e-ScriptaRomanica
vol. 5 (2018): 28-38
MONOGRÁFICO

http://www.e-scripta.uni.lodz.pl/ojs/index.php/escriptaromanica | ISSN 2392-0718
Received: 22/03/2016 | Accepted: 14/06/2016
http://dx.doi.org/10.18778/2392-0718.05.04

Construyendo un corpus oral para el gallego. El proyecto CORILGA

Building a corpus of spoken Galician language. The CORILGA project

José Manuel Dopazo Entenza
Instituto da Lingua Galega (Universidade de Santiago de Compostela), España

Resumen: El CORILGA (Corpus Oral Informatizado da Lingua Galega) es un corpus de grabaciones alineadas con su transcripción y anotadas en distintos niveles (ortográfico, fonético, morfológico…). Una codificación completa y minuciosa de los datos de las grabaciones y de los informantes permite, mediante un buscador *online* abierto al público, conseguir unos resultados de búsqueda muy precisos. Esta información se podría utilizar para la realización de estudios de variación y cambio lingüístico, así como para crear materiales para la enseñanza o desarrollar tecnologías del habla.

Palabras clave: CORILGA, variación y cambio lingüístico, corpus oral, reconocimiento de voz

Abstract: The CORILGA (Corpus Oral Informatizado de la Lengua Gallega) is a corpus of recordings aligned with their transcription and annotated at different levels (spelling, phonetic, morphological, syntactic...). A complete and thorough recordings and participants data allows, through an online open search engine, to get very accurate search results. This information could be used in language variation and change studies and to create materials for teaching or developing speech technology.

Keywords: CORILGA, language variation and change, oral corpus, speech recognition

1. Introducción

En 1967, la publicación de la obra *Computational Analysis of Present-Day American English* de Henry Kucera y Nelson Francis supuso un hito para la lingüística, al ser el primer estudio de una lengua basado en un corpus lingüístico. Esto no sería posible sin el desarrollo de la tecnología informática.

Los primeros corpus estaban formados por textos escritos, por su mayor accesibilidad, ya que la realización de transcripción de las grabaciones presentaba problemas mucho más complejos. Con el tiempo, la realización de grabaciones fue haciéndose más asequible y fue a partir de entonces que empezaron a aparecer los primeros corpus orales. El Proyecto Francés de Montreal, creado en 1971 (Sankoff & Sankoff, 1971: 7–64), se considera el primero de ellos. Hasta ese momento o bien no se trabajaba con lengua oral o, en su defecto, se hacía a través de encuestas que se transcribían *in situ*. Todavía a día de hoy el número de corpus orales en relación a los escritos es verdaderamente reducido.

La utilización de corpus se ha extendido para diferentes tipos de estudios y finalidades (enseñanza de lenguas, variación y cambio lingüístico, etc.) y, desde hace algunos años, es habitual hablar de la lingüística de corpus como una subdisciplina diferenciada, con metodologías y problemas específicos. Debido a la gran cantidad de datos que son precisos para este tipo de estudios, la lingüística de corpus echa mano de los avances tecnológicos en materia computacional. Cada vez es posible almacenar más datos y procesarlos a mayor velocidad.

El Instituto da Lingua Galega (ILG), centro dependiente de la Universidade de Santiago de Compostela, en colaboración con el grupo AtlanTIC, de la Universidade de Vigo, ha emprendido la creación de un corpus oral alineado para el gallego. De este modo, en 2012 se puso en marcha el proyecto CORILGA (Corpus Oral Informatizado de la Lengua Gallega), que se concibe como una herramienta web de libre acceso con posibilidad de búsquedas combinadas (ortográfica, fonética, morfológica, léxica). Hasta el momento no se ha abierto al público, pero está prevista su presentación para este año 2016.

2. Estructura del corpus

2.1. Composición

El ILG reúne grabaciones desde la década de 1960 hasta la actualidad, realizadas para diferentes proyectos, extraídas de tesis doctorales y cedidas por particulares. Se trata, sobre todo, de entrevistas semidirigidas que corresponden a hablantes de zonas rurales y con modelos de lengua conservadores, por lo que hay muy pocas conversaciones, así como textos de estilos más formales y poca lengua urbana. Con este fondo documental de producciones orales se comenzó a trabajar en las transcripciones que ahora integran CORILGA.

