El corpus ESLORA de español oral
Diseño, desarrollo y explotación

Mario Barcala°, Eva Domínguez*, Alba Fernández*, Raquel Rivas*, M.ª Paula Santalla*, Victoria Vázquez*, Rebeca Villapol*[1]
°NLPgo Technologies S.L., *Universidade de Santiago de Compostela

ESLORA is a corpus of Spanish made up of semi-directed interviews and spontaneous conversations recorded in Galicia between 2007 and 2015. The design and construction of the corpus meets three objectives: to register the use of a variety of Spanish which to date has been scarcely documented, to gain additional insight into the methods for the construction of spoken corpora, and to develop computational tools for corpus search. The paper presents the main characteristics of ESLORA and the criteria followed in the corpus building process. It also includes a brief description of the tools used to build the corpus and how they work together to achieve the project needs and, moreover, it shows that the decisions taken at various stages of the compilation of the corpus are closely related to the wide range of possibilities for retrieving the lexical, grammatical and contextual information provided by the materials.

Keywords: spoken corpus, semi-directed interview, conversation, Galician Spanish, POS-tagging

1. Objetivos

El diseño y elaboración del corpus ESLORA <http://eslora.usc.es/>, integrado por grabaciones en español con hablantes de Galicia, persigue tres objetivos generales. En primer lugar, ESLORA incrementa la amplitud y variedad de los materiales orales disponibles, con la particularidad de que los registros que

[1] En el desarrollo del corpus ESLORA participan también como miembros del equipo de investigación Marta Blanco, Francisco García, Sol López, Montserrat Recalde, Guillermo Rojo y Susana Sotelo.

CHIMERA. Romance Corpora and Linguistic Studies 5.2 (2018), 217-237. ISSN 2386-2629

forman el corpus representan el uso del español en una comunidad con lengua propia. Si la documentación y el estudio del español hablado en los territorios con lengua propia sigue siendo en conjunto una asignatura pendiente de la dialectología y la sociolingüística hispánica, en el caso del español hablado en Galicia la falta de datos empíricos es especialmente notoria. Además, el corpus aporta materiales obtenidos en dos contextos de situación diferentes, el contexto conversacional coloquial y el contexto de la entrevista semidirigida, lo cual permite abordar el estudio de la variación diafásica en el habla, en sintonía con una línea consolidada de investigación de la variación a través de los registros que se desarrolla a partir de los trabajos de Douglas Biber y colaboradores (Biber 1995; Biber & Conrad 2009).

En segundo lugar, el corpus proporciona las bases para realizar un análisis metodológico de los procedimientos de recopilación de registros orales. El objetivo es evaluar las dos técnicas más usadas en la obtención de registros de habla informal: la realización de entrevistas semidirigidas y la grabación secreta de conversaciones espontáneas. El contraste entre ambos subcorpus confirma que el empleo de una u otra técnica condiciona en diversos aspectos las características de los datos obtenidos. Dado el interés de los datos conversacionales para estudios y aplicaciones lingüísticas de diversa índole, es preciso determinar en qué aspectos divergen las muestras registradas con ambos métodos y en qué medida es posible equiparar los dos tipos de datos.

El tercer objetivo general del proyecto ESLORA es el desarrollo de herramientas informáticas que permiten el análisis, la consulta y la explotación de los materiales del corpus. Se trata tanto de diseñar instrumentos que facilitan al equipo la realización del propio proyecto, como de generar un recurso lo más enriquecido posible, un recurso que, con la ayuda de una herramienta de explotación adecuada, permita a los investigadores y al conjunto de la comunidad desarrollar estudios y actividades relacionadas con el español oral.

2. Características generales

El corpus ESLORA está formado por entrevistas semidirigidas y conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2015. El subcorpus de entrevistas (60 horas) se integra en el proyecto PRESEEA, lo que garantiza su comparabilidad con los corpus recopilados por otros cuarenta equipos de diferentes países hispanohablantes (Moreno Fernández 2006, 2016). No obstante, hay que señalar que las entrevistas de PRESEEA difieren en cuanto a la estructuración en módulos temáticos más o menos rígida que aplican los distintos

equipos. Los módulos se asocian a las diferentes secuencias discursivas (narración, descripción, argumentación...) y estas a diferentes unidades y construcciones lingüísticas, de modo que para asegurarse la homogeneidad de los materiales algunos equipos aplican un guion temático-secuencial estricto incluso en su distribución temporal. La contrapartida de una estructuración rígida es la menor espontaneidad de la interacción, y con ella la dificultad de obtener estilos de habla próximos al 'vernáculo', que como observó Labov (1972, 1981, 2001) es el registro idóneo para los estudios variacionistas. Así, lo que se gana en homogeneidad de las muestras se pierde en su representatividad sociolingüística.

El subcorpus de conversaciones (20 horas) es una iniciativa independiente de PRESEEA que, por una parte, reúne materiales de habla espontánea de español de Galicia y, por otra, mediante su comparación con el subcorpus de entrevistas, permite determinar el efecto que tiene el uso de cada una de estas dos técnicas de obtención de muestras de habla en las características de los datos registrados.

Si bien la entrevista tiene indudables ventajas como instrumento para registrar muestras amplias de habla estratificadas sociolingüísticamente y con la calidad sonora requerida, el hecho mismo de crear un contexto artificial limita su validez como reflejo del habla conversacional, pese a los intentos de salvar la "paradoja del observador". Por su parte, la técnica de grabación no intrusiva de interacciones coloquiales reales, que sí representan el uso espontáneo, dificulta la obtención de una muestra estratificada según variables sociales (grupos etarios, hombres / mujeres, niveles educativos...) y resulta más problemática éticamente y más compleja técnicamente (Recalde Fernández & Vázquez Rozas 2009).

