Carcelén Guerrero, Andrea (2024): Aspectos éticos y legales en la construcción de un corpus oral de conversación conflictiva. *Cultura, Lenguaje y Representación*, Vol. XXXV, 13-51
ISSN 1697-7750 · E-ISSN 2340-4981
DOI: https://doi.org/10.6035/clr.7907

Aspectos éticos y legales en la construcción de un corpus oral de conversación conflictiva

Ethical and legal aspects in the construction of an oral corpus of conflictual conversation

ANDREA CARCELÉN GUERRERO
UNIVERSIDAD DE HELSINKI[1]
https://orcid.org/0000-0003-3454-8544

Artículo recibido el / *Article received*: 2024-01-30
Artículo aceptado el / *Article accepted*: 2024-06-21

RESUMEN: La creación de un corpus oral, definido como una colección de grabaciones de habla natural dispuestas para su tratamiento computacional, presenta varios desafíos que deben ser considerados cuidadosamente para garantizar su calidad y utilidad. En un corpus de las características del que se presenta a continuación, el corpus ESPRINT de conversación conflictiva entre personas con una relación íntima, el reto principal tiene que ver con atender a los principios ético-legales que implica tanto su recopilación, como su posterior procesamiento para el análisis lingüístico. En todo caso, debe garantizarse el derecho a la privacidad de los participantes y su anonimato. El corpus ESPRINT presenta una doble naturaleza, por un lado conversacional (ESPRINT-Conversación), por otro, terapéutica (ESPRINT-Terapias), hecho que implica dos modos de recogida, así como de tratamiento de los datos, de gestión de la privacidad y del almacenamiento y análisis, puesto que las terapias se graban en un contexto clínico sociosanitario con especificidades en cuanto al acceso a los datos, que obliga a un estricto protocolo de almacenamiento y una política clara sobre quién puede acceder al corpus, con qué fines y bajo qué condiciones, respetando los derechos de intimidad y los acuerdos de consentimiento. Además, en ambos

[1] Esta publicación es parte del proyecto de I+D+i «Estrategias pragmático-retóricas en la interacción conversacional conflictiva entre íntimos y conocidos: intensificación, atenuación y gestión interaccional (ESPRINT)» (ref. PID2020-114805GB-I00), financiado por MICIU/AEI/10.13039/501100011033/ y de la «Red Temática sobre comunicación conflictiva y mediación: interacción, vínculos relacionales y cohesión social (CoCoMInt)», financiada por la ayuda RED2022-134123-T de la convocatoria «Redes de Investigación 2022», MICIU/AEI /10.13039/501100011033.

DOI: https://doi.org/10.6035/clr.7907

casos, deben extremarse las precauciones dado el contenido sensible de las grabaciones. Las actuaciones llevadas a cabo desde la dirección del corpus ESPRINT en materia de protección de datos garantizan el cumplimiento de los preceptos legales, así como la salvaguarda del derecho a la intimidad y el tratamiento ético de los datos.

Palabras clave: lingüística de corpus, corpus orales, conversación espontánea, tratamiento ético-legal, protección de datos, terapia de pareja, conflicto

ABSTRACT: The creation of an oral corpus, defined as a collection of natural speech recordings arranged for computational processing, presents several challenges that must be carefully considered to ensure its quality and usefulness. In a corpus of the characteristics of the one presented below, the ESPRINT corpus of conflicting conversation between people with an intimate relationship, the main challenge has to do with attending to the ethical and legal principles involved in both its collection and its subsequent processing for linguistic analysis. In any case, the participants's right to privacy and anonymity must be guaranteed. The ESPRINT corpus has a dual nature, on the one hand conversational (ESPRINT-Conversation), on the other hand therapeutic (ESPRINT-Therapies), which implies two modes of collection, as well as data processing, privacy management, storage and analysis, since therapies are recorded in a socio-health context with specificities in terms of data access, which requires a strict storage protocol and a clear policy on who can access the corpus, for what purposes and under what conditions, respecting privacy rights and consent agreements. Furthermore, in both cases, extreme precautions must be taken given the sensitive content of the recordings. The actions carried out by the corpus management in the area of data protection guarantee compliance with the legal precepts in this regard, as well as the safeguarding of the right to privacy.

Key words: corpus linguistics, corpus orals, spontaneous conversation, ethical-legal treatment, data protection, couple therapy, conflict

1. INTRODUCCIÓN

En los últimos años ha habido un gran desarrollo de la lingüística de corpus en general y, en el caso del español, en particular. Así lo demuestran trabajos como los de Moreno Fernández (2005), Briz y Albelda (2009), Enghels, Vanderschueren y Bouzouita (2015), Rojo (2016), Solís (2018), Parodi y Burdiles (2019), Briz y Carcelén (2019), Llisterri (2021) y Carcelén (2024) que recogen, a modo de panorámica, los corpus orales del español más relevantes desde sus orígenes.

Sin embargo, como han señalado autores como Briz (2012: 124) o Vázquez y Recalde (2009: 52), existe una menor representación en este panorama de corpus orales del español de géneros conversacionales, frente a otros géneros discursivos, como la entrevista semidirigida, que sí cuentan con materiales en abundancia[2]. Este hecho puede estar motivado por las particularidades metodológicas que operan sobre la recogida de

[2] Otros tipos de corpus cuya creación ha experimentado un incremento en las últimas décadas han sido los corpus para el desarrollo de las tecnologías del habla y síntesis de voz o los corpus de aprendices de español.

material conversacional espontáneo, a saber, se deben recoger las grabaciones preferiblemente de manera secreta si es posible, en entornos que favorezcan la naturalidad de los datos, con una calidad de audio lo suficientemente óptima para su posterior procesamiento y transcripción y, además, se deben recoger los consentimientos informados de cada uno de los participantes y garantizar un tratamiento ético y legal de los datos obtenidos, así como salvaguardar la privacidad de los hablantes. Al ser interacciones que suceden de manera espontánea y no planificada, el control de las situaciones de grabación es más complicado. Además, la elaboración de corpus de lengua oral implica llevar a cabo un cuidadoso proceso de planificación y reflexión anticipada con el fin de garantizar que los datos recopilados sean representativos y adecuados para su posterior análisis (Carcelén, 2024: 82).

