En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

CODIFICACIÓN Y ANOTACIÓN DEL HABLA EN UN CONTEXTO BILINGÜE: EL CORPUS ESLORA DE ESPAÑOL DE GALICIA*

Speech codification and annotation in a bilingual context: The ESLORA corpus of Galician Spanish

VICTORIA VÁZQUEZ ROZAS
Universidade de Santiago de Compostela

MARIO BARCALA
NLPgo Technologies, S.L.

EVA DOMÍNGUEZ NOYA
Centro Ramón Piñeiro para a Investigación en Humanidades, Universidade de Santiago de Compostela

ALBA FERNÁNDEZ SANMARTÍN
Universidade de Santiago de Compostela

GUILLERMO ROJO
Universidade de Santiago de Compostela

MARÍA PAULA SANTALLA
Universidade de Santiago de Compostela

* El corpus ESLORA fue financiado por el Ministerio de Economía y Competitividad a través de los proyectos de investigación ESLORA (FFI2010-17417) y ESLORA2 (FFI2014-52287-P), y actualmente por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER) a través del proyecto ESLORA+ (FFI2017-86379-P). El equipo del proyecto forma parte del grupo de investigación Gramática del español de la Universidad de Santiago de Compostela, beneficiario de una ayuda para «Consolidación e estruturación de Grupos con Potencial de Crecemento 2017» de la Consellería de Cultura, Educación e Ordenación Universitaria de la Xunta de Galicia (ED431B 2017/39). El equipo de ESLORA está asimismo integrado en la Red Temática en Estudios de Análisis del Discurso, financiada por el Ministerio de Ciencia, Innovación e Universidades a través de la AEI (FFI2017-90738-REDT).

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

Resumen
El artículo ofrece en primer lugar una caracterización general del diseño y composición del corpus ESLORA y muestra su utilidad para el análisis de la variación social y situacional. El corpus supone asimismo una aportación al estudio de los procesos de cambio ligados a la variación geográfica del español, puesto que registra su uso en un territorio con lengua propia, lo que facilita el reconocimiento del español hablado en Galicia como objeto de investigación de la dialectología hispánica y permite su comparación en pie de igualdad con otras variedades geográficas. En el núcleo del trabajo se describen algunas de las dificultades que surgen en la transcripción, codificación y anotación de registros de habla en un contexto de contacto lingüístico y se exponen los argumentos que sustentan las soluciones adoptadas.

***Palabras clave*:** corpus oral, español, variación, transcripción, anotación

Abstract
This article provides an overview of the design and composition of the corpus ESLORA and shows its usefulness in analysing social and situational variation. The corpus also contributes to the study of the processes of change related to the geographical variation of Spanish, since it records its use in a region with its own distinctive language. This aspect facilitates the recognition of the Spanish spoken in Galicia as a research object in the field of Hispanic dialectology, allowing for its comparison with other geographical varieties on an equal footing. The main focus of the paper is on some difficulties encountered during transcription, codification, and annotation of spoken recordings as well as with the arguments that justify the solutions taken by the research team.

***Keywords*:** spoken corpus, Spanish, variation, transcription, annotation

1. INTRODUCCIÓN

El desarrollo de corpus orales en español se inició hace más de medio siglo con el *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades de Iberoamérica y de la Península Ibérica* (*cfr.* Lope Blanch 1986), una empresa colectiva que supuso un cambio radical en la consideración de los registros de habla como base empírica de la investigación lingüística. Superando limitaciones técnicas, materiales y humanas, el proyecto logró con creces sus objetivos, pues no solo amplió el conocimiento de la realidad del español en América y en España sino que abrió nuevas perspectivas teóricas y metodológicas para el estudio de la variación en los diferentes niveles de análisis.

La propuesta de la *Norma culta* recogía algunos parámetros de variación que se han incorporado en buena parte de los corpus orales del español desarrollados desde entonces. Además de la variable geográfica, que permitió abordar estudios comparativos, los materiales documentan de forma equilibrada el habla de hombres y mujeres de diferentes franjas etarias mediante muestras de discurso formal e informal. Por otra parte, frente a la investigación dialectal tradicional, basada en el uso lingüístico de hablantes rurales de edad avanzada y con una escolarización limitada, la *Norma culta* se centró en los hablantes urbanos «cultos». No se contemplaba, por tanto, la variable estrática, común en la investigación variacionista, que sí se incluyó en el diseño de otros corpus orales elaborados posteriormente.

Frente al interés creciente por el registro del español hablado, sea con propósitos sociolingüísticos (PRESEEA, por ejemplo) o dialectológicos (COSER)[1], se echa en falta

[1] Las referencias completas a los corpus y otros recursos electrónicos mencionados en el texto se encuentran al final

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

una mayor atención al uso oral del español en comunidades bilingües. Ciñéndonos a la situación en España, el retraso en documentar adecuadamente el español hablado en Galicia, Cataluña, Valencia o el País Vasco es, en parte, el resultado del sesgo normativo que ha dominado en los estudios sobre estas variedades y, en parte, la consecuencia de su exclusión como objetos de estudio con entidad propia en la dialectología tradicional. No obstante, en los últimos años algunas iniciativas han empezado a llenar ese vacío (Vila Pujol 2001, Sinner 2001, Vann 2009, Gómez Molina 2001, Briz y grupo Val.Es.Co 2002, Paasch-Kaiser 2015, además de los equipos correspondientes del proyecto PRESEEA, así como diversos estudios basados en materiales procedentes de COSER, entre otros, Fernández Ordóñez 2007, De Benito 2015, Camus & Gómez Seibane 2015, Gómez Seibane 2015).

En este contexto se hizo patente la necesidad de documentar el uso oral del español en Galicia, ya que el estudio de esta variedad se reducía prácticamente a la identificación y caracterización de sus peculiaridades («galleguismos»), casi siempre a partir de su empleo en obras literarias o textos periodísticos de autores gallegos y observaciones aisladas de la lengua hablada (*cfr*. Rabanal 1967, García 1986, Acín 1996, Rojo 2004, Fernández-Ordóñez 2016)[2]. La creación de un corpus de habla permitiría ofrecer una base empírica adecuada para fundamentar las descripciones de una variedad apenas investigada y facilitar su comparación con otras variedades, al tiempo que reivindicaba su legitimidad como objeto de estudio lingüístico más allá de consideraciones prescriptivas.

El diseño de un corpus oral y los procesos de registro, codificación y tratamiento de los materiales para construir un recurso útil para el análisis lingüístico traen consigo múltiples decisiones teóricas y metodológicas, que en el corpus ESLORA están asimismo condicionadas por el objetivo de documentar el uso del español en un contexto bilingüe. A las dificultades inherentes al trabajo con muestras orales se añade, pues, el hecho diferencial de elaborar un corpus del español hablado en Galicia, un territorio con lengua propia.

En el presente trabajo se exponen las principales características de ESLORA, desde su diseño hasta las múltiples posibilidades de consulta y explotación que ofrece (*cfr*. también Barcala *et alii* 2018). En sus diferentes apartados se describen y justifican los distintos aspectos y fases de construcción del corpus, pero se desarrollan y fundamentan con más detalle aquellos fenómenos que presentan un interés especial en el ámbito de la variación lingüística. Se destaca así la aportación imprescindible de los corpus de variedades distintas de la estándar a la investigación lingüística en general y a la lingüística de corpus en particular.

El contenido del capítulo está estructurado de la siguiente manera. En la sección 2 se describen la composición y características generales del corpus. En la sección 3 se discuten los aspectos más relevantes de la construcción, codificación, anotación y consulta del recurso. El apartado 3.1 está dedicado al proceso de construcción (obtención del consentimiento informado, grabación, herramientas de transcripción y alineación, y anonimización); en el 3.2 se exponen y justifican las particularidades del sistema de codificación ortográfica adoptado; el 3.3. se centra en el sistema de anotación, con especial atención a la etiquetación morfosintáctica; y en el 3.4 se detallan las características y opciones que ofrece la aplicación de consulta en línea. Por último, la sección 4 aborda aspectos cuantitativos del corpus y ofrece un ejemplo ilustrativo de la

del trabajo.

[2] Los trabajos de Celia Pollán (2001, 2002) constituyen una excepción a esta tendencia, puesto que están basados en el análisis del contenido de un corpus de lengua oral, el de la lengua hablada en A Coruña (*cfr*. Fernández Rodríguez no publicado).

rentabilidad del corpus para el estudio de las frecuencias léxicas. El capítulo se cierra con un apartado final de síntesis y proyección hacia el trabajo futuro.

2. EL CORPUS ESLORA

El objetivo general del proyecto ESLORA es poner a disposición de investigadores y personas interesadas un corpus de español hablado en Galicia en las condiciones más adecuadas para su uso y explotación en la investigación lingüística y especialmente en el estudio de la variación. La finalidad del corpus ha guiado su diseño y características, que se manifiestan en tres aspectos nucleares: (i) la composición del corpus, (ii) su condición de recurso de acceso abierto, (iii) su enriquecimiento mediante herramientas de PLN.

(i) El corpus está formado por 54 entrevistas semidirigidas, de una hora aproximadamente, y 20 horas de conversación espontánea. Las entrevistas corresponden a 27 mujeres y 27 hombres divididos en tres grupos de edad (de 19 a 34 años, de 35 a 54 y de 55 o más años) y tres niveles de estudios (primarios, medios y universitarios). En la versión accesible actualmente en internet (1.2.2, de noviembre de 2018) están disponibles 53 entrevistas y 3 conversaciones, que suman un total de 647 758 formas ortográficas (medida poco adecuada para un corpus de este tipo) y 776 260 elementos gramaticales[3].

Los informantes de la muestra de entrevistas residen en Santiago de Compostela y su entorno, y en su mayoría son originarios de la ciudad y su área de influencia. Los participantes en las conversaciones proceden de diferentes puntos de Galicia y, en casos contados, de fuera de la comunidad. De todos ellos se recopila la información sociológica básica (mujer/hombre, edad, nivel educativo), que constituye el núcleo de los metadatos. Se registran además el papel comunicativo (entrevistador, informante o audiencia en las entrevistas), la relación o conocimiento previo entre los participantes y las circunstancias, localización y tiempo del encuentro.

Pero para documentar adecuadamente el uso del español por parte de hablantes gallegos no basta con registrar entrevistas y conversaciones y anotar los metadatos habituales para llevar a cabo estudios de corte sociolingüístico. Para alcanzar una imagen más completa de la realidad lingüística reflejada en el corpus, los participantes entrevistados respondieron a un cuestionario detallado sobre sus usos y actitudes lingüísticas en relación al español y al gallego y realizaron un test de inseguridad lingüística. Tanto el cuestionario como el test fueron grabados con el fin de analizar sus ventajas y desventajas metodológicas como instrumentos para el estudio de ideologías y actitudes lingüísticas (*cfr*. Recalde 2012).

(ii) Muchas de las características del corpus ESLORA están relacionadas con el objetivo prioritario de su puesta a disposición pública en condiciones apropiadas para ser realmente útil a las personas interesadas. Para poder ofrecer el acceso abierto a los materiales, los participantes firmaron su consentimiento informado con la condición de preservar su privacidad mediante la anonimización de las transcripciones y los audios, un proceso que se llevó a cabo de forma meticulosa (*cfr*. *infra* apdo. 3.1.4).

La accesibilidad y legibilidad del corpus se plasma también en el tipo de transcripción elegida, basada en la convención ortográfica y con un reducido componente interpretativo. La alineación de los textos transcritos con el sonido permite

[3] Llamamos elementos gramaticales a los que resultan de llevar a cabo los procesos de lematización y análisis morfosintáctico habituales en el procesamiento de corpus. Así, en una forma como *diciéndomelo* se reconoce la existencia de tres elementos (el gerundio *diciendo*, el pronombre *me* y el pronombre lo), en *al* hay dos elementos, mientras que *Instituto Nacional de Estadística* es considerado como una unidad.

la consulta inmediata del fragmento de audio correspondiente a cada búsqueda. Está disponible asimismo la descarga del corpus completo en formato textual y, bajo petición, se pueden obtener las grabaciones completas, el texto etiquetado y los datos sociolingüísticos recogidos en los cuestionarios mencionados. Para facilitar la recuperación de la información contenida en el corpus se diseñó una completa aplicación de consulta en línea con su correspondiente guía de uso, en la que se detallan las amplias posibilidades de búsqueda que ofrece el sistema (*cfr. infra* apdo. 3.4).

(iii) Dado que el corpus fue concebido como un instrumento para el análisis del español oral, el diseño general del recurso y las decisiones específicas de transcripción y codificación estuvieron dirigidas a facilitar el posterior enriquecimiento de los materiales mediante herramientas de PLN. Así, se optó por un sistema ortográfico de representación textual que permitiera la anotación morfosintáctica de las transcripciones con los programas de lematización y etiquetación disponibles, pensados para textos escritos estandarizados. Se evitó, por tanto, el empleo de elementos gráficos convencionales, como subrayados, cursivas, comillas, guiones, etc., y se optó por signos y etiquetas unívocas y procesables por medios automáticos. Otro elemento central en el diseño del corpus fue el empleo de programas de alineación de la transcripción con el audio, que permite el tratamiento conjunto de texto y habla y abre nuevas vías de aprovechamiento y explotación de los materiales.

3. Construcción, codificación y anotación de ESLORA

3.1. El proceso de construcción

3.1.1. Grabaciones

Todas las entrevistas se registraron en archivos WMA con grabadoras digitales Olympus DS-40 con micrófono integrado y sonido estéreo de extra alta calidad (ST XQ). En la mayor parte de los casos, la grabadora fue accionada antes del encuentro con el informante, de manera que quedaron registrados los saludos, presentaciones y, en general, charlas espontáneas previas a la entrevista propiamente dicha. Durante la entrevista la grabadora estaba a la vista, sin obstáculos físicos que pudiesen entorpecer la grabación. En general, se obtuvieron audios de buena calidad y libres de ruido de ambiente, pues las entrevistas se realizaron en casas particulares y, ocasionalmente, en los lugares de trabajo de los participantes.

