Valeš, Miroslav

Recopilación de datos primarios para la descripción y documentación de la lengua

Études romanes de Brno. 2020, vol. 41, iss. 1, pp. 87-98

ISSN 1803-7399 (print); ISSN 2336-4416 (online)

Stable URL (DOI): https://doi.org/10.5817/ERB2020-1-6
Stable URL (handle): https://hdl.handle.net/11222.digilib/142574

ÉTUDES ROMANES DE BRNO 41/2020/1
HTTPS://DOI.ORG/10.5817/ERB2020-1-6

Recopilación de datos primarios para la descripción y documentación de la lengua

Primary Data Collection for Language Description and Documentation

Miroslav Valeš [miroslav.vales@tul.cz]
Technická univerzita v Liberci, República Checa

Resumen
En septiembre de 2018, CIDLeS (Centro Interdisciplinar de Documentação Linguística e Social, Minde, Portugal), en cooperación con la Universidad Técnica de Liberec, República Checa, inició el proyecto: *MSCA TUL: Documentación y Descripción de A Fala impulsada por la comunidad*. La metodología utilizada en el proyecto se fundamenta en la recopilación de datos primarios y su uso, tanto para los fines de descripción como para la documentación. El objetivo del trabajo es presentar el diseño del corpus de datos primarios, que es la base de todo el proyecto. Los datos primarios tienen una variedad de formas: grabaciones de audio y video, textos escritos publicados o no publicados, recursos lingüísticos existentes y también los datos creados o recopilados por la comunidad de hablantes. En el trabajo se discuten varios aspectos a tener en cuenta al recopilar y procesar los datos. Uno de estos aspectos es el equilibrio entre las tres variedades principales de A Fala, lagarteiru, mañegu y valverdeñu. Otro aspecto a tener en cuenta es la selección de temas para las entrevistas y también la selección de participantes para lograr una muestra equilibrada respecto a edad y género. En el caso de los textos escritos, los derechos del autor deben respetarse y resolverse en los casos en que no sea posible obtener el consentimiento de los autores o editores. Por último, pero no menos importante, el tamaño del corpus también fue uno de los temas a considerar junto con la posibilidad de ampliar la base de datos fácilmente en el futuro. El documento expone la experiencia adquirida en el curso de la recopilación de datos y también la brecha entre las soluciones ideales y las soluciones viables.

Palabras clave
A Fala; datos primarios; recopilación de datos; corpus

Abstract
In September 2018 CIDLeS (Centro Interdisciplinar de Documentação Linguística e Social, Minde, Portugal) in cooperation with Technical University of Liberec, Czech Republic, launched a project: *MSCA TUL: Community-Driven Documentation and Description of A Fala*. The methodology used in the project is based on primary data collection and its usage for both description and documentation purposes. The objective of the paper is to introduce the design of primary data corpus, which is the basis of the whole project. The primary data have variety of forms: audio and video recordings, written texts published or unpublished, existing linguistic resources, and also the data created or collected by the community of speakers. The paper discusses

various aspects to be considered while collecting and processing the data. One of these aspects is the balance between the three main varieties of A Fala, Lagarteiru, Mañegu and Valverdeñu. Another aspect to take into account is the selection of topics for the interviews and also the selection of participants, to achieve age and gender balanced sample. In case of written texts copy rights have to be respected and resolved in cases when it is not possible to get the consent from authors or editors. Last but not least, the size of the corpus was also one of the issues to be considered together with the possibility to enlarge the database easily in the future. The paper exposes the experience gained in the course of data collection and also the gap between the ideal solutions and the viable solutions.

Keywords
A Fala; primary data; data collection; corpus

RECIBIDO 2019–11–11; **ACEPTADO** 2020–03–18

El texto fue escrito con el apoyo del proyecto: *Mobilita MSCA Technická univerzita v Liberci* (CZ.02.2.69/0.0/0.0/17_050/0008461)

1. Situación sociolingüística de A Fala

A Fala es una lengua minoritaria hablada en tres pueblos en la Extremadura española. Estos tres pueblos de la Sierra de Gata, en la frontera entre España y Portugal tienen unos 4 300 habitantes y la mayoría de ellos son capaces de hablar en su lengua vernácula. Existen tres variedades principales de A Fala correspondientes a las tres localidades. La variedad de Valverde del Fresno se llama valverdeñu, la variedad de Eljas se conoce como lagarteiru y la de San Martín de Trevejo se llama mañegu. Las diferencias entre las tres variedades son bastante considerables, sin embargo, al mismo tiempo son fácilmente inteligibles, ya que siempre ha existido una intensa relación entre los habitantes de los tres pueblos.