En el caso que nos ocupa, el corpus ESPRINT de conversación conflictiva (Albelda y Estellés, dirs.), este proceso es más complejo, ya que se trata de conversaciones con un alto grado de intimidad y de confidencialidad, con la presencia continuada de datos sensibles que ponen en riesgo la imagen de los participantes. Para estudiar el conflicto, por tanto, deben establecerse criterios para la recogida y la posterior clasificación de las muestras de modo que se obtengan que presenten discusiones e interacciones no armoniosas.

Además, como veremos más adelante, este corpus no solo está compuesto de grabaciones de conversaciones conflictivas, sino que también trabaja con grabaciones de terapias realizadas en un entorno clínico sociosanitario. Sobre estos materiales opera un grado de confidencialidad elevado, es decir, están sometidos a unas medidas de protección, encriptación y anonimización máximamente restrictivas.

En este sentido, el compilador de corpus se enfrenta a un doble reto metodológico: por un lado, debe asegurar que la recogida de las grabaciones se dé en un entorno de naturalidad de los datos y que, a su vez, permita el acceso a interacciones problemáticas; por otro, en el procesamiento de las grabaciones para su posterior análisis lingüístico, se debe proteger la privacidad de los participantes. Esto implica no solo la adopción de un sistema de anonimización eficaz que impida la reidentificación de los participantes, sino también la recogida de aquellos documentos legales que permitan el tratamiento ético de los datos obtenidos, esto es, el consentimiento informado de los hablantes y los compromisos de confidencialidad, tanto del personal investigador que va a trabajar con los datos, como de las personas encargadas de transcribir y anonimizar las grabaciones.

En última instancia, la recogida de un corpus de estas características requiere de la aprobación del Comité de Ética en la Investigación de la institución en la que se enmarque el proyecto, el cumplimiento de los preceptos contenidos en la legislación sobre derecho a la intimidad y a la protección de datos para la investigación en territorio español (Ley Orgánica 1/1982, de 5 de mayo, de Protección civil y derecho al honor, la intimidad personal y a la propia imagen y Ley Orgánica 3/2018, de 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales), así como las indicaciones de la Agencia Española de Protección de Datos (2016).

La primera ley mencionada, la Ley Orgánica 1/1982, de 5 de mayo, de Protección civil y derecho al honor, la intimidad personal y a la propia imagen, establece como derechos fundamentales el honor, la intimidad personal y familiar, así como la propia imagen. Por lo tanto, cualquier intrusión en el ámbito privado que no cuente con una autorización explícita de la ley o no haya obtenido un consentimiento claro por parte de la persona implicada sería considerada como un acto punible (artículo 2.2). En este

 DOI: https://doi.org/10.6035/clr.7907

sentido, según se contempla en el Código Penal (artículo 197), la recogida de grabaciones sin el consentimiento de las personas involucradas puede constituir un delito grave contra la privacidad. Con respecto a la Ley Orgánica 3/2018 de Protección de datos personales y garantía de los derechos digitales, los preceptos que afectan a la recogida de corpus orales tienen que ver con la manera en la que se van a gestionar y utilizar los datos una vez obtenidos. De esta manera, los principios de la protección de datos deben aplicarse a toda información relativa a una persona física identificada o identificable, aunque los datos se encuentren anonimizados. Se considera que los datos están 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: 48)

Por otra parte, 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 en los que se vean involucrados menores de edad[3] (Ley 3/2018, artículos 6 y 7). En la redacción de esta autorización, la regulación establece el *principio de transparencia* (Ley 3/2018, art. 11). Esto significa que se debe informar a los participantes 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). En este caso, las responsables legales del tratamiento son las directoras −también gestoras− del proyecto ESPRINT.

A modo de resumen, los participantes en este estudio deben saber qué uso se le va a dar a sus datos, de qué manera van a ser tratados y qué medidas se aplicarán para su protección, así como deberán conocer que tienen la opción de retirar su permiso en cualquier momento (derecho al desistimiento) y deben saber ante quienes deberán dirigirse para ejercer este derecho.

Los investigadores deben asegurar el cumplimiento de los principios éticos que fomenten una gestión responsable de los datos obtenidos durante la recopilación del corpus (Rock, 2001, Adolphs y Knight, 2010, McEnery y Hardie, 2011, Schneider, 2018, Childs *et al*., 2011, D'Arcy y Bender, 2023, Carcelén, en prensa).

Con el objetivo de detallar los aspectos éticos y legales a los que se ha debido hacer frente para la compilación de este corpus, presentamos a continuación la caracterización del corpus ESPRINT (sección 2), donde se explicará, en primer lugar, el origen del proyecto (sección 2.1.) y, en segundo lugar, los subcorpus que lo componen, el corpus ESPRINT-Conversación y el corpus ESPRINT-Terapias (sección 2.2.). Seguidamente, en la sección 3, se abordan los retos que se han debido superar para un tratamiento ético y legal de los datos personales, tanto en la recogida de los datos (sección 3.1.) como en su procesamiento para el posterior análisis lingüístico (3.2.), así como los

[3] Como veremos en la sección 3, se han dado casos en los que se han registrado las intervenciones de los hijos e hijas de algunas de las parejas que participan en la recogida en el corpus.

documentos legales que han sido necesarios recoger para garantizar una recogida y tratamiento ético de los datos. En la sección 4 expondremos las consideraciones finales acerca de los obstáculos que hay que salvar a la hora de plantearse la construcción de un corpus oral de conversación conflictiva entre parejas recogido en dos entornos situacionales diferentes.