En el caso de las conversaciones, la gran mayoría se grabó en formato WAV y, más raramente, en MP3, mediante aplicaciones de grabación para dispositivos electrónicos, casi siempre teléfonos móviles. En casi todos los casos se empleó la aplicación *Tape-a-talk Voice Recorder*, disponible para dispositivos Android de manera gratuita. Dicha aplicación permite efectuar grabaciones de buena calidad en formatos mp3, wav o 3gp. En estos casos, la disminución de calidad con respecto a la grabadora profesional se ve compensada por el hecho de que los móviles, al formar parte de la vida cotidiana de todos los participantes, podían situarse en el lugar más conveniente para la grabación sin levantar ningún tipo de sospecha, mientras que la grabadora profesional tenía que permanecer oculta en algún lugar no visible, lo cual podía obstaculizar la grabación.

En cualquier caso, todos los archivos que forman finalmente parte del corpus tienen una calidad de grabación media-alta, ya que fueron descartados todos aquellos que presentaban cualquier tipo de problema que dificultase la transcripción y la posterior escucha por parte del usuario del corpus (ruido de fondo, interferencias…).

3.1.2. Transcripción y alineación

Todas las entrevistas fueron transcritas y alineadas de forma manual empleando el programa Transcriber[4], que proporcionaba un entorno cómodo y rápido para codificar interacciones de dos participantes con un esquema de pregunta-respuesta. Transcriber cuenta además con un sistema de introducción de etiquetas fácilmente personalizable y de manejo sencillo.

Sin embargo, Transcriber presenta serias limitaciones a la hora de transcribir interacciones espontáneas con más de dos participantes. Por poner un ejemplo, la interfaz no permite introducir solapamientos en los turnos de más de dos hablantes. A ello se suma su falta de mantenimiento y actualización en los últimos años, que da lugar a continuos problemas técnicos. Por estos motivos, en la segunda parte del proyecto se decidió transcribir las conversaciones utilizando el programa ELAN[5], que, con un sistema de transcripción por niveles o *tiers* (el denominado sistema «en pentagrama»), y una gran cantidad de posibilidades de personalización y exportación, permite reflejar más fielmente la estructura de la conversación espontánea.

3.1.3. Permisos

El consentimiento informado de los participantes es un requisito legal y ético para el registro de los datos, que adquiere especial relevancia en el caso de las conversaciones, ya que son registradas sin el conocimiento de los participantes en ese momento concreto. Así, mientras que en las entrevistas se consideró suficiente la obtención de un consentimiento verbal anterior, y de uno escrito posterior, en el caso de las conversaciones se obtuvo de los participantes un doble permiso por escrito: uno previo, en el que aceptaban participar en el proyecto y ser grabados en cualquier momento, y otro posterior, en el que daban su autorización para emplear las grabaciones realizadas. En ambos casos, el equipo investigador se comprometía a restringir el uso de las grabaciones y transcripciones a los ámbitos de la investigación y la docencia, así como a la preservación del anonimato de los participantes.

3.1.4. Anonimización

La anonimización de los materiales, además de responder a un compromiso adquirido a través del formulario de consentimiento, constituye una obligación ética del investigador, entre cuyos deberes está evitar a los participantes cualquier tipo de perjuicio derivado de su participación en el proyecto.

El proceso no estuvo exento de complicaciones, ya que los elementos que pueden conducir a la identificación de una persona son múltiples y complejos. De este modo, preservar el derecho a la intimidad de los participantes y su entorno sin comprometer la validez y coherencia de los datos necesitó de toda la atención, buen juicio y en algunos casos ingenio de los transcriptores y revisores.

La anonimización se llevó a cabo sobre lo que la página web del *UK Data Archive*[6] denomina *identificadores directos*, que se clasifican en tres categorías: nombres**,** apellidos y apodos de personas; nombres de lugares (ciudades, pueblos, calles, etc.), y nombres de instituciones, agrupaciones o asociaciones (colegios, partidos políticos, etc.). Puesto que el corpus pone a disposición del investigador tanto la transcripción de las grabaciones como el audio, la anonimización se efectuó también en ese doble plano.

[4] http://trans.sourceforge.net

[5] https://tla.mpi.nl/tools/tla-tools/elan

[6] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

Las opciones a la hora de anonimizar la transcripción son muy variadas. Algunas de las más frecuentes consisten en sustituir los nombres propios por iniciales –opción empleada, por ejemplo, en el *Corpus de habla culta de Salamanca* (CHCS)–, por nombres genéricos, del tipo *topónimo* o *nombre propio* –o simplemente *name*, como en el *British National Corpus* (BNC)–, por símbolos (como en el *Corpus de lengua hablada de la ciudad de A Coruña*, *cfr*. Vázquez Veiga 2003: 9) o incluso por códigos alfanuméricos. En el caso de ESLORA la opción elegida fue la empleada, entre otros, en el *Santa Barbara Corpus of Spoken American English*, consistente en sustituir aquellos nombres que se deseaba anonimizar por otros nombres del mismo tipo. Sin duda, este es el sistema cuya aplicación resulta más compleja, pero presenta una serie de ventajas para el usuario del corpus. Entre otras, permite captar con más facilidad la coherencia interna de las referencias, así como recibir gran parte de la información que los nombres propios originales proporcionan acerca de sus referentes.

En primer lugar, los nombres, apellidos y apodos de personas fueron sustituidos por otros nombres reales, tratando, no solo de que fuesen métricamente equivalentes, sino que también lo fuesen social y culturalmente. Como señala Sampson (2000), en muchos casos los nombres pueden estar asociados a un determinado grupo de edad, clase social o procedencia. En la misma línea, Agha (2007: 65-66) sostiene que los nombres propios pueden proporcionar una gran cantidad de información acerca de sus referentes, concerniente a su género, lugar o circunstancias de nacimiento, afiliación, religión o pertenencia a un grupo. Toda esa información se pierde irremediablemente si los nombres se sustituyen por símbolos, códigos o iniciales.

Por este motivo, en ESLORA se llevó a cabo un proceso de sustitución que no solo respeta la oposición de género, las diferencias entre nombres y apellidos o los diminutivos, sino que también refleja otros aspectos, como por ejemplo, las pistas acerca de las características demográficas de los referentes. Para empezar, dado que se trata de un corpus recogido en Galicia, aparecen en él multitud tanto de nombres (Breixo,[7] Xoán, Catuxa) como de apellidos (Ferreiro, Barreiro, Calviño) típicamente gallegos. En todos los casos se sustituyeron por otros nombres (Antón, Anxo, Sabela) o apellidos (Piñeiro, Cunqueiro, Mariño) que conservan la referencia al origen de los referentes. Del mismo modo se actuó con nombres extranjeros como Richard, sustituidos por equivalentes como por ejemplo Peter.

Además de la información geográfica, algunos nombres también proporcionaban pistas sobre otros aspectos, como la edad de la persona aludida o su grado de cercanía con el emisor. Así, en las grabaciones aparecen nombres muy frecuentes en el pasado pero que no lo son tanto en la actualidad, y que por tanto suelen corresponder a personas de edad avanzada. Es el caso de nombres como Brígida, Casilda, Socorro o Celestino que fueron sustituidos por otros con esa misma característica como Amelia, Luciana, Angustias o Edelmiro. En el caso de los nombres abreviados, como Tito, Chema, Cuca, Concha o Paqui, se reemplazaron por otros que conservan su carácter cercano y familiar, como Fito, Berto, Charo, Espe o Puri. Este proceso de anonimización supuso un laborioso trabajo que en muchos casos sobrepasó los límites de la lingüística y se introdujo de lleno en la sociología. En este sentido, resultó de gran utilidad la página del Instituto Nacional de Estadística, que, en una de sus secciones, ofrece la posibilidad de averiguar la edad

[7] Los ejemplos de nombres sustituidos no corresponden en ningún caso con los nombres reales que aparecen en las grabaciones originales, ya que esto constituiría una infracción del acuerdo de confidencialidad. Se trata de nombres inventados similares a los eliminados. Los nombres que se presentan para ejemplificar las sustituciones sí que forman parte de los que se emplean realmente en el corpus.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

media de las personas que tienen un determinado nombre o apellido, además de su procedencia geográfica[8].

Cabe señalar, para terminar con los nombres de persona, que aquellos pertenecientes a personajes de la esfera pública, como actores, deportistas, políticos, escritores, etc., se excluyeron del proceso de anonimización, siempre que las referencias no atañesen a su vida privada y no resultaran lesivas para su imagen ni atentasen contra su honor (Sampson 2000). En estos casos la conservación de las referencias puede tener interés documental, más allá del ámbito estrictamente lingüístico, ya que reflejan la visión que los distintos informantes tenían de la sociedad en la que vivían en un momento histórico concreto.

Sí se cambiaron algunos nombres de personajes públicos cuando la referencia podía llevar a la identificación del hablante, como en un caso en el que el informante menciona durante la entrevista que ganó un certamen literario cuyo premio consistía en unas clases con un famoso escritor. Su nombre hubo de ser cambiado por el de otra figura literaria, ya que a través del nombre del escritor se podría identificar el premio y por tanto a su ganador.

Otro tipo de identificadores directos son los nombres de ciudades, pueblos, aldeas, etc., muy frecuentes en la mayor parte de las grabaciones. En estos casos, la sustitución no se llevó a cabo de manera sistemática, sino valorando en cada caso cuáles eran las ventajas y los riesgos de conservar la referencia original. En primer lugar, se conservaron las abundantísimas alusiones a la ciudad de Santiago, ya que todos los receptores del corpus saben de antemano que los informantes son naturales de dicha ciudad o residentes en ella. Tampoco se modificaron sistemáticamente los nombres de otras ciudades cuando aparecían en alusiones a viajes, opiniones, descripciones, etc. En síntesis: solamente se cambiaron las referencias que pudiesen conducir a la identificación del informante o de alguna persona de su entorno. Se operó de modo similar con los nombres de barrios, calles, plazas, etc., que solo se modificaron en caso de hacer referencia a la localización exacta del domicilio del hablante o de otra persona aludida en la interacción, así como a su lugar de trabajo o cualquier otro elemento identificador. Y otro tanto con los nombres de colegios, institutos, instituciones, comercios o asociaciones.

En los casos en los que se consideró oportuno modificar los topónimos, se llevó a cabo el mismo procedimiento que con los nombres de personas, tratando de conservar, en la medida de lo posible, la información que cada denominación proporcionaba acerca de su referente. Así, por supuesto, los nombres de barrios se cambiaron por nombres de barrios, los pueblos por pueblos, ciudades por ciudades y así sucesivamente. Pero además, se trató de escoger nombres que compartiesen con los originales sus características más definitorias. Por ejemplo, si el informante mencionaba sus vacaciones en un pueblo típico de veraneo de la costa gallega como Sanxenxo, la sustitución se hizo con otro de características similares como por ejemplo Baiona.

En todos los casos se trató de mantener la coherencia interna que permitiese el seguimiento de la continuidad referencial del discurso, tratando además de resultar verosímil en lo que se refiere a la información geográfica y demás indicaciones contextualizadoras.

Los elementos modificados aparecen en la aplicación de consulta resaltados en color amarillo, lo cual indica al lector que el nombre que figura en la transcripción no es el que aparece en el audio original.

[8] https://www.ine.es/widgets/nombApell/index.shtml.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

Por su parte, la anonimización del audio consistió en introducir un ruido en todos los fragmentos anonimizados en la transcripción, empleando para ello el programa de tratamiento de sonido Audacity[9].

3.2. El sistema de codificación

La creación de una transcripción a partir de una grabación de habla implica la plasmación textual selectiva y parcial de un registro sonoro del evento comunicativo original. De la misma manera, tanto los archivos de audio como las transcripciones de las entrevistas y conversaciones que integran el corpus ESLORA constituyen una selección de datos relevante para el análisis del español de Galicia, pero no dejan de ser representaciones indirectas y fragmentarias del objeto de estudio como tal. Como apuntó Elinor Ochs (1979), las decisiones que el analista adopta en el proceso de transcripción no son neutrales, sino que están condicionadas por sus presupuestos teóricos e ideológicos y por el propósito de su investigación. Duranti (2006: 308) afirma que «transcription is a cultural activity used for creating and sustaining a science of the particular slice of the universe that interests us (…)». Son por tanto esos intereses los que orientan el proceso de registro y representación del habla.

El objetivo de documentar y analizar el uso oral del español en Galicia ha guiado el diseño y construcción del corpus ESLORA, tanto en la composición de la muestra –selección de hablantes y géneros discursivos–, como en la configuración de las transcripciones y de los metadatos recogidos a través de cuestionarios durante su elaboración.

Para la transcripción de las grabaciones se optó por un único nivel de representación que sigue la ortografía convencional excepto en dos aspectos. Uno de ellos es la puntuación, limitada a los signos de interrogación y admiración. Se prescinde, por tanto, de puntos, comas, etc., dado que su uso tiene una función convencional de estructuración de enunciados y organización textual propia de la escritura, mientras que en la codificación del habla interesa dejar constancia de las pausas, que se clasifican dependiendo de su duración en pausa breve, pausa larga y silencio.

El segundo aspecto diferencial es el empleo de mayúsculas, que se reduce a los nombres propios puesto que no hay mayúsculas dependientes de la puntuación. Como la distinción entre nombre común y nombre propio es problemática en ciertos casos, para garantizar un tratamiento homogéneo de las mayúsculas se adoptó como criterio básico de transcripción la norma del *Diccionario panhispánico de dudas* (DPD), que era el texto de carácter normativo más reciente de la Real Academia Española en el momento en que arrancó el proyecto. El DPD no disipa sin embargo todas las dudas, pues admite ambas posibilidades, por ejemplo, para los nombres de marcas comerciales, dependiendo de si la referencia es más o menos específica, una distinción no siempre determinable en las muestras del corpus.[10] En todo caso, para solventar la dificultad planteada por alguna discordancia que subsista en el corpus, en el sistema de consulta se ofrece también la opción de recuperar las formas con independencia del uso de mayúsculas o minúsculas (*cfr. infra* apdo. 3.4)

En relación con las mayúsculas, cabe indicar que se han usado en el término gallego *Rúa* ‘calle’, presente en denominaciones como *Rúa Nueva* (gal. *Rúa Nova*), *Rúa del*

[9] https://www.audacityteam.org/.

[10] «Las marcas comerciales son nombres propios, de forma que, utilizados específicamente para referirse a un producto de la marca, han de escribirse con mayúscula: *Me gusta tanto el Cinzano como el Martini*; *Me he comprado un Seat*; pero cuando estos nombres pasan a referirse no exclusivamente a un objeto de la marca en cuestión, sino a cualquier otro con características similares, se escriben con minúscula: *Me aficioné al martini seco en mis años de estudiante* (al vermú seco, de cualquier marca)» (DPD, *s.v. mayúsculas*, apdo. 4.22.).