La lengua no ha recibido el reconocimiento oficial hasta ahora, ya que el *Estatuto de Autonomía de Extremadura* no dice nada al respecto, sin embargo, en el año 2001 A Fala fue proclamada "Bien de Interés Cultural" (Gobierno de Extremadura 2001: 2 860):

> "A Fala" forma parte del Patrimonio Histórico y Cultural de Extremadura, siendo necesario que las distintas Instituciones y Administraciones Públicas coordinen sus actividades para garantizar su defensa y protección de modo que "Lagarteiru", "Mañegu" y "Valverdeiru" sigan siendo una realidad mientras sus hablantes así lo quieran.

A pesar de esta proclamación, que prepara el marco institucional de apoyo, A Fala no se enseña en los colegios de las tres localidades. De esta forma la lengua solo queda relegada al ámbito familiar llegando a una situación diglósica. Según la definición de Ferguson (1959) A Fala cumple con la variedad L (low), siendo usada entre los miembros de la comunidad, mientras que el español es la variedad H (high) que sirve para la comunicación fuera de la comunidad. A pesar de esta

situación desventajosa para A Fala, la lengua se sigue transmitiendo de generación en generación y es probablemente una de las lenguas minoritarias con más vitalidad en Europa. Aun así, existen estudios, por ejemplo, Ramallo (2011), que indican el peligro que corre la lengua y documentan el decrecimiento de su uso en ciertas situaciones. Otros estudios (Dondelewski 2020) indican el gradual desplazamiento de las palabras vernáculas por las castellanizadas o castellanas. Es natural, que una lengua minoritaria en contacto intensivo con la lengua nacional necesite apoyos de todo tipo, institucional, educacional y también lingüístico.

2. El proyecto MSCA TUL y sus objetivos

Los objetivos y la metodología de documentación lingüística y de la descripción lingüística son diferentes, sin embargo, muchos lingüistas, por ejemplo (Austin 2016; Good 2018), ven estos dos procesos como estrechamente relacionados. De hecho, la dificultad de separar la descripción de la documentación aparece ya en los escritos clásicos: "Despite the differences, listed in Table 1, the two activities are also closely interrelated and partially overlap for varios epistemological, methodological, and practical reasons" Himmelmann (1998: 162). Es evidente que la documentación suministra datos empíricos que sirven para fines académicos de descripción de la lengua. Por esta razón, las lenguas bien documentadas, con datos que incluyen un amplio rango de situaciones comunicativas y contextos sociales, consiguen una descripción más fiel y pormenorizada. Esta descripción luego contribuye a la emancipación de estas lenguas y su mayor prestigio en comparación con otras lenguas usadas en el mismo territorio. En caso de las lenguas minoritarias y amenazadas, los datos de la documentación pueden contribuir a la revitalización de la lengua. En el caso de A Fala, que siempre ha sido una herramienta de comunicación oral, la documentación detallada y consiguiente descripción puede potenciar el ámbito de la comunicación escrita incluyendo la educación escolar en la lengua materna.

El proyecto *MSCA TUL: Documentación y Descripción de A Fala impulsada por la comunidad* se está llevando a cabo con la colaboración de la Universidad Técnica de Liberec, República Checa (TUL) y el Centro Interdisciplinar de Documentação Linguística e Social, Minde, Portugal (CIDLeS) a partir del septiembre de 2018. El objetivo del proyecto es documentar y describir A Fala con el uso de datos primarios que reflejan el uso del lenguaje natural en la comunidad de habla. Estos datos se recopilan en forma de grabaciones de audio y video, así como textos publicados y no publicados, que comprenden las tres variedades del idioma. A base del corpus de los datos primarios se creará una base de datos multimedia que permitirá la compilación de un diccionario, una gramática básica y posteriormente servirá para varios estudios relacionados con la lengua. Las grabaciones también servirán para los fines de documentación y por lo tanto se guardarán en un repositorio especializado.

El uso de los datos primarios, provenientes de la documentación lingüística, es la característica fundamental del proyecto, siendo la otra característica fundamental la cooperación con los miembros de la comunidad de habla. En el marco del proyecto se intenta crear un corpus formado en mitad por las grabaciones literalmente transcritas, en su mayor parte grabaciones video, y la otra mitad proveniente de los textos escritos en alguna de las tres variedades de A Fala, es decir, que el corpus resultante tendrá un 50% de origen oral y en el restante 50% de origen escrito. Estas

dos fuentes esenciales de datos primarios se complementarán con otros datos relacionados con la lengua, ya que algunas palabras son muy poco frecuentes por su naturaleza y así no se puede suponer que aparezcan ni en las grabaciones ni en los textos. Por lo tanto, se incluirá también el vocabulario aportado por los miembros de la comunidad de habla, especialmente el léxico relacionado con la cultura material, nombres de herramientas, plantas etcétera. Además, se considerarán los vocabularios existentes de A Fala, por ejemplo, el de Rey Yelmo (1999) o Román Domínguez (2008), y otros vocabularios no publicados, por ejemplo, el cuaderno de Alfonso Berrio que recopila palabras valverdeñas o vocabularios de las palabras mañegas que aparecen en internet (Diccionario virtual de Extremadura 2019). Sin embargo, como ya se he dicho, estas fuentes solo tendrán el carácter complementario. El corpus resultante de la recopilación contendrá más que 200 000 palabras y estas se almacenarán en una base de datos que, además, puede complementarse con el material gráfico y audiovisual relacionado con las palabras individuales.