2. CARACTERÍSTICAS DEL CORPUS ESPRINT

2.1. ORIGEN DEL PROYECTO

El corpus ESPRINT se enmarca dentro del proyecto ESPRINT, *Estrategias pragmático-retóricas en la interacción conversacional conflictiva entre íntimos y conocidos: intensificación, atenuación y gestión interaccional*, PID2020-114805GB-I00 (IP: Marta Albelda y Maria Estellés), trabajo que continúa las investigaciones iniciadas en dos proyectos previos, Es.Var.Atenuación, *La atenuación pragmática en el español hablado: su variación diafásica y diatópica*, MINECO FFI2013-40905-P (IP: Marta Albelda) y 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 (IP: Marta Albelda y Maria Estellés).

El proyecto tiene como objetivo principal profundizar en el estudio de los fenómenos pragmático-retóricos y de gestión interaccional presentes en la conversación espontánea. La novedad de este trabajo radica en la particularidad de que estas muestras se han recogido en entornos de comunicación problemática y conflictiva entre personas con un vínculo relacional íntimo, en concreto, se ha trabajado con parejas sentimentales. En cambio, en los proyectos previos se había trabajado con conversación espontánea armoniosa obtenidas en entornos vivenciales de familiaridad y cercanía a través del corpus de conversación Ameresco (Albelda y Estellés, en línea).

Para el caso de ESPRINT, por tanto, se hacía necesario contar con otras disciplinas –como la psicología clínica, social y de la comunicación, o el análisis de la conversación– que pudieran ayudar a dar solución a la superación de los problemas lingüísticos de comunicación que provocan desencuentros en las relaciones personales, o los problemas relacionales que se manifiestan en hostilidades y malentendidos en la comunicación.

Si bien, existen corpus orales en los que podía localizarse conflicto en la interacción comunicativa (como, por ejemplo, trabajos de corpus sobre las sesiones parlamentarias, de programas televisivos, entre otros), estos se enmarcan en un contexto de realización público. Para nuestros intereses de investigación, en cambio, era necesario trabajar con situaciones de conflicto sucedidas en la esfera privada. No obstante, obtener segmentos conflictivos es de por sí una tarea compleja, ya que implica un nivel muy alto de exposición personal de los participantes, esto es, en este tipo de grabaciones se revelan datos íntimos de su comportamiento y su historia personal. Este hecho puede justificar la escasez de los estudios lingüísticos en español en estas situaciones (no así en el ámbito de la psicología donde sí cuenta con desarrollo), puesto que puede darse cierto pudor del hablante a la hora de compartir información tan sensible o puede haber sesgo por parte del investigador o investigadora a favor de mantener la armonía.

DOI: https://doi.org/10.6035/clr.7907

2.2. Los corpus ESPRINT-Conversación y ESPRINT-Terapia

Como hemos señalado anteriormente, hasta donde llega nuestro conocimiento, no existen trabajos previos que nos permitieran estudiar la conversación conflictiva entre personas con una relación íntima sucedidas en entornos de familiaridad, por lo que ha sido necesario recopilar nuestros propios materiales para elaborar el corpus ESPRINT. Así, se han recogido materiales a partir de dos fuentes: en primer lugar, se ha construido un corpus de conversación espontánea problemática y no armoniosa (ESPRINT-Conversación), en el que las interacciones están lesionadas y/o presentan conflicto; por otro, hemos trabajado con un corpus cedido de sesiones de terapia de pareja (ESPRINT-Terapias) recogidas en un entorno clínico sociosanitario gracias a la colaboración del proyecto E(f)FECTS, *Emotionally Focused Couple Therapy in Spanish* (Martiño Rodríguez, dir.) del Instituto Cultura y Sociedad de la Universidad de Navarra.

Además, el proyecto ESPRINT cuenta con un convenio de colaboración con las Clínicas de Psicología Lluís Alcanyís de la Universitat de València para la futura incorporación de nuevos materiales de estudio.

Veremos, a continuación, las particularidades técnicas de cada uno de estos subcorpus.

2.2.1. El corpus ESPRINT-Conversación

En el corpus ESPRINT-Conversación la recogida de las grabaciones ha sido coordinada por el propio proyecto ESPRINT. En este caso, los informantes que participan son parejas a las que se les exige el requisito de haber reportado problemas de pareja continuados en el tiempo y que han recibido una compensación económica a cambio de su colaboración en la recogida de los datos.