Villar (gal. *Rúa do Vilar*), porque se considera parte del nombre de la calle, a diferencia de los sustantivos españoles *calle* y *avenida*, que se han tratado como comunes siguiendo el DPD. Por otra parte, en la escritura de los topónimos se reproduce la elección de cada participante, sea la forma original gallega (*Ourense, Vilagarcía, Conxo*, etc.), sea la adaptación al español (*Orense, Villagarcía, Conjo*).

Frente a una posible representación fonética, el uso de un sistema ortográfico aligera el trabajo de transcripción, permite abordar la construcción de un corpus más amplio y tiene ventajas de legibilidad y comparabilidad; además, la escritura convencional facilita la búsqueda y recuperación de información y posibilita la aplicación de herramientas estándar de PLN para la anotación textual (*vid.* Poplack 1989: 430; 1993: 265-266; Edwards 2001: 324; Torres Cacoullos & Travis 2018: 46-47). No obstante, la codificación ortográfica puede verse también como un obstáculo para la documentación de la variación, especialmente en un corpus que tiene entre sus objetivos visibilizar usos que se alejan de los considerados normativos.

La cuestión de la representación escrita de las variantes lingüísticas no canónicas ha sido objeto de debate al menos desde que Dennis Preston llamó la atención sobre el efecto de los *respellings* utilizados en la tradición folclorista. Preston (1982, 1985, 2000) muestra cómo la «escritura dialectal» (*eye dialect*) estigmatiza a los hablantes representados, sobre todo si, como suele ser el caso, la codificación de las variantes menos prestigiosas no afecta por igual a todos los participantes (*cfr.* también Bucholtz 2000; Jaffe & Walton 2000; Beal 2005). La alternativa, si se requiere la codificación detallada de las particularidades del habla, es recurrir a un sistema de transcripción fonética como el *Alfabeto Fonético Internacional*. En ESLORA, al igual que en otros corpus de construcción reciente, el acceso inmediato al audio alineado deja abierta la posibilidad de añadir a la representación ortográfica una línea o nivel de representación fonética (un nuevo *tier* en el sistema de ELAN). No obstante, la puesta a disposición libre de los datos sonoros puede no ser la opción más adecuada en otros casos. Torres-Cacoullos & Travis (2018: 48) advierten de las interpretaciones erróneas a que pueden dar lugar los registros de audio si los investigadores no están familiarizados previamente con el uso lingüístico representado, además de las consecuencias no deseables que derivan del reforzamiento de estereotipos negativos hacia una comunidad que habla una variedad no estándar.

Menos problemática para la imagen de los hablantes, pero tampoco exenta de dificultades, es la representación textual de otros fenómenos característicos del habla para los que no disponemos una convención gráfica fija, como los alargamientos vocálicos y consonánticos, las palabras truncadas o la risa. El fragmento de (1) muestra varias etiquetas XML de alargamientos, palabras cortadas, énfasis, pausa corta y pausa larga, marcas que si bien enriquecen la transcripción y permiten refinar las opciones de recuperación automática de información, presentan el inconveniente de impedir la lectura fluida del texto. En la aplicación de consulta del corpus ESLORA se ha resuelto este problema de visualización mediante indicaciones que se activan al pasar el cursor por los elementos afectados (*cfr. infra*, apdo. 3.4).

(1) <alargamiento>pues</alargamiento><pausa/> eeh mi madre había <alargamiento>sido</alargamiento> <palabra_cortada>me</palabra_cortada> eh mecanógrafa <pausa/> y yo tenía nociones <alargamiento>de</alargamiento> <pausa_larga/> <palabra_cortada>taquin</palabra_cortada> de <énfasis_inicio/>mecanografía<énfasis_fin/> (SCOM_M22_034)

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

El fragmento incluye también la representación de las vocalizaciones *ehh* y *eh*, que manifiestan vacilación y funcionan como prolongadores. En esta categoría se incluyen elementos interjectivos, unidades denominadas «cuasi-léxicas», «pausas sonoras», «apoyos vocálicos», que en muchos casos no poseen una forma escrita estandarizada, aunque sí se representan de forma convencional con más o menos variantes en algunos tipos de textos, como en las viñetas de *cómics* y novelas gráficas. Algunos protocolos de transcripción restringen las opciones de representación ortográfica a unas pocas formas preestablecidas. Por ejemplo, Tagliamonte (2004: 5-6) admite solo tres «hesitation words» (*um, ah, oh*) y establece una lista cerrada de «legal fillers», cada uno con su correspondiente significado.

En el ámbito del español, el macroproyecto PRESEEA también propone codificaciones convencionales para unidades como *ah, ay, aha, mmm, eeh, pff, bah*, y para la representación de onomatopeyas (*zas, bum, plas*), aunque no limita explícitamente el repertorio de variantes posibles[11]. Prueba de la convencionalidad de algunas de esas expresiones es su reconocimiento como entradas de diccionario. El DLE, por ejemplo, incluye *ah, ay, bah, zas, bum, plas*, pero no *mmm, ehh, pff y aha* (aunque sí *ajá*).

Una de las tareas actualmente en curso en el proyecto ESLORA tiene como objetivo ofrecer una codificación consistente y homogénea de tal tipo de unidades, que por una parte represente la variedad y particularidades de su forma y uso en las muestras del corpus y que por otra parte introduzca un cierto nivel de abstracción y generalización que permita contar con un conjunto limitado de opciones. Teniendo en cuenta que el acceso al audio alineado es inmediato y está a disposición de los estudiosos para análisis más delicados, el nivel básico de transcripción ortográfica debe acotar la dispersión que contiene la versión actual, con variantes como *eh, ee, eeh, ehh, eeeh*, o como *buf, buff, bufff, buuf*, entre otras.

Sin embargo, la unificación de la forma escrita de interjecciones y vocalizaciones requiere especial cautela, ya que la estandarización ortográfica tiende a favorecer la representación y el reconocimiento de las formas propias de variedades asociadas con el estándar y en cambio puede ignorar las formas y usos propios del habla considerada «dialectal». En ESLORA se documentan formas propias, y quizá exclusivas, del español de Galicia (y del gallego), que deben ser identificadas ortográficamente. Véase, por ejemplo el uso de *ho* en (2), que no debe confundirse con la interjección *oh* recogida en (3):

(2) lo malo que hay eh pue pue <pausa/> nos pues bueno nosotros con el repertorio que tenemos ya nos llega *ho* <pausa/> (SCOM_H31_046)

(3) en esa excursión fue como muy divertido ¿ sabes ? en plan <pausa/> españoles se encuentran a un español <pausa/> famoso y vas en plan *oh* no sé qué ¿ sabes ? <pausa/> fue como muy <risa/> <pausa/> muy gracioso sí <pausa/> (SCOM_M13_008)

En (4) se observa la forma *boh*, que hay que distinguir de *bah*, ejemplificada en (5):

[11] http://preseea.linguas.net/Portals/0/Metodologia/Marcas_etiquetas_minimas_obligatorias_1_2.pdf.

(4) otro de mis hermanos <pausa/> que hizo Químicas <pausa/> salió de casa <pausa/> eeh todo convencido de que iba a hacer Medicina <pausa/> y iba ya a matricularse en Medicina <pausa/> dio una vuelta por la Alameda <pausa/> porque era temprano y no sé qué <pausa/> y se encontró a unos mmm amigos de él <pausa/> que <pausa/> que habían hecho el bachillerato <pausa/> iban para Químicas <pausa/> y dijo ¡ *boh* ! pues mira <pausa/> voy para Químicas <pausa/> y se marchó a Químicas <pausa/> (SCOM_M33_005)

(5) (…) cuando veo que son las ocho de la tarde <pausa/> o las nueve que ya llevo doce horas conduciendo <pausa/> digo *bah* pues ahora el próximo hotel que vea <pausa_larga/> paro (SCOM_M23_018)

Otra interjección registrada es *mima* (6), probablemente una abreviación de *mi madriña*, expresión ponderativa también presente en el corpus con un valor similar a *madre mía* o *mi madre* (7).

(6) (…) buff la Navidad ¡*mima* ! yo pongo siempre ya pues el árbol el belén <pausa/> pongo ahí todo el misterio <pausa/> que lo hice yo de manualidades (SCOM_M23_004)

(7) no hay peor cosa que te idealicen una cosa <pausa/> después llegas allí <pausa/> mira <pausa_larga/> (…) llevé una decepción tan grande tan grande tan grande que digo yo *mi madriña* ¿quién me mandaría a mí venir aquí? (SCOM_M31_045)

Se ha hecho referencia a las ventajas de la ortografía estandarizada que prescinde de variantes fónicas al representar los registros orales. Con ella se facilita la búsqueda y recuperación de las unidades lingüísticas y el enriquecimiento de los materiales al aplicarles los mismos recursos de lematización y anotación morfosintáctica disponibles para textos escritos. Sin embargo, de forma contradictoria, el objetivo de documentar adecuadamente los hechos de variación léxica y morfológica justifican la opción de reflejar las opciones de los hablantes que no coinciden con las convenciones normativas (*cfr*. Preston 1982: 323; Tagliamonte 2004; Nagy & Sharma 2013: 240; Torres Cacoullos & Travis 2018: 46-47). La decisión no está exenta de problemas, ya que, como señala Tagliamonte (2004: 2), los límites entre el componente fónico y el morfológico no siempre están claros: «interpreting what is phonological and what is morphological variation is often not clear-cut!».

En la línea de los autores citados, en ESLORA, por una parte, se unifica la transcripción de las variantes de pronunciación manteniendo la ortografía establecida, lo cual implica, por ejemplo, que se escriba *entonces*, *es decir*, *para* o *para adelante*, independientemente de que estas y otras expresiones se reduzcan con frecuencia en la pronunciación. Un caso especial es el de los lapsus de dicción, que se representan tal como se han pronunciado, pero se marcan con la etiqueta <sic> para evitar que se interpreten como errores de transcripción:

(8) mmm tampoco fff siento una gran <sic>actración</sic> <pausa/> atracción por caer en eso ¡eh! <pausa/> realmente <pausa_larga/> (SCOM_H11_052)

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

Pero, por otra parte, se adopta el criterio de no estandarizar las variantes morfológicas empleadas por los hablantes. La consecuencia es que en el corpus se representen formas no canónicas como *estes* (9) y *eses* (10) para los demostrativos de plural.

(9) Al Aquasella por ejemplo llevo yendo todos los años desde que tengo diecisiete <pausa_larga/> y solo tengo fotos de *estes* dos últimos años <pausa/> (SCOM_H11_047)

(10) es diferente <pausa/> en Cuba no tanto pero bueno <pausa/> Venezuela <pausa/> Colombia ya no te cuento y todos *eses* países sí sí <pausa/> (SCOM_M12_020)

También se registran formas verbales no estándares para la segunda persona de singular del pretérito de indicativo (*dijistes, estuvistes, hicistes,* etc.) y para el presente de subjuntivo del verbo estar (*estea*, *esteamos*). La identificación de estos casos permite su incorporación al sistema de formas reconocibles por el etiquetador morfosintáctico. Retomaremos esta cuestión en el apdo. 3.3.

Requiere asimismo un tratamiento específico la sufijación apreciativa en -iño, que se da con sustantivos y adjetivos tanto de lema español (*hijiño*) como gallego (*filliño*). En el apdo. 0 se detalla la casuística del fenómeno y se explican las soluciones de anotación adoptadas.

Como es de esperar, se documentan en el corpus palabras originalmente gallegas integradas en secuencias en español sin que quepa hablar propiamente de cambio de código sino más bien de un repertorio léxico compartido por ambas variedades. Así ocurre con formas como *colo* 'regazo o disposición adoptada por los brazos para sostener a un niño' (11), *cacharela* 'hoguera' (12), *latar* 'hacer novillos' (13), entre otras.

(11) si tengo que ir a algún sitio <pausa/> hombre a lo mejor un día me apetece ir de compras <pausa/> y no me la llevo <pausa/> porque sé que a los diez minutos me dice mamá *colo* <pausa/> mamá <pausa/> no entres no pruebes más <risa/> me dice no ahí a esa tienda no <pausa/> más comercios no <pausa_larga/> (SCOM_M12_030)

(12) hacemos ahí la *cacharela* y montamos una cantina <pausa_larga/> (...) eeh regalamos las sardinas y el pan <pausa_larga/> (SCOM_M23_001)

(13) yo pues sí eh con los diecisiete años en el instituto pues sí las hacías <pausa/> *latabas* <pausa/> te ibas para el bar <pausa/> (SCOM_H12_027)

Lo mismo ocurre con palabras de otras lenguas, sobre todo del inglés, formas como *hall, freelance, realities, jazz, funky*, etc., que están integradas en el uso habitual de diferentes variedades del español (y del gallego) y para las que tampoco se ha considerado conveniente utilizar una etiqueta de lengua en la transcripción.

Por el contrario, sí se marca el cambio de código cuando la alternancia entre español y gallego (u otra lengua) va más allá de una sola palabra, tanto si representa la sucesión entre dos hablantes, con frecuencia en discurso referido (14), como si el cambio se produce en el interior de un enunciado de un hablante (15).

(14) *<gl> e logo <pausa_larga/> non vamos dar una voltiña?</gl>* <pausa/> no mamá porque Sonia no está <pausa/> y podemos caer <pausa_larga/> (SCOM_M31_045)

(15) tienes que estar así todo el día <pausa/> o sea no <pausa/> pues yo cojo y yo me largo para casa vamos *<gl>e non traballo</gl>* (SCOM_H21_039)

3.3. El sistema de anotación

El corpus ESLORA ha sido morfosintácticamente anotado y lematizado al modo habitual en este tipo de recursos, de modo que cada elemento del corpus es adscrito a un lema, es asignado a una clase de palabras y se indica el valor de cada una de las subcategorías gramaticales que le son de aplicación. Las dos últimas informaciones se expresan en lo que se conoce habitualmente como «etiqueta», en la que, de modo sintético, se sitúa en primer lugar la clave que corresponde a la clase de palabras[12] y luego se van añadiendo los símbolos que se refieren a las subcategorías. Así, por ejemplo, en el caso de una forma verbal, la etiqueta VIP1S indica que se trata de una forma de un verbo (V) perteneciente al modo indicativo (I), el tiempo presente (P), primera persona (1) del singular (S). Es decir, el número y el carácter de las clases de palabras, subclases y categorías gramaticales identificados en el sistema de anotación de ESLORA está plenamente en línea con las recomendaciones presentes en los estándares de anotación morfosintáctica (EAGLES en primer lugar) y en la práctica más común de anotación de corpus en español.

