consentimiento informado de los hablantes (sección 5.1.), así como del proceso de anonimización de datos personales que se ha establecido (sección 5.2.).

2. La recogida de corpus: la particularidad de los corpus orales

Cuando se aborda la planificación de un corpus, el primer principio fundamental se relaciona con el propósito de investigación, es decir, cuál es el objetivo de su creación, qué tipo de análisis lingüístico se busca realizar y cuáles son las preguntas de investigación que guían su recopilación. En segundo lugar, es necesario establecer de qué manera el corpus será representativo, definiendo los criterios de selección para obtener una muestra que refleje la diversidad en términos de variación diatópica, diastrática y/o diafásica. La reflexión sobre el tamaño del corpus, considerando su objetivo y las características de representatividad elegidas, también es crucial. Por ejemplo, la construcción de un corpus

para una investigación específica, como una tesis doctoral con recursos limitados y una persona a cargo de todo el proceso, difiere significativamente de la construcción de un corpus respaldado por un grupo de investigación o una institución con mayores recursos técnicos, humanos y económicos.

Otro aspecto fundamental se refiere a los materiales que se recopilarán y las implicaciones asociadas a estos. En el caso de corpus escritos, es necesario considerar cuestiones relacionadas con los derechos de autor de los archivos que conformarán la muestra. Para los corpus orales, especialmente aquellos que incluyen grabaciones de individuos particulares, surge una complejidad adicional en términos de conocimiento y cumplimiento de la legislación sobre privacidad y protección de datos, como veremos más adelante. En ambos casos, los investigadores deben asegurar el cumplimiento de los principios éticos que garanticen una conducta responsable en el tratamientos de estos datos durante la recogida de corpus (Schneider, 2018, McEnery & Hardie, 2011, Adolphs & Knight, 2010, Childs et al., 2001).

Finalmente, se deben abordar aspectos relacionados con la conservación y almacenamiento de los datos, así como la decisión sobre cómo se accederá y distribuirá el corpus. Esto implica establecer cómo el público podrá acceder al corpus y en qué condiciones, determinando si los materiales recopilados estarán disponibles de manera abierta o si se restringirá el acceso según sea necesario.

Al centrarnos específicamente en los corpus orales, surge una consideración adicional. A diferencia de la recopilación de materiales escritos, la incorporación de materiales a un corpus oral implica una fase adicional de transcripción y etiquetado del material antes de llegar al paso final de su inclusión en la página electrónica, si así se ha dispuesto. Por lo tanto, es necesario plantearse cómo representar lingüísticamente el habla a través de la escritura, la exhaustividad de la transcripción y codificación, así como la recopilación de metadatos e información contextual sobre las circunstancias de la grabación y los participantes, incluyendo los permisos y consentimientos informados de las personas involucradas en las grabaciones.

La tarea detallada es ardua, pero en la actualidad, la lingüística de corpus cuenta con un contexto más propicio para su desarrollo y ejecución, del que carecía hace apenas unos años. Recopilar un corpus es una tarea que se puede emprender con cierta facilidad, ya que los avances en tecnología y las nuevas herramientas informáticas ponen al alcance de los investigadores una gran cantidad de datos que puede servirles como base para construir su propio corpus lingüístico. Esta es una labor relativamente sencilla si el género discursivo objeto de la investigación es textual y no oral. Si es oral, el proceso es más productivo cuando las grabaciones proceden de medios públicos (televisión, radio, redes sociales) o cuando el trabajo se basa en géneros discursivos con mayor grado de planificación, como entrevistas o conversaciones semidirigidas.

En el caso que nos ocupa en este trabajo, el de los corpus orales de conversación coloquial espontánea grabados secretamente, el panorama se complica porque no es posible planificar de antemano las condiciones de la grabación y, además, porque el hecho de grabar secretamente conlleva la adopción de un protocolo para obtener el consentimiento informado que es diferente al de otros géneros discursivos, ya que, como veremos más adelante (sección 3), grabar a otra persona sin que esta lo sepa y difundir el contenido de la grabación, aunque sea con fines de investigación, puede acarrear problemas legales que deriven en sanciones penales.

Si bien en cualquier campo de investigación o práctica profesional existen métodos de anonimización válidos, en el campo de la lingüística no parece existir ningún estándar. A esta conclusión llega Rock (2001, p. 2) tras recopilar los puntos de vista de distintos investigadores, creadores y usuarios de corpus del panorama científico internacional para recabar las diferentes actitudes y prácticas existentes en la actualidad respecto a cuestiones éticas sobre tratamiento de datos, anonimización, etc.

McEnery & Hardie (2011, p. 60), por su parte, aluden a que mientras que sí existe información sobre cuestiones jurídicas relacionadas con la construcción de corpus, no se ha prestado suficiente atención a las cuestiones éticas que implican estos trabajos. Esto último se comprueba en los primeros corpus orales, creados cuando aún no estaban vigentes los requisitos legales actuales, donde no se planteaban cuestiones como que se pudiera estar atentando contra la intimidad de los participantes. En estos corpus, por ejemplo, sí que se pedía el consentimiento de los hablantes. En el caso del corpus PRESEEA, que recoge entrevistas semidirigidas, la petición de autorización previa a la grabación no presenta ningún inconveniente. Sin embargo, en el corpus de conversación coloquial Val.Es.Co. (2002), la petición de autorización previa podía condicionar las circunstancias de producción y afectar a los rasgos primarios y, por ende, a la prototipicidad de la conversación. Por esta razón, se solicitaba únicamente una autorización posterior, tras haberse realizado la grabación. Por lo tanto, si bien se tuvieron en cuenta algunas cuestiones éticas para la recogida y el tratamiento de la información y se tomaron medidas concretas (entre ellas, la anonimización de nombres propios), estas se aplicaron sin que hubiera una metarreflexión sobre la motivación que hay detrás de las decisiones tomadas. McEnery & Hardie (2011, p. 68), sin embargo, concluyen que buena parte de los errores cometidos en el pasado a este respecto en la lingüística de corpus se ha reparado hoy día.