El consentimiento informado de los participantes es un requisito legal y ético para el registro de los datos, que en el caso de las conversaciones implica un doble permiso, previo y posterior a la grabación. El compromiso por parte del equipo investigador es restringir el uso de las grabaciones y transcripciones a los ámbitos de la investigación y la docencia, y preservar el anonimato de los hablantes. La anonimización de las transcripciones se ha conseguido sustituyendo las menciones de los nombres y localizaciones que pudieran revelar la identidad de los informantes y de otras personas aludidas por denominaciones métrica y socialmente equivalentes (Sampson 2000). El proceso es laborioso pues la versión anonimizada no solo ha de mantener la coherencia interna que permita el seguimiento de la continuidad referencial del discurso sino que además ha de resultar verosímil en lo que se refiere a la información geográfica y demás indicaciones contextualizadoras. En el audio las referencias personales se eliminan y se sustituyen por un ruido, como se muestra en (1).

(1) <cita> ese es mi amigo </cita> y le digo <cita> ese no es tu amigo ~*Emilio* ese es un conocido </cita> <cita> para mí es mi amigo </cita> y claro esa distinción de amigo y conocido él no la tiene / entonces un tiene muchos amigos <risa=todos> (SCOM_M13_010_hab1) [link to audio_1_eslora.mp3]

3. La elaboración del corpus

3.1 Metadatos

Un componente fundamental del corpus es la información recopilada sobre las características de los hablantes y las condiciones de cada interacción. En ESLORA se registra de forma sistemática información sobre

a. la situación del evento grabado: fecha, localización y circunstancias concretas
b. los participantes (edad, sexo, estudios, profesión, lugar de nacimiento), su rol en el intercambio, en el caso de la entrevista semidirigida (entrevistador/a, informante, audiencia) y la relación entre ellos (desconocidos previamente, relación de amistad o parentesco)
c. el tipo de interacción: entrevista, conversación
d. las propiedades del archivo y el proceso de transcripción: formato, archivo de audio, transcriptor/a y fecha de transcripción, revisores y fechas de revisión

Los datos contextuales recopilados no solo sirven para identificar cada interacción sino que son imprescindibles para facilitar la posterior recuperación de la información, ya que permiten combinar las propiedades del evento y los elementos de la transcripción al formular los criterios de búsqueda. Además, dado el objetivo de documentar la variedad de español de Galicia, se recoge también información de interés sociolingüístico sobre el uso que los hablantes hacen del español y del gallego. En el proceso de recopilación del subcorpus de entrevistas, se incluyó un cuestionario sociolingüístico que permitió registrar con un grado de detalle considerable los datos sociológicos y las declaraciones de los hablantes sobre sus usos y actitudes lingüísticas.

La muestra de entrevistas incorpora también una prueba de inseguridad lingüística con el objetivo de examinar críticamente este instrumento clásico del método variacionista (Labov 1966; Preston 2013). Además de la anotación por escrito, las respuestas al cuestionario y a la prueba de inseguridad se registraron

también en audio, registro que constituye un material especialmente valioso tanto desde el punto de vista sociolingüístico como desde la perspectiva metodológica. Un primer análisis del método y los datos obtenidos puede consultarse en Recalde Fernández (2012), trabajo que aborda el estudio de las representaciones sociales sobre el español de Galicia a partir de las manifestaciones metalingüísticas de los hablantes en sus respuestas a los cuestionarios citados.

3.2 Grabación y transcripción

Todas las entrevistas y una parte de las conversaciones se registraron en archivos WMA con una grabadora digital Olympus DS-40 con micrófono integrado y sonido estéreo de extra alta calidad (ST XQ). Otra parte de las conversaciones se grabó en formato WAV, y ocasionalmente en MP3, mediante aplicaciones de grabación para dispositivos electrónicos.

La transcripción de las grabaciones se realizó de forma manual con ayuda de los programas Transcriber <http://trans.sourceforge.net> y ELAN <https://tla.mpi.nl/tools/tla-tools/elan/>, que permiten la alineación automática del sonido con texto codificado en formato XML. En la primera fase del proyecto, en la que se recogieron y transcribieron las entrevistas, se usó Transcriber, pero posteriormente, ante ciertas limitaciones de este programa (su falta de mantenimiento y actualización y los problemas de incompatibilidad derivados del peculiar tratamiento de los solapamientos), se eligió ELAN para transcribir las conversaciones. En ambos programas la alineación texto-audio tiene en cuenta los límites de los turnos de hablante y, dentro de cada turno, fragmentos menores delimitados por pausas.

Los materiales se transcribieron siguiendo las convenciones ortográficas básicas, sin excluir por ello la posibilidad de añadir en el futuro otras opciones de transcripción aprovechando las múltiples capas o niveles de anotación (*tiers*) que ofrece ELAN. No se utilizan signos de puntuación, a excepción de los signos de interrogación y admiración, aunque sí se marcan las pausas y los silencios. El uso de las mayúsculas está restringido a las iniciales de los nombres propios.

La representación ortográfica tiene ventajas en el proceso de transcripción porque, por una parte, simplifica notablemente la toma de decisiones por su carácter uniformador y, por otra, facilita el desarrollo de aplicaciones automáticas de recuperación de información a partir del texto transcrito. Además, en el caso de un corpus alineado como ESLORA, el acceso al registro sonoro es inmediato, lo cual compensa la falta de una transcripción más próxima a la realización fónica. No obstante, la convención ortográfica supone imponer la fijación de la tradición escrita a un uso oral diferente de la variedad legitimada por la escritura, el llamado

estándar. Dentro de los márgenes de la norma escrita, en ESLORA se establecieron criterios para garantizar la homogeneidad de las transcripciones. En general no se reflejan soluciones fónicas que alteran los límites entre palabras (contracciones y realizaciones fonéticas abreviadas), pero sí se transcriben realizaciones morfológicas ajenas al estándar (diminutivos como *bueniña* o las formas demostrativas *eses*, *estes*, por ejemplo).