Aunque la recogida de datos todavía no ha sido finalizada, los siguientes apartados de este trabajo exponen la experiencia adquirida durante la recopilación de la mayor parte de los datos en la comunidad de habla y dificultades esperadas e inesperadas con las cuales nos enfrentamos. Las dificultades relacionadas con la recogida de datos retan al investigador a balancear entre el diseño ideal, el originalmente propuesto, y el diseño viable, el que más se corresponde a la realidad lingüística de la comunidad estudiada.

3. Grabaciones

Las grabaciones son definitivamente una fuente de datos muy rica y es indudable que su recopilación es uno de los méritos más importantes del proyecto. El lenguaje oral es la forma principal y primaria de cualquier idioma, es también la forma más natural y por eso merece su representación significativa en el corpus. Sin embargo, también debemos tener en cuenta las limitaciones relacionadas con este tipo de datos. Si decidiéramos usar solo los datos grabados, sin los textos escritos, pronto nos encontraríamos con el reducido rango de vocabulario que usamos en la expresión oral. En un discurso hablado informal las personas usan un vocabulario bastante limitado, especialmente si se trata de conversaciones cotidianas sin preparación anterior. En cambio, los textos escritos contienen vocabulario más amplio, más elaborado y por lo tanto el equilibrio entre las dos fuentes principales de datos es esencial.

Como el diseño del corpus suponía que la mitad de los datos primarios serían de origen oral, provenientes de grabaciones, en su mayor parte vídeo, hacía falta recopilar más que 100 000 palabras. En teoría eso significaba realizar el mínimo de 33 entrevistas, 11 en cada una de las localidades estudiadas. En práctica se han grabado y transcrito, hasta el momento, 36 entrevistas, 12 en lagarteiru, 11 en mañegu y 13 en valverdeñu. El pequeño desbalance entre las tres variedades está causado por el proyecto escolar en el cual participaron más estudiantes de Valverde. En las 36 entrevistas contribuyeron en total 64 participantes con diferentes papeles, algunos como entrevistadores otros como entrevistados.

Es natural que nuestra intención fuera crear una muestra equilibrada no solo respecto a las tres variedades sino también respecto al género y la edad de los participantes. Afortunadamente, no hubo muchos problemas con el reclutamiento de participantes de las en-

trevistas. Los miembros de la comunidad de habla son generalmente bastante cooperativos y normalmente se sienten orgullosos de compartir su lengua y felices de que alguien se interese por ella.

Conseguir el equilibrio de géneros y de la edad de los participantes causaba complicaciones de poca gravedad. Con respecto al género, se han entrevistado 24 mujeres y 25 hombres, sin contar los estudiantes entrevistadores, lo cual crea una muestra muy equilibrada. Por lo que se refiere a la edad, los estudiantes entrevistadores tenían de 15 a 16 años, mientras que el mayor participante de las entrevistas tenía 90 años. El amplio rango de edad asegura que la muestra contendrá tanto el habla de los jóvenes como de los mayores. Sin embargo, los participantes "ideales" a los que se buscaba generalmente eran personas de 45 a 75 años de edad, ya que estos usan el lenguaje bastante tradicional con pocos castellanismos y son suficientemente elocuentes para responder de manera más elaborada las preguntas del entrevistador. De esta forma la edad promedia de los participantes, sin contar los entrevistadores jóvenes del proyecto escolar, oscila entre 52 y 61 años, siendo un poco más alta en San Martín y más baja en los dos pueblos restantes. Este desequilibrio está causado por el tamaño reducido de la muestra, en la que uno o dos participantes muy mayores cambian fácilmente el promedio de la edad. A pesar de las diferencias existentes se puede afirmar con certeza que la muestra de participantes que contribuyeron en la parte oral del corpus era equilibrada respecto a la edad y género.