El corpus está compuesto por grabaciones que han realizado los propios miembros de la pareja en entornos físicos familiares, generalmente sus propios domicilios o durante trayectos largos en el coche. En este momento, contamos con la participación de ocho parejas procedentes de seis ciudades distintas (Alicante, Burgos, Coruña, Madrid, Málaga y Valencia). Como se muestra en la Tabla 1, el corpus está compuesto de unas 27 horas de grabación que se distribuyen de la siguiente manera:

Tabla 1. Cómputo total de horas del corpus ESPRINT-Conversación

CORPUS ESPRINT-Conversación	
ALICANTE	6h 05' 07"
BURGOS	8h 06' 00"
CORUÑA	3h 50' 27"
MADRID 1	5h 15' 47"
MADRID 2	0h 13' 00"
MADRID 3	1h 27' 05"
MÁLAGA	0h 33' 35"
VALENCIA	1h 15' 38"
TOTAL	**26h 46'39"**

2.2.2. El corpus ESPRINT-Terapias

En este caso, el corpus ha sido cedido por el proyecto E(f)FECTS mencionado anteriormente, proyecto enmarcado en el campo de la psicología que constituye el primer ensayo clínico aleatorizado en terapia de pareja focalizado en las emociones en países de habla hispana. Si bien desde este proyecto se han recogido materiales procedentes de cinco países distintos (España, México, Guatemala, Costa Rica y Argentina), para el caso del corpus ESPRINT se han utilizado únicamente las grabaciones del español peninsular, por ser este su objeto de estudio.

Se han seleccionado cuatro parejas (dos de Madrid y dos de Málaga) que han participado en este ensayo clínico. Del total de grabaciones obtenidas en un entorno clínico sociosanitario por el proyecto E(f)FECTS para estas cuatro parejas (20 sesiones por pareja con una duración aproximada de entre 70-80 minutos), se han seleccionado varias sesiones sucedidas en momentos iniciales, intermedios y finales del ensayo, con un total de 29 horas de grabación como puede verse en la Tabla 2:

Tabla 2. Cómputo total de horas del corpus ESPRINT-Terapias

CORPUS ESPRINT-Terapias							
MADRID 1		**MADRID 2**		**MÁLAGA 1**		**MÁLAGA 2**	
Madrid 1.1	59' 20"	Madrid 2.1	53' 44"	Málaga 1.1	1h 15' 57"	Málaga 2.1	1h 12' 27"
Madrid 1.3	1h 25' 09"	Madrid 2.3	1h 20' 21"	Málaga 1.2	1h 14' 42"	Málaga 2.2	1h 16' 57"
Madrid 1.9	1h 19' 47"	Madrid 2.7	1h 02' 09"	Málaga 1.3	1h 12' 53"	Málaga 2.3	1h 17' 06"
Madrid 1.15	1h 21' 24"	Madrid 2.9	1h 19' 04"	Málaga 1.9	1h 16' 05"	Málaga 2.9	1h 10' 26"
Madrid 1.20	1h 24' 41"	Madrid 2.15	1h 18' 22"	Málaga 1.15	1h 18' 55"	Málaga 2.15	1h 17' 36"
		Madrid 2.20	1h 24'10"	Málaga 1.20	1h 14' 24"	Málaga 2.20	1h 15' 19"
TOTAL	6h 30' 21"		7h 17' 50"		7h 32' 56"		7 h 29' 51"
28 h 50' 58"							

3. RETOS EN EL TRATAMIENTO DE LOS DATOS PERSONALES DEL CORPUS ESPRINT

Como se mencionaba en la introducción, la compilación de un corpus de conversación conflictiva debe enfrentarse a ciertos retos relacionados con el proceso de recogida de las grabaciones, así como con el protocolo de tratamiento de los datos personales. Tratar estos materiales de manera cuidadosa y ética es crucial para la creación de un corpus oral valioso y confiable que pueda utilizarse para investigaciones lingüísticas, sociolingüísticas, antropológicas y otros campos relacionados, pero se hace especialmente necesario en la compilación de corpus con presencia de datos sensibles, como es el caso del corpus ESPRINT, en el que la imagen de los hablantes está en especial riesgo, dadas las situaciones de discusión, de conflicto y de desacuerdo que aparecen en la interacción comunicativa.

A continuación, se explicitan los protocolos de recogida de materiales y de tratamiento de los datos adoptados en el corpus ESPRINT, tanto en el corpus ESPRINT-Conversación, como en el corpus ESPRINT-Terapias, de acuerdo con los requisitos éticos y legales aplicables en investigación vistos en la sección 1. Estos protocolos se aplican en dos ámbitos de actuación: por un lado, en la fase de recogida de los materiales y, por otro, en la fase del procesamiento de los datos para su posterior análisis lingüístico.

DOI: https://doi.org/10.6035/clr.7907

3.1. Recogida de los datos: grabaciones y consentimientos informados

Cuando se plantea la recogida de los datos, cabe señalar que esta va más allá de la realización de las grabaciones; esto es, se requiere, además, la obtención previa a la grabación de los consentimientos informados de todas las personas que participan en el estudio.

En el caso del corpus ESPRINT-Conversación, los dos miembros de la pareja participante firman un consentimiento informado individual en el que, además, se explica la política de protección de datos, así como el protocolo ético de anonimización y preservación de los datos adoptados en la investigación. Asimismo, cuando durante las horas de grabación se recojan las voces de otras personas que no son los sujetos de experimentación, al entrar en contacto con ellas espontáneamente, bien porque aparezcan en escena (es el caso, por ejemplo, de otros miembros de la familia, generalmente los hijos e hijas), bien porque se atiendan llamadas telefónicas o mensajes de audio a través de mensajería instantánea, como Whatsapp o Telegram, sería conveniente que estas personas también firmaran el consentimiento *a posteriori*. Si no lo hacen, se contemplan actuaciones como la anonimización y el borrado de los datos, como explicaremos en la siguiente sección. En caso de que aparezcan menores de edad, además del consentimiento individual, la pareja deberá firmar el apartado correspondiente a este respecto en el consentimiento como responsables legales de sus hijos e hijas. Para el corpus ESPRINT-Terapias, además de los miembros de la pareja, se ha recogido el consentimiento del profesional en psicología y psicoterapia que guía la sesión.