El sistema es peculiar, sin embargo, en un aspecto de su desarrollo: para las palabras que, en una determinada subcategoría, puedan presentar más de un valor *x* o *y* en razón del contexto, hay una etiqueta que les atribuye el valor *x*, otra que les asigna el valor *y,* y una tercera que las identifica como con uno de los dos valores *x* o *y*. Cuando el contexto considerado permite identificarlo, a la palabra en cuestión se le asignará en el corpus una de las etiquetas con valores *x* o *y*. Cuando el contexto considerado no lo permite, a la palabra en cuestión se le asignará la etiqueta que se refiere a la indeterminación entre *x* o *y* para la categoría implicada. Este planteamiento, obviamente, extiende el número de etiquetas. Así, un adjetivo como *interesante* tendrá, por causa del género, al menos 5 etiquetas posibles, puesto que, dependiendo del contexto, puede ser masculino [*un libro interesante*], femenino [una exposición interesante], neutro [*Eso es interesante*], masculino o femenino [*Cualquier artista interesante*] y masculino, femenino o neutro [*Resulta interesante*]. El coste que supone el aumento de etiquetas se compensa, sin embargo, con las ventajas que esta aproximación produce en el proceso automático de etiquetación.

Como resultado de todo lo anterior, el sistema de anotación aplicado a ESLORA consta de 455 etiquetas morfosintácticas distintas: 198 de pronombres, 136 de determinantes, 78 de verbos, 15 de sustantivos, 13 de adjetivos, 1 de adverbio, 1 de preposición, 1 de conjunción, 1 de interjección y 1 de puntuación. Las 8 etiquetas restantes sirven para elementos tales como fechas, símbolos o cifras, elementos que, en realidad, no se esperan en el corpus, bien por su carácter oral (es el caso de los símbolos), bien por haber sido transcritos ortográficamente (caso de las fechas o las cifras).

Para anotar el corpus se ha utilizado el etiquetador XIADA, el cual, aunque en principio diseñado para el gallego en el *Centro Ramón Piñeiro para a investigación en humanidades*, se ha podido adaptar con facilidad para etiquetar español en el marco del proyecto ESLORA. Si bien se trata de un etiquetador fundamentalmente estadístico,

[12] En nuestro caso, adjetivo, adverbio, conjunción, determinante, interjección, preposición, pronombre, sustantivo y verbo.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

permite también la introducción de reglas lingüísticas que contribuyen a incrementar sensiblemente su nivel de acierto. Aparte de por sus buenos resultados, comprobados para el gallego (*cfr*. Domínguez *et alii* 2009), el etiquetador XIADA fue el elegido para la anotación de ESLORA porque era un etiquetador que se podía adaptar de manera mucho más simple que el resto de los considerados (Freeling en primer lugar) para la gestión simultánea de la etiquetación morfosintáctica y las marcas de oralidad empleadas en el corpus. Todo ello ha tenido como resultado que el producto final, la aplicación web de explotación del corpus, pueda manejar y proporcionar ambos tipos de información, etiquetación y marcas de oralidad, de manera mucho más eficaz y rentable desde el punto de vista del usuario.

Con la herramienta XIADA, la anotación del corpus se ha hecho entonces de manera automática, aunque la parte del mismo que luego sirvió para entrenar el sistema automático implicado (50 000 formas aproximadamente), ha sido también revisada a mano. Para desarrollar ese corpus de entrenamiento, creamos en primer lugar un minicorpus con el tamaño estrictamente necesario para que cada etiqueta del sistema de anotación estuviera representada al menos una vez. Ese minicorpus incluía secuencias reales tomadas del corpus siempre que era posible, y secuencias inventadas en caso contrario. Con ese recurso se etiquetó automáticamente un corpus de 25 000 palabras de entrevistas semidirigidas. Este corpus se revisó manualmente y se utilizó después para etiquetar automáticamente otro corpus de 25 000 palabras de conversaciones, corpus que también fue después revisado manualmente. Para la revisión manual, en cada una de sus fases, los revisores trabajaban con un editor XML personalizado[13] en el que se veían todas las etiquetas posibles de cada elemento (no solo las seleccionadas por el etiquetador), de manera que se podía, de forma cómoda, aprobar la etiqueta propuesta por el sistema, cambiarla por alguna de las rechazadas o incluso proponer una nueva.

Veamos ahora algunos de los problemas surgidos en este proceso, tanto manual como automático, de aplicación de las etiquetas al corpus, en concreto los que tienen que ver con la situación de contacto de lenguas en que se encuentran español y gallego en Galicia. En comparación con los que aparecen en cualquier proceso de anotación automática, se entrecruzan aquí dos tipos de problemas adicionales. Por una parte, están todos aquellos que son propios de la lengua oral y también los que, aunque puedan estar presentes en la lengua escrita, se presentan en la oralidad de un modo diferente y, por tanto, requieren un tratamiento distinto. En este punto, conviene tener en cuenta que la tradición descriptiva suele dar soluciones para los textos escritos, pero no para los orales. Por otra parte, aparecen aquí los problemas causados por las características específicas de una variedad del español, el español de Galicia, surgida, directa o indirectamente, como consecuencia del contacto entre español y gallego[14]. Los rasgos propios de esta variedad pueden, en efecto, requerir actuaciones sobre los recursos utilizados para la etiquetación (el lexicón o módulos de reconocimiento específicos, por ejemplo) o bien exigir la toma de decisiones que afectan a la relación de las etiquetas con las formas y comportamientos gramaticales a los que se refieren (casos como *medio* o formas verbales en *-ra*, *vid. infra*).

En el primer grupo, se encuentran, por ejemplo, los problemas que plantea la frecuente formación de los diminutivos a partir de la forma que es propia del gallego, el sufijo *-iño*. En su forma más simple, se trata del empleo de este sufijo sobre una base del español estándar:

[13] XMLmind XML Editor, URL: http://www.xmlmind.com/xmleditor/.

[14] Las características propias del español de Galicia no son un tema suficientemente tenido en cuenta en el estudio de las variedades del español, tal como señala Rojo (2004: 1091 y ss.). De esta presentación nos servimos en este apartado para la organización de buena parte de los problemas de etiquetación aquí tratados.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

(16) eh <pausa/> y comemos <pausa/> a la vuelta comimos en el <pausa/> en el *barquiño* que que íbamos <pausa_larga/> (SCOM_H31_042)

Complicando un poco el problema, la forma en diminutivo es construida ya no a partir de una palabra del español, sino del gallego (*grolo* 'sorbo'), una palabra cuyo uso puede estar más o menos normalizado en el español de Galicia:

(17) si no tomo ahora un *groliño* de café estoy medio (SCOM_M13_008)

A veces, incluso, encontramos en el corpus las formas diminutivas en *-iño* asociadas tanto a una palabra en español como a su equivalente de traducción en gallego:

(18) y yo *filliño* <pausa/> además lo comprara en Portugal por nada por <pausa/> por nada (SCOM_M31_045)

(19) y yo le dije a mi marido dije oy *hijiño* no mejor es no ir porque pft (SCOM_M33_009)

Y, en ocasiones, la palabra en cuestión es también propia del español común, pero su diminutivo solo es posible en la variedad usada en Galicia y tiene además con esa forma un significado muy específico ('amable, entrañable' en el caso de *riquiño*):

(20) ¿sabes? <pausa/> y estás en el reservado hasta las cuatro de la mañana <pausa/> que eso es lo bueno que tienen <pausa/> en el Carretas que no tienen pro ¿sabes? <pausa/> tienen mm <pausa/> son muy *riquiños* en ese sentido (SCOM_M13_008)

A veces también sucede que la forma en diminutivo se constituye en (parte de) una unidad discursiva lexicalizada:

(21) y mi *madriña* <pausa/> las fiestas de Ribeira colega pero pfff (SCOM_H12_027)

En nuestro conjunto de etiquetas no está previsto que las formas apreciativas, ya sean diminutivos u otras modalidades, dejen ningún rastro específico: un sustantivo diminutivo es etiquetado exactamente igual que la forma no diminutiva de la que procede. Pero, naturalmente, el proceso de anotación requiere que ese diminutivo sea reconocido como tal y reciba la etiqueta que le corresponde. Ese objetivo puede ser alcanzado mediante dos caminos distintos. Algunos diminutivos se reconocen como tales a partir del diccionario al que acude la herramienta de etiquetación y otros por medio de un módulo de asignación de etiquetas a formas desconocidas sobre la base del contexto (este módulo se está desarrollando actualmente para la reconstrucción del lema original y su contraste con el lexicón). Ambas estrategias pueden utilizarse para la anotación de los diminutivos formados a partir del sufijo gallego *-iño*. Por supuesto, el problema de que el lema reconstruido no se halle en el diccionario siempre existe[15], tanto si la forma base

[15] Hay diminutivos previsibles, por ser de uso recurrente en español de Galicia, pero hay otros de aparición aislada. En la versión actual de ESLORA hemos hallado, en diferentes géneros y números, los siguientes: *airiño, barquiño, besiño, bueniño, cabeciña, casiña, cojiño, escueliña, filliño, gentiña, groliño, grupiño, hambriña, hijiño, hombriño, madriña, maliño, mujeriña, neniño, normaliño, pachuchiño, pesadiño, pitufiño, pobriño, poquiño, puertiña, riquiño, santiño, tranquiliño, viejiño*.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

es española como si es gallega, pero, con una herramienta diseñada para el español estándar, es obvio que el problema se incrementa enormemente en el segundo caso. De hecho, consideramos perfectamente razonable la posibilidad de enriquecer las herramientas de etiquetación para diferentes variedades de una lengua (en situación de contacto con otras o no) con cuantos más y más detallados recursos para cada una de ellas. Con independencia de ello, resulta además muy conveniente la identificación inequívoca en los sistemas de etiquetación de los recursos específicos utilizados para el análisis de características propias de una variedad, de modo que sea posible activarlos o desactivarlos según sea aconsejable en cada caso. Entre estos recursos se encontrarán, sin duda, los que puedan servir para la identificación y anotación correcta de las formas diminutivas en *-iño* en el español hablado en Galicia.

Sin que podamos entrar en detalles al respecto de cada una de ellas, de naturaleza similar son los problemas que para la etiquetación plantea la presencia en español de Galicia de palabras o unidades fraseológicas de distinto tipo procedentes del gallego. Dado que es preciso, por supuesto, asignarles la etiqueta morfosintáctica que les corresponde, es necesario habilitar los recursos para ello. Hay que decidir qué palabras y unidades fraseológicas deben entrar en los recursos generales (el diccionario), identificadas como formas gallegas propias del español de Galicia, y qué otras formas, de uso esporádico o fortuito, han de ser reconocidas bien por el módulo de asignación de etiquetas a palabras desconocidas, bien mediante un diccionario de la lengua en contacto. Con algunos ejemplos de ESLORA, parece claro que el uso de la forma *estes* como demostrativo plural masculino en lugar de *estos* se encuentra en el primer grupo (para empezar es un determinante, con una función como palabra más gramatical que léxica), así como el uso de las formas *dea*, *estea* para los presentes de subjuntivos de los verbos *dar* y *estar*:

(22) ¿sabes? <pausa/> y cosas así ¿sabes? son <pausa/> no en plan <pausa/> como a los <pausa/> tipos *estes* que <pausa/> acaban locos y ven y escuchan cosas en plan <pausa/> ¿qué te decían las voces? ¿sabes? porque yo tuve un colega que le decían unas voces que matase a su madre (SCOM_H11_047)

(23) ya esta semana ya la estás preparando a las nueve para que *estea* <pausa_larga/> durmiendo para irla acostumbrando ya (SCOM_M12_030)

Sin embargo, las cosas dejan de estar tan claras si pensamos en integrar en el diccionario palabras como *home, neno, alá, alí, bico, pola, mallados*, etc., que también encontramos en ESLORA. Con más facilidad, creemos, se tiende a la integración de las unidades fraseológicas que, siendo calco de las gallegas correspondientes o incorporándolas tal cual, se utilizan en español de Galicia, y ello porque las unidades fraseológicas, al fin y al cabo, lo son, en buena medida, por el uso recurrente que se hace de ellas: *de aquella* (gallego *daquela* 'por aquel entonces')[16]*, si cuadra* (gallego *se cadra* 'tal vez', *cfr*. Rodríguez Espiñeira 2019)*, y más* (gallego *e mais* 'y, y sin embargo')*, de carallada* ('de juerga, de broma'), etc. son algunas de las que hemos encontrado en nuestro corpus.

(24) bueno fui al cuartel y tal <pausa/> hice la mili que *de aquella* había que hacer la mili (SCOM_H21_039)

[16] Esta forma no es exclusiva del español de Galicia, pues se encuentra también en las variedades de influencia asturleonesa. *Cfr*., entre otros, Le Men (2002-2012: *s.v. aquel*).

(25) supuestamente hay un santo <pausa/> enterrado <pausa/> para quien crea *si cuadra* es un cerdo no lo sé (SCOM_H13_012)

(26) crisis evidentemente hay <pausa/> yo la noto <pausa/> *y más* no me quitaron <pausa/> a mi marido no le quitaron <pausa/> nada de la nómina (SCOM_M11_040)

(27) dijo no no no te lo tomes *de de carallada* (SCOM_H21_053)

En el segundo tipo de problemas, los de más calado y concernientes a la relación entre etiquetas, formas y comportamientos gramaticales para los que estas están previstas, podemos mencionar, en primer lugar, la utilización de *medio/a/os/as* concordado con un adjetivo a continuación, allí donde en español estándar se utilizaría la forma invariable *medio*[17]. Así ocurre en (28) y (29):

(28) pero después así que eran ya *medios* adolescentes (SCOM_M33_005)

(29) aunque sean así <pausa/> *medias* así <pausa/> *medias* raras (SCOM_M22_019)

En español estándar la etiqueta que correspondería, a nuestro modo de ver, a *medio* como forma invariable en esa circunstancia sería la de adverbio (W). Pero, obviamente, en español de Galicia la forma que aparece, que es variable, no puede ser etiquetada de esa manera. Aun teniendo a nuestra disposición las etiquetas previstas en el sistema para *medio* como determinante o pronombre partitivo (DP?? o PP??)[18], no creímos, por diferentes razones, que su aplicación fuera apropiada en el contexto de los ejemplos (28) y (29), por lo que decidimos la introducción de una nueva etiqueta, la de adjetivo (A??), para este ítem en tales circunstancias. Respondimos, pues, en este caso, al problema planteado en español de Galicia por una asociación no prevista por nuestro sistema de anotación para español estándar entre una palabra y un comportamiento gramatical concreto creando esa asociación y etiquetándola, y eso lo hicimos porque tal asociación solo afectaba a un ítem léxico (era, por lo tanto, económico crearla) y porque la etiqueta correspondiente y su aplicación no se prestaban a discusión o duda.