Sorprendentemente más complicada fue la clasificación de los participantes respecto a la delimitación del concepto del hablante nativo o natural de A Fala. La dificultad de etiquetar a alguien como hablante nativo la encontramos especialmente en San Martín, pero no solamente allí. Tras realizar algunas entrevistas surgió la pregunta sobre la lengua materna y ocurrió en más de una ocasión que los participantes respondieron que en su familia, en realidad, se hablaba castellano, normalmente por causa de que alguien de la familia vino de fuera del pueblo, mientras que el mañegu era la lengua que usaban para comunicar con el resto del pueblo. A veces el uso de las dos lenguas era todavía más complicado. Así que una de las participantes respondió que hablaba castellano con sus padres, mañegu con su marido y también con toda su familia y mañegu y castellano con sus hijos. En otra ocasión nos hemos encontrado con una persona que hablaba el castellano con uno de sus hermanos y el valverdeñu con su otro hermano. Estos casos ejemplifican solo un fragmento de la complejidad del uso de las dos lenguas en las tres localidades. Es cierto que el bilingüismo y preferencias hacia una de las dos lenguas merecerían un estudio más detallado en esta comunidad de habla, ya que este podría aportar datos muy interesantes sobre el uso de la lengua en la sociedad y su relación con a la identidad. De todas maneras, para los fines de nuestro estudio era necesario definir participantes elegibles para aportar datos al corpus oral. Considerando que en casi todas las familias aparecen personas de los otros dos pueblos de A Fala, a causa de matrimonios mixtos, y en muchas también personas de pueblos castellanohablantes sería un error excluir la mayor parte de la población y crear una isla artificial de los hablantes con "ocho apellidos" mañegus, lagarteirus o valverdeñus. Otra consideración a favor de la inclusión de los participantes que hablan de manera natural una de las variedades a pesar de que su lengua materna es el castellano es la perspectiva futura. Como se ha demostrado en el ejemplo arriba descrito, hasta los hablantes que recibieron de sus padres el castellano como su lengua materna hablan en A Fala con sus hijos, y de esta forma transmiten la lengua a la siguiente generación. Por estas razones hemos rechazado la solución "purista" e incluimos en la base de datos también las entrevistas con hablantes naturales de A Fala, completamente bilingües,

aunque algunos marcaron el castellano como su lengua materna. La experiencia que hemos ganado comprueba que los participantes que aprendieron A Fala "en la calle", es decir, de sus amigos pero no en su familia, no usan más palabras castellanas o castellanizadas que otros participantes. Desgraciadamente, la influencia del castellano se puede observar en el habla de todos los participantes, aún en el habla de los más mayores y los más arraigados en los tres pueblos.

Por lo que se refiere a los aspectos técnicos de las grabaciones, casi todas ellas eran grabaciones video, menos dos realizadas antes del inicio del proyecto. La mayoría de las grabaciones eran estáticas, es decir, con cámara fija y los participantes hablando en un espacio antes preparado. En algunos casos, cuando se intentaba grabar también el ambiente del que se hablaba, por ejemplo, una bodega de vinos, se hicieron grabaciones con cámara en la mano para poder acercarse más a los objetos de los que se hablaba. El formato original .MOV fue transformado a .mp4 para su mejor manejo y sobre todo para la posibilidad de archivarlo a largo plazo en alguno de los repositorios especializados. El audio fue grabado directamente en el formato .wav y por lo tanto no necesitaba ninguna conversión.

Una vez realizadas las grabaciones estas se transcribieron y luego se incluyeron en la base de datos. Tanto el video como el audio se insertaron primero en el programa ELAN que sirvió para la segmentación de la pista de audio y su consiguiente transcripción. La ventaja del programa ELAN, con los archivos en el formato .eaf, es su versatilidad que permite futuras modificaciones, por ejemplo se puede añadir transcripción fonética, traducción, crear subtítulos, etcétera, búsquedas rápidas de palabras individuales y posibilidad de exportar e importar los datos al programa FLEx (Fieldworks Language Explorer). En este programa se guarda y procesa la base de datos y por lo tanto es importante la intercomunicación entre estos dos programas. También es importante subrayar que ambos programas son no-propietarios, o sea, de libre difusión y no necesitan compra de licencia para su uso.

La transcripción de las grabaciones era la tarea más importante y también la más trabajosa ya que de su exactitud dependía el resultado final de la base de datos. El ambiente de las entrevistas intentaba conseguir el uso más natural posible de la lengua para asegurar la mayor autenticidad del corpus. Las conversaciones informales son, no obstante, difíciles de transcribir de manera exacta, porque en el habla natural pronunciamos algunos sonidos de manera menos cuidadosa. Esto se manifiesta especialmente en la pronunciación de las vocales átonas y por lo tanto fue a veces difícil decidir si el participante pronunció una /e/ o una /i/, *escuela* o *iscuela*, del mismo modo era difícil diferenciar las vocales /o/ y /u/, *oliveira* o *uliveira*, ya que las vocales átonas prácticamente no se pronuncian. Hay que destacar que las grabaciones se transcribieron con absoluta meticulosidad, a pesar del consumo extensivo del tiempo, los borradores se consultaron muchas veces con los participantes y también con otros miembros colaboradores de la comunidad. Por esta razón creemos que el resultado tiene una precisión superior a 97% y por consiguiente el corpus será una fuente muy fiable de datos primarios, reflejando con precisión la lengua en la forma que se usa hoy en día.