En ambos corpus, según las instrucciones dadas por los diferentes comités de ética y las delegaciones de protección de datos de ambas instituciones (Universitat de València y Universidad de Navarra) y cumpliendo con el principio de transparencia exigido por la legislación señalado en la sección 1, el consentimiento informado incorpora cláusulas en torno a los siguientes bloques de información: los objetivos de la investigación, las condiciones de la grabación y la compensación que se dará a las personas colaboradoras en el ensayo.

En primer lugar, se da a conocer el objetivo del proyecto, así como el personal investigador responsable del mismo y la manera de contactar con este en el caso de querer ejercer su derecho al desistimiento, en cualquier fase del proyecto incluso cuando ya se hayan recogido y procesado las grabaciones. En este sentido, se explica que la finalidad del corpus es detectar los problemas de comunicación entre hablantes íntimos, con especial atención a los problemas de naturaleza pragmática, así como identificar los esquemas y patrones de interacción que generan el desarrollo de un conflicto comunicativo. Asimismo, se especifican los requisitos que deben cumplir quienes estén interesados en participar, esto es, haber reportado problemas de comunicación de pareja continuados en el tiempo y ser de nacionalidad española de forma nativa.

En segundo lugar, se les informa sobre el modo de recoger las grabaciones. En el caso del corpus ESPRINT-Terapias, estas se realizaron en formato audiovisual colocando un dispositivo de grabación en la consulta del terapeuta. Para el corpus ESPRINT-Conflicto, se les facilitó a las parejas participantes una minigrabadora de voz digital espía con encriptación que cada miembro de la pareja debía colocarse en la solapa para registrar las conversaciones. En este caso, la pareja recibió instrucciones precisas sobre la metodología de obtención de sus intervenciones, esto es, debían llevar las grabadoras encendidas durante varias horas consecutivas a lo largo de una o dos semanas y actuar con naturalidad. Dentro de lo posible, se les pidió que cuando se fuera a realizar

la recogida, hubiera un ambiente tranquilo y con poco ruido externo en el espacio físico donde se lleve a cabo la grabación.

Como hemos advertido, estas grabadoras poseen un sistema de encriptación con el que se salvan dos obstáculos. El primero tiene que ver con garantizar la naturalidad de las intervenciones; en este sentido, los participantes no pueden acceder al contenido recogido ni borrarlo o alterarlo. El segundo obstáculo está vinculado a la protección de los datos, ya que, una vez recogido el material, los participantes deben enviar las grabadoras con las conversaciones a las responsables del corpus. Téngase en cuenta que los datos recogidos contienen información sensible y privada y si ocurriera que los dispositivos se extraviaran, terceras personas no involucradas en el proyecto podrían acceder a esta información. Al emplear dispositivos con esta posibilidad de encriptación, el contenido queda totalmente protegido ante estos posibles problemas.

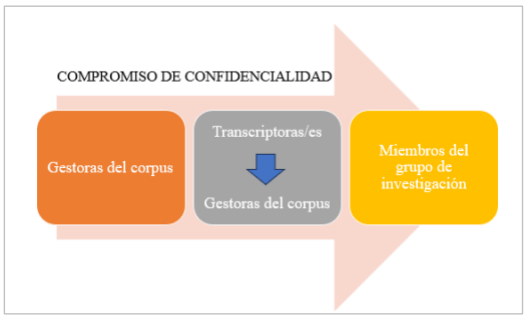
Para la recogida de las grabaciones se estableció una primera fase de reclutamiento que fue posible gracias a la colaboración de psicólogas/os, terapeutas y abogadas/os de familia, quienes conocían a personas con problemas de pareja. Posteriormente, se procedió a una segunda fase de recogida experimental de las grabaciones, momento en el que se informa de las pautas para recoger las interacciones que hemos referido arriba. De las treinta parejas preseleccionadas, finalmente solo ocho han completado el ciclo completo de grabaciones establecidas. El resto o bien decidió retirarse del estudio, o bien no presentaban tanto conflicto en interacción como pensaban.

3.2. Procesamiento de los datos: transcripción, anonimización y compromiso de confidencialidad

Con respecto a la fase de procesamiento de los datos, se describe a continuación cómo ha sido el protocolo de transcripción y anonimización utilizado en el corpus ESPRINT, así como el compromiso de confidencialidad que se debe firmar para trabajar con los datos resultantes.

En la fase de recogida de los materiales las personas implicadas eran cada uno de los miembros de la pareja, así como el/la terapeuta para el corpus de terapia. En cambio, en esta segunda fase las figuras que entran en juego son el personal de investigación –incluidas las gestoras responsables del corpus– y el personal encargado de transcribir y anonimizar los materiales, como puede verse en la Figura 1. Todos ellos deben firmar un compromiso de confidencialidad, que describiremos en detalle más adelante.

Figura 1. Flujo de trabajo en la fase de procesamiento de los datos del corpus ESPRINT



DOI: https://doi.org/10.6035/clr.7907

Así, una vez que las gestoras del corpus han recibido todos los materiales (gracias a la cesión del corpus de terapias y a las grabaciones realizadas por las parejas en sus hogares), se inician las labores de procesamiento de las conversaciones.

Los participantes han sido informados de que todos los materiales de habla grabados serán tratados con absoluta confidencialidad y de que serán sometidos a un proceso de anonimización, borrando tanto cualquier identificador directo (como nombres propios, de lugares y de instituciones), como identificadores indirectos, es decir, cualquier otra información que pudiera servir para identificar a los hablantes.