Además de lo señalado, la inclusión del subcorpus de entrevistas en el proyecto PRESEEA obliga a adoptar unas convenciones mínimas de transcripción y etiquetación de los textos que aseguren su compatibilidad con los materiales de las demás variedades representadas en el proyecto[2]. Pero aun partiendo de los "mínimos" de PRESEEA, el proceso de codificación de materiales orales implica una toma de decisiones constante en la aplicación de las directrices de transcripción a datos o fenómenos no previstos. Entre los elementos lingüísticos que plantean más dudas de transcripción están los marcadores de tipo fático e interjectivo, que cumplen una función interactiva fundamental para el desarrollo normal de las conversaciones, por lo que deben registrarse adecuadamente en texto. En (2) y (3) se ofrece la representación de algunos de estos segmentos junto con el audio correspondiente.

2

(2) ahora <pausa/> pues <pausa/> bueno <pausa/> *mmm eeh eh* la cuestión es que que <pausa/> *mm* en cuanto lo que tu decías de si era *mm* <ruido tipo="chasquido boca"/> / si <alargamiento>estaba</alargamiento> de acuerdo con con (SCOM_M32_023_hab1) [link to audio_2_eslora.mp3]

3

(3) en la <palabra_cortada>z</palabra_cortada> toda la zona esta bueno <pausa/> *oye* <pausa/> lo que es precioso la casa de Galicia ¡ *eh* ! <pausa/> *ah* <pausa/> ¡ qué salones ! (SCOM_M31_045_hab1) [link to audio_3_eslora.mp3]

No obstante, aunque una transcripción cuidadosa de los elementos cuasi-léxicos e interjectivos es imprescindible, cualquier comparación entre transcripción y grabación pone de manifiesto que la categorización discreta inherente a la representación escrita oculta con frecuencia la variación formal y funcional que caracteriza el uso de los elementos cuasi-léxicos e interjectivos. De nuevo, el registro de una variedad poco estudiada como es el español de Galicia muestra la necesidad de reconocer funciones y elementos propios también en este ámbito. Las decisiones de transcripción adoptadas son una respuesta a esta necesidad,

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

pues tratan de reflejar lo más fielmente posible el uso registrado en las muestras, con las limitaciones inherentes a una representación de tipo ortográfico.

3.3 Marcas y etiquetas de oralidad

Una parte crucial del tratamiento de los textos para su posterior utilización como fuente de datos lingüísticos consiste en su anotación (*markup*) y etiquetación (*tagging*). Las posibilidades de anotación (tomando el término en su sentido más abarcador) son múltiples, como ilustraron ya en su momento Garside, Leech y McEnery (1997). En el corpus ESLORA el texto plano que representa el componente verbal básico se ha enriquecido con marcas que aportan información de diverso tipo, entre las que destacan, por un lado, las indicaciones relacionadas específicamente con el modo oral, en coherencia con los objetivos del propio corpus, y por otro, la información sobre el lema y la categoría y subcategoría gramatical de cada una de las formas que integran los textos.

La Figura 1 representa una secuencia con anotaciones de risas, alargamientos y pausas seguida de un pequeño fragmento inicial de la etiquetación morfosintáctica[3].

```
<fragmento hablante="habl" comienzo="108219.0" fin="115088.0">
   <expresión>¿sabes? no <risa/> <pausa/> yo no quiero saber <alargamiento>si</alargamiento>
   una cosa es un adverbio otra cosa <risa_inicio/>es un pronombre <risa_fin/> ¿sabes? <pausa/>
   que no te van a valer <alargamiento>mucho</alargamiento> <pausa/></expresión>
       <análisis>
         <análisis_unidad>
           <unidad>¿</unidad>
           <constituyente>
             <forma>¿</forma>
             <etiqueta>Q</etiqueta>
             <lema>¿</lema>
           </constituyente>
         </análisis_unidad>
         <análisis_unidad>
           <unidad>sabes</unidad>
           <constituyente>
             <forma>sabes</forma>
             <etiqueta>VIP2S</etiqueta>
             <lema>saber</lema>
           </constituyente>
         </análisis_unidad>
```

Figura 1. Secuencia de ESLORA anotada y etiquetada (fragmento)

Aunque en su formato actual no se establece un nivel específico de anotación prosódica, se ha hecho referencia a la representación de las pausas y silencios, elementos que sirven también como delimitadores de unidades de texto para el etiquetador estadístico. El sistema de transcripción de ESLORA incluye, además

[3] El Apéndice 1 muestra un ejemplo completo de secuencia etiquetada.

de la representación ortográfica de las unidades verbales, un conjunto limitado de marcas y etiquetas que informan sobre algunas características lingüísticas, paraverbales y contextuales comunicativamente relevantes:

Tabla 1. Lista de marcas y etiquetas usadas en ESLORA

¿?	Enunciado interrogativo
¡!	Enunciado exclamativo
~*Nombre*	Nombre ficticio
<alargamiento></alargamiento>	Aumento de cantidad que afecta a algún sonido de la palabra marcada
<cita_inicio/> <cita_fin/>	Inicio y fin de fragmento en estilo directo
<énfasis_inicio/> <énfasis_fin/>	Segmento pronunciado con especial intensidad
<ininteligible>	Sustituye un fragmento no comprensible y por tanto no transcrito
<lengua_inicio nombre="xx"/> <lengua_fin/>	Fragmento en una lengua distinta del español (gl: gallego, en: inglés, pt: portugués, it: italiano, fr: francés, el: griego)
<palabra_cortada> </palabra_cortada>	Fragmento de una palabra
<pausa/>	Pausa breve
<pausa_larga/>	Pausa más larga pero inferior a un segundo
<risa/>	Risa de un/a hablante
<risa_inicio/> <risa_fin/>	Fragmento pronunciado entre risas
<ruido tipo="">	Entre las comillas se especifica el tipo de ruido
<sic_inicio/> <sic_fin/>	Lapsus de dicción que no debe confundirse con un error de transcripción
<sigla_inicio/> <sigla_fin/>	Sigla
<silencio>	Pausa de más de un segundo
<transcripción_dudosa_inicio/> </transcripción_dudosa_fin/>	La transcripción es problemática
<vacilación>	Sustituye fragmentos similares a palabras cortadas pero imposibles de transcribir

El uso del programa Transcriber facilita la indicación del habla solapada entre dos participantes (casi siempre suficiente para entrevistas, pero no así para conversaciones de más de dos hablantes), y proporciona un sistema cómodo para la introducción de etiquetas compatibles con PRESEEA, bien modificando o sustituyendo las que ofrece por defecto, bien creando directamente otras nuevas. La interfaz de trabajo de Transcriber se adecua bien a un sistema de codificación

diferencial con anotaciones verbales. Además, el programa agiliza el proceso de transcripción con el uso de combinaciones de teclas para añadir las marcas de oralidad, lo que ahorra mucho tiempo y evita errores tipográficos. Dado el formato XML de los archivos .TRS creados en Transcriber, el sistema requiere un documento .DTD que define la estructura y contenidos de los archivos de transcripción y garantiza su buena formación.