En primer lugar se empezó con los materiales de otros proyectos del ILG, entre los que se encuentran los del Atlas Lingüístico Galego (ALGA), realizadas en la década de 1970, de las cuales se han escogido dos, de 21 minutos de duración en conjunto. Del proyecto Arquivo do Galego Oral (AGO), se han extraído 29 grabaciones de la década de 1990, con un total de 7 horas y 12 minutos. El Proxecto Prosodia da Lingua Galega contiene grabaciones realizadas en conferencias y discursos durante los años 1999, 2000 y 2001, de las cuales se han utilizado 23, con una duración de 5 horas y 48 minutos. Del proyecto Atlas Multimedia Prosódico do Espazo Románico (AMPER-Galicia) se han transcrito 6 grabaciones de entre el 2003 y el 2012, con una duración total de casi una hora.

A continuación, se han recuperado archivos de tesis doctorales como la de Francisco Dubert García[1], sobre el habla de Santiago de Compostela, 21 grabaciones de la primera mitad de la década de 1990, un total de 20 horas y 29 minutos. La tesis doctoral de Xosé Luís Regueira Fernández[2] sobre el habla de la comarca de Vilalba consta de 50 archivos de la década de 1980, un total de 15 horas.

Más tarde se ha contado con aportaciones de corpus particulares de gran interés como el de Gustav Henningsen. Este antropólogo danés llegó a Galicia en

[1] Estas grabaciones están transcritas, pero pendientes de ser introducidas en la base de datos.
[2] Estas grabaciones están transcritas, pero de momento solo se han introducido en la base de datos 12 (2 horas y 9 minutos).

los años 60 para realizar su tesis doctoral y recogió 205 grabaciones de toda la geografía lingüística gallega (Galicia y comarcas colindantes de habla gallega) con un total de 69 horas y 12 minutos, de las cuales hay introducidas en CORILGA 9 horas y 45 minutos. Y el corpus de Manuel Rico, 23 grabaciones realizadas en el programa de entrevistas Noroeste-Terra e Xente, que se emitía en los años 1980 a través de Radio Nacional de España para Galicia. De estas entrevistas tenemos un total de casi 2 horas transcritas hasta el momento, con intención de seguir introduciendo más.

Este proyecto no solo se nutre de archivos propios, sino que también hace acopio de materiales públicos de los medios de comunicación o de Internet. De este género contamos con 31 archivos transcritos, unas 7 horas y media.

Todo esto se complementa con cesiones particulares como las de Miguel Abraira, Noemi Basanta, Xabier Iglesias o Eduardo Louredo, con lo que el número de archivos transcritos asciende a 203, con una duración total de 85 horas. Este es el estado actual, pero se está ampliando, realizando grabaciones con atención a aquellas modalidades y perfiles sociolingüísticos menos representados (conversaciones, lengua urbana, gente joven…). CORILGA es un corpus en continua construcción.

2.2. Criterios de selección

El criterio básico para entrar a formar parte de este corpus es que la grabación esté, total o parcialmente, realizada en gallego. A continuación, seleccionamos las grabaciones y, para ello, debemos establecer unos criterios. En este caso, se ha realizado una clasificación de los archivos existentes, dando como resultado tres grandes grupos: oralidad informal, oralidad formal y medios de comunicación. Por supuesto, estas tres categorías constan de sus correspondientes subgrupos.

Pertenecen a la **oralidad informal** las subcategorías de conversaciones, entrevistas semidirigidas, monólogos y literatura oral. Dentro de la **oralidad formal** podemos encontrar el discurso formal (oral y leído), lectura literaria y teatro. Y, por último, en **medios de comunicación** discernimos si son informativos, magazines, entrevistas, debates, conversaciones, series, cine y doblaje.

Teniendo en cuenta que la mayoría de las comunicaciones humanas se pueden clasificar dentro de lo que hemos llamado oralidad informal y dado que esta variedad de lengua es la que siempre ha suscitado la atención del personal investigador, en CORILGA se ha reservado para esta categoría un porcentaje mayor del total de grabaciones del corpus (un 50%). Para la oralidad formal, que goza de mayor prestigio social y es la más próxima a la norma, se ha considerado que suponga el 30% del total del corpus. Por último, los medios de comunicación (televisión, radio, Internet…), que constituyen un espacio discursivo diferente por sus finalidades y usos concretos y planificados, suponen un 20% del corpus. Dentro de estos grandes grupos se ha intentado compensar el tiempo de grabación transcrita en cada una de las subcategorías para que el corpus ofrezca una visión holística de la lengua gallega en todos sus usos.