Asimismo, se desvinculan las grabaciones de los datos personales e identificativos de los participantes, almacenándose estos de manera encriptada, y con acceso restringido únicamente de las gestoras del corpus. Además, se contemplan medidas más restrictivas como, por ejemplo, en el caso de que los participantes así lo manifiesten, se podrían distorsionar las voces con programas de edición de voz (por medio de cambios de tono utilizando un *software* de edición de sonido como, por ejemplo, Audacity®).

Las grabaciones son entregadas a las personas encargadas de realizar la transcripción, quienes deberán firmar un compromiso de confidencialidad antes de comenzar el procesamiento del corpus. De este modo, se garantiza que (1) se va a transcribir con suma confidencialidad los datos, (2) el transcriptor o transcriptora declara ser plenamente consciente de la sensibilidad de estos datos (sonoros y audiovisuales) y de la responsabilidad de tratarlos con plena confidencialidad; (3) no revelar ninguna información contenida en este corpus de conversaciones a ninguna persona y a no realizar un uso improcedente de los archivos y contenidos; (4) no compartir ni mostrar estos datos a ninguna otra persona; (5) borrar de su ordenador y de sus archivos electrónicos las grabaciones y transcripciones cuando termine y entregue su trabajo y, especialmente para el corpus ESPRINT-Terapia, (6) realizar este trabajo en un espacio físico en el que no haya más personas pues, aunque pudiera realizar la tarea adjudicada con auriculares, los documentos son visuales (terapias), y el compromiso contiene la doble confidencialidad, visual y sonora. No obstante, para el corpus ESPRINT-Conversación se recomienda también el cumplimiento de esta cláusula, aunque existe cierta flexibilidad siempre y cuando se trabaje con auriculares.

Con respecto al sistema de anonimización para ambos corpus, tanto ESPRINT-Terapias como ESPRINT-Conversación, han sido sometidos a un proceso de anonimización tanto del material sonoro, como la propia transcripción. Se ha realizado, por tanto, una anonimización[4] en dos capas —textual y oral— atendiendo tanto a los identificadores directos como a los indirectos señalados anteriormente, siguiendo la metodología adoptada por trabajos previos como el corpus Ameresco (Briz *et al*., 2019, Carcelén y Uclés, 2019, Carcelén, 2024, Carcelén, en prensa).

En la anonimización textual, los nombres propios son sustituidos en la transcripción por otros ficticios, respetando si es posible, las características socioculturales y diatópicas. En el caso de los identificadores directos, se introduce una

[4] Hablaremos de datos anonimizados y no seudonimizados, ya que se han sustituido por otros ficticios sin que haya posibilidad de reconstruir la identidad de los participantes (Agencia Estatal de Protección de Datos, 2016: 2). A diferencia de estudios clínicos sanitarios donde sí podría darse esta posibilidad si, por ejemplo, el estudio permitiera mejorar las condiciones de salud del participante, en cuyo caso se trabajaría con datos seudonimizados. En nuestro caso, el objetivo de estudio es puramente de análisis lingüístico, por lo que esta opción no se contempla.

marca de transcripción que señala que ese segmento ha sido anonimizado, como puede verse en el ejemplo 1:

Ej. 1.

H: no <anonimizado>Maribel</anonimizado> estás muy negada por comprar el babi desde el primer momento y no pasa nada por comprar un babi vas a tener más

(1_ALI_H)

Asimismo, como puede verse en el ejemplo 2, para la identificación interna en el programa de transcripción alineada ELAN, se utilizan los códigos T o P (terapeuta o psicoterapeuta), M (mujer), H (hombre) y N (niño/a) para identificar las líneas de intervención de cada hablante:

Ej. 2.

H: ya ves ((RISAS))
P: entonces
M: ((es)) que se me olvida está diciendo cosas que yo quiero [(( ))]
P: [que son importantes]
¿verdad?

(Málaga 1.9)

Para la anonimización del audio se ha utilizado un sistema semiautomático en el que tras una primera selección del fragmento o fragmentos que se quieran anonimizar en ELAN, el archivo se traslada a un *software* de edición de audio que borra el contenido seleccionado. En este caso, hemos elegido Audacity® por ser un programa de acceso libre y con una interfaz sencilla de manejar.

ELAN (Max Planck Institute for Psycholinguistics, 2023) es un *software* para la transcripción y anotación de archivos de audio y vídeo que permite la creación de diferentes líneas de trabajo en las que no solo pueden crearse diferentes líneas de anotación para la transcripción de las intervenciones de cada uno de los participantes, sino que también permite incluir otras líneas para otros fines de anotación. En este caso, se utiliza una de ellas para la selección y anotación del fragmento que es necesario anonimiza; una vez localizados todos los cortes en el audio, el nuevo archivo resultante se exporta utilizando la opción de *exportar texto tabulado*, momento en el que además deben excluirse los nombres de las líneas del *output* y excluir los nombres de los participantes. Posteriormente, este archivo se importa al *software* de edición de audio, donde se reconocen las anotaciones que hemos exportado de ELAN del audio original y se eliminan. Se ha elegido la opción *silenciar* para prescindir de estos fragmentos por ser más amable que la aparición de un indicador sonoro. No queda más que generar un nuevo archivo en formato .wav en el que se han eliminado las secuencias necesarias y vincular este archivo de nuevo con ELAN.