Las transcripciones de las conversaciones fueron realizadas con ELAN e incluyen prácticamente el mismo conjunto de etiquetas que las usadas en Transcriber[4]. No obstante, dadas las características del programa ELAN, para la fase de transcripción se optó por una representación simbólica de las marcas en algunos casos, como puede verse en las equivalencias recogidas en la Tabla 2. Posteriormente, mediante tratamiento informático, se procedió a la conversión y unificación de la codificación XML procedente de Transcriber (formato .TRS) y ELAN (formato .EAF) para obtener unos nuevos XML adaptados a las condiciones de los sistemas de lematización y etiquetación morfosintáctica.

Tabla 2. Correspondencia de marcas entre el sistema unificado de ESLORA y ELAN

ESLORA	ELAN
<alargamiento></alargamiento>	=
<cita_inicio/> <cita_fin/>	<cita> </cita>
<énfasis_inicio/> <énfasis_fin/>	<enf> </enf>
<ininteligible>	<inint/>
<lengua_inicio nombre="xx"/> <lengua_fin/>	<xx> </xx> (xx= gl: gallego, en: inglés, pt: portugués, it: italiano, fr: francés, el: griego)
<palabra_cortada> </palabra_cortada>	-
<pausa/>	/
<pausa_larga/>	//
<risa/>	<@@/>
<risa_inicio/> <risa_fin/>	<@> </@>
<sic_inicio/> <sic_fin/>	<sic> </sic>
<transcripción_dudosa_inicio/> </transcripción_dudosa_fin/>	<dud> </dud>
<vacilación>	<vac/>

[4] El Apéndice 2 contiene el sistema de marcas usado en Transcriber.

4. La anotación morfosintáctica del corpus

El corpus ESLORA está anotado morfosintácticamente y lematizado. Tras evaluar las características de dos de los etiquetadores que estimamos más adecuados para realizar el procesamiento automático de los textos, FreeLing <http://nlp.lsi.upc.edu/freeling> y XIADA <http://corpus.cirp.gal/xiada>, nos decantamos por este último.

Si bien la tasa de acierto de la etiquetación de textos escritos resultaba ser similar en ambos casos, el etiquetador XIADA resultó más conveniente para el proyecto. Básicamente, porque el manejo de las marcas de oral conjuntamente con las etiquetas morfosintácticas resulta extremadamente complejo en FreeLing, mientras que la gestión de ficheros XML de manera nativa por parte del etiquetador XIADA simplifica de forma significativa esta tarea. La adaptación del etiquetador XIADA, originalmente desarrollado para el análisis del gallego en el *Centro Ramón Piñeiro para a investigación en Humanidades*, ha dado buenos resultado frente a los intentos fallidos por alinear los dos marcajes (el morfosintáctico y el oral) en FreeLing.

El sistema de etiquetas aplicado al corpus trabaja con 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 restantes 8 etiquetas son para elementos residuales tales como símbolos, fechas o cifras, elementos que, en realidad, no aparecen en el corpus, bien por las características del mismo, bien por la transcripción que se ha elegido hacer de tales elementos en los escasos casos en que se utilizan). De acuerdo con su comportamiento gramatical, las etiquetas correspondientes a cada una de las clases de palabras identificadas se extienden, junto al símbolo que identifica la clase de palabras, con símbolos que se refieren al valor que cada palabra del corpus muestra para las categorías gramaticales de género, número, persona, tiempo, modo y caso. Así, por ejemplo, las etiquetas que corresponden a los determinantes posesivos de primera persona son las indicadas en (4) (incluimos junto a ellas ejemplos de contextos en los que se aplica cada una de esas etiquetas):

(4) DS1AS, Det. Pos. 1.ª-Pers.-Sing. Masc./Neut. Sing., *Es* mío.
DS1EP, Det. Pos. 1.ª-Pers.-Sing. Fem./Masc. Plur., Mis *guardaespaldas me protegerán.*
DS1ES, Det. Pos. 1.ª-Pers.-Sing. Fem./Masc. Sing., Mi *guardaespaldas me protegerá.*
DS1FP, Det. Pos. 1.ª-Pers.-Sing. Fem. Plur., *mías;* Mis *hermanas me*

ayudarán.
DS1FS, Det. Pos. 1.ª-Pers.-Sing. Fem. Sing., *mía;* Mi *hermana me ayudará.*
DS1MP, Det. Pos. 1.ª-Pers.-Sing. Masc. Plur., *míos;* Mis *hermanos me ayudarán.*
DS1MS, Det. Pos. 1.ª-Pers.-Sing. Masc. Sing., *Ese es* mío*;* Mi *hermano me ayudará.*
DS1NS, Det. Pos. 1.ª-Pers.-Sing. Neut. Sing., *Eso es* mío.

El número y carácter de las clases de palabras y categorías gramaticales identificados está de acuerdo con las recomendaciones recogidas en los tradicionales estándares de anotación morfosintáctica (EAGLES principalmente) y en la práctica más habitual de la anotación de corpus en español. El sistema de anotación es un tanto particular, sin embargo, en la aplicación de esas etiquetas en el aspecto siguiente. Como se puede observar en los ejemplos de (1), el sistema de anotación prevé que, para las palabras que pueden referirse a más de un valor para una categoría gramatical, el valor que se asigne sea el más específico que permite el contexto de la palabra en cuestión (la secuencia entre pausas). De ahí que *mi* tenga un valor de género E (Fem./Masc.) junto a *guardaespaldas* y uno de M (Masc.) junto a *hermano*. La práctica más extendida, por el contrario, hace que *mi* tenga bien un valor de género E en todas las circunstancias, bien uno de los dos, M o F, en relación con el contexto.