Otro criterio a tener en cuenta para la selección de las grabaciones que se van a transcribir es el año y el lugar de procedencia. Como se ha dicho en el punto anterior, existen archivos desde la década de 1960 hasta la actualidad, por ello, la

introducción de material al corpus debe irse compensando en lo que a fechas se refiere. Es cierto que las grabaciones más antiguas gozan de importancia por ser pocas y es preciso introducir una cierta abundancia para poder estudiar el cambio, pero también necesitamos completar el corpus con grabaciones actuales para tener una imagen del presente e incluso una visión del cambio en tiempo real. Algo semejante ocurre con el lugar de procedencia, pues hay que tener en cuenta la población rural, semiurbana y urbana y transcribir en consecuencia. Las grabaciones existentes en el Instituto da Lingua Galega pertenecían en su mayoría a un perfil de hablante rural y de edad avanzada, por tanto, ahora se procura completar con informantes urbanos y gente joven.

Con un carácter menos estricto, pero que también se valora en la medida de lo posible, se toma como criterios de selección el sexo y la edad de las personas grabadas. Estas pautas contribuyen a que el corpus sea una muestra de la variación lingüística de la lengua gallega.

2.3. Transcripción

Las transcripciones se realizan con el programa ELAN[3], de acceso libre, que permite alinear el audio con el texto. Con este programa se pueden seleccionar fragmentos de la grabación, que contienen información del momento concreto en que empieza y termina. De este modo, la transcripción estará ligada al audio de forma exacta.

Además, ELAN permite crear numerosas líneas dependientes e independientes. Con las líneas independientes podemos transcribir separadamente lo que dice cada una de las personas que hablan en la grabación. Las dependientes permiten incorporar los niveles de anotación que se quieren realizar para cada una de esas personas (fonético, morfológico, léxico).

La línea de transcripción principal de CORILGA es la ortográfica, acompañada de otras con diferentes anotaciones (fonéticas, morfológicas, léxicas). En esta línea se sigue la representación escrita establecida por la normativa ortográfica y se presenta una transcripción fiel del texto oral, lo cual supone que no se estandariza la lengua del texto, aunque se evita reflejar las variantes ocasionales de pronunciación que aparecen en las grabaciones, para facilitar la recuperación de la información.

Cabe mencionar que hay una serie de casos que se desvían de la representación ortográfica habitual. Es el caso de los pronombres personales clíticos, que en gallego no se separan por guiones (como en el caso del portugués), pero que en CORILGA se ha optado por hacerlo (por ejemplo: *fála-me, déixa-me-lle*, *achegóu-se-nos, fun po-lo ver*). Lo mismo podemos decir de las formas alomórficas del artículo determinado *-la(s), -lo(s), -na(s), -no(s)* (por ejemplo: *eu e mai-lo outro, po-lo río abaixo, tamén-o viu, collen-o tren*). Por último, también se separa por guion la vocal paragógica *-e*, como en *facer-e, muller-e, eu-e, tamén-e*.

Como es habitual en este tipo de corpus, existen una serie de convenciones a la hora de transcribir que marcan determinados rasgos de la lengua oral. Es el caso de los turnos de palabra, para lo cual se emplean los siguientes signos:

[3] Para más información sobre el programa: https://tla.mpi.nl/tools/tla-tools/elan/

&	Continuación del turno de palabra.
%	Rapidez en la sucesión de los turnos de palabra.
[*texto afectado*]	Solapamientos.
texto afec#	Corte abrupto de una palabra.

También se emplean signos para marcar algunos rasgos prosódicos:

: :: :::	Alargamientos de un sonido.
¿?	Entonación interrogativa.
¡!	Entonación exclamativa.
TEXTO AFECTADO	Énfasis.

Utilizamos signos también para otros fenómenos lingüísticos que resulten de utilidad:

{<*valor*> *expresión*} Onomatopeyas, interjecciones y elementos paralingüísticos. Ejemplo: *{<asentimento> ahá}, {(risas) ¡eu tamén!}*.

{<*incomprensíbel*>} Fragmentos incomprensibles.

Para la línea de transcripción fonética se utiliza el Alfabeto Fonético Internacional. En el estado actual de CORILGA muchas de las transcripciones están realizadas de forma aproximada y necesitan ser revisadas. Pese a ello, las grabaciones correspondientes a los corpus de Dubert (CDUB) y de Regueira (CREG), entre algunas otras, si se encuentran totalmente revisadas.

Para agilizar la labor de transcripción, el grupo AtlanTIC ha incorporado a CORILGA un transcriptor automático basado en un reconocedor de voz que habían creado para el gallego. Con él, se puede contar con una primera transcripción sobre la cual trabajar manualmente hasta afinar el resultado.