Retomando lo señalado en la sección anterior, en esta fase deben considerarse los casos referidos a personas que aparezcan de manera repentina en escena (como las hijas e hijos de la pareja), u otras voces procedentes de llamadas telefónicas o mensajes de audio a través de mensajería instantánea. Dado que la grabadora está colocada cerca

DOI: https://doi.org/10.6035/clr.7907

de la boca de los participantes, es posible reconocer e identificar las intervenciones telefónicas de estas terceras personas. Ante esta situación, se debe valorar la sensibilidad del contenido de estas intervenciones para decidir si es necesaria su eliminación completa o si, por el contrario, podría dejarse la transcripción. Desde la dirección del corpus se ha tomado la decisión operativa de eliminar el fragmento de audio en todos los casos, si bien, el contenido transcrito y anonimizado en la capa textual podría mantenerse, siempre y cuando su contenido no sea clasificado como sensible o delicado. Para ejemplarizar el caso con un ejemplo ficticio, si en el audio enviado por mensajería instantánea la persona externa pregunta a uno de los miembros de la pareja sobre la posibilidad de ir a comer juntos el próximo fin de semana, se entiende que no hay datos sensibles. En cambio, si en el audio esta persona relata que está teniendo un problema personal en su trabajo, esta información sí que se considera comprometida y, en consecuencia, no solo se borraría el fragmento de audio, sino que también se omitiría en la transcripción. En su lugar, aparecerá una observación en la transcripción en la que se informa de la supresión del fragmento.

La última parte del proceso consiste en el traslado del material transcrito y anonimizado a los miembros del equipo de investigación. Los investigadores e investigadoras de este proyecto, incluidas las gestoras del corpus, deben firmar un compromiso de confidencialidad en el que se comprometen a no ceder estos materiales aunque estén anonimizados y protegidos, a ningún otro investigador o investigadora, o persona en general, ajena al grupo de investigación autorizado, a tratar confidencialmente las conversaciones (en audio y transcritas), a no revelar ninguna información contenida en el corpus, a no hacer uso improcedente de la información contenida y a no publicar ninguna conversación completa en ningún trabajo, ni reproducir ninguna conversación públicamente, limitándose el uso de ejemplos a un máximo de ocho intervenciones consecutivas. En el caso de necesitar emplear un número mayor, se requiere la autorización de las directoras del proyecto.

Además, dada la sensibilidad de los datos contenidos en las grabaciones, tanto del corpus ESPRINT-Terapia, como del ESPRINT-Conversación, el acceso a ambos corpus está restringido exclusivamente al equipo de investigación registrado como personal autorizado, aprobado por el Comité de Ética en la Investigación, y no se contempla ni se admite la posibilidad de hacerlos accesibles a personal ajeno a este, ni su inclusión en motores de búsqueda o plataformas en línea que proporcionen alojamiento público al corpus.

Desde la dirección del corpus se ha decidido aplicar las medidas más restrictivas tanto al corpus de terapias como al de conversación, a pesar de que existen trabajos previos de conversación espontánea recogidos con una metodología similar a la utilizada para ESPRINT-Conversación que son puestos a disposición pública –piénsese, por ejemplo, en el corpus Val.Es.Co. (Pons, en línea) o el corpus Ameresco (Albelda y Estellés, en línea), ambos de conversación coloquial espontánea grabados secretamente, donde también pueden aparecer datos sensibles[5]–. No obstante, con afán de actuar acorde a la legislación y de respetar la protección de datos personales, se han igualado los

[5] Si bien es cierto que los datos sensibles que pueden aparecer son de otra naturaleza –no conflictiva– y que, además, los participantes de estos corpus cuentan con la posibilidad de escuchar las grabaciones después de su recogida y decidir si quieren retirar su participación totalmente o se debe borrar algún fragmento que consideren demasiado personal. Esta última opción no es posible en el corpus ESPRINT, como hemos señalado en la sección 3.1.

protocolos de tratamiento y protección de los datos personales, así como el derecho a la intimidad de los hablantes que participan en el estudio.

4. CONSIDERACIONES FINALES

A modo de conclusión, a lo largo de este trabajo hemos repasado las características técnicas de la construcción de un corpus de conversación conflictiva, el corpus ESPRINT, centrando la atención en las cuestiones relacionadas con el tratamiento y la protección de datos personales. Tras desgranar cuáles son las obligaciones establecidas en la legislación en cuanto a protección y tratamiento de datos personales en el ámbito de la investigación, hemos concretado los principios éticos que operan en un corpus como el corpus ESPRINT, tanto para la obtención de las conversaciones en contextos de familiaridad, como en para los datos extraídos de las sesiones de terapia en entornos clínicos sociosanitarios.

En este sentido, cabe destacar la novedad que presenta este corpus, ya que constituye un trabajo sobre un tipo de discurso, las interacciones problemáticas y lesionadas entre parejas sentimentales, hasta ahora escasamente representado en los estudios lingüísticos y de análisis del discurso. Esta infrarrepresentación se debe a las propias dificultades intrínsecas para acceder a este tipo materiales y los obstáculos que hay que salvar con respecto a la recogida y el tratamiento responsable del material obtenido, esto es que garantice el derecho a la intimidad y la privacidad de los sujetos que participan en el estudio, así como la confidencialidad en el acceso a las grabaciones y transcripciones resultantes.

En el caso del corpus ESPRINT, el acceso a estas interacciones conflictivas ha sido posible, especialmente, gracias al trabajo interdisciplinar y la colaboración con profesionales de la psicología, a través de los cuáles hemos accedido al registro de sesiones de terapia de pareja (corpus ESPRINT-Terapia), así como contactado con parejas interesadas en participar de la recopilación de las conversaciones (corpus ESPRINT-Conversación), que han accedido a grabar y compartir aspectos tan íntimos de su relación como son los momentos de conflicto y comunicación problemática que se registraron durante las grabaciones. Para ambos corpus, como resultado del análisis en profundidad de las legislaciones aplicables en materia de derecho a la intimidad y tratamiento y protección de los datos personales, desde el proyecto ESPRINT se han adoptado unos modelos de consentimiento informados que, con sumo grado de detalle, atienden a las implicaciones éticas y legales más abarcadoras.