Como ya se indicó, la anotación morfosintáctica del corpus se ha llevado a cabo automáticamente, aunque una parte de la misma, la que luego se utilizó para entrenar el sistema automático, se ha revisado manualmente (hasta ahora 26.000 formas, que esperamos que lleguen al menos a 50.000). Este proceso de revisión manual se sirvió de un entorno desarrollado *ad hoc* que aprovecha la funcionalidad del etiquetador que permite generar el resultado de la etiquetación en formato XML sin eliminar las alternativas no elegidas como la mejor solución por la herramienta. El resultado se integra después en una personalización del estilo de visualizado de un editor XML configurable (XMLMind editor <http://www.xmlmind.com/xmleditor/>) en el que los anotadores pueden de manera muy simple, bien sancionar como válida la etiqueta propuesta por el etiquetador automático, bien modificarla escogiendo alguna otra de las etiquetas posibles para la palabra en cuestión. Como en la Figura 2, en la que vemos que de entre las tres etiquetas posibles para *gustaría* el etiquetador ha seleccionado como adecuada VIC3S (verbo, indicativo, pospretérito, 3.ª, singular), que aparece en verde, mientras que VIC1S (verbo, indicativo, pospretérito, 1.ª, singular) y VIC4S (verbo, indicativo, pospretérito, 1.ª o 3.ª, singular) no han sido seleccionadas y aparecen en rojo. Si el anotador está de acuerdo con este estado de cosas,

simplemente lo respeta, y si no, modifica el recuadro de *Etq lema Valido,* marcándolo en alguna de las otras etiquetas posibles.

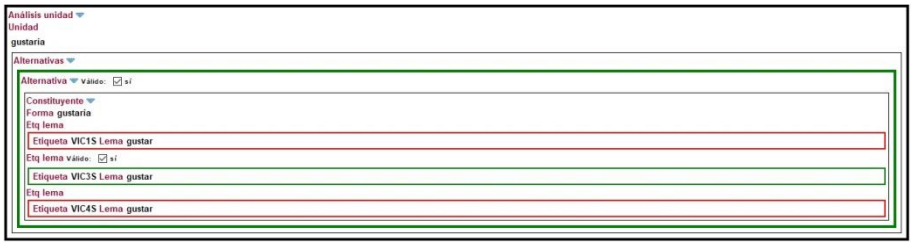

Figura 2. Visualización de la etiquetación en XMLMind para posedición manual

En este proceso de aplicación de las etiquetas al corpus, el trabajo que entraña más dificultad ha sido y sigue siendo la determinación de las etiquetas que deben corresponder a los ítems del corpus propios y exclusivos, o casi exclusivos, de la lengua prototípicamente oral que este codifica (para cuya descripción nos ampara mucho menos la tradición) o a los ítems compartidos por lengua oral y lengua escrita, pero con usos muy distintos en cada una de ellas (la tradición suele ampararnos en la descripción del uso escrito, pero no tanto del oral). Entre los primeros se cuentan sobre todo los elementos fáticos no lingüísticos (*mmm, er*, etc.), así como expresiones de valoración (*coño, Dios, vaya, venga, vamos, jolín, joder, guau, madre mía*), de saludo (*hola, buenas tardes*), de agradecimiento (*gracias, muchas gracias*), de felicitación (*enhorabuena*), de disculpa (*perdón*) o de asentimiento más o menos entusiasta (*bueno, vale, claro*), que en general hemos etiquetado como interjección (I) y donde ha sido necesario como expresiones multipalabra. Entre los segundos, uno de los elementos más representativos es *tal*, que ha sido etiquetado como pronombre o determinante demostrativos en los usos descritos en las gramáticas (*tal día, es tal que…*), y que se ha etiquetado como pronombre demostrativo en cualquier género y en singular (PDIS) en los usos específicamente orales (mucho más frecuentes) en los que sirve como elemento de relleno o transición (en expresiones como *y tal, un poco tal*, o directamente aislado entre sucesivos centros discursivos relacionados).

Por fin, y enlazando ya con la aplicación de explotación, siendo conscientes de la dificultad que entraña para los usuarios hacerse con el control de 455 etiquetas, en la herramienta de búsqueda se diseñó un menú amigable para la introducción de la etiqueta deseada, de modo que el usuario no tiene que conocer el sistema de etiquetas para expresar la búsqueda gramatical que quiere hacer en el corpus. Así, si lo desea, el usuario accede a un menú de opciones en cascada que organiza referencias tradicionales a clases de palabras y categorías, y que lo va orientando en la elección sucesiva de los valores de subclases o categorías

posibles en cada momento de acuerdo con las elecciones efectuadas hasta entonces.

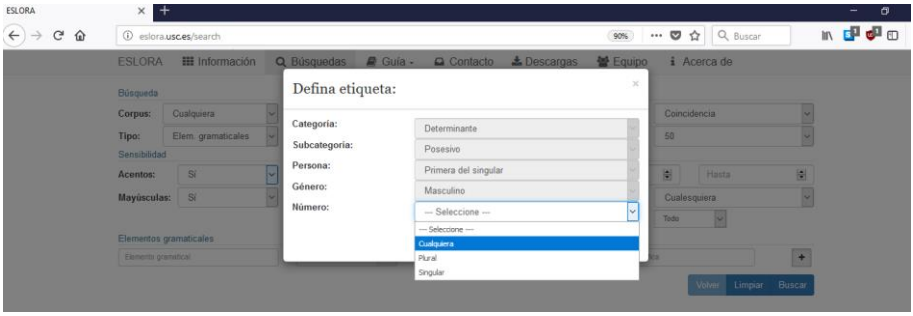

Figura 3. Ejemplo de búsqueda gramatical en ESLORA

Como se ve en la Figura 3, el usuario tras seleccionar *determinante-posesivo-primera del singular-masculino* puede detenerse o seleccionar uno de los valores *Cualquiera*, que equivale a no elegir nada en relación con el número, *Plural* o *Singular*. De seleccionar este último desembocaría en la etiqueta **DS1MS** vista antes para los posesivos. De no seleccionar nada o seleccionar *Cualquiera*, la etiqueta objeto de búsqueda acabaría siendo **DS1A***, que deja abierto el último valor relativo al número. Las etiquetas solo son visibles para el usuario en la vista de resultados que las evidencia, y su conocimiento en este momento sí creemos que contribuye, obviamente, a mejorar la interpretación que de los datos que le proporciona el corpus puede extraer el usuario.