Actualmente, en cambio, hay unanimidad en la literatura a la hora de aceptar que la metodología de recogida de corpus orales conlleva necesariamente la autorización y firma del consentimiento informado por parte de los hablantes, así como la obligatoriedad de proteger a estos participantes y los datos que proporcionen a la investigación por medio de un sistema de anonimización más o menos complejo, dependiendo de las características del corpus y la investigación. Así, se deberán contemplar diferentes estándares dependiendo del tipo de datos recogidos, sean solo auditivos, con contenido multimedia, etc. (Rock, 2001, Childs et al., 2011, Schneider, 2018, D'Arcy & Bender, 2023, McEnery & Hardie, 2011, Adolphs & Knight, 2010).

En las líneas que siguen se resumirá la parte de la legislación vigente en materia de protección y tratamiento de datos que afecta a la recogida de muestras de habla, anonimización y política de protección de datos en la investigación lingüística de corpus; seguidamente, se propondrá una metodología que trata de buscar un compromiso entre la naturalidad de la muestra y el cumplimiento de la legalidad, a través del modelo de autorización en tres fases propuesto en el corpus Ameresco (Albelda & Estellés, en línea, Carcelén & Uclés, 2019, Briz et al., 2020, Carcelén, 2024b).

3. El tratamiento de datos personales según la legislación española y europea

Para manejar cualquier tipo de datos personales, es necesario conocer las obligaciones que se desprenden de las legislaciones reguladoras sobre protección de datos y el derecho al honor y la intimidad. La metodología de trabajo de corpus no es una excepción, y debe asentarse sobre pilares éticos y de buenas prácticas en la investigación.

En particular, para el territorio español operan la Ley Orgánica 1/1982, de 5 de mayo, de Protección civil y derecho al honor, la intimidad personal y a la propia imagen, el Reglamento (UE) 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 y su aplicación en España a través de la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales.

El primer texto, la Ley Orgánica 1/1982, de 5 de mayo, establece en su artículo 1 el derecho al honor, la intimidad personal y familiar y la propia imagen como derechos fundamentales recogidos por la Constitución. Por tanto, cualquier intromisión ilegítima en este ámbito que no esté autorizada expresamente por ley o para la que no se haya otorgado un consentimiento expreso por parte del afectado sería penalizable (art. 2.2). Según el artículo 7 de esta ley, por intromisión ilegítima se entiende:

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. [...].

Nótese que la recogida de muestras de conversaciones coloquiales espontáneas, que preferentemente han de obtenerse de manera secreta para que no se vea alterado el comportamiento lingüístico de los participantes, requiere precisamente de la utilización de aparatos de grabación en contextos físicos que, de acuerdo con lo estipulado arriba, podría atentar contra la intimidad de las personas. Cabe señalar, además, que la grabación de una conversación sin haber obtenido el permiso de las personas afectadas puede constituir

un delito grave contra la intimidad según el Código Penal (artículo 197). Parece ser lícita la grabación sin consentimiento si se da una casuística como la siguiente: la persona que graba participa activamente de la conversación y esta podría ser utilizada como prueba judicial. Sin embargo, en cualquier otro caso, como la grabación de conversaciones ajenas o conversaciones en las que dicha persona apenas participe, con el agravante de que el contenido no solo se grabe, sino que también se difunda a terceros, sí se tipifica como delito grave contra la intimidad. Si lo trasladamos a la metodología de recogida de conversaciones del corpus Ameresco, 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 & 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, como hemos detallado en el apartado 1, 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 da 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 seudonimizados o anonimizados (véase apartado 4.2.).[2] 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 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*

[2] En este punto debe tenerse en cuenta que el significado de *datos anonimizados* y *datos seudonimizados* en el ámbito legal difiere en parte del concepto de *anonimización* en lingüística. En lingüística de corpus se utiliza *anonimización de datos* para referirse al proceso por el cual o bien se elimina o bien se sustituye por un término ficticio o por una codificación todo dato que pudiera permitir la identificación del hablante (Rock, 2001, McEnery & Hardie, 2011, Schneider, 2018, Adolphs & Knight, 2010, Childs et al., 2001). Mientras que, legalmente, *anonimizar* implica la eliminación total de datos personales e impide cualquier tipo de reidentificación, y *seudonimizar* conlleva el uso de un sistema de encriptación en base a ficheros que permitiría, en casos particulares recogidos por la Ley 3/2018, realizar la reidentificación del usuario. Piénsese, por ejemplo, en una investigación en la que se persigue la cura de una enfermedad y que tras realizar los ensayos pertinentes los resultados podrían mejorar la calidad de vida del paciente sometido al estudio.

> *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; énfasis nuestro).

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, p. 48; énfasis nuestro).

Si bien el Reglamento señala que los datos que hayan sido anonimizados con éxito ya no se consideran datos personales y, por tanto, los principios de protección de datos no deben aplicarse (esto es, podrían hacerse públicos), en el caso de los corpus orales se hace necesaria la adopción de las medidas más restrictivas posibles para evitar problemas legales, al no existir una anonimización completa. Si el corpus se concibe para ser abierto, se tendrá acceso al audio de la grabación y, por tanto, a la voz, que es un rasgo identificador directo de cada persona. Junto a esta premisa, cabe prestar atención a la parte destacada en cursiva, según la cual "deben tenerse en cuenta todos los factores objetivos [...] teniendo en cuenta tanto la tecnología disponible en el momento del tratamiento como los avances tecnológicos." Es decir, dado que el desarrollo de herramientas informáticas de tratamiento de sonido y voz es cada vez más rápido y más accesible a equipos caseros, la precaución en este sentido debe extremarse con vistas al futuro.