Frente a lo decidido respecto a *medio*, se optó por la solución contraria en la etiquetación de las formas verbales en *-ra* (*amara*), que en la variedad que nos ocupa asumen valores modotemporales que en español estándar están ligados a otras formas, pero que en gallego son propios de las formas en *-ra* (aparte de los valores de anterioridad al origen en subjuntivo, los de anterioridad al origen o a una referencia anterior al origen en indicativo, Rojo & Vázquez Rozas 2014). Así ocurre con la forma *comprara* en (18), o con otras formas en *-ra* en los ejemplos (30) y (31).

(30) y estábamos Laura y yo y <pausa/> además *fuera* <pausa/> *fuera* curioso porque <pausa/> <ruido tipo="chasquido boca"/> <pausa/> eran eeh mmm mogollón de ellos (SCOM_H21_039)

[17] Esta utilización de *medio* concordado con los adjetivos a los que acompaña no es, de todos modos, exclusiva del español de Galicia. Se da también en otras variedades del español (*vid*. NGLE, § 19.4k y ss.). Cabe, pues, considerar que no sea un rasgo del español de Galicia debido al contacto de lenguas o que al menos solo resulte un rasgo reforzado por ese hecho.

[18] En la posición de los signos de interrogación aparecen los caracteres referentes al género y número correspondiente.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

(31) y el otro <pausa/> eh este *estudiara* <pausa_larga/> este mmm <pausa/> ¿ cómo se llama ? <pausa_larga/> de Empresariales (SCOM_H31_042)

Se decidió en estos casos que se etiquetaría siempre de acuerdo estrictamente con la forma, asociada al lugar que se asigna habitualmente en el paradigma verbal a las formas en *-ra* en español estándar europeo: pretérito de subjuntivo [19]. Es decir, adoptamos aquí una solución contraria a la que hemos descrito para *medio* y no añadimos una etiqueta adicional para una asociación de forma y comportamiento gramatical que no es propia del español estándar y que, por tanto, no teníamos prevista. Hay que tener en cuenta que estaríamos hablando en este caso de una etiqueta que habría de ser añadida para todos los verbos. Y, sobre todo, que, aunque procedente de la situación de contacto con el gallego, no dejaría de ser esta en español de Galicia una etiquetación no de formas, sino de valores de las formas verbales, tarea que podría, con toda probabilidad, dar lugar a discusión, duda, discrepancia e inconsistencia ya en la anotación manual y, desde luego, tarea muy difícil de llevar a cabo en una anotación automática.

Finalmente, hay que aludir a aspectos que solo suponen una cierta diferencia entre la consideración que merece un fenómeno en español de Galicia y la que recibe en otras variedades del español, como sucede con los casos de segunda persona de singular del pretérito de indicativo terminadas en *-s*. Estas formas, que aparecen con cierta frecuencia en los registros orales de otras variedades del español[20], surgen también en el español de Galicia quizá aquí reforzadas por la influencia del gallego, donde terminan regularmente en *-s*[21].

(32) eso nunca lo *vistes* jugar ¿no? (SCOM_H21_053)

(33) *estuvistes* pendiente de alguien <pausa/> bueno yo y todos <pausa/> más o menos toda mi familia pero <pausa/> a cada uno le afecta a su manera (SCOM_M11_040)

Con más urgencia que para la etiquetación de español, se hace entonces necesario en el español de Galicia habilitar algún recurso o estrategia para la etiquetación automática de estas formas. De momento, la etiquetación que se les atribuye en ESLORA no incluye ninguna marca adicional con respecto a las etiquetas que reciben las formas estándares sin *-s* final.

3.4. La aplicación de consulta

Para facilitar la recuperación de información que ofrece ESLORA se ha diseñado una aplicación de consulta que se organiza en varias pestañas, de las cuales las más destacables son *Información, Guía, Descargas* y *Búsquedas*, esta última el núcleo de la aplicación.

En *Información* se proporcionan tanto los datos globales del corpus (número de versión, fecha de esta, número de documentos y número de palabras ortográficas o

[19] Que no es el único posible: los verbos modales conservan usos indicativos de las formas en *-ra* en español estándar peninsular (*debiera* equivalente a *debería*, *quisiera* a *querría* y semejantes). También son indicativos los usos del tipo *el libro que publicara en 1998* (*cfr.* Rojo 2011b).

[20] Como en el caso de *medio* (*vid*. nota 17), podría dudarse de que sea un fenómeno atribuible o siquiera reforzado por la situación de contacto.

[21] En ESLORA hasta ahora se han identificado en total 28 casos de 17 formas distintas: *abaratastes, anduvistes, dijistes, empezastes, enterastes, estuvistes, fuistes, hicistes, llamastes, nacistes, oístes, pasastes, pensastes, pusistes, salistes, tomastes, vivistes.*

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

elementos gramaticales que constituyen el corpus), como el número de palabras ortográficas y elementos gramaticales que posee el corpus en función de cada uno de los parámetros de clasificación establecidos (grupo de edad, papel del hablante, sexo, nivel de estudios y tipo de interacción).

Por su parte, *Guía* ofrece una descripción detallada del sistema de consultas, así como la relación de marcas aplicadas en la transcripción, que se reflejan en los resultados mostrando las palabras afectadas por la marca en cuestión sobre fondo amarillo. Las marcas empleadas son las siguientes:

Alargamiento: aumento de cantidad que afecta a algún sonido de la palabra marcada.

Cita: el fragmento resaltado reproduce estilo directo.

Énfasis: señala casos de pronunciación especialmente acentuada.

Ficticio: el nombre ha sido cambiado para preservar el anonimato de los hablantes.

Lengua [nombre=xx]: el fragmento está en una lengua diferente al español; xx puede ser *gl*: gallego, *en*: inglés, *pt*: portugués, *it*: italiano, *fr*: francés, *el*: griego.

Palabra_cortada: el segmento marcado representa un fragmento de una palabra.

Risa: marca un segmento en que se ríe un hablante.

Sic: señala errores de dicción para que no se interpreten como errores de transcripción.

Sigla: indica que una cierta forma es una sigla.

Dado que el corpus se enriqueció con su etiquetado morfosintáctico automático, como ya hemos visto en el apdo. 3.3, en *Guía* se ofrece asimismo el etiquetario que se emplea en una descripción que, organizada por clases de palabra, muestra cada etiqueta, su significado y un ejemplo de uso.

A su vez, la sección *Descargas* facilita la descarga del corpus en formato textual y proporciona un formulario para, previa justificación, disponer del corpus en formato ya etiquetado, los audios correspondientes o la información sociolingüística de los hablantes.

El sistema de consultas, al que se accede a través de la pestaña *Búsquedas,* constituye el núcleo de la aplicación y presenta el aspecto que se observa en la figura 1:

FIGURA 1. Pantalla inicial de captación de datos

Grosso modo, el sistema de consultas permite:

i. realizar consultas combinando variables sociales con variables léxicas y gramaticales;

ii. acceder a los fragmentos de audio que corresponden al resultado de la consulta;

iii. descargar el resultado de la consulta en formato TSV (*Tab Separated Values*).

Las consultas pueden realizarse sobre la totalidad del corpus, opción por defecto que muestra la figura 1, o bien sobre uno de los tipos específicos de interacción incluidos bajo el ítem *Corpus*: entrevistas o conversaciones. Además, es posible referir las búsquedas a la totalidad del texto, que es la opción habitual, o bien, mediante la selección de una de las alternativas presentes en el bloque *Buscar en*, centrarlas en fragmentos que han recibido cierta etiquetación (entre risas, con marca de énfasis o alargamiento, reproducción de una cita o identificación de siglas).

Asimismo, el sistema ofrece la posibilidad de combinar libremente valores de las diferentes variables sociales empleadas en la clasificación del corpus, de modo que el usuario crea subcorpus virtuales a la medida. Los parámetros clasificatorios de los que dispone y sus valores son los siguientes:

i. Edad del hablante: Cualquiera, 19-34, 35-54, >54, Desconocida.

ii. Papel o rol en la comunicación: Cualquiera, Audiencia, Entrevistador, Informante.

iii. Sexo: Cualquiera, Hombre, Mujer.

iv. Nivel de estudios: Cualesquiera, Universitarios, Medios, Primarios, Desconocidos.

A su vez, la búsqueda textual se organiza en el ítem *Tipo* del bloque *Búsqueda* por palabras ortográficas (*ir, del, yéndome*) o por elementos gramaticales (*voy, de* –incluyendo los casos de la contracción *del*–, *yendo* –incluyendo los casos del tipo *yéndome*, *yéndonos*–, etc.). La primera modalidad restringe las consultas a una cadena

formalmente coincidente con la grafía, lo que resulta de utilidad en consultas léxicas pero limita en grado sumo las opciones de búsqueda cuando interesa trabajar con información gramatical. La segunda modalidad permite dar un salto cualitativo en la recuperación de información y posterior análisis gramatical al poder introducir en los parámetros de consulta información de carácter gramatical, no solo ya la forma léxica, sino también lemas (todas las formas del verbo *ir*), clases de palabras (verbo, por ejemplo) o valores de las subcategorías gramaticales aplicables en cada caso (persona, por ejemplo, en el caso de los posesivos y verbos).

La importancia que adquieren las búsquedas gramaticales en el análisis de la lengua queda patente en el modo de obtener los datos para estudiar el influjo del gallego en la perífrasis *ir + infinitivo*. La consulta por palabras ortográficas es inútil en este caso debido, por un lado, a la variación e irregularidad del verbo auxiliar (*iba*, *íbamos*, *ir*, *vayamos*, *yendo*, *yéndome*, *váyase*, *irte*, etc.) y, por otro, debido a la ingente relación de infinitivos que pueden aparecer como auxiliado (*hacer, jugar, trabajar, ver*...). Sin embargo, el etiquetado morfológico automático del corpus y la implementación en la aplicación de consultas de las variables relativas a elemento gramatical, etiqueta y lema, combinables entre sí, y combinables asimismo con las variables sociales, no lo olvidemos, junto con la posibilidad de trabajar con hasta cinco elementos sucesivos, hacen que baste realizar la búsqueda que muestra la figura siguiente: lema *ir* en la línea correspondiente al primer elemento y etiqueta *VN*,* equivalente a verbo en infinitivo para el segundo elemento concurrente.

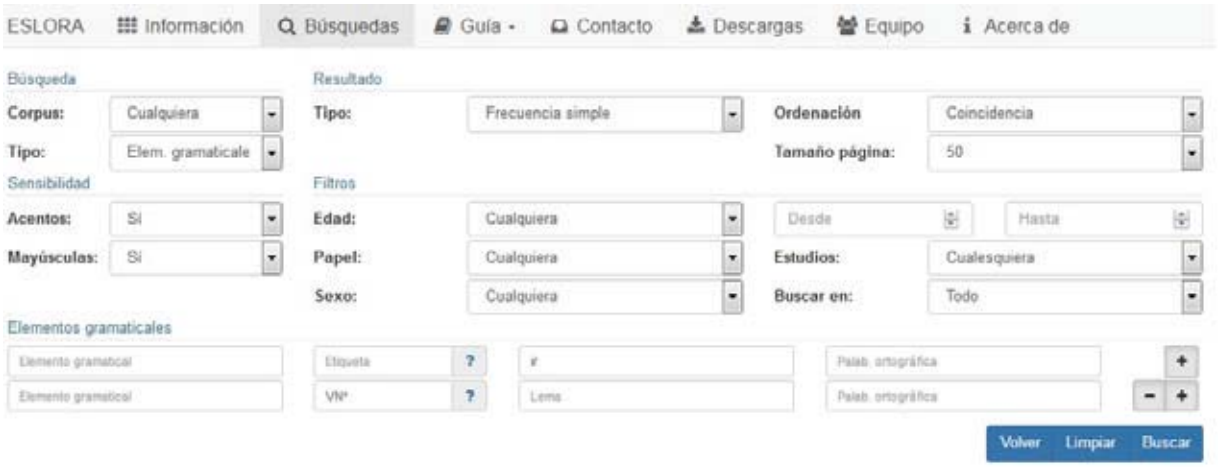


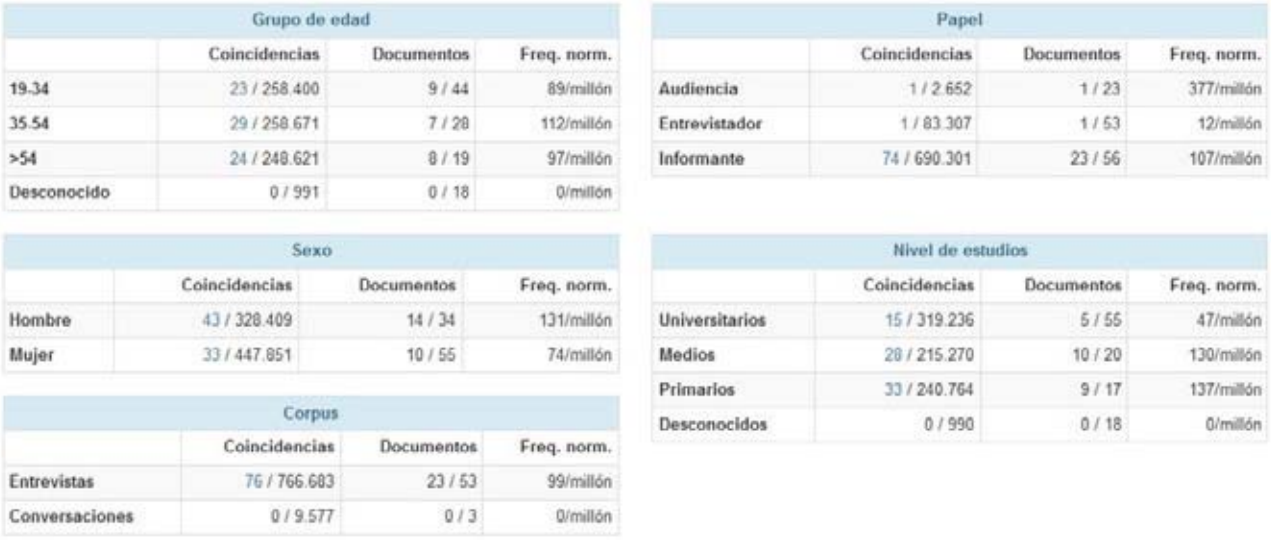
FIGURA 2. Pantalla de captación de datos gramaticales