5. Recuperación de información: aplicación de consulta

Las características de las transcripciones, metadatos y anotación del corpus descritas en los apartados anteriores amplían notablemente las posibilidades de recuperación de la información que contiene. Para acceder a toda la variedad y detalle de datos que ofrece ESLORA se ha diseñado una aplicación de consulta, disponible en <http://eslora.usc.es/>, que permite[5]

[5] Para una descripción detallada de la aplicación de consulta véase http://eslora.usc.es/guide_description

a. realizar búsquedas que combinan variables sociales (edad, nivel de estudios, sexo y papel del hablante) con variables léxicas y gramaticales (lema, clase de palabra, categorías morfológicas)
b. acceder directamente a los fragmentos de audio correspondientes a los resultados de las consultas
c. descargar el resultado de las búsquedas en formato TSV (Tab Separated Values)

Por lo que se refiere a las variables sociales, la opción de combinar libremente valores de los distintos parámetros que estructuran el corpus abre la posibilidad de trabajar con corpus virtuales a la medida del usuario. Además, cualquiera de esas selecciones se cruza con la información lingüística que interese en cada momento. Se puede elegir búsqueda de palabras ortográficas (también con algunas expresiones regulares) o elementos gramaticales, y estos permiten la recuperación por lema, forma ortográfica y clase de palabras con sus correspondientes categorías morfológicas. Especialmente interesante para el análisis gramatical es la posibilidad de recuperar cadenas y construcciones complejas combinando diferentes niveles de abstracción en una misma consulta, jugando también con la diferencia entre elementos contiguos o distantes en la secuencia.

Los resultados de las búsquedas incluyen, entre otras opciones, los datos de frecuencia absoluta y normalizada, la visualización del contexto en formato KWIC (*Key Word in Context*) con el acceso al segmento de audio alineado, la posibilidad de descarga de las concordancias o la ampliación del contexto con la representación de las etiquetas morfosintácticas, como muestra la Figura 4:

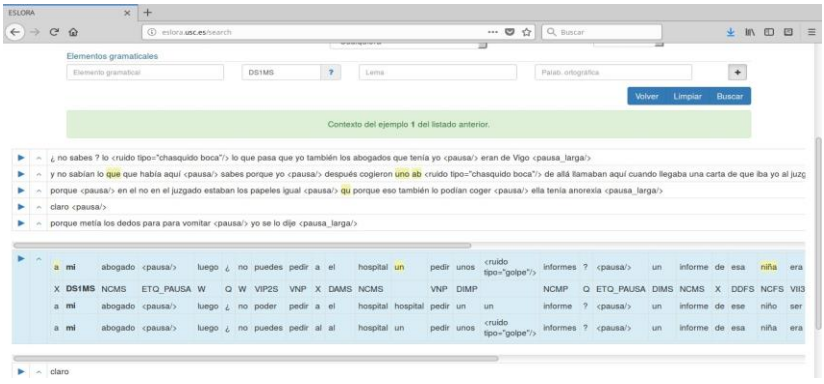

Figura 4. Visualización de anotación morfosintáctica en ESLORA

El corpus está disponible para su descarga en formato textual y, si se justifica la necesidad de acceso, se facilita asimismo el corpus en formato ya etiquetado, los audios correspondientes y la información sociolingüística de los hablantes.

Agradecimientos

El proyecto de investigación ESLORA+ (*El corpus ESLORA de español oral: enriquecimiento, análisis lingüístico y extracción de recursos*, ref. PFFI2017-86379-P) está financiado por la Agencia Estatal de Investigación (AEI) y por el Fondo Europeo de Desarrollo Regional (FEDER). 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 (ref. ED431B 2017/39).

Bibliografía

Biber, D. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge & New York: Cambridge University Press.

Biber, D. & Conrad S. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

EAGLES. 1996. *Recommendations for the Morphosyntactic Annotation of Corpora*. EAGLES Document EAG–TCWG–MAC/R. http://www.ilc.cnr.it/EAGLES96/browse.html (consultado el 20 de junio de 2018).

Garside, R., Leech, G. & McEnery, T. (eds) 1997. *Corpus annotation. Linguistic information from computer text corpora*. London & New York: Routledge.

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center of Applied Linguistics.

Labov, W. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, W. 1981. Field methods of the project on linguistic change and variation, *Sociolinguistic Working Paper nr. 81*, Southwest Educational Development Laboratory, Austin, Texas.

Labov, W. 2001. The anatomy of style-shifting. En P. Eckert & J. R. Rickford (eds), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, 85-108.

Moreno Fernández, F. 2006. Información básica sobre el Proyecto para el Estudio Sociolingüístico del Español de España y de América – PRESEEA (1996-2010). *Revista Española de Lingüística* 36: 385-392.

Moreno Fernández, F. 2016. En torno a preseea: Notas de investigación y de sociología de la ciencia. *Boletín de filología* 51(2): 369-376.

Preston, D. 2013. Linguistic Insecurity Forty Years Later. *Journal of English Linguistics* 41(4): 304-331

Recalde Fernández, M. 2012. Aproximación a las representaciones sociales del español de Galicia. En 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, Universidade de Santiago de Compostela, 667-680.

Recalde Fernández, M. & Vázquez Rozas V. 2009. Problemas metodológicos en la formación de corpus orales". En P. Cantos Gómez & A. Sánchez Pérez (eds), *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO, 37-49. https://www.um.es/lacell/aelinco/contenido/pdf/4.pdf.