Por otro lado, volviendo al concepto de *consentimiento informado*, se señala la necesidad de que los usuarios, en este caso, los informantes, concedan su autorización para el uso de los datos recogidos de manera libre, informada e inequívoca, incluyendo las especificidades de aquellos casos en los que aparezcan menores de edad (Reglamento 2016/679 apartados 32 y 42 y Ley 3/2018 artículos 6 y 7; véase también, más abajo, el apartado 4.1). Para la elaboración de este consentimiento, la normativa plantea el *principio de transparencia* (Reglamento 2016/679, 39, Ley 3/2018 art. 11), por el cual se establece que debe informarse al usuario de forma inteligible, de fácil acceso y en un lenguaje claro y sencillo de los fines explícitos y legítimos para los que se utilizarán estos datos. Además, nombra 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).

Esto es, la persona afectada debe saber que puede retirar su consentimiento en cualquier momento y saber también a quién o a qué institución debe dirigirse para ejercer su derecho a retirarlo.

La normativa anterior se aplica a todos los casos en que se manejan datos personales, desde grabaciones de cámaras de seguridad hasta ficheros comerciales; sin embargo, cabe preguntarse si la ley contempla alguna especificidad para el caso de fines de investigación científica.

Volviendo al Reglamento Europeo 2016/679, en su apartado 159 se señala que su ámbito de aplicación concierne también al tratamiento de datos personales que se realice con fines de investigación científica entendida de manera amplia; esto es, comprendería el desarrollo tecnológico y la demostración, la investigación fundamental, la investigación aplicada y la investigación financiada por el sector privado. Además, se hace una mención especial a las investigaciones en el ámbito de la salud y, especialmente, en el ámbito biomédico; de hecho, hay disposiciones adicionales en la Ley 3/2018, que desarrollan específicamente las características del tratamiento de datos para la investigación sanitaria (disposiciones adicionales sexta y decimoséptima).

De todo lo anterior, se derivan dos aspectos: primero, existe poca legislación específica que regule las condiciones de la recogida de datos en investigación en general, probablemente por la multiplicidad de fines que puede tener esta, y se centra fundamentalmente en el ámbito de las ciencias de la salud, que implica datos altamente sensibles de los pacientes y requiere un proceso de anonimización y almacenamiento de los datos que se hace innecesario en el caso de los corpus orales espontáneos.[3] Segundo, se deriva que debe quedar meridianamente clara en el consentimiento que firman los sujetos cuál es la finalidad de la recogida, de modo que cabe detallar qué interesa especialmente al grupo de investigación.

Sin embargo, volvemos en este punto a una paradoja: explicitar el objeto de investigación, que en nuestro caso es lingüístico, puede llevar al sujeto a modificar significativamente su forma de expresión, puesto que se le hace consciente de que es la manera en la que habla lo que centra el interés de la investigación y, por tanto, su propio discurso pasa a ser su centro de atención. En este sentido, el protocolo que presentaremos en la sección 5 trata de minimizar el impacto de esta medida introduciendo un lapso de tiempo entre el consentimiento y la grabación efectiva que permita al sujeto olvidar la grabación y, con ella, quitar el foco de la forma de habla.

En resumen, a la hora de trabajar con datos personales, bien sea en el ámbito de la investigación científica o en cualquier otro ámbito (comercial, administrativo, financiero, publicitario, etc.), se hace necesario, por un lado, (a) obtener el *consentimiento* informado previo por escrito de los usuarios, los informantes o hablantes en el caso de la recogida de corpus orales y (b) someter estos datos a un proceso de *anonimización* que impida su

[3] No obstante, véase Carcelén (2024a) para analizar las especificidades de un corpus de conflicto entre íntimos y parejas, cuyo tratamiento se asemeja al del ámbito sanitario por la sensibilidad del contenido.

reidentificación. En la siguiente sección trataremos sobre las diferentes metodologías de anonimización encontradas en los principales corpus orales del español afectados por la legislación española y europea.[4]

4. Prácticas de tratamiento de datos en corpus orales del español: el consentimiento informado y la anonimización

De acuerdo con los preceptos contenidos en la legislación española y europea presentados arriba, queda claro que para una buena praxis en la investigación se necesita obtener el consentimiento informado de los participantes y adoptar un sistema de anonimización de datos. Así mismo, los corpus orales que se recogen en la actualidad deben incluir en su metodología de trabajo los objetivos relacionados con sus intereses de investigación. Su construcción requiere de un trabajo previo de reflexión y concreción de las medidas que se van a tomar para hacer frente a posibles implicaciones éticas y legales. Además, se debe de garantizar la protección de los participantes de la manera más estricta posible.

4.1 El consentimiento informado

Conseguir el consentimiento informado es un proceso fundamental en la investigación científica. En nuestro caso, los hablantes deben conceder al investigador de manera voluntaria y libre su permiso para usar las grabaciones que conforman el corpus.

Como se ha mencionado en el apartado 2, las especificidades de este consentimiento variarán dependiendo del género discursivo que se desee analizar en base a muestras de habla reales.