Para obtener la transcripción, hay que incorporar un audio a la plataforma web y parte de la transcripción ya realizada manualmente. El reconocedor utilizará la parte transcrita como modelo para reconocer el resto de la grabación. También se puede introducir solo el audio, de este modo el reconocedor actuará basándose en los modelos lingüísticos que tiene incorporados el gestor web. En este caso le será más compleja la tarea de transcribir y la calidad de la transcripción será más baja. Cabe apuntar que esta herramienta solo funciona con resultado aceptable en las grabaciones formales.

Basado en el reconocedor, AtlanTIC ha desarrollado un alineador que ajusta una transcripción ya realizada con el audio correspondiente. Esta herramienta permite introducir no solo la transcripción ortográfica, sino también la fonética. El alineador se encarga de segmentar la grabación y atribuir a cada segmento la parte correspondiente de las transcripciones.

Además, se ha incorporado un analizador morfológico, que se encarga de hacer las anotaciones morfológicas automáticamente. El resultado que se obtiene con estas herramientas es un archivo de ELAN con los niveles de anotación ortográfico, fonético y morfológico. La persona encargada de revisión debe comprobar la segmentación realizada, corregir las transcripciones y los análisis y validar todo ello.

Estas herramientas están disponibles únicamente para los administradores y no para el público usuario externo.

2.4. Base de datos

El proyecto CORILGA necesita una base de datos que permita refinar la búsqueda de la información lingüística. La base de datos de CORILGA está creada en Microsoft Access, con la extensión .mdb, y consta de siete tablas que recogen toda la información de que se dispone en relación a las grabaciones y a las personas que en ellas intervienen.

De este modo, se anota información acerca de las grabaciones como son: **tipo de texto** (ya mencionados en §2.2.), **hábitat** (rural, urbano o semiurbano), **procedencia del corpus**, **lugar de grabación** (lugar, parroquia, *concello* y provincia), **fecha** (día, mes y año), **contexto**, **minutos y segundos de la grabación**, **minutos y segundos transcritos**, **responsable(s) de la grabación**, **responsable(s) de la transcripción**, **responsable(s) de la revisión** y **notas**.

En la tabla de hablantes se almacena la información relativa a las personas que intervienen en la grabación, como es: **sexo** (hombre o mujer), **edad** (años en el momento en que se realizó la grabación), **tramo de edad** (TI1, de 0-14 años; TI2, de 15-24 años; TI3, de 25- 49 años; TI4, de 50-69 años; y TI5, 70 años o más), **profesión**, **nivel de estudios** (sin estudios, primarios, secundarios o universitarios), **lugar de nacimiento** (lugar, parroquia y *concello*), **lugar de residencia** (lugar, parroquia y *concello*), **lengua inicial** (gallego, castellano o ambas), **lengua de la grabación** (gallego, castellano o ambas) y **observaciones**.

En las otras tablas se contemplan los **temas** de que tratan las grabaciones, que son un número cerrado de ellos: 1. Historias de vida (infancia, enamoramiento, matrimonio, muerte...), 2. Lengua y comentarios sociolingüísticos, 3. Literatura, 4. Historia, filosofía y antropología, 5. Otras ciencias (conocimientos muy específicos), 6. Trabajos y productos del mar y del río, 7. Trabajos y productos del agro, 8. Otros oficios (tradicionales o modernos), 9. Política y actualidad, 10. Deportes, juegos y ocio, 11. Religión, mitos y creencias, 12. Música tradicional y cantares, 13. Fiestas y gastronomía.

Las tablas de la base de datos también se emplean para mostrar al personal de CORILGA información sobre el corpus. Es el caso de la tabla denominada ***Gravacións_temas***, que muestra el número de grabaciones que se han transcrito de cada uno de los temas, así como el tiempo de transcripción que hay realizado de cada uno de ellos.

2.5. Gestor web y búsquedas

El gestor web fue desarrollado por el grupo AtlanTIC de la Escola de Enxeñería de Telecomunicación de la Universidade de Vigo. Este grupo es el que se encarga de la puesta a punto y el buen funcionamiento del buscador, en base a las exigencias y necesidades del objeto para el que se está realizando este corpus.

Los archivos de audio y las transcripciones, así como la base de datos, están alojados en ese gestor web. A través de una página de búsqueda creada para CORILGA, se puede ir seleccionando la información que se precise tanto de las

grabaciones como de las personas que intervienen en ellas (§2.4.) antes de realizar la búsqueda. Esta página web guía a la usuaria o usuario a través de tres pasos que actúan a modo de filtro para la selección de esa información, como se ve en las imágenes 1 a 3.