Las grabaciones abarcan una gran variedad de temas desde los más tradicionales como el contrabando, cocina tradicional o agricultura, hasta los más actuales como el deporte, el ocio, la despoblación o el uso de los fondos europeos para el desarrollo local. Además, algunas de las entrevistas están grabadas en ambientes relacionados con el tema de la entrevista y así los participantes, por ejemplo, explican el uso de herramientas y historias relacionadas con ellas en el sitio donde las almacenan.

En conclusión, los datos obtenidos en las grabaciones resultan muy valiosos y serán útiles para su posterior procesamiento formando uno de los pilares fundamentales de la base de datos, especialmente si consideramos la selección de los participantes, la amplia gama de los temas discutidos, la meticulosidad de la transcripción de las entrevistas y la participación de la comunidad de habla en su recopilación.

4. Datos escritos

En el apartado anterior se comentaban las limitaciones de los datos recopilados en forma de grabaciones. Es cierto que los textos compensan los límites del lenguaje oral aportando una gama más amplia del vocabulario. Los textos usan también registros más formales y cubren los temas que difícilmente aparecerían en una conversación informal. Otro aspecto positivo es que los autores de los textos escritos no pueden escribir ciertos sonidos "a medias" u omitir la escritura de las vocales tal como sucede en el discurso habado. Por todas estas razones forman los textos escritos aproximadamente la mitad del corpus. Se suponía recopilar 100 000 palabras y esta cifra ha sido superada en la cuenta final.

Los textos escritos tienen sus aspectos positivos, como el vocabulario más amplio, sin embargo, conllevan también todo un conjunto de complicaciones. La limitación más importante es la menor "espontaneidad" de los textos escritos en comparación con un discurso hablado, en el cual usamos la lengua vernácula sin pensar y reflexionar sobre ella. En los textos escritos a menudo aparecen traducciones literales del español, erratas de escritura, y también castellanismos, porque los autores están acostumbrados a escribir en castellano y la forma gráfica de las palabras que tienen grabada en la mente es la española. Así, fácilmente escriben algo que nunca pronunciarían en un discurso natural informal.

La incorporación de los textos en el corpus suponía entonces como mínimo dos pasos importantes. Primero, hacía falta unificar la forma de escribir para no tener escritas palabras iguales en formas diferentes. Este paso era importante para las búsquedas que facilita la base de datos y, en general, para asegurar su coherencia. Con este fin se usaron las reglas ortográficas publicadas por la *Asociación Cultural A Nosa Fala* en el año 2017. Aunque no se trata de una ortografía oficial las líneas fundamentales de esta han sido diseñadas a base de consultas con numerosos miembros de la comunidad de habla y han surgido así de un consenso común. El otro paso era la corrección de errores que han cometido los autores. En este punto hay que admitir que se trataba hasta cierto punto de cambios arbitrarios de los textos sin saber a ciencia cierta cuál fue la intención de los autores. A pesar de esta desventaja intentamos de la manera más prudente y cuidadosa corregir los errores más llamativos como por ejemplo las terminaciones castellanas -e, -o: *disfrute* por *disfruti*, *contrabando* por *contrabandu*, etcétera. En caso de las sibilantes escritas de manera inconsistente intentamos siempre guiarnos por las soluciones mayoritarias en cada pueblo y por las consultas con los hablantes.

Otra complicación relacionada con los textos era la búsqueda de textos adecuados en cada una de las tres variedades, ya que no todos los textos cumplían con un criterio fundamental que era: la intención del autor escribir en su variedad vernácula. En A Fala existen textos escritos con la intención de manifestar su relación de origen con el portugués, con el gallego o con el astur-leonés.

A estos textos sería demasiado difícil hacer la corrección antes descrita, ya que no podríamos averiguar donde el autor usó su variedad vernácula y donde ya pasó al portugués u otra lengua para acentuar el parentesco. En consecuencia, el texto necesitaría o una corrección demasiado profunda que lo alejaría demasiado de la versión original o correríamos el riesgo de implementar en el corpus palabras completamente ajenas a A Fala. Lo mismo sucedía con los textos escritos en una variedad intermedia, aunque sea más cercana a una de las tres variedades de A Fala. Ya que los textos y sus respectivas palabras siempre llevan la etiqueta de la variedad original, en caso de los textos indefinidos esta no se le podría adscribir de manera satisfactoria. Desgraciadamente, este criterio elemental limitó la elegibilidad de algunos textos causando su escasez en una de las tres variedades. Los retos relacionados a los textos eran específicos para cada una de las tres variedades.