En cuestión de segundos, modificando la opción *Tipo* del bloque *Resultado*, se obtienen:

1) los datos relativos a su frecuencia simple (figura 3):

FIGURA 3. Información sobre la frecuencia simple

2) las frecuencias completas según los distintos parámetros de clasificación del corpus (figura 4):



Grupo de edad	Coincidencias	Documentos	Freq. norm.
19-34	23 / 258.400	9 / 44	89/millón
35-54	29 / 258.671	7 / 28	112/millón
>54	24 / 248.621	8 / 19	97/millón
Desconocido	0 / 991	0 / 18	0/millón

Papel	Coincidencias	Documentos	Freq. norm.
Audiencia	1 / 2.652	1 / 23	377/millón
Entrevistador	1 / 83.307	1 / 53	12/millón
Informante	74 / 690.301	23 / 56	107/millón

Sexo	Coincidencias	Documentos	Freq. norm.
Hombre	43 / 328.409	14 / 34	131/millón
Mujer	33 / 447.851	10 / 55	74/millón

Nivel de estudios	Coincidencias	Documentos	Freq. norm.
Universitarios	15 / 319.236	5 / 55	47/millón
Medios	28 / 215.270	10 / 20	130/millón
Primarios	33 / 240.764	9 / 17	137/millón
Desconocidos	0 / 990	0 / 18	0/millón

Corpus	Coincidencias	Documentos	Freq. norm.
Entrevistas	76 / 766.683	23 / 53	99/millón
Conversaciones	0 / 9.577	0 / 3	0/millón

FIGURA 4. Datos de las frecuencias completas

3) las muestras con todos los casos de la perífrasis (figura 5) en formato KWIC (*Key Word in Context*):

FIGURA 5. Pantalla con algunas de las muestras finales

Además, y no menos importante dado que se trata de un corpus oral, el usuario puede acceder al audio del segmento que se corresponde con el objeto de la búsqueda y comprobar que no está presente la preposición *a*.

Los resultados de las búsquedas incluyen, también, la posibilidad de descarga de las concordancias o la ampliación del contexto con la representación de las etiquetas morfosintácticas, como muestra la figura 6:

FIGURA 6. Visualización de anotación morfosintáctica

Conscientes de la dificultad que entraña conocer el complejo sistema de etiquetación, en la herramienta de consulta de ESLORA hemos habilitado la introducción de la etiqueta a través de un menú intuitivo que va guiando al usuario en la formulación de la búsqueda gramatical. De este modo, como se muestra en la imagen siguiente, el usuario formula su consulta apoyándose en un menú desplegable que a modo de cascada va mostrando los valores posibles, organizados en un primer nivel en clases de palabras y posteriormente en los valores disponibles para las diferentes categorías gramaticales asociadas a cada clase de palabra, siempre en función de las elecciones previamente realizadas:

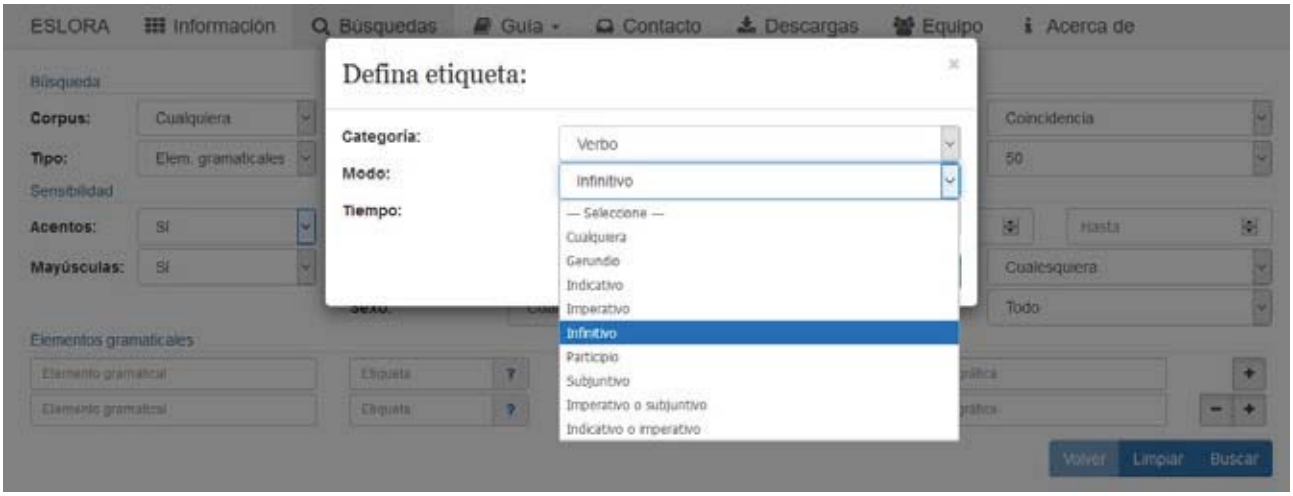


FIGURA 7. Ejemplo de consulta por etiqueta

Asimismo, en la observación de los resultados, ya sea en las muestras ya sea a través del contexto de estas, siempre que esté explícita una etiqueta, la herramienta desarrolla dicha etiqueta al situar encima el puntero del ratón pues, al igual que sucede con las marcas de codificación, emerge un cuadro de texto que la desglosa proporcionando la caracterización morfosintáctica de la unidad en cuestión en un lenguaje comprensible para cualquier usuario.

Esta doble actuación, menú desplegable para formular las consultas y cuadro de texto emergente con el desarrollo de la etiqueta en las concordancias o en el análisis completo de la secuencia visible desde el contexto, permite que un usuario no

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

familiarizado con el sistema de etiquetación aplicado pueda realizar consultas a través de la modalidad elementos gramaticales y comprender fácilmente los resultados.

La posibilidad de utilizar directamente rasgos gramaticales, independientes de las formas que los soportan, ilustra la potencia de la recuperación de información y su utilidad para el trabajo gramatical, pero el sistema de explotación de ESLORA cuenta además con otras opciones que permiten refinar todavía más las búsquedas. Existe la opción de realizar consultas acerca de la copresencia tanto de palabras ortográficas como de elementos gramaticales que no están limitadas a la aparición de una de ellas inmediatamente antes o después de la otra, sino que admiten la existencia de varios elementos (hasta diez) entre ambos. Gracias a esta posibilidad, es posible recuperar, por ejemplo, todos los casos del lema *olvidar* –información presente en el Elemento 1– que concurran con la preposición *de* –información aportada para el Elemento 2– a una distancia máxima de 4 o menos elementos –hay que indicar un valor de 1 a 10, pudiendo ser ese valor distancia exacta o inferior a la expresada–, con independencia de que existan casos de enclisis pronominal o interpolación de elementos varios.

De otra parte, el sistema cuenta con la posibilidad de emplear en los campos de búsqueda los comodines habituales: el cierre de interrogación (?) sustituye a un carácter en la posición que ocupe en la cadena y el asterisco (*) sustituye a ningún carácter, uno o varios en la posición que ocupe en la cadena, de forma que *p?so* devolverá los casos que contengan la palabra *paso*, *peso*, *piso*, *poso y puso*, mientras que **miento* devolverá todas las secuencias que incluyan *-miento*, incluida la primera persona del singular del presente de indicativo de *mentir*.

Asimismo, el usuario puede recurrir en la formulación de la consulta a dos operadores booleanos: el cierre de la admiración (!) equivalente a NO, y la barra vertical (|) equivalente a O, de modo que, como muestra la figura 8, es posible recuperar todas las formas verbales –etiqueta *V** en el campo *Etiqueta*– pertenecientes a la segunda y tercera conjugaciones salvo las de los verbos *ser* e *ir* –en el campo *Lema *er|*ir!ser!ir*–.

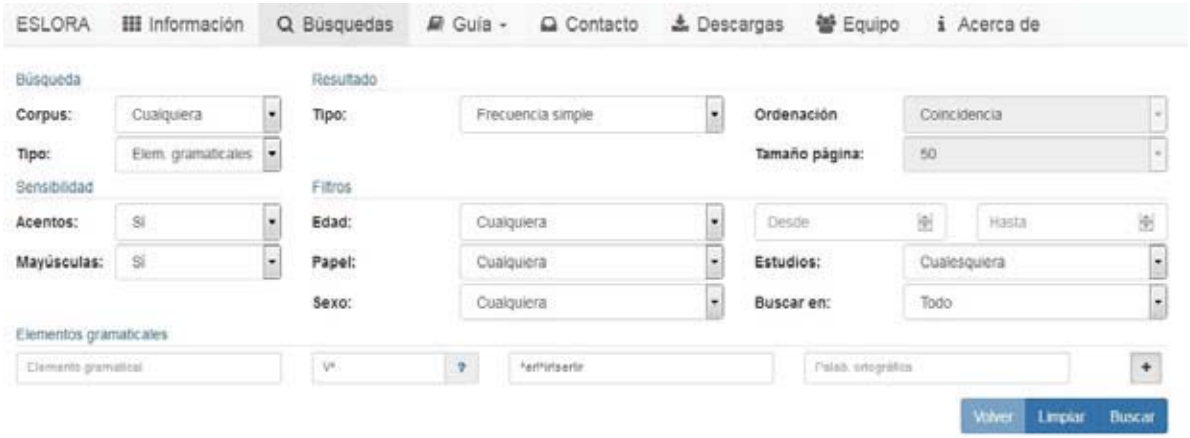


FIGURA 8. Pantalla de captación de datos con comodines y operadores booleanos

Naturalmente, existe la posibilidad de combinar en una misma consulta comodines, operadores booleanos, sensibilidad a acentos o mayúsculas y variables sociales con los distintos tipos de modalidad de búsqueda, por palabras ortográficas o elementos gramaticales, bien sucesivos bien discontinuos, lo que convierte al corpus ESLORA en una herramienta muy útil para obtener datos del español de Galicia.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

4. ASPECTOS CUANTITATIVOS

Se ha aludido en numerosas ocasiones al escaso interés que los estudios acerca de la frecuencia de elementos o fenómenos lingüísticos ha despertado en la lingüística española. Si bien es cierto que las frecuencias léxicas han sido las menos desatendidas, también lo es que son muy pocos los estudios referidos a la lengua oral[22]. El más destacado de ellos es el que Ávila Muñoz (1999) dedicó al análisis de la lengua hablada en Málaga. Está planteado en el estilo tradicional de los diccionarios de frecuencias, de modo que resulta sencillo examinar lo referente a un lema y a las formas integradas en él, pero no está pensado para estudios que manejen elementos más abstractos como, por ejemplo, las clases de palabras o las subcategorías verbales.

El hecho de que ESLORA haya sido concebido desde el principio como un corpus de lengua oral y las transcripciones se hayan lematizado y analizado morfosintácticamente permite enfocar el análisis de las frecuencias, no solo las léxicas, desde una óptica más amplia y ambiciosa. Las 53 entrevistas que hemos podido procesar arrojan un total de 669 502 elementos lingüísticos. Nótese que aquí no se habla de palabras ortográficas (que, aunque comprensible, resultaría un tanto chocante cuando se trata de transcripciones de lengua oral). En esa cifra se incluyen también los escasos signos ortográficos que se han utilizado en la transcripción (*cfr*. *supra* apdo. 3.2), pero no entran en cambio, indicaciones de pausa, silencio, ruido, etc.[23]. Si se eliminan los signos ortográficos, operación precisa para poder hacer comparaciones con textos escritos, la cifra anterior se reduce a 647 574 elementos lingüísticos. En un segundo recorte, podemos eliminar las 9699 apariciones de nombres propios (el 1,5 % del total de elementos sin signos ortográficos). Por tanto, deducidos estos dos bloques, que no resultan de interés para el análisis de las frecuencias léxicas, queda un total de 637 875 elementos. Ciertamente, no son muchos para los volúmenes a los que estamos acostumbrados en los corpus, pero superan a los que han servido para la confección de obras de gran utilidad como el FDSW y también al construido sobre el español hablado en Málaga, que tiene 523 639 palabras (Ávila Muñoz 1999: 93).

Las cifras anteriores se refieren al total de apariciones de los elementos lingüísticos, esto es, a los *tokens* contenidos en las entrevistas. Los elementos distintos, es decir, los *types*, son 25 891, que se reducen a 23 074 una vez descontados los correspondientes a signos ortográficos y nombres propios. Para posibles comparaciones con los datos procedentes de otros corpus, es preciso tener en cuenta que lo que aquí consideramos elemento único (*type*) es más complejo y abstracto que lo que se maneja cuando se diferencia entre *types* y *tokens* trabajando con formas ortográficas. En efecto, el *type* en nuestro caso implica también una cierta etiqueta morfosintáctica y la pertenencia a un lema determinado. Por tanto, la palabra ortográfica *vacío* corresponde a tres elementos distintos (la forma singular del sustantivo *vacío*, la forma masculina singular del adjetivo *vacío* y la forma de primera persona del singular del presente de indicativo del verbo *vaciar*). Por otro lado, las contracciones *al* y *del* son sistemáticamente analizadas en dos elementos cada una de ellas y, en sentido contrario, las expresiones multipalabra del tipo *sin embargo*, *a pesar de*, lo mismo que los nombres propios formados por varias palabras ortográficas, se agrupan en un elemento único.