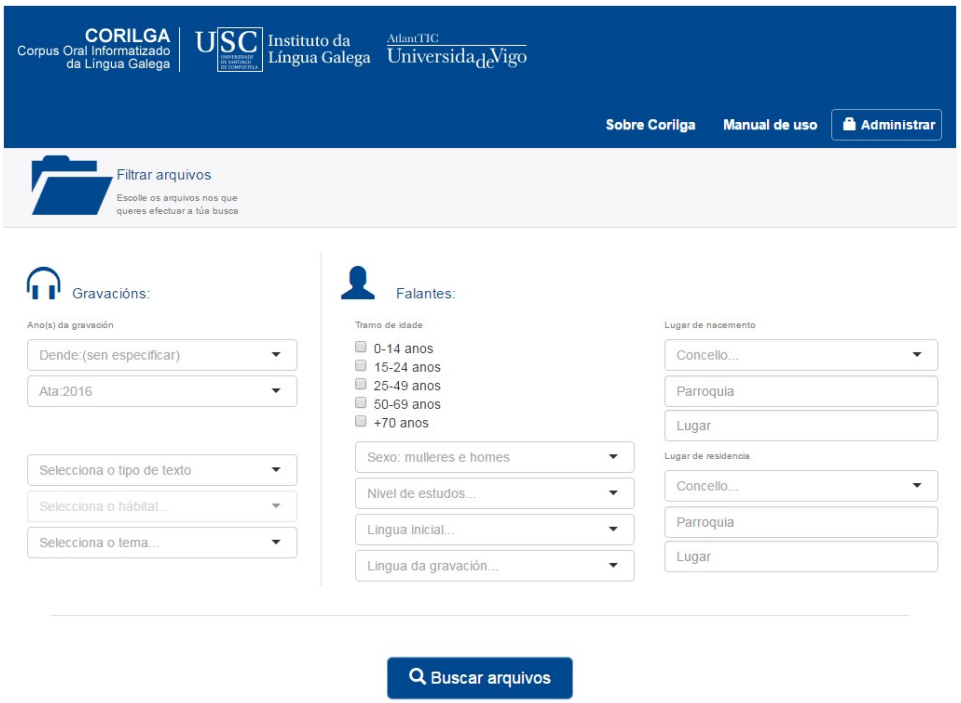

Imagen 1: selección de la información de grabaciones y hablantes.