Sampson, G. 2000. CHRISTINE Corpus: Documentation. Disponible en http://www.grsampson.net/ChrisDoc.html

Apéndice 1

Ejemplo de secuencia de ESLORA anotada y etiquetada

```
<fragmento hablante="hab2" comienzo="3307989.0" fin="3312460.0">
  <expresión>no <palabra_cortada>si</palabra_cortada> no sirve de nada
    <alargamiento>no</alargamiento> <pausa> </pausa> </expresión>
  <análisis>
    <análisis_unidad>
     <unidad>no</unidad>
     <constituyente>
      <forma>no</forma>
      <etiqueta>W</etiqueta>
      <lema>no</lema>
     </constituyente>
    </análisis_unidad>
    <análisis_unidad>
     <unidad>si</unidad>
     <información_adicional>palabra_cortada</información_adicional>
    </análisis_unidad>
    <análisis_unidad>
     <unidad>no</unidad>
     <constituyente>
      <forma>no</forma>
      <etiqueta>W</etiqueta>
      <lema>no</lema>
     </constituyente>
    </análisis_unidad>
    <análisis_unidad>
     <unidad>sirve</unidad>
     <constituyente>
      <forma>sirve</forma>
      <etiqueta>VIP3S</etiqueta>
      <lema>servir</lema>
     </constituyente>
    </análisis_unidad>
    <análisis_unidad>
     <unidad>de</unidad>
     <constituyente>
      <forma>de</forma>
      <etiqueta>X</etiqueta>
      <lema>de</lema>
     </constituyente>
    </análisis_unidad>
    <análisis_unidad>
```

```
      <unidad>nada</unidad>
      <constituyente>
       <forma>nada</forma>
       <etiqueta>PNNS</etiqueta>
       <lema>nada</lema>
      </constituyente>
     </análisis_unidad>
     <análisis_unidad>
      <unidad>no</unidad>
      <información_adicional>alargamiento</información_adicional>
      <constituyente>
       <forma>no</forma>
       <etiqueta>W</etiqueta>
       <lema>no</lema>
      </constituyente>
     </análisis_unidad>
     <análisis_unidad>
      <unidad>
       <pausa></pausa>
      </unidad>
      <constituyente>
       <forma>
        <pausa></pausa>
       </forma>
       <etiqueta>ETQ_PAUSA</etiqueta>
       <lema>
        <pausa></pausa>
       </lema>
      </constituyente>
     </análisis_unidad>
  </análisis>
</fragmento>
```

Apéndice 2

Lista de etiquetas utilizada en Transcriber

<Event desc="alargamiento" type="pronounce" extent="instantaneous"/>
<Event desc="ca" type="language" extent="begin"/>
<Event desc="ca" type="language" extent="end"/>
<Event desc="cita" type="lexical" extent="begin"/>
<Event desc="cita" type="lexical" extent="end"/>
<Event desc="el" type="language" extent="begin"/>
<Event desc="el" type="language" extent="end"/>
<Event desc="en" type="language" extent="begin"/>
<Event desc="en" type="language" extent="end"/>
<Event desc="fr" type="language" extent="begin"/>
<Event desc="fr" type="language" extent="end"/>
<Event desc="fático=A1" type="lexical" extent="instantaneous"/>
<Event desc="fático=A2" type="lexical" extent="instantaneous"/>
<Event desc="fático=E" type="lexical" extent="instantaneous"/>
<Event desc="fático=I" type="lexical" extent="instantaneous"/>
<Event desc="gl" type="language" extent="begin"/>
<Event desc="gl" type="language" extent="end"/>
<Event desc="ininteligible" type="pronounce" extent="instantaneous"/>
<Event desc="it" type="language" extent="begin"/>
<Event desc="it" type="language" extent="end"/>
<Event desc="palabra cortada" type="pronounce" extent="instantaneous"/>
<Event desc="pt" type="language" extent="begin"/>
<Event desc="pt" type="language" extent="end"/>
<Event desc="risa=A1" type="lexical" extent="instantaneous"/>
<Event desc="risa=E" type="lexical" extent="instantaneous"/>
<Event desc="risa=I" type="lexical" extent="instantaneous"/>
<Event desc="risa=todos" type="lexical" extent="instantaneous"/>
<Event desc="risas=A1" type="lexical" extent="begin"/>
<Event desc="risas=A1" type="lexical" extent="end"/>
<Event desc="risas=E" type="lexical" extent="begin"/>
<Event desc="risas=E" type="lexical" extent="end"/>
<Event desc="risas=I" type="lexical" extent="begin"/>
<Event desc="risas=I" type="lexical" extent="end"/>
<Event desc="risas=todos" type="lexical" extent="begin"/>
<Event desc="risas=todos" type="lexical" extent="end"/>
<Event desc="ruido de fondo" type="noise" extent="begin"/>
<Event desc="ruido de fondo" type="noise" extent="end"/>
<Event desc="ruido=boca" type="noise" extent="instantaneous"/>
<Event desc="ruido=carraspeo" type="noise" extent="instantaneous"/>
<Event desc="ruido=chasquido boca" type="noise" extent="instantaneous"/>
<Event desc="ruido=chasquido dedos" type="noise" extent="instantaneous"/>
<Event desc="ruido=estornudo" type="noise" extent="instantaneous"/>