La variedad con más abundancia de textos era lagarteiru. En lagarteiru se redacta anualmente la revista *Anduriña* que lleva ya 16 ediciones. Los contribuyentes son varios autores originales de Eljas, algunos publican de forma regular mientras que otros solo escriben de manera puntual. Otra fuente de textos redactados en lagarteiru era el autor de numerosas obras teatrales, que representa el grupo *U Lagartu Cómicu*, Furén Moreno Blanco. Además, existen dos libros publicados por otro autor local: Severino López Fernández, uno con el título: *Arreidis: Palabras y Ditus Lagarteirus* (1999) y el otro titulado *Topónimus d'as Ellas y rimas en lagarteiru* (1992). No obstante, la abundancia de textos en lagarteiru y la cantidad de autores no significa que no enfrentáramos ningún desafío. De ejemplo pueden servir las contribuciones a la revista *Anduriña* y la forma final de los artículos incluidos en el corpus. El proceso de publicación comienza cuando el autor envía su artículo. Los editores intentan modificarlo para que sea coherente respecto a la ortografía, y también corrigen errores y reemplazan algunos castellanismos. El resultado es un trabajo colaborativo entre el autor y los editores que, sin embargo, no está libre de erratas y por lo tanto todavía pasa por la adaptación para poder figurar en el corpus. Podemos ver que son muchos sujetos los que intervienen en el resultado final. Sin embargo, este proceso es inevitable y hasta tiene sus aspectos positivos, ya que el texto que se incluye en el corpus ha sido verificado varias veces.

La recopilación de datos escritos en valverdeñu se enfrentaba con la escasez de textos disponibles publicados en esta variedad. Más aún, uno de los libros clásicos titulado *Seis Sainetes Valverdeiros*, escrito por Isabel López Lajas en la primera mitad del siglo XX, resultó inadecuado porque el lenguaje que se usa en este libro no corresponde con la realidad lingüística actual. No es cierto hasta que punto el libro documenta el habla del siglo XIX y principios del siglo XX o hasta que punto se trata de una licencia literaria folclorizante de la autora, pero sería difícilmente sostenible proclamar que el texto refleja el valverdeñu aunque nos refiriésemos a su uso por parte de las personas más mayores. Otro libro *Alcaldis de Valverdi*, escrito por David Carrasco, planteó la cuestión de los derechos de autor y la ética de la recopilación de los datos. En el texto del libro se dice explícitamente que este fue escrito para la familia y no se supone su publicación. Además, es cierto que el libro contiene información sensible cuya publicación podría ser problemática. En consecuencia, el libro puede servir para la ampliación del léxico en la base de datos, pero no puede formar parte del corpus. A causa de esto algunas palabras en la base de datos no serán comprobables en su contexto original y así se disminuirá parcialmente su funcionalidad. Desgraciadamente este no es el único texto con esta limitación, ya que conseguir consentimiento para la publicación de todos los textos en la base de datos fue un reto que no siempre se ha conseguido. Los demás textos escritos en valverdeñu, por ejemplo, el libro *Istu lo se le*, publicado por el Ayuntamiento de

Valverde, solo requirieron una adaptación habitual para conseguir la compatibilidad ortográfica y naturalmente también planteaba decisiones respeto a los errores. Sin embargo, esta interferencia, aunque no deseable, era inevitable para asegurar la coherencia del corpus y de la base de datos.

Los datos referentes a mañegu suponían el reto más grande entre las tres variedades. Aunque existen varios libros publicados en mañegu, quizás más que en valverdeñu, el inconveniente común es que solo tienen un autor: Domingo Frades Gaspar. Gracias a él tenemos tantos recursos en mañegu, pero por otro lado la parte mañega del corpus casi carece otras fuentes de datos escritos ya que de las 25 000 palabras en mañegu 22 000 fueron escritas por Domingo Frades. Se trata de un autor culto que publicó poesía, libros muy conocidos como *Vamus a falal* (1996) y también traducciones, pero como cada hablante tiene su idiolecto particular. En el *Prefacio* de la *Biblia* que tradujo, podemos leer las palabras de Carrasco González:

> Domingo Frades ha optado por reproducir las características propias del mañego hablado en su localidad natal, San Martín de Trevejo, y dentro de ellas se ha decidido por las soluciones personales que más le satisfacen. Anuncio en este respecto que se trata de una persona culta, con conocimiento de idiomas (incluido el latín), que ha estudiado, trabajado y residido fuera de San Martín. Su mañego no será el que esperaríamos de una persona escogida por la dialectología tradicional para recuperar las formas más ‚puras' y antiguas, menos ‚contaminadas' por la lengua oficial. [...] Quizás lo primero que llame la atención es la incorporación de castellanismos al mañego (Carrasco González 2015: pp. VIII-IX).