La recogida del consentimiento informado es una práctica habitual en la investigación lingüística actual. Aplicados al ámbito de la investigación, los requisitos implican que el interesado debe, en la línea de lo señalado por D'Arcy & Bender (2023, pp. 53-54) comprender qué se les está pidiendo, cómo se recopilarán sus datos, qué ocurrirá con ellos una vez obtenidos, tanto a corto como a largo plazo, cómo se protegerán estos datos y quién tendrá acceso a ellos, así como los derechos de los participantes a retirarse. No obstante, en el caso de la recogida de corpus orales existen consideraciones adicionales derivadas de la situación de grabación, en la que pueden verse involucradas terceras personas, bien por mención explícita, bien por participación accidental.

McEnery & Hardie (2011) plantean que hay cuatro áreas relacionadas con la ética en lingüística de corpus: la que afecta a los informantes, a los investigadores que construyen un corpus, a los distribuidores y a los usuarios, y mencionan la necesidad de que haya un consentimiento informado, apelando además al derecho de terceros, esto es, personas de las que se habla en la grabación y cuyo consentimiento no es posible obtener. Ambos autores se plantean qué ocurre cuando en la conversación se relatan acciones, nombres,

[4] En este artículo se ha trabajado con corpus orales del español dirigidos desde diferentes universidades españolas en los que, por tanto, operan la legislación española y europea. No se han tenido en cuentra otros corpus de español elaborados fuera de estas fronteras por tener estos que responder a otras legislaciones. No podemos, a su vez, hablar de corpus de español de España, ya que PRESEEA y Ameresco son corpus de carácter panhispánico.

etc. que pueden vulnerar su intimidad, e incluso ayudar a identificar al propio sujeto por correlación con estos. Sin embargo, no ofrecen una respuesta a esta inquietud. Ante esta situación, Adolphs & Knight (2010, pp. 42-43) afirman que hay que asegurar el consentimiento formal por escrito de *todos* los participantes implicados antes de llevar a cabo la grabación y, aunque esta práctica es ideal y deseable, no resuelve situaciones sobrevenidas en los entornos no preparados, como el hecho de encontrar grabaciones en espacios públicos en los que el sujeto conversa paralelamente con camareros, viandantes, etc., o en las que queda registrada una conversación ajena durante un silencio.

A modo de resumen, los mínimos que parecen imperar en la recogida de corpus lingüísticos orales, según lo visto anteriormente, pasan por solicitar el consentimiento de manera previa a la realización de la grabación (puesto que el sujeto debe saber que va a ser grabado y para qué), por redactar dicho consentimiento en un lenguaje claro y sencillo, que garantice la inteligibilidad y por incluir, al menos, los siguientes bloques de información:

- Consentimiento claro y explícito del informante (previo a la grabación –y posterior en el caso de los corpus que así lo requieran–).
- Naturaleza y objetivos concretos de la investigación.
- Propósito o finalidad de la recogida de estos datos.
- Tratamiento que se le dará a los datos para garantizar su privacidad (por ejemplo, la anonimización).
- Derechos que amparan al informante frente a la cesión de datos (derecho al desestimiento, así como los datos de contacto para que pueda ejercerlo).

4.2 La anonimización

De acuerdo con la opinión general en la bibliografía, no existe un estándar en lingüística para realizar la anonimización (Rock, 2001, Childs et al., 2011).

Recordemos que, según lo comentado en el apartado 3, en la recogida de corpus orales podrían considerarse datos personales todas aquellas informaciones relativas a una persona identificada o identificable según el Reglamento Europeo 2016/679 y la Ley Orgánica 3/2018. Por tanto, en el diseño del sistema de anonimización han de tenerse en cuenta todas las posibles variables que pudieran verse afectadas, desde los identificadores directos (nombres propios, apodos, direcciones) de las personas que aparecen en la grabación, es decir, todas aquellas características que por sí mismas permiten la identificación de una persona; hasta los identificadores indirectos (datos de historias que se relatan, fechas) que, si bien no permiten la reidentificación *per se*, si se cruzaran diferentes variables, podrían permitirla. Para evitar esta identificación o reidentificación, la legislación contempla los procedimientos de anonimización y seudonimización: los primeros ya no incluyen

CC BY

identificadores, mientras que, en los segundos, estos identificadores están encriptados de alguna manera, con diferentes grados de sofisticación dependiendo de las características de la investigación (ver sección 3).

En el caso de la recogida de datos que incluya registro de voz o de imagen, como el de los corpus orales, la Agencia Española de Protección de Datos (2016, p. 20) de manera genérica propone la posibilidad de realizar una transcripción previa con su respectiva eliminación de identificadores, así como la reproducción de las transcripciones mediante dispositivos sintetizadores de voz en el caso necesario. Sin embargo, en investigaciones lingüísticas este procedimiento afectaría a posibles análisis fónicos o prosódicos (Adolphs & Knight, 2010, p. 43), del mismo modo que el pixelado de imágenes para corpus con vídeo afectaría, por ejemplo, a estudios sobre kinésica.

4.3 El tratamiento de datos en corpus orales del español

Si nos centramos en el panorama actual de corpus orales del español en el territorio nacional, la recogida de muestras de habla naturales se ha hecho respetando las prácticas mencionadas anteriormente, es decir, se han recogido los consentimientos informados pertinentes y se ha aplicado un protocolo de anonimización en los casos en los que se ha planteado necesario. Para ejemplificar dicho panorama se han seleccionado los corpus orales PRESEEA, COSER y ESLORA. Todos ellos son de acceso público y están en formato digital, han sido informatizados, son sincrónicos y están en activo. Además, permiten el acceso directo a textos con su contextos, tanto a la transcripción escrita como al fragmento de audio correspondiente a la búsqueda realizada. Esta última característica es destacable no solo por su importancia a la hora de realizar estudios prosódicos, sino también porque la consulta del audio es imprescindible a la hora de desambiguar algunas cuestiones lingüísticas y referenciales, para entender la estructura de la información, la sintaxis y posibles significados pragmáticos adicionales que de otro modo se perderían.