[22] Como es previsible dada la época en que fue confeccionado, no hay análisis de textos orales en el FDSW (Juilland & Chang 1964). En el diccionario de frecuencias de Davies (2006) aparecen indicaciones acerca de si una palabra muestra una predisposición especial a aparecer en textos orales, pero hay que tener en cuenta que el corpus sobre el que se hacen los cálculos es el primero de los elaborados por Davies para el español (el formado por cien millones de formas), donde la consideración de texto oral se aplica a, por ejemplo, entrevistas publicadas en periódicos (*cfr*. Rojo 2010).

[23] Eso explica la discrepancia entre estas cifras y las que se dan en la página de la aplicación o el apdo. 2.

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

La relación entre el número total de elementos (*tokens*) y el número de elementos distintos (*types*) permite calcular la TTR (*type-token ratio*), que ha sido utilizada con cierta frecuencia como un índice de la riqueza léxica de un texto o un conjunto de textos. Aunque la fórmula más sencilla admite bastantes mejoras (*cfr*. Torruella & Capsada 2013, 2017), es siempre una medida poco refinada y, lo que es más grave, muy sensible al tamaño del corpus, de modo que desciende de forma muy marcada con el aumento del volumen. El núcleo básico de los elementos (esto es, sin signos ortográficos ni nombres propios) del corpus ESLORA arroja una TTR de 0,036. Para poder situar estas cifras con relación a las que pueden corresponder a textos de otro tipo hemos extraído del *Corpus del español del siglo XXI* (CORPES) dos muestras de tamaño similar al que tiene ESLORA en la actualidad: una de ellas está constituida por noticias de prensa publicadas en España en 2002 y la otra está formada por libros (tanto de ficción como de otras clases) del mismo país y año. Los datos procedentes de los tres subconjuntos aparecen en la tabla 1:

	ESLORA		Muestra CORPES prensa		Muestra CORPES libros	
	Total elementos (tokens)	Total elementos distintos (types)	Total elementos (tokens)	Total elementos distintos (types)	Total elementos (tokens)	Total elementos distintos (types)
Elementos identificados	669 502	25 891	667 805	58 420	807 174	61 237
Signos Ortográficos	21 928	7	77 804	35	102 920	23
Nombres propios	9699	2810	24 323	10 020	20 520	56
Elementos sin signos ortográficos ni nombres propios	637 875	23 074	565 678	48 365	683 684	55 606
TTR	0,036		0,085		0,081	
1 elemento nuevo cada	27,645		11,696		12,295	

TABLA 1. Frecuencias de elementos y elementos distintos en tres subcorpus diferentes. Fuentes: ESLORA y CORPES. Elaboración propia

Puede observarse que la «riqueza léxica» en los textos orales es considerablemente menor que la que se aprecia en los textos escritos. Cabe pensar que una agrupación de noticias periodísticas, de escasa extensión, con los cambios de tema que suponen, trae consigo una variedad temática que forzosamente tiene que reflejarse en variedad léxica. Sin embargo, el resultado del análisis de textos procedentes de libros, no solo de ficción, muestra que las cifras son similares a las obtenidas en textos de prensa y ambas muy superiores a las que se observan en textos orales: con una ilustración clara, en corpus de tamaño similar, en los textos orales aparece uno nuevo cada 27,6 elementos, mientras que en los textos escritos es suficiente con unos 12. Es una diferencia muy evidente que, por otro lado, confirma lo esperable.

Una medida mucho más adecuada de la riqueza léxica pasa por utilizar las ventajas que proporciona la anotación de ESLORA. Dado que la anotación se ha hecho de forma

En Ángel Gallego & Francesc Roca (eds.), Dialectología digital del español. Santiago de Compostela: USC, 2020, 189-224.

automática, hay que tener en cuenta siempre la posibilidad de que la información manejable contenga un cierto número de errores, además de elementos no identificados. Incluyendo en la caracterización del lema la pertenencia a una determinada clase de palabras, la versión actual de ESLORA (la 1.2.2) está formada por un total de 11 147 «lemas» distintos, que se reducen a 9368 si eliminamos los signos ortográficos, los nombres propios y los no identificados[24]. Esos 9368 lemas suman en total 639 160 elementos, lo cual arroja una frecuencia media de 67,7 elementos por lema, cifra bastante elevada. Para poder valorar adecuadamente estas cifras, hemos calculado también las correspondientes a las dos muestras de tamaño similar extraídas del CORPES. Los resultados figuran en la tabla 2.

	ESLORA		Muestra CORPES prensa		Muestra CORPES libros	
	Número de lemas distintos	Frecuencia total	Número de lemas distintos	Frecuencia total	Número de lemas distintos	Frecuencia total
Total «lemas»	11 747	669 502	31 142	667 805	29 913	807 174
Signos ortográficos	7	21 928	25	77 804	28	102 970
Nombres propios	2366	9699	9842	24 323	5623	20 520
No identificados	6	3715				
Total lemas	9368	634 160	21 275	565 678	24 262	683 684
Frec./lemas	67,7		26,6		28,2	

TABLA 2. Número de lemas y frecuencia total en ESLORA y dos subcorpus del CORPES.
Fuentes: ESLORA y CORPES. Elaboración propia

Como era de esperar, los textos escritos contienen un número de lemas considerablemente superior al que encontramos en ESLORA. La muestra de prensa, de menor tamaño que el corpus oral, tiene, a pesar de ello, más del doble de lemas diferentes. La divergencia entre los tres conjuntos textuales se hace muy evidente en la frecuencia media: 67,7 apariciones por lema en ESLORA y 26,6 y 28,2 en los dos subcorpus escritos.

La configuración resumida en la ley de Zipf se cumple también en los textos orales: los 25 elementos más frecuentes suponen el 38,97 % del total, mientras que los 25 lemas más frecuentes alcanzan el 44,53 %. Por otra parte, el porcentaje de hápax (es decir, elementos con frecuencia igual a 1) es del 53,43 % del total de elementos distintos y el 30,70 % de lemas. La tabla 3 muestra estos datos en comparación con los obtenidos de las dos muestras escritas que estamos usando para la comparación[25].

[24] Entrecomillamos «lemas» para señalar la peculiaridad de aplicar este término a signos ortográficos y nombres propios. Por otra parte, dado que aquí se usa lema para la unión de un término y una clase de palabras, se comprenderá que, como se muestra en la tabla 2, los lemas no identificados sean solo seis, tantos como clases de palabras distintas son asignadas a lemas no identificados. Los elementos distintos (elemento y etiqueta morfosintáctica) no identificados en la versión actual son 1694, con una frecuencia total conjunta de 3715, indicada en la tabla 2.

[25] Los porcentajes han sido obtenidos sobre el número de elementos o lemas ya sin signos ortográficos ni nombres propios ni unidades no identificadas.

	Totales y % sobre el total de ESLORA	Totales y % sobre el total del subconjunto de CORPES (prensa)	Totales y % sobre el total del subconjunto de CORPES (libros)
25 elementos lingüísticos más frecuentes	248 588 (38,97 %)	217 829 (38,48 %)	252 110 (36,84 %)
25 lemas más frecuentes	301 214 (44,99 %)	252 099 (44,53 %)	301 908 (44,02 %)
Elementos con frecuencia = 1 (hápax)	12 330 (53,43 %)	29 323 (60,54 %)	27 829 (49,96 %)
Lemas con frecuencia = 1	9440 (40,91)	14 872 (30,70 %)	9398 (31,41 %)

TABLA 3. Totales y porcentajes de elementos lingüísticos y lemas (sin signos de puntuación ni nombres propios) en ESLORA y en una muestra del CORPES escrito. Elaboración propia

Los porcentajes que suponen sobre los totales respectivos los 25 lemas o elementos más frecuentes son muy próximos entre sí y también a los que se obtienen en corpus de tamaño mucho mayor. En cuanto a los elementos con frecuencia igual a 1 (hápax), los tres porcentajes difieren, pero no demasiado y el hecho de que resulten más altos que los obtenidos para muestras de mayor tamaño puede explicarse por el escaso volumen de los corpus que estamos examinando aquí. Por fin, el porcentaje de lemas con frecuencia igual a 1 es congruente con lo que se observa en corpus más grandes en las muestras escritas, mientras que resulta más alto en el corpus oral (*cfr*. Rojo 2008 y 2017 para datos de este tipo correspondientes a CREA y CORPES).

Más allá de estos resultados, todavía bastante toscos, relacionados con las frecuencias generales, las cuestiones realmente importantes se refieren a la posible existencia de diferencias en la configuración gramatical de los textos orales del tipo al que corresponden las entrevistas semidirigidas en comparación con (ciertos tipos de) textos escritos. Como primera aproximación, forzosamente muy superficial y tentativa, a un problema que requiere análisis mucho más profundos, hemos examinado el peso de las clases de palabras de contenido más léxico en ESLORA y en dos muestras procedentes del CORPES, tanto en el inventario como en los textos[26]. Los porcentajes del número de adjetivos, adverbios, sustantivos comunes y verbos sobre el total de los lemas (de nuevo sin signos ortográficos ni nombres propios ni lemas no identificados) es el que muestra la tabla 4. Llama la atención el hecho de que los textos orales y los escritos procedentes de libros están muy próximos en cuanto a lo que la suma de estas cuatro clases de palabras supone con respecto al total del inventario de lemas, mientras que los textos de prensa se quedan ocho puntos porcentuales por debajo de los otros dos. Con respecto a la distribución de las cuatro clases de palabras, lo más destacable es, sin duda, que los

[26] Se trata de la distinción establecida en Rojo (2011a) entre frecuencia de inventario y frecuencia de uso, que es reelaboración de la propuesta por Bybee (2007), entre otros, entre *type frequency* y *token frequency*, pero en una formulación bastante más general. La frecuencia de inventario es el número de elementos distintos de un cierto tipo (en nuestro caso, por ejemplo, adjetivos o verbos) que se localizan en un determinado corpus, es decir, en su lemario. La frecuencia de uso, en cambio, es la que arrojan *todas* las apariciones de un determinado tipo de elemento en un corpus (en nuestro caso, por ejemplo, el número total de adjetivos o de verbos, contando todas y cada una de las apariciones de elementos de la clase correspondiente en los textos que integran el corpus). Los artículos, las preposiciones o las conjunciones, por ejemplo, pesan muy poco en el inventario de los elementos de una lengua, pero, en cambio, tienen una frecuencia muy alta si nos referimos al total de sus apariciones en los textos.

textos orales muestran un menor porcentaje de adjetivos, compensado con una diferencia de signo contrario en los verbos.

	Inventario				Textos		
	ESLORA	Muestra CORPES prensa	Muestra CORPES libros		ESLORA	Muestra CORPES prensa	Muestra CORPES libros
Adjetivos	16,35	21,02	22,04	Adjetivos	1,78	7,65	6,87
Adverbios	4,23	3,25	3,74	Adverbios	14,48	4,63	5,35
Sustantivos comunes	55,63	52,23	56,65	Sust. com.	12,06	25,10	22,61
Verbos	19,81	12,08	14,26	Verbos	20,64	13,53	15,22
Totales	96,02	88,58	96,70	Totales	48,95	50,91	50,05

TABLA 4. Porcentajes de cuatro clases de palabras sobre el total de los lemas y el total de los textos en ESLORA y dos subcorpus del CORPES. Fuentes: ESLORA y CORPES. Elaboración propia

Más diferencias muestra la frecuencia en los textos. Aquí las dos muestras escritas resultan mucho más próximas entre sí y las entrevistas de ESLORA se diferencian de ambas con bastante claridad. Todas las clases examinadas presentan diferencias importantes en el corpus oral con respecto a los textos escritos: los porcentajes de adjetivos y sustantivos comunes son muy inferiores en los textos orales, mientras que los correspondientes a adverbios y verbos resultan muy superiores. No se puede descartar la posibilidad de que una parte de estas diferencias procedan de los diferentes programas de anotación automática que se han aplicado (uno en el caso de ESLORA y otro en las dos muestras del CORPES), pero no son estas clases de lemas las que pueden variar en mayor medida y, por otro lado, las diferencias son de una entidad considerable. Todo indica, por consiguiente, que aquí hay una diferencia importante que será necesario investigar con mayor profundidad.

Veamos, por último, algunos datos procedentes del cruce del número de lemas distintos con los parámetros utilizados habitualmente en los estudios sociolingüísticos. Con las eliminaciones habituales y sin tener en cuenta tampoco las intervenciones de los encuestadores, en las entrevistas con hombres como informantes documentamos 6371 lemas diferentes (sobre 266 035 elementos totales) y 5879 lemas diferentes en las protagonizadas por mujeres (sobre un total de 300 239 elementos). Con estas cifras, la «riqueza léxica» parece mayor en los hombres que en las mujeres. En la tabla 5 pueden verse estos datos y también los correspondientes a los otros dos parámetros.

Sexo	Lemas distintos	Nivel educativo	Lemas distintos	Edad	Lemas distintos
Hombre	6371	Bajo	4709	19-34	4688
Mujer	5879	Medio	5032	35-54	5042
		Alto	5300	>54	5300

TABLA 5. Número de lemas distintos en las entrevistas de ESLORA en relación con diferentes parámetros sociolingüísticos. Fuente: ESLORA. Elaboración propia

El número de lemas diferentes documentados en las entrevistas aumenta con el nivel educativo y también con la edad, pero las diferencias no parecen realmente significativas.

5. CONCLUSIONES

En este trabajo se ha presentado el corpus ESLORA, un corpus de lengua oral del español de Galicia. En la introducción hemos reflexionado sobre la necesidad y la utilidad de un corpus de estas características en especial para los estudios de dialectología. ESLORA viene, en efecto, a sumarse a otros corpus que documentan las variedades del español propias de las zonas en las que hay lenguas en contacto. El artículo revisa en primer lugar la composición del corpus desde el punto de vista de los tipos de interacciones incluidas y de las características sociolingüísticas de los informantes. Se explica a continuación el proceso de recogida (grabación) y preparación (transcripción, alineación, anonimización y codificación) de los datos, así como el sistema de anotación morfosintáctica utilizado en ellos. Todo ello desde la perspectiva de las peculiaridades o problemas específicos que presenta el desarrollo de un corpus oral para documentar una variedad de lengua en situación de contacto de lenguas. Se ha descrito también la aplicación de consulta en línea que pone el corpus a disposición de los investigadores y se ha hecho un estudio cuantitativo de frecuencias léxicas (*types* en relación con *tokens*, lemas y clases de palabras) en ESLORA en comparación con corpus de lengua escrita.