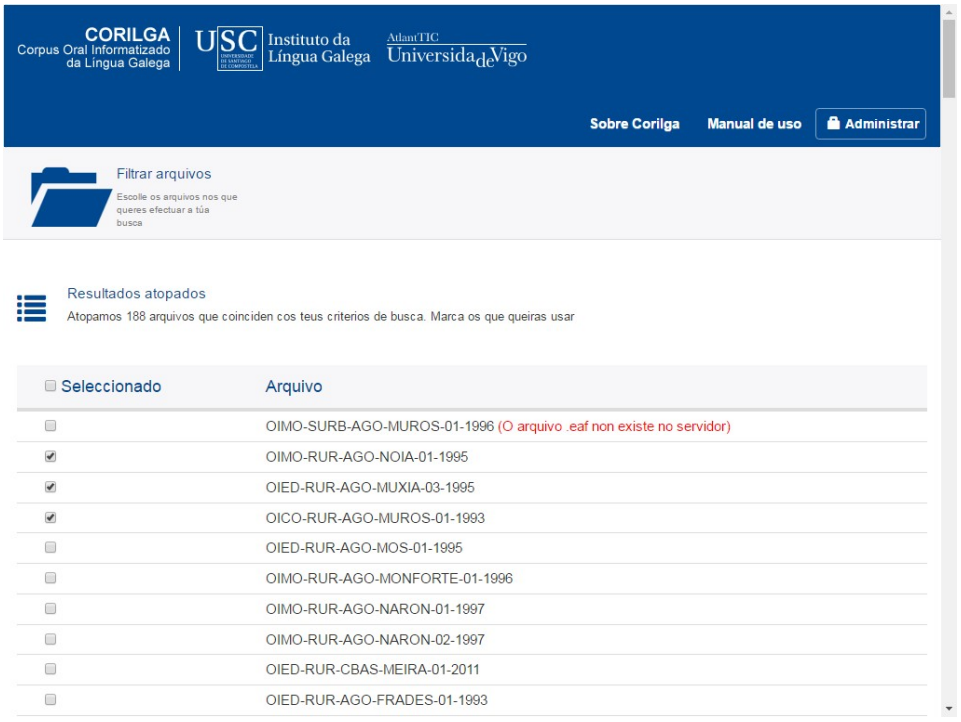

Imagen 2: selección de los archivos en los que se quiere hacer la búsqueda

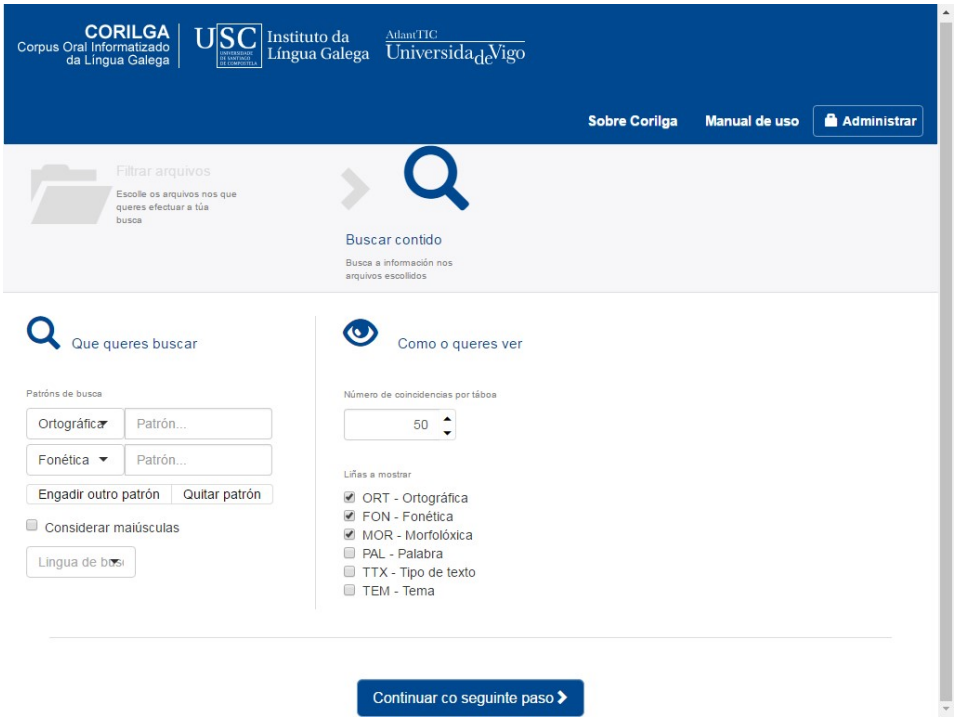

Imagen 3: selección de las líneas de anotación y búsqueda del ítem lingüístico.

En la imagen 4 se puede ver el resultado de una búsqueda en CORILGA, en el cual aparecen las líneas de transcripción ortográfica, fonética y morfológica alineadas con el audio, que se puede escuchar al mismo tiempo que se ve el resultado.

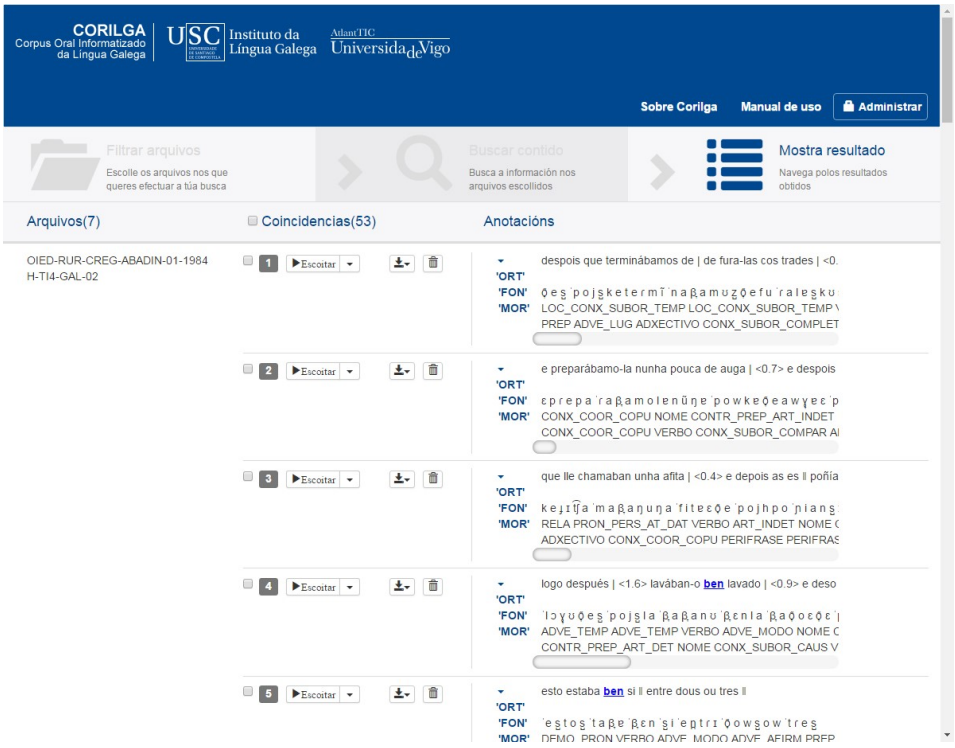

Imagen 4: Resultado de la búsqueda de "ben".