Por otro lado, en el protocolo de procesamiento de los datos, también se ha adoptado una metodología que garantiza la confidencialidad de los datos. Esto es, tanto las personas encargadas de transcribir las grabaciones, como los miembros del grupo de investigación en el que se inscribe el corpus, han firmado compromisos de confidencialidad en el que se comprometen a no revelar ninguna información de la contenida en ambos corpus. Además, en el caso de los transcriptores y transcriptoras, se comprometen a no trabajar en lugares públicos con estos materiales y a eliminar los archivos una vez transcritos, en el caso de los transcriptores y transcriptoras. Se han contemplado, en último lugar, cláusulas con respecto a la difusión de los resultados y ejemplos obtenidos del corpus; en particular, el personal investigador no puede utilizar más de ocho intervenciones consecutivas en sus trabajos, tanto científicos como académicos con el objetivo de imposibilitar la reconstrucción del contexto discursivo y una posible reidentificación de los hablantes.

 DOI: https://doi.org/10.6035/clr.7907

Dado el alto grado de contenido sensible que es susceptible de aparecer en la conversación conflictiva, desde el corpus ESPRINT hemos centrado nuestros esfuerzos en el cumplimiento, en el grado más estricto posible, de las consideraciones éticas y legales aplicables a su construcción y gestión.

REFERENCIAS BIBLIOGRÁFICAS

Adolphs, Svenja y Knight, Dawn (2010). Building a spoken corpus. What are the basics? En Anne O'Keeffe y Michael 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, Marta y Estellés, Maria (coords.) (En línea). Corpus Ameresco, Universitat de València, ISSN: 2659-8337, en línea, www.corpusameresco.com

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

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

Briz Gómez, Antonio *et al*. (2019). *Protocolo de trabajo para los equipos Ameresco*. En línea https://esvaratenuacion.es/protocolo-de-trabajo (versión actualizada enero de 2020).

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

Carcelén Guerrero, Andrea (en prensa). ¿Es posible elaborar corpus orales espontáneos y cumplir legislación? El modelo en tres fases del corpus Ameresco. *Revista Española de Lingüística Aplicada*.

Carcelén Guerrero, Andrea y Uclés Ramada, Gloria (2019). Diseño y construcción de un corpus oral multidialectal. El corpus Ameresco. *Normas: Revista de Estudios Lingüísticos Hispánicos*, 9 (1), pp. 17-36.

Childs, Becky, Van Herk, Gerard y Thorburn, Jennifer (2011). Safe harbour: Ethics and accesibility in sociolinguistic corpus building. *Corpus Linguistics and Linguistic Theory* (7-1), pp. 163-180.

D'Arcy, Alexandra y Bender, Emily (2023). Ethics in Linguistics, *Annual Review of Linguistics*, 9 (1), pp. 49-69.

ELAN (Version 6.7) [*Software* informático]. (2023). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Obtenido de https://archive.mpi.nl/tla/elan

Enghels, Renata, Vanderschueren, Clara y Bouzouita, Miriam (2015). Panorama de los corpus y textos del español peninsular contemporáneo. En Maria Iliescu y Eugeen Roegiest (Ed.), *Manuel des anthologies, corpus et textes romans*, pp. 147-170, De Gruyter.

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 10/1995, de 23 de noviembre, del Código Penal.

Ley Orgánica 3/2018, de 5 de diciembre, de Protección de datos personales y garantía de los derechos digitales.

Llisterri, Joaquín (2021). Corpus para investigar sobre el componente fónico en español como LE/L2. En Mar Cruz y Javier Muñoz (Eds.), *e-Research y español LE/L2.: investigar en la era digital*, pp. 164-196, Routledge.

McEnery, Tony, y Hardie, Andrew (2011). *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press.

Moreno Fernández, Francisco (2005), Corpora of Spoken Spanish Language. The Representativeness Issue. En Yuji Kawaguchi *et al*. (Eds.), *Linguistic Informatics, State of the Art and the Future*, pp. 120-144, John Benjamins.

Parodi, Giovanni y Burdiles, Gina (2019). Corpus y bases de datos (Corpora and databases). En Javier Muñoz *et al*. (coord.), *The Routledge Handbook of Spanish Language Teaching: metodologías, contextos y recursos para la enseñanza del español*, pp. 596-612, Routledge.

Pons Bordería, Salvador (dir.) (En línea). Corpus Val.Es.Co 3.0. http://www.valesco.es

Rock, Frances (2001). Policy and Practice in the Anonymisation of Linguistic Data, *International Journal of Corpus Linguistics* 6(1), pp.1-26.

Rojo, Guillermo (2016), Citius, maius, melius. Del CREA al CORPES XXI. En Johannes Kabatek, Carlota de Benito Moreno (coords.), *Lingüística de corpus y lingüística histórica iberorrománica*, pp. 197-212, De Gruyter.

Schneider, Klaus P. (2018). Methods and ethics of data collection. En Andreas H. Jucker, Klaus P. Schneider y Wolfram Bublitz (ed.), *Methods in Pragmatics*, pp. 37-93, De Gruyter.

Solís García, Inmaculada (2018). Corpus españoles dialógicos para el análisis de la conversación. *CHIMERA: Romance Corpora and Linguistic Studies*, 5 (1), pp. 117-129.

Vázquez, Victoria y Recalde, Monserrat (2009). Problemas metodológicos en la formación de corpus orales. En Pascual Cantos y Aquilino Sánchez (Eds.), *A survey of corpus-based research*, pp. 51-64. Recurso electrónico https://www.um.es/lacell/aelinco/contenido/index.html