Desde el punto de vista del equilibrio del corpus no es ideal que una variedad dependa de las "soluciones personales" de un autor, pero tampoco podemos ignorar la realidad y hay que admitir que en mañegu no escribieron más autores que Domingo Frades. Para compensar esta desventaja, tratamos de confirmar todas palabras sospechosas con los miembros de la comunidad de habla.

La falta de textos en mañegu y valverdeñu nos hizo usar las traducciones como fuente de datos primarios. Las traducciones son naturalmente más propensas a incluir castellanismos, ya que se trata de traducciones del español hechas por traductores no profesionales. Por otro lado, la traducción es un proceso de toma de decisiones como afirma por ejemplo Levý (2012) y por eso el traductor piensa más sobre el idioma y selecciona las formas con más cuidado, lo cual es un factor enriquecedor.

Como queda evidenciado de lo anteriormente escrito, los textos que forman parte del corpus tienen sus limitaciones, pero esta circunstancia no sorprende con una lengua minoritaria con poca tradición escrita. De hecho, la tradición escrita podría más bien etiquetarse como emergente, ya que uno de los factores positivos que se pueden observar, es que se escribe cada vez más en A Fala, se organizan concursos literarios, se empiezan a traducir obras literarias y por lo tanto es bastante probable que en el futuro sea posible extender la parte escrita del corpus con más facilidad.

5. Participación de la comunidad de habla

Uno de los aspectos fundamentales del corpus y de la base de datos es la participación de la comunidad de habla en su recopilación. En general, los miembros de la comunidad de habla no sirvieron solo como participantes en las entrevistas, su incorporación en el proceso de la recopilación de

datos fue mucho más amplia. Primero, hay que destacar su intervención en las transcripciones de las entrevistas. Algunos de los entrevistados transcribieron su propia entrevista o la de su familiar, con lo cual contribuyeron a la exactitud de los datos. En otros casos se consultaban los enunciados dudosos con los participantes y así fueron ellos mismos quienes aclararon la verbalización exacta. Segundo, toda la parte del corpus escrito proviene de los miembros de la comunidad de habla. En el apartado anterior se discutieron las ventajas y desventajas de esta forma de datos sin destacar la amplia gama de contribuidores lagarteirus y valverdeñus, de esta forma el corpus refleja idiolectos de muchos hablantes y documenta así la rica variedad de A Fala.

Otra forma de colaboración con la comunidad fue a través de la aplicación Whatsapp. Se han creado tres grupos de colaboradores, un grupo para cada una de las variedades, en los cuales se consultaron dudas y a veces los participantes mismos sugerían palabras que deberían incluirse en la base de datos. En esos casos se trataba normalmente de palabras poco usadas, las que "se usaban antes" y los hablantes las conocían por estar en contacto con personas mayores. Como los contribuidores no tenían formación lingüística, a veces no entendían bien que la variación es una cualidad natural de las lenguas, al revés, por la influencia de la escolarización estaban acostumbrados a etiquetar la formas como correctas o incorrectas con lo cual podía ocurrir que alguien sintiera su idiolecto como menosvalorizado. Sin embargo, los grupos de Whatsapp aportaron mucha información valiosa, además, comprobada por más que un solo hablante, y de esta forma enriquecieron notablemente la base de datos.

La participación de la comunidad en la recopilación de datos incluía también hablantes jóvenes. En cooperación con el Instituto *Ieso Val de Xálima* en Valverde del Fresno se realizó el proyecto "Xálima Fala / A vo de Xálima". En este proyecto los estudiantes mismos eligieron a los participantes para ser entrevistados, prepararon la entrevista y la llevaron a cabo. Después transcribieron la grabación y así aprendieron hacer todo el proceso de documentación lingüística. De esta forma se llevaron a cabo ocho entrevistas: cinco en valverdeñu, dos en lagarteiru y una en mañegu, lo cual desequilibró un poco la representación de las variedades y tenía que ser compensado posteriormente. Los estudiantes naturalmente aprendieron mucho más que el proceso de la documentación, transcribiendo las entrevistas podían observar la diferencia entre el lenguaje hablado y escrito, pero sobre todo se dieron cuenta de que su lengua materna está en peligro. La valoración positiva de su lengua vernácula fue entonces uno de los mayores logros de esta actividad.

Con los miembros de la comunidad también se consultaban personalmente los aspectos individuales de la lengua y el uso de las palabras. Algunos hasta crearon sus propias listas de palabras, en este punto hay que destacar que los miembros de la comunidad de habla están preocupados por la castellanización de A Fala y el reemplazo de palabras vernáculas por las castellanas. Por lo tanto, las listas normalmente incluían palabras en gradual desuso y poco frecuentes. En la creación de la base de datos participaron, en total, unas 130 personas que es el 3% de la población de los tres pueblos.