Además, con el fin de ampliar contexto, el audio puede extenderse desde 5 a 20 segundos. De este modo, se facilita la comprensión de algunos casos que puedan ser dudosos. Asimismo cuenta con un botón para eliminar los resultados que no interesen (ejemplo: buscamos *que* en función de nexo y aparece uno interrogativo; ese resultado lo podemos eliminar).

Las casillas situadas al lado de los resultados permiten seleccionar aquellos que interesen para su posterior descarga. Las descargas de CORILGA pueden realizarse en EXCEL, ELAN y PRAAT, dependiendo del interés del estudio.

3. Posibilidades de utilización

CORILGA nació con la intención de tener un corpus de lengua oral gallega en todos sus usos. El objetivo principal que se persigue con él es que las personas investigadoras de la variación y cambio lingüístico del gallego no tengan que crear pequeños corpus particulares cada vez que quieran realizar un estudio. Gracias a las posibilidades de filtrado de búsquedas, CORILGA se puede utilizar en investigaciones en el ámbito de la variación en tiempo aparente (distintas generaciones en un momento concreto), así como en tiempo real (mismas generaciones a lo largo del tiempo, al haber sido posible incorporar material de audio desde los años 60 hasta la actualidad).

Otra de las posibilidades que ofrece por su composición es hacer estudios comparativos tanto en materia de usos habituales como de contacto de variedades. Es decir, hasta qué punto influyen las variedades geolingüísticas en la variedad estándar y viceversa. Esto es posible porque se han introducido no solo materiales de lengua informal, sino también de lengua formal e incluso de los medios de comunicación.

Puede también ser útil para estudios de sociolingüística y de interferencia entre el gallego y el castellano gracias a las anotaciones de **lengua inicial** y **lengua de grabación**. Teniendo en cuenta estas dos variables, se pueden hacer estudios de dominio bilingüe, de influencia de una lengua sobre otra. De esos estudios puede derivar la planificación y normalización lingüística, entre otros fines posibles.

Este corpus ha sido un elemento fundamental para el desarrollo de tecnologías del habla como un reconocedor de voz y un alineador audio-texto. Pero un corpus oral transcrito y anotado del estilo de CORILGA es un potencial elemento de desarrollo de otras muchas, tal que convertidores texto-voz, voz-texto; reconocimiento y diálogo; etc.

4. Estado actual y proyecto de futuro

En la actualidad, CORILGA cuenta con 203 grabaciones y 85 horas transcritas. El trabajo que estamos desarrollando en estos momentos es, básicamente, el de ampliar el número de transcripciones, tanto de las ya existentes desde los años 60, como de grabaciones actuales que se están incorporando a este proyecto. Del mismo modo, se están adaptando algunas otras ya hechas para otros proyectos y tesis a los criterios de CORILGA y al programa ELAN.

Todas las transcripciones ya incorporadas se están sometiendo a revisión por otros miembros del proyecto. En muchos casos, también se están ampliando los niveles de anotación en aquellos casos en que no poseen todos (fonético y morfológico). El

proceso de revisión no solo se realiza archivo por archivo, sino que, periódicamente, se realizan búsquedas aleatorias en el repositorio web con el fin de identificar posibles errores puntuales que se hayan escapado a una primera inspección, así como de comprobar el buen funcionamiento del mismo.

Visto el funcionamiento de CORILGA y los criterios seguidos a la hora de realizar las transcripciones, parece obvio que, en un futuro, lo que se pretende es seguir ampliando el corpus en cuanto a variedades y registros, es decir, continuar equilibrándolo en los tres grupos en que se divide.

Con 85 horas de transcripción, el corpus es todavía pequeño y precisa de mucho más material y más diverso, así como una revisión en todos los niveles. Por tanto, los planes de futuro para este corpus son continuar ampliando las transcripciones y revisando las ya hechas, así como aumentar los niveles de anotación incluyendo el sintáctico, prosódico y uno con la lematización de todas las formas.