Actualmente continuamos trabajando en la introducción de conversaciones, así como en la mejora de algunos aspectos de la codificación y la anotación morfosintáctica. Próximamente iniciaremos también la anotación sintáctica del corpus, que pretendemos facilitar en dos versiones: constitutiva y dependencial, de acuerdo con las directrices del proyecto Universal Dependencies.[27] Esto por lo que respecta al enriquecimiento del corpus. Por lo que toca a su explotación, aparte de los estudios para los que esperamos que sea útil a otros usuarios, es nuestra intención trabajar en distintos aspectos, entre ellos, la elaboración de diccionarios de frecuencias léxicas y gramaticales de lengua oral, el estudio de unidades, construcciones y formas verbales en lengua oral y la alternancia de código español-gallego. Por último, también está abierta una línea de trabajo para utilizar el corpus en ELE, conectando con la elaboración de diccionarios de frecuencias, para la investigación en secuenciación de contenidos de aprendizaje de destrezas orales y también desde el punto de vista de la elaboración de materiales para ese aprendizaje.

RECURSOS ELECTRÓNICOS MENCIONADOS

BNC: *The British National Corpus*, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/>

CORPES: *Corpus del español del siglo XXI*. Real Academia Española <http://rae.es/recursos/banco-de-datos/corpes-xxi>

COSER: *Corpus oral y sonoro del español rural.* <http://www.corpusrural.es/>

ESLORA: *Corpus para el estudio del español oral* <http://eslora.usc.es>, versión 1.2.2 de noviembre de 2018, ISSN: 2444-1430.

[27] URL: https://universaldependencies.org/.

PRESEEA: *Proyecto para el Estudio Sociolingüístico del Español de España y América*. <http://preseea.linguas.net/>

XIADA: Etiquetador/Lematizador do Galego Actual. Centro Ramón Piñeiro para a investigación en humanidades. <http://corpus.cirp.gal/xiada>

REFERENCIAS BIBLIOGRÁFICAS

ACÍN VILLA, E. (1996): "Galleguismos en la prensa gallega escrita en castellano", *in* M. Casado Velarde *et alii* (eds.): *Scripta Philologica in memoriam Manuel Taboada Cid*. A Coruña: Universidad de A Coruña, vol. 1, pp. 267-277.

AGHA, A. (2007): *Language and social relations*. New York: Cambridge University Press. Disponible en https://epdf.tips/language-and-social-relations-studies-in-the-social-and-cultural-foundations-of-.html

ÁVILA MUÑOZ, A. (1999): *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

BARCALA, M, E. DOMÍNGUEZ, A. FERNÁNDEZ, R. RIVAS, Mª P. SANTALLA, V. VÁZQUEZ & R. VILLAPOL (2018): "El corpus ESLORA de español oral: diseño, desarrollo y explotación", *CHIMERA. Romance Corpora and Linguistic Studies* 5/2, pp. 217-237. https://doi.org/10.15366/chimera2018.5.2.003

BEAL, J. (2005): "Dialect representation in texts", in *The Encyclopedia of Language and Linguistics*. Amsterdam / London: Elsevier, 2.ª ed., pp. 531-538. https://doi.org/10.1016/B0-08-044854-2/00504-6

BRIZ, A. & Grupo Val.Es.Co (2002): *Corpus de conversaciones coloquiales*. Madrid: Arco Libros.

BUCHOLTZ, M. (2000): "The politics of transcription", *Journal of Pragmatics* 32, pp. 1439-65. https://doi.org/10.1016/S0378-2166(99)00094-6

BYBEE, J. (2007): *Frequency of use and the organization of language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

CAMUS, B. & S. GÓMEZ SEIBANE (2015): "La diversidad del español en Álava: Sistemas pronominales a partir de las encuestas del COSER", *Revista de Filología Española* XCV, pp. 279-206. https://doi.org/10.3989/rfe.2015.11

DAVIES, M. (2006): *A frequency dictionary of Spanish. Core vocabulary for learners*. Nueva York: Routledge.

DE BENITO, C. (2015): *Las construcciones con «se» desde una perspectiva variacionista y dialectal*. Tesis doctoral. Universidad Autónoma de Madrid.

DOMÍNGUEZ NOYA., E. Mª, F. Mario BARCALA RODRÍGUEZ & M. Á. MOLINERO (2009): "Avaliación dun etiquetador automático estatístico para o galego actual: Xiada", *Cadernos de Lingua* 30-31, pp. 151-193.

DU BOIS, J. W., W. L. CHAFE, C. MEYER, S. A. THOMPSON, R. ENGLEBRETSON & N. MARTEY (2000-2005): *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.

DURANTI, A. (2006): "Transcripts, like shadows on a wall", *Mind, Culture, and Activity* 13/4, pp. 301-310. https://doi.org/10.1207/s15327884mca1304_3

EAGLES (1996): *Recommendations for the morphosyntactic annotation of corpora*. EAGLES Document EAG-TCWG-MAC/R. http://www.ilc.cnr.it/EAGLES96/browse.htlm [última consulta: 20/6/2018].

EDWARDS, J. A. (2001): "The transcription of discourse", in D. Tannen, D. Schiffrin, & H. E. Hamilton (eds.): *Handbook of discourse analysis*. Oxford: Blackwell, pp. 321-348.

FERNÁNDEZ JUNCAL, Mª C. (2005): *Corpus de habla culta de Salamanca* (CHCS). Burgos: Fundación Instituto Castellano y Leonés de la Lengua.

FERNÁNDEZ RODRÍGUEZ, M. (dir.) (no publicado): "Formación de un corpus de lengua hablada en la ciudad de A Coruña". Proyecto financiado por la Universidad de A Coruña, la Xunta de Galicia (XUGA10402A90) y la DGICYT (PB90-0324).

FERNÁNDEZ-ORDÓÑEZ, I. (2007): "El 'neutro de materia' en Asturias y Cantabria. Análisis gramatical y nuevos datos", *in* I. Delgados Cobos & A. Puigvert Ocal (eds.), *Ex admiratione et amicitia. Homenaje a Ramón Santiago*. Madrid: Ediciones del Orto, pp. 395-434.

FERNÁNDEZ-ORDÓÑEZ, I. (2016): "Dialectos del español peninsular", *in* J. Gutiérrez Rexach (ed.): *Enciclopedia lingüística hispánica*. Londres & New York: Routledge, vol. 2, pp. 387-404.

GARCÍA, C. (1986): "El castellano en Galicia", in V. García de la Concha *et alii*: *El castellano actual en las comunidades bilingües de España*. Salamanca: Junta de Castilla y León, pp. 49-64

GÓMEZ MOLINA, J. R. (2001): *El español hablado en Valencia: Materiales para su estudio*. Valencia: Universidad de Valencia.

GÓMEZ SEIBANE, S. (2017): "Español en contacto con la lengua vasca: datos sobre la duplicación de objetos directos posverbales", *in* A. Palacios (ed.), *Variación y cambio lingüístico en situaciones de contacto*. Madrid: Iberoamericana/Vervuert, pp. 143-159. https://doi.org/10.31819/9783954876648-008

JAFFE, A. & S. WALTON (2000): "The voices people read: Orthography and the representation of non-standard speech", *Journal of Sociolinguistics* 4/4, pp. 561-587. https://doi.org/10.1111/1467-9481.00130

JUILLAND, A. & E. CHANG-RODRÍGUEZ (1964). *Frequency dictionary of Spanish words*. La Haya: Mouton.

LE MEN, J. (2002): *Léxico del leonés actual*. León: Centro de estudios e investigación *San Isidoro*, 2002-2012, 6 vols.

LOPE BLANCH, J. M. (1986): *El estudio del español hablado culto. Historia de un proyecto*. México: UNAM.

NAGY, N. & D. SHARMA (2013): "Transcription", *in* R. J. Podesva & D. Sharma, (eds.): *Research methods in linguistics*. Cambridge: Cambridge University Press, pp. 235-256. https://doi.org/10.1017/CBO9781139013734.014

OCHS, E. (1979): "Transcription as theory", in E. Ochs & B. Schieffelin (eds.): *Developmental pragmatics*. New York: Academic Press, pp. 43-72.

PAASCH KAISER, C. (2015): *El castellano de Getxo: estudio empírico de aspectos morfológicos, sintácticos y semánticos de una variedad del castellano hablado en el País Vasco*. Berlin: De Gruyter. https://doi.org/10.1515/9783110366518

POLLÁN, C. (2001): "The expression of pragmatic values by means of verbal morphology: A variationist study", *Language Variation and Change* 13, pp. 59-89. https://doi.org/10.1017/S0954394501131030

POLLÁN, C. (2002): "The morphological expression of pragmatic values in oral and written Galician", in M. Fernández Ferreiro & F. Ramallo (eds.): *Sociolinguistics in Galicia: Views on diversity, a diversity of views* [= *Estudios de Sociolingüística*, 3/2 (2002) & 4/1 (2003)], pp. 113-138.

POPLACK, S. (1989): "The care and handling of a mega-corpus: The Ottawa-Hull French project", in R. Fasold & D. Schiffrin (eds.): *Language change and variation*. Amsterdam: John Benjamins, pp. 411-451. https://doi.org/10.1075/cilt.52.25pop

POPLACK, S. (1993): "Variation theory and language contact", in D. R. Preston (ed.): *American dialect re-search: An anthology celebrating the 100th anniversary of the American Dialect Society*. Amsterdam: John Benjamins, pp. 251-286. https://doi.org/10.1075/z.68.13pop

PRESTON, D. R. (1982): "'Ritin' Fowklower Daun 'Rong: Folklorists' failures in phonology", in *Journal of American Folklore* 95, pp. 304-326. https://doi.org/10.2307/539912

PRESTON, D. R. (1985): "The Li'l Abner syndrome: Written representations of speech", *American Speech* 60/4, pp. 328-336. https://doi.org/10.2307/454910

PRESTON, D. R. (2000): "Mowr and mowr bayud spellin': Confessions of a sociolinguist", *Journal of Sociolinguistics* 4/4, pp. 614-621. https://doi.org/10.1111/1467-9481.00132m

RABANAL, M. (1967): "Gramática breve del castellano hablado en Galicia y otros temas", in M. Rabanal: *Hablas hispánicas. Temas gallegos y leoneses*. Madrid: Ed. Alcalá, pp. 11-69.

RAE - ASALE (2009): *Nueva gramática de la lengua española*. Madrid, Espasa-Calpe.

RECALDE FERNÁNDEZ, M. (2012): "Aproximación a las representaciones sociales del español de Galicia", in T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas & A.

Veiga (eds.): *Cum corde et in nova grammatica: estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, Universidad de Santiago de Compostela, pp. 667-680.

RODRÍGUEZ ESPIÑEIRA, M. J. (2019): "La expresión epistémica *si cuadra* en español de Galicia", *Estudos de Lingüística Galega* 11, pp. 197-231. http://ojs3usc.devxercode.es/index.php/elg/article/view/5343. https://doi.org/10.15304/elg.11.5343

ROJO, G. (2004): "El español de Galicia", in R. Cano Aguilar (coord.): *Historia de la lengua española*. Barcelona: Ariel, pp. 1087-1101. 2.ª ed., 2005.

ROJO, G. (2008): "Lingüística de corpus y lingüística del español", ponencia plenaria en el XV congreso de la *Asociación de Lingüística y Filología de América Latina* (Montevideo, 18-21 de agosto de 2008). Montevideo. Edición en CD. [ISBN 978-9974-8002-6-7]

ROJO, G. (2010): "Sobre codificación y explotación de corpus textuales: Otra comparación del *Corpus del español* con el CORDE y el CREA", *Lingüística* 24, pp. 11-50.

ROJO, G. (2011a): "Frecuencia de inventario y frecuencia de uso", *Revista española de lingüística* 41/1, pp. 5-43.

ROJO, G. (2011b): "Me pidieron que reseñara~reseñase el libro que ?publicara / *publicase Bosque en 1980", in Mª V. Escandell Vidal, M. Leonetti & C. Sánchez López (eds.): *60* problemas *de gramática dedicados a Ignacio Bosque*. Akal: Madrid, pp. 213-219.

ROJO, G. (2017): "Sobre la configuración estadística de los corpus textuales", *Lingüística* 33/1, pp. 121-134. https://doi.org/10.5935/2079-312X.20170008

ROJO, G. & V. VÁZQUEZ Rozas (2014): "Sobre las formas en *-ra* en el español de Galicia", in A. Enrique-Arias, M. J. Gutiérrez, A. Landa & F. Ocampo (eds.): *Perspectives in the study of Spanish language variation. Papers in honor of Carmen Silva-Corvalán*. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 237-270. DOI: dx.doi.org/10.15304/va.2014.701

SAMPSON, G. (2000): "CHRISTINE Corpus: Documentation". http://www.grsampson.net/ChrisDoc.html.

SINNER, C. (2001): *Corpus oral de profesionales de la lengua castellana en Barcelona*. Accesible en http://www.carstensinner.de/castellano/corpusorales/index.html

TAGLIAMONTE, S. A. 2004. *Transcription protocol*. https://www.cambridge.org/gb/files/3713/6689/9690/2847_APPENDIX_C.pdf. [Última consulta: 19/01/2019].

TORRES CACOULLOS, R. & C. TRAVIS (2018): *Bilingualism in the community: Code-switching and grammars in contact*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108235259

TORRUELLA, J. & R. CAPSADA (2013): "Lexical statistics and typological structures: A measure of lexical richness", *Procedia. Social and Behavioral Sciences* 95, pp. 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668

TORRUELLA, J. & R. CAPSADA (2017): "Métodos para medir la riqueza léxica de un texto. Revisión y propuesta. Aplicación en el Corpus Informatizado del Catalán Antiguo", *Verba* 44, pp. 347-408. https://doi.org/10.15304/verba.44.3155

VANN, R. E. (2009): *Materials for the sociolinguistic description and corpus-based study of Spanish in Barcelona. Toward a documentation of colloquial Spanish in naturally occurring groups*. Lewiston, NY: The Edwin Mellen Press.

VÁZQUEZ VEIGA, N. (2003): *Marcadores discursivos de recepción*. Santiago de Compostela: Universidade de Santiago de Compostela.

VILA PUJOL, Mª R. (2001): *Corpus del español conversacional de Barcelona y su área metropolitana*. Barcelona: Universitat de Barcelona.