<Event desc="ruido=golpe" type="noise" extent="instantaneous"/>
<Event desc="ruido=golpes" type="noise" extent="instantaneous"/>
<Event desc="ruido=grabadora" type="noise" extent="instantaneous"/>
<Event desc="ruido=indeterminado" type="noise" extent="instantaneous"/>
<Event desc="ruido=inspiración" type="noise" extent="instantaneous"/>
<Event desc="ruido=jadeo" type="noise" extent="instantaneous"/>
<Event desc="ruido=moto" type="noise" extent="instantaneous"/>
<Event desc="ruido=palmada" type="noise" extent="instantaneous"/>
<Event desc="ruido=palmas" type="noise" extent="instantaneous"/>
<Event desc="ruido=pasos" type="noise" extent="instantaneous"/>
<Event desc="ruido=pitido" type="noise" extent="instantaneous"/>
<Event desc="ruido=puerta" type="noise" extent="instantaneous"/>
<Event desc="ruido=reloj" type="noise" extent="instantaneous"/>
<Event desc="ruido=resoplido" type="noise" extent="instantaneous"/>
<Event desc="ruido=silbido" type="noise" extent="instantaneous"/>
<Event desc="ruido=sillas" type="noise" extent="instantaneous"/>
<Event desc="ruido=suspiro" type="noise" extent="instantaneous"/>
<Event desc="ruido=teléfono" type="noise" extent="instantaneous"/>
<Event desc="ruido=timbre" type="noise" extent="instantaneous"/>
<Event desc="ruido=tos" type="noise" extent="instantaneous"/>
<Event desc="ruido=voces" type="noise" extent="instantaneous"/>
<Event desc="sic" type="pronounce" extent="begin"/>
<Event desc="sic" type="pronounce" extent="end"/>
<Event desc="siglas" type="lexical" extent="end"/>
<Event desc="siglas=[aena]" type="lexical" extent="begin"/>
<Event desc="siglas=[ave]" type="lexical" extent="begin"/>
<Event desc="siglas=[bebecé]" type="lexical" extent="begin"/>
<Event desc="siglas=[bemeúve]" type="lexical" extent="begin"/>
<Event desc="siglas=[benegá]" type="lexical" extent="begin"/>
<Event desc="siglas=[bup]" type="lexical" extent="begin"/>
<Event desc="siglas=[cao]" type="lexical" extent="begin"/>
<Event desc="siglas=[cap]" type="lexical" extent="begin"/>
<Event desc="siglas=[ce e e]" type="lexical" extent="begin"/>
<Event desc="siglas=[cedé]" type="lexical" extent="begin"/>
<Event desc="siglas=[cedés]" type="lexical" extent="begin"/>
<Event desc="siglas=[cou]" type="lexical" extent="begin"/>
<Event desc="siglas=[dat]" type="lexical" extent="begin"/>
<Event desc="siglas=[dea]" type="lexical" extent="begin"/>
<Event desc="siglas=[diyéi]" type="lexical" extent="begin"/>
<Event desc="siglas=[diyéis]" type="lexical" extent="begin"/>
<Event desc="siglas=[efepé]" type="lexical" extent="begin"/>
<Event desc="siglas=[egebé]" type="lexical" extent="begin"/>
<Event desc="siglas=[eme pe tres]" type="lexical" extent="begin"/>
<Event desc="siglas=[erre]" type="lexical" extent="begin"/>
<Event desc="siglas=[eso]" type="lexical" extent="begin"/>
<Event desc="siglas=[eta]" type="lexical" extent="begin"/>

<Event desc="siglas=[fol]" type="lexical" extent="begin"/>
<Event desc="siglas=[geteí]" type="lexical" extent="begin"/>
<Event desc="siglas=[icona]" type="lexical" extent="begin"/>
<Event desc="siglas=[ies]" type="lexical" extent="begin"/>
<Event desc="siglas=[inserso]" type="lexical" extent="begin"/>
<Event desc="siglas=[insálud]" type="lexical" extent="begin"/>
<Event desc="siglas=[jo]" type="lexical" extent="begin"/>
<Event desc="siglas=[joc]" type="lexical" extent="begin"/>
<Event desc="siglas=[mec]" type="lexical" extent="begin"/>
<Event desc="siglas=[oenegé]" type="lexical" extent="begin"/>
<Event desc="siglas=[oenegés]" type="lexical" extent="begin"/>
<Event desc="siglas=[oequis]" type="lexical" extent="begin"/>
<Event desc="siglas=[oté]" type="lexical" extent="begin"/>
<Event desc="siglas=[pas]" type="lexical" extent="begin"/>
<Event desc="siglas=[peodé]" type="lexical" extent="begin"/>
<Event desc="siglas=[pepé]" type="lexical" extent="begin"/>
<Event desc="siglas=[pesoe]" type="lexical" extent="begin"/>
<Event desc="siglas=[petedé]" type="lexical" extent="begin"/>
<Event desc="siglas=[pre / se / gal]" type="lexical" extent="begin"/>
<Event desc="siglas=[prosegur]" type="lexical" extent="begin"/>
<Event desc="siglas=[sergas]" type="lexical" extent="begin"/>
<Event desc="siglas=[seu]" type="lexical" extent="begin"/>
<Event desc="siglas=[seus]" type="lexical" extent="begin"/>
<Event desc="siglas=[sida]" type="lexical" extent="begin"/>
<Event desc="siglas=[tóefl]" type="lexical" extent="begin"/>
<Event desc="siglas=[ucedé]" type="lexical" extent="begin"/>
<Event desc="siglas=[uci]" type="lexical" extent="begin"/>
<Event desc="siglas=[uesecé]" type="lexical" extent="begin"/>
<Event desc="siglas=[uned]" type="lexical" extent="begin"/>
<Event desc="siglas=[upegá]" type="lexical" extent="begin"/>
<Event desc="siglas=[urs]" type="lexical" extent="begin"/>
<Event desc="siglas=[uve hache ese]" type="lexical" extent="begin"/>
<Event desc="silencio" type="pronounce" extent="instantaneous"/>
<Event desc="tentativa de turno=A2" type="lexical" extent="instantaneous"/>
<Event desc="tentativa de turno=E" type="lexical" extent="instantaneous"/>
<Event desc="tentativa de turno=I" type="lexical" extent="instantaneous"/>
<Event desc="transcripción dudosa" type="pronounce" extent="begin"/>
<Event desc="transcripción dudosa" type="pronounce" extent="end"/>
<Event desc="término" type="lexical" extent="begin"/>
<Event desc="término" type="lexical" extent="end"/>
<Event desc="vacilación" type="pronounce" extent="instantaneous"/>
<Event desc="énfasis" type="pronounce" extent="begin"/>
<Event desc="énfasis" type="pronounce" extent="end"/>