Las herramientas de transcripción y alineado todavía necesitan perfeccionarse. Se nutren de las transcripciones ya realizadas manualmente y de las realizadas automáticamente (tras haber sido revisadas y validadas) y, precisamente por ello, la ampliación del corpus contribuye a que estas herramientas se vayan adiestrando y trabajando cada vez mejor. A pesar de todo, todavía queda mucho por mejorar y por avanzar en este aspecto.

De cara a un futuro próximo, pretendemos que todo el corpus esté revisado por, como mínimo, una persona. Además, se está trabajando en pulir errores de procesado de las búsquedas y de la visualización de los datos en la interfaz. Esto se pretende que esté listo antes de la presentación de este proyecto al público, prevista para finales del presente año.

5. Conclusiones

Ante la necesidad de un corpus oral para el gallego y con la abundancia de grabaciones reunidas en el Instituto da Lingua Galega, se proyectó la creación de CORILGA. A medida que se perfilaba la idea, se fue ampliando el número de grabaciones y el tipo de registro de las mismas hasta un total de 203, con una duración total de 85 horas.

El avance que estamos viviendo en la informática nos da facilidades para que este corpus sea cada día mayor. Además, contribuye a facilitar no solo el almacenamiento y búsqueda de datos en las transcripciones, sino también a automatizar y agilizar la transcripción mediante herramientas construidas para tal fin.

Como se trata de un proyecto en construcción y todavía incipiente, queda mucho por trabajar en él y por mejorar antes de que llegue a ser un corpus como el que está ideado, con una gran cantidad de textos de todo tipo (variedades, registros, etc.) completamente transcritos y anotados. Con todo, estaría listo para usar en un breve período de tiempo, siempre teniendo en cuenta su limitado volumen y que no se ha revisado totalmente.

Bibliografía

AGO = FERNÁNDEZ REI, F. (dir.) (2010-): *Arquivo do Galego Oral*. Santiago de Compostela: Instituto da Lingua Galega. http://ilg.usc.es/ago/ [Consultado: <08/03/2016>]

AMPER-Galicia = FERNÁNDEZ REI, E. (dir.) (2006-): *Atlas Multimedia Prosódico do Espazo Románico*. Santiago de Compostela: Instituto da Lingua Galega. http://ilg.usc.es/amper/ [Consultado: <08/03/2016>]

BRUGMAN, H. y A. RUSSEL (2004): "Annotating Multimedia/ Multi-modal resources with ELAN". *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. https://tla.mpi.nl/tools/tla-tools/elan/citing_elan/ [Consultado: <14/03/2015>]

CORILGA = REGUEIRA FERNÁNDEZ, X. L. (dir.) (2012-): *Corpus Oral Informatizado da Lingua Galega*. Santiago de Compostela: Instituto da Lingua Galega. http://ilg.usc.es/corilga/ [Consultado: <08/03/2016>]

DUBERT GARCÍA, F. X. (1998): *A fala de Santiago de Compostela. Estudio xeolingüístico*. Tesis doctoral dirigida por Rosario Álvarez Blanco. Universidade de Santiago de Compostela (inédita)

GARCÍA MATEO, C., A. CARDENAL LÓPEZ, X. L. REGUEIRA FERNÁNDEZ, E. FERNÁNDEZ REI, M. MARTÍNEZ MAQUIEIRA, R. SEARA DOPAZO, R. VARELA FERNÁNDEZ & N. BASANTA LLANES (2014): "CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis". *Proc. 9th Language Resurces and Evaluation Conference (LREC2014)*. Reykjavik, 26-31 May 2014.

KUCERA, H. & N. FRANCIS (1967): *Computational Analysis of Present-Day American English*. Michigan: Brown University Press

REGUEIRA FERNÁNDEZ, X. L. (1989): *A fala do norte da Terra Cha: estudio descritivo*. Tesis doctoral dirigida por Antón L. Santamarina Fernández. Universidade de Santiago de Compostela (inédita).

SANKOFF, D. & G. SANKOFF (1973): "Sample Survey Methods and Computer-Assisted Analysis in the Study of Grammatical Variation". *Canadian Languages in Their Social Context Edmonton: Linguistic Research Incorporated*. R. DARNELL (ed.). Edmonton: Linguistic Research Incorporated, pp. 7–64.

SANTAMARINA FERNÁNDEZ, A. L., R. ÁLVAREZ BLANCO, F. FERNÁNDEZ REI & M. GONZÁLEZ GONZÁLEZ (1990-2015): *Atlas Lingüístico Galego*. Vol. I-VI. A Coruña: Fundación Pedro Barrié de la Maza.