6. Conclusión

Uno de los aspectos positivos del corpus y la base de datos es la posibilidad de su futura ampliación. Cuantas más palabras contenga el corpus tanto más precisa es la información sobre el idioma, por lo que se supone que el proceso de complementación del corpus continuará aún después

de que se acabe el proyecto. En primer lugar, el corpus todavía no incluye todo el material disponible en la forma escrita, sobre todo en lagarteiru, existen muchos textos no incluidos. Segundo, hay otras fuentes de datos, mencionadas en el apartado dos, como vocabularios publicados o sin publicar, páginas web y otros recursos que pueden complementar la base de datos. Otro aspecto que no se ha cubierto de manera satisfactoria es el etnográfico. La base de datos está adaptada para la inclusión de material fotográfico y audiovisual muy propenso al almacenamiento de la información etnográfica. También, la incorporación de los participantes jóvenes debería continuar para que los hablantes de la lengua se den cuenta que el destino de su lengua materna depende de ellos.

En la descripción de las fuentes de datos se puede notar que todas tienen sus ventajas pero también sus limitaciones. Los problemas y decisiones a los que nos enfrentamos al recopilar los datos grabados: vocabulario limitado, decisiones sobre el hablante nativo, el equilibrio de edad y género, el equilibrio de variedades y las transcripciones, deben considerarse en el momento del uso de la base de datos. Igualmente hay que considerar las limitaciones de los datos recopilados en forma de textos. En general, no existen datos ideales, ni en forma grabada ni escrita, ya que no hay hablantes ideales ni textos ideales, solo existen datos primarios auténticos y estos están fielmente reflejados en el corpus, que fue diseñado de forma ideal; sin embargo, su compilación tenía que respetar las soluciones viables y la situación sociolingüística de los tres pueblos. De todas maneras, el resultado final, el corpus y la base de datos elaborada con el esfuerzo compartido entre el lingüista y la comunidad de habla será una herramienta muy útil para una amplia gama de estudios sobre A Fala, lo cual naturalmente no excluye sus futuras correcciones y ampliaciones.

Referencias bibliográficas

Asociación Cultural A Nosa Fala. (2017). *Proposta de ortografía de A Fala*. Eljas: ACNAF

Austin, P. K. (2016). Language documentation 20 years on. In L. Filipović & M. Pütz (Eds.), *Endangered Languages and Languages in Danger* (pp. 147–170). Amsterdam: John Benjamins Publishing.

Carrasco González, J. M. (2015). El mañego de Domingo Frades. In *Novu Testamentu en Fala* (pp. viii-x). Madrid: Sociedad Bíblica de España.

Diccionario virtual de Extremadura. (2019). <https://diccionariovirtualextremadura.blogspot.com/2019/01/vocabulario-de-san-martin-detrevejo.html?spref=fb&m=1&fbclid=IwAR0qxvCh94M1TGhAPYu5CR-6HMHlxqG9IZEhNsFPoP3nUSP1uvasyIRgmnDk>

Dondelewski, B. (2020). *La identidad y el cambio en a fala (Cáceres, España)*. Krakow: Uniwersytet Jagiellońsk w Krakowie (disertación inédita).

Ferguson, C. (1959). Diglossia. *Word,* 15, 325–340.

Frades Gaspar, D. (2000). *Vamus a falal*, 2ª edición. Mérida: Editora regional de Extremadura.

Gobierno de Extremadura. (2001). *Diario Oficial de Extremadura* (DOE), *36*, 2859–2860. <http://doe.gobex.es/pdfs/doe/2001/360o/01040052.pdf>

Good, J. (2018). Reflections on the scope of language documentation. In B. McDonnell & A. L. Berez-Kroeker & G. Holton (Eds.), *Reflections on Language Documentation 20 Years after Himmelmann 1998* (pp. 13–21), Honolulu: University of Hawai'i at Mānoa.

Himmelmann, N. (1998). Documentary and descriptive linguistics. *Linguistics*, *36*, 161–195.

Levý, J. (2012). *Umění překladu,* 4ª ed. Praha: Miroslav Pošta - Apostrof.

López Fernández, S. (1992). *Topónimus d'as Ellas y rimas en lagarteiru.* Eljas.

————. (1999). *Arreidis. Palabras y ditus lagarteirus*. Mérida: Editora regional de Extremadura.

Ramallo, F. (2011). O enclave lingüístico de Xálima: unha análise sociolingüística. *Estudios de lingüística Galega*, *3*, 111–135.

Rey Yelmo, J. C. (1999). *La fala de San Martín de Trevejo: O mañegu*. Mérida: Editora regional de Extremadura.

Román Domínguez, A. (2008). *Contribución ao léxico do galego exterior: O val do río Ellas*. Universidade de Vigo (Trabajo de investigación inédito).