PRESEEA, el corpus del Proyecto para el Estudio Sociolingüístico del Español de España y América,[5] constituye uno de los corpus más extensos para el estudio sociolingüístico del español. Recopila entrevistas semidirigidas de más de 40 países y, tal y como se recoge en su protocolo de trabajo (Moreno Fernández, 2021a, p. 13) no hay marca específica para señalar la anonimización, a excepción de la indicación de uso de mayúscula en la inicial de nombres propios. Al realizar consultas en el corpus, a través de su página general, puede observarse la sustitución de nombres propios por la inicial correspondiente (Figura 1) cuando se trata de nombres propios o nombres de lugares 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

[5] Acceso en línea en: http://preseea.linguas.net/Inicio.aspx

identificadores, ya que queda a criterio de la persona encargada de la transcripción la pertinencia de realizarla o no. En el audio, aquellos fragmentos que se han anonimizado se han borrado y sustituido por un indicador sonoro.

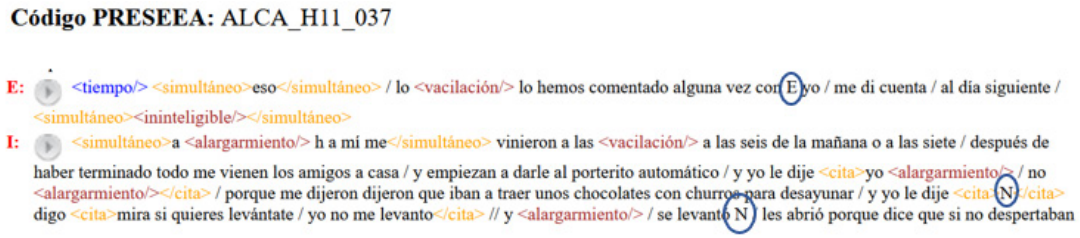


Figura 1. Captura de pantalla del corpus PRESEEA.

El Corpus Sonoro del Español Rural (COSER),[6] desde el ámbito de la dialectología, se centra en recoger muestras de lengua en enclaves rurales también por medio de entrevistas semidirigidas y dispone de 1744 horas de grabación. En este corpus, la anonimización se realiza por medio de la marca Nombre propio [NP] a nivel textual (Figura 2). Esta convención oculta el nombre propio de los informantes, para evitar su divulgación en la publicación de los materiales. Respecto a la anonimización del audio, se ha optado por la sustitución del nombre en cuestión por un indicador sonoro. Su etiquetado se engloba dentro de la sección *marcas de la conversación* y viene recogido en su página web.

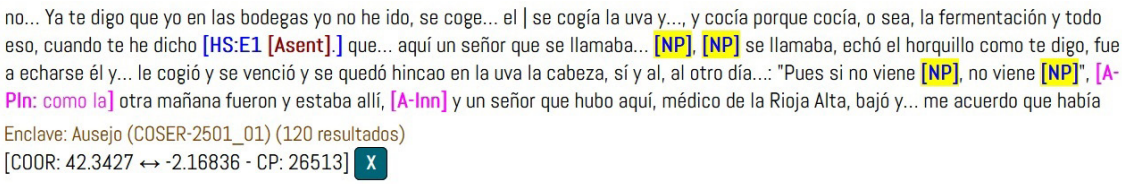


Figura 2. Captura del corpus COSER.

Esta marca presenta cierta problemática, ya que, aún 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 correlativa junto a la marca.

Por último, el corpus ESLORA[7] contiene 60 horas de entrevistas semidirigidas (integradas en PRESEEA) y 20 horas de conversaciones de hablantes de español de Galicia. En su versión actual, 2.1, de junio de 2022, el corpus consta de 83 documentos que incluyen 768 023 palabras ortográficas, de las cuales están accesibles para búsqueda

[6] Acceso en línea en http://www.corpusrural.es/index.php
[7] Acceso en línea en http://eslora.usc.es/

751 970. Quedan excluidas de las consultas las formas gráficas de fragmentos en lenguas distintas del español, las palabras cortadas y los nombres ficticios (3041), necesarios para garantizar el anonimato de los hablantes.

Barcala et al. (2018) ofrecen información explícita sobre la obtención del consentimiento y la metodología de anonimización, así como una metarreflexión sobre el tratamiento de datos y la privacidad de los hablantes. En concreto, se detalla que las conversaciones espontáneas recogidas implican un doble permiso, previo y posterior a la grabación. Además, manifiestan el compromiso por parte del equipo investigador de restringir el uso de las grabaciones y transcripciones a los ámbitos de la investigación y la docencia, y preservar el anonimato de los hablantes.

Respecto a la metodología de anonimización que han implantado para este corpus, en las transcripciones se han sustituido las menciones de los nombres y localizaciones que pudieran revelar la identidad de los informantes y de otras personas aludidas por denominaciones métrica y socialmente equivalentes (Sampson, 2000). Es un proceso laborioso pues, como indican estos autores, la versión anonimizada debe mantener la coherencia interna que permita el seguimiento de la continuidad referencial del discurso (algo que resulta más dificultoso en sistemas como el mostrado en la Figura 2) y, además, ha de resultar verosímil en lo que se refiere a la información geográfica y demás indicaciones contextualizadoras. En el audio, las referencias personales se eliminan y se sustituyen por un ruido.

En la Figura 3 puede observarse la aparición de nombres propios, en este caso ficticios, sin ninguna marca identificadora de haber sido sustituidos, pero al reproducir el fragmento de audio correspondiente a la búsqueda se comprueba que el original ha sido borrado y sustituido por un sonido.



Figura 3. Captura del corpus ESLORA.

De esta breve caracterización del tratamiento de datos en corpus orales del español se pueden extraer múltiples buenas prácticas desde el punto de vista ético y legal. Algunas de estas han sido incorporadas a la metodología del corpus Ameresco, junto con otras implementaciones, dando como resultado un consentimiento informado en tres pasos que, como se verá a continuación, busca lograr un compromiso entre la naturalidad del producto y el seguimiento de la legislación en materia de privacidad y protección de datos.

5. El corpus Ameresco

El corpus Ameresco (Albelda & 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, 2001, 2016). Se centra en conversaciones coloquiales grabadas secretamente en las principales ciudades de Hispanoamérica y España, y se encuentra todavía en construcción, ya que actualmente aparecen representados 9 países, con un total de 14 ciudades, y cerca de 750 000 palabras.

Su metodología de trabajo parte de los principios establecidos por PRESEEA (Moreno Fernández, 2005, 2016, 2021a, 2021b) y por Val.Es.Co. (Briz et al., 1995, 2002) en cuanto a la representatividad sociolingüística y el sistema de transcripción y procesado informático. El corpus se encuentra anotado y etiquetado en XML, así como alineado audio-texto por medio del programa de anotación ELAN (Carcelén & Uclés, 2019, Briz et al., 2020, Carcelén, 2024b). El corpus es de acceso abierto y cuenta con un motor de búsqueda que permite diferentes métodos de consulta, y tanto las conversaciones como los documentos de trabajo pueden consultarse en línea.[8]

5.1 El consentimiento informado en Ameresco

Como hemos visto el apartado 3, 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 (Briz, 1995, 2001) 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. Por ello, para el desarrollo del corpus Ameresco se diseñó un modelo de consentimiento informado en tres pasos que garantiza una recogida de datos conforme a la legalidad y pretende no alterar, o modificar en el menor grado posible, el comportamiento lingüístico de los hablantes.

A continuación se detalla la triple estructura del consentimiento informado en el corpus Ameresco, 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, los participantes rellenarán sus datos personales (nombre completo y número de identidad – DNI, cédula o el correspondiente según el país de recogida–). Es aquí donde 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. Sin embargo, no se especifica en qué momento exacto

[8] Acceso en línea en www.corpusameresco.com

se realizará. 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 una última sección sobre cómo se va a proceder al tratamiento de los datos personales. Esta práctica hoy día, como hemos visto en el epígrafe 3, 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. Según la Agencia Español de Protección de Datos,[9] ante cualquier intercambio de datos personales, en el ámbito comercial, en el ámbito de administración y servicios, etc., los usuarios deben poder ejercer el derecho de rectificación, oposición y supresión de los datos personales que deseen ante el responsable del tratamiento de esos datos.

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 Estellés y Albelda, 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 segundo 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.

[9] Disponible en línea: www.aepd.es/es/derechos-y-deberes/conoce-tus-derechos

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 3: obtención del consentimiento previo, del consentimiento posterior a la grabación y del permiso para el tratamiento de datos. La innovación más importante, que aplica a la recolección de corpus orales de conversación coloquial, 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 de 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. Se lleva utilizando en la recogida de grabaciones del corpus Ameresco desde ese momento (ver Apéndice 1). 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 aparición de la Ley 3/2018 se incorpora, además, el apartado específico sobre la política de protección de datos.

5.2 Anonimización

En el corpus Ameresco se ha desarrollado un proceso de anonimizazión en dos capas: a nivel textual (transcripciones) y a nivel oral (audio).

1. Nivel textual

   Para la anonimización de las transcripciones se ha optado por el reemplazo del identificador directo por otro ficticio.[10] En la medida de lo posible se ha intentado que la sustitución respete el patrón fónico del término original (Sampson, 2000, Vázquez Rozas, 2014, Barcala et al., 2018). Además, debe de atenderse 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. Otro factor a

[10] 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.

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. Esto es, 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 de que el término sustituto fuera un nombre propio característico de una zona catalanoparlante o un nombre actualmente típico de mujeres mayores de 60 años.

Los identificadores directos que han sido anonimizados se marcan con la etiqueta XML<anónimo></anónimo>.[11] En el caso de los identificadores indirectos, en algunos casos se ha optado por el empleo de esta etiqueta; en otros, ha sido necesaria la eliminación del fragmento con la indicación en línea de observaciones de que se ha producido un corte en la grabación. En la Figura 4 puede verse un extracto del corpus en el que se ejemplifica la anonimización realizada por medio de la etiqueta XML mencionada anteriormente.

Conversación	Ciudad	Edad	Intervención	Nivel	Sexo	Audio	Contexto
MEX_001_03_19	Ciudad de México	Más de 56	B: pues di- tenía una tienda en el centro/ vivía en el centro y tenía una tienda/ y que ese día fue<alargamiento/>// fue su hermano de <anonimo>Miguel</anonimo>/ o sea de <anonimo>Rubén</anonimo> [también son- var-] son cuatro hombres	Medio	Mujer	Escuchar	Ver contexto

Figura 4. Captura del corpus Ameresco

En algunos casos, no se ha encontrado una equivalencia válida que respete el contexto discursivo, por tanto, se ha optado no por buscar un término similar, sino por usar una observación con un título general. Por ejemplo, ante la imposibilidad de encontrar un nombre sustituto ficticio de una empresa, se ha optado por la siguiente codificación: <anónimo><obs t="Empresa1"/></anónimo>.[12]

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 automático. De momento, no se ha logrado automatizar esta primera capa, 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, si bien habría opciones como la búsqueda automatizada que distinguiera mayúsculas y minúsculas para discriminar 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

[11] El sistema de transcripción y etiquetado completo puede consultarse en línea https://esvaratenuacion.es/protocolo-de-trabajo

[12] En el caso de que esta codificación aparezca varias veces, es decir, haya que anonimizar nombres de varias empresas, la observación incluirá numeración consecutiva para evitar que haya ambigüedad referencial: <anónimo><obs t="Empresa 1"/></anónimo>, <anónimo><obs t="Empresa 2"/></anónimo>.

CC BY

no es viable. La persona encargada de realizar la transcripción deberá valorar cada identificador de este tipo y decidir la pertinencia o no de su anonimización. Por ejemplo, conlleva diferentes implicaciones decir que se pidió la cena a una cadena de hamburgueserías que decir que se trabaja en esa cadena. Solo en el segundo caso podría considerarse un identificador indirecto. Como se ha señalado en el apartado 4, queda en mano del transcriptor evaluar la necesidad de anonimización en cada caso particular.

2. Nivel oral

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. Para la eliminación de estos fragmentos se ha optado por silenciar el audio, en lugar de añadir un indicador sonoro.

El programa ELAN permite acotar los fragmentos que posteriormente se van a silenciar. En este caso, se crea una línea llamada *anónimo* donde se marcarán las secciones de tiempo seleccionadas. Una vez marcados en todo el audio, este *software* 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*.

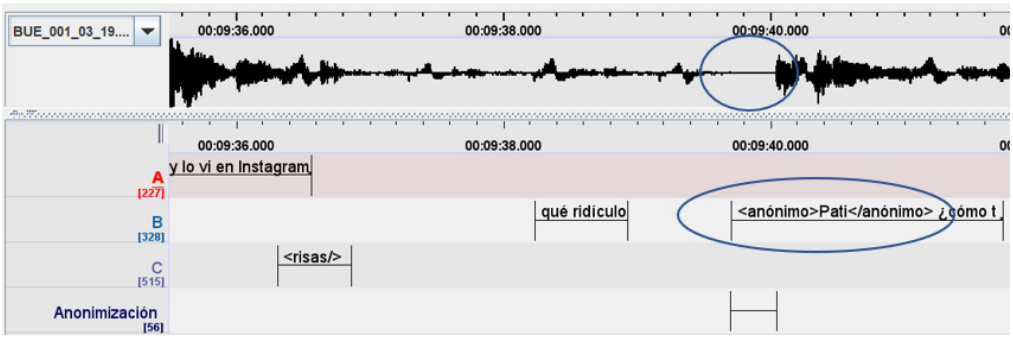


Figura 5. Captura de pantalla *software* ELAN.

Una vez realizada la exportación, el archivo resultante se carga en el *software* de edición de audio Audacity. Al realizar la importación se selecciona 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 Figura 5 podemos ver en primer lugar el borrado del segmento en el oscilograma de sonido (capa de audio), la sustitución que se ha hecho del nombre original por un nombre ficticio, 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.

6. Conclusiones

En este trabajo se han analizado en profundidad los retos que supone la recopilación de corpus orales del español. Concretamente, se abordan la fase de recogida de grabaciones, de transcripción y difusión, que conllevan la obtención del consentimiento de los participantes previamente a la recogida de la muestra de habla y la anonimización en la fase de tratamiento de datos personales para garantizar la privacidad y el derecho a la intimidad de los hablantes.

Si bien existen preceptos claros en la legislación actual sobre las implicaciones que conlleva la salvaguarda del derecho a la intimidad y la privacidad y al tratamiento de datos personales, su aplicación particular en el ámbito de la investigación lingüística cuenta con menos desarrollo. Es por esto por lo que, como investigadores, debemos ser cautos a la hora de recopilar corpus orales espontáneos planteando desde el principio un sistema de anonimización y tratamiento de datos personales lo más estricto posible para garantizar buenas prácticas en la investigación.

La legislación establece la necesidad de la recogida de los consentimientos informados, la claridad y el derecho a ejercer la retirada, así como someter audio y transcripción a un proceso de anonimización. Estos mínimos se mantienen en los corpus españoles actuales con algunas diferencias, pero no existe en general un estándar que sea aplicable a todos ellos.

El mantenimiento de la legalidad parece contradictorio con la garantía de la naturalidad. En este sentido, el modelo de Ameresco presenta un sistema en tres pasos que tiene como principal novedad el hecho de que la petición de autorización previa se realiza bajo la premisa de que la grabación tendrá lugar en el lapso de un tiempo determinado, entre cuatro y ocho semanas aproximadamente, pero sin especificar cuándo exactamente. Por lo tanto, se garantiza que el hablante se haya olvidado de que va a ser grabado y que no se produce la paradoja laboviana. De esta manera se obtienen muestras de habla reales y naturales, pero también se respeta la legislación vigente con la firma de la autorización previa, posterior y de la política de tratamiento de datos.

CC BY

Los retos son cada vez mayores porque la multimodalidad de los corpus, que recogen audio e imagen, gana presencia en la investigación lingüística. El rápido desarrollo de la tecnología y de las herramientas informáticas podrían, a su vez, facilitar el reconocimiento de las hablantes, por tanto los trabajos de identificación y anonimización de los datos personales presentes en el corpus deben asegurar de manera escrupulosa la privacidad y estos deben aún realizarse bajo la supervisión humana.

CRediT Author contribution / *Contribución de los autores*
Conceptualization / *Conceptualización:* Andrea Carcelén
Formal Analysis / *Análisis formal:* Andrea Carcelén
Methodology / *Metodología:* Andrea Carcelén
Writing / *Redacción:* Andrea Carcelén
Research dataset / *Datos de investigación:* Andrea Carcelén

Funding, data availability, and copyright / *Financiación, disponibilidad de datos y derechos de autoría*
Funding / *Financiación:* Este trabajo ha contado con el apoyo del 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 MCIN/AEI/ 10.13039/501100011033.
Image use consent / *Consentimiento para el uso de imágenes:* No procede.
Conflict of interest / *Conflicto de intereses:* No hay conflicto de intereses.
Data availability statement/ *Declaración de disponibilidad de datos:* Acceso abierto
License/ *Licencia:* This article is published under the CC BY 4.0 License / *Este artículo se publica bajo la Licencia CC BY 4.0.*

Editorial history / *Fechas del proceso editorial*
Received / *Recibido:* 03/07/2023
Accepted / *Aceptado:* 30/01/2024
Published / Publicado: 15/11/2026

Referencias

Adolphs, S., & Knight, D. (2010). Building a spoken corpus. What are the basics? En A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38–52). Routledge. https://doi.org/10.4324/9780203856949.ch4

Agencia Española de Protección de Datos. (2016). *Orientaciones y garantías en los procedimientos de anonimización de datos personales* [Archivo PDF]. 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 & Consejo de Europa. (2014). *Manual de legislación europea en materia de la protección de datos*. Oficina de Publicaciones de la Unión Europea.

Albelda, M., & Estellés, M. (2012–.). *Corpus Ameresco* [en línea]. Recuperado de www.corpusameresco.com

Barcala, M., Domínguez, E., Fernández, A., Rivas, R., Santalla, M. P., Vázquez, V., & Villapol, R. (2018). El corpus ESLORA de español oral: diseño, desarrollo y explotación. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, 5*(2), 217–237. https://doi.org/10.15366/chimera2018.5.2.003

Briz Gómez, A. (1995). *La conversación coloquial (Materiales para su estudio)*. Anejo XVI de la Revista Cuadernos de Filología. Cuadernos de Filología.

Briz Gómez, A. (2001). *El español coloquial en la conversación. Esbozo de pragmagramática*. Ariel.

Briz Gómez, A. (2016). El proyecto Ameresco: la idea de un corpus de conversaciones coloquiales del español de América. En A. M. Bañón et al. (Eds.), *Oralidad y análisis del discurso: homenaje a Luis Cortés* (pp. 81–104). Universidad de Almería.

Briz Gómez, A., et al. (2020). *Protocolo de trabajo para los equipos Ameresco* [Archivo PDF]. https://esvaratenuacion.es/protocolo-de-trabajo

Briz Gómez, A., & Grupo Val.Es.Co. (2002). *Corpus de conversaciones coloquiales*. Anejos Oralia. Arco Libros.

Carcelén Guerrero, A. (2024a). Aspectos éticos y legales en la construcción de un corpus oral de conversación conflictiva. *Cultura, Lenguaje Y Representación, 35*, 13–51. https://doi.org/10.6035/clr.7907

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

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

Childs, B., Van Herk, G., & Thorburn, J. (2011). Safe harbour: Ethics and accesibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory, 7*(1), 163–180. https://doi.org/10.1515/cllt.2011.008

D'Arcy, A., & Bender, E. (2023). Ethics in linguistics. *Annual Review of Linguistics, 9*(1), 49–69. https://doi.org/10.1146/annurev-linguistics-031120-015324

ESLORA. (2007–). *Corpus para el estudio del español oral* [En línea]. Recuperado de http://eslora.usc.es

Fernández Ordóñez, I. (dir.) (2005–). *Corpus Oral y Sonoro del Español Rural* [En línea]. Recuperado de http://www.corpusrural.es

Labov, W. (1983). *Modelos sociolinguisticos*. Cátedra.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Moreno Fernández, F. (2005). Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA. *Oralia: Análisis del discurso oral, 8*, 123–140. https://doi.org/10.25115/oralia.v8i1.8333

Moreno Fernández, F. (2016). En torno a PRESEEA: Notas de investigación y de sociología de la ciencia. *Boletín de filología: Universidad de Chile, 51*(2), 369–376. https://doi.org/10.4067/S0718-93032016000200014

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

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

PRESEEA. (2014–). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* [En línea]. Recuperado de http://preseea.uah.es

Rock, F. (2001). Policy and practice in the anonymisation of linguistic data. *International Journal of Corpus Linguistics, 6*(1), 1–26. https://doi.org/10.1075/ijcl.6.1.01roc

Sampson, G. (2000). *CHRISTINE Corpus: Documentation* [En línea]. Recuperado de https://www.grsampson.net/ChrisDoc.html

Schneider, K. P. (2018). Methods and ethics of data collection. En A. H. Jucker, K. P. Schneider, & W. Bublitz (Ed.), *Methods in Pragmatics* (pp. 37–93). De Gruyter Mouton. https://doi.org/10.1515/9783110424928-002

Vázquez Rozas, V. (2014). ESLORA: diseño, codificación y explotación de un corpus oral del español de Galicia. En *II Workshop de Procesamiento Automatizado de Texto y Corpus (WOPATEC-2014)*. Pontificia Universidad Católica de Valparaíso, 13–14 de noviembre de 2014. [Archivo PDF] https://gramatica.usc.es/~vvazq/pdf_publ/corpus_eslora_pres.pdf

Legislación

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).

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.

Apéndice 1

Modelo de autorización en tres pasos 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 subcorpus 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.*