ESLORA: Diseño, codificación y explotación de un corpus oral de español de Galicia

Victoria Vázquez Rozas

Universidade de Santiago de Compostela
victoria.vazquez@usc.es

Resumen

ESLORA es un corpus informatizado integrado por entrevistas semidirigidas y conversaciones espontáneas registradas en Galicia entre 2007 y 2014 como parte del *Proyecto para el Estudio Sociolingüístico del Español de Galicia* (*PRESEGAL*) que se desarrolla en la Universidad de Santiago de Compostela[1]. El diseño y elaboración del corpus cumple un doble objetivo: (i) recopilar y codificar muestras orales representativas de una variedad infradocumentada del español, y (ii) avanzar en el conocimiento de los métodos de construcción de corpus orales y en el desarrollo de recursos computacionales para la recuperación de la información. En esta presentación se exponen algunos aspectos de la construcción del corpus más directamente relacionados con sus posibilidades de explotación.

Palabras Clave: corpus oral, entrevista semidirigida, conversación, español de Galicia.

Introducción

El avance en los estudios sobre la lengua oral depende estrechamente de los conocimientos obtenidos mediante análisis de corpus. Por ello es necesario contar con corpus orales accesibles, adecuadamente documentados, transcritos y preferiblemente anotados.

La atención creciente al discurso oral en las últimas décadas del siglo XX por parte de las corrientes ligadas a la investigación social y antropológica confluyó con la expansión de la lingüística de corpus, lo que impulsó el registro y transcripción del habla y favoreció los análisis cualitativos de orientación funcional e interaccional en diversas situaciones de uso (conversación cotidiana, encuentros de servicios y comerciales, comunicación terapéutica, etc.). No obstante, las posibilidades de acceso a materiales orales son aún bastante limitadas si se comparan con las actuales condiciones de recuperación de información en corpus textuales escritos. La posición de desventaja de lo oral frente a lo escrito en lingüística de corpus es común a todas las lenguas y variedades, pero en la medida en que afecta especialmente a aquellas variedades y registros alejados de los usos considerados estándar constituye una traba insalvable para acceder a su (re)conocimiento y a su análisis, y en general al estudio de los hechos de variación y cambio desde un enfoque basado en el uso.

Algunos corpus generales del español ofrecen una parte de materiales orales, aunque en proporciones relativamente pequeñas: lo oral supone solo un 10% en CREA, y un 20% en el Corpus del español (pero este incluye como "oral" entrevistas y discursos en formato escrito). Más peso tendrá el habla en el CORPES XXI, que se enriquece con respecto al CREA con nuevas posibilidades de recuperación de información, si bien de momento no da acceso a la parte oral[2].

Por otro lado, hay corpus específicamente orales, entre otros el *Macrocorpus de la Norma Lingüística Culta*, el *Corpus Oral de Referencia del Español*

[1] Participan actualmente en el desarrollo del corpus ESLORA Mario Barcala, Marta Blanco, Eva Domínguez, Alba Fernández, Manuel Fernández, Pablo Gamallo, Francisco García, Marlén González, Sol López, Pia Poulsen, Montserrat Recalde, Guillermo Rojo, Paula Santalla, Carme Silva, Susana Sotelo y Victoria Vázquez.
El proyecto ha recibido financiación del Ministerio de Economía y Competitividad de España (FFI2010-17417). Página web: http://gramatica.usc.es/proxectos/eslora/

[2] Cf. http://web.frl.es/CORPES/view/inicioExterno.view.

Contemporáneo (CORLEC)[3], el *Corpus oral y sonoro del español rural* (COSER)[4], el *Proyecto de Estudio Sociolingüístico del Español de España y América* (PRESEEA)[5], el corpus *Valencia, Español Coloquial* (Val.Es.Co)[6], el *Corpus Oral de Lenguaje Adolescente* (COLA)[7], y el *Corpus oral didáctico anotado lingüísticamente* (C-Or-DiAL)[8], con diferentes características en cuanto a variedades representadas, amplitud de las muestras y posibilidades de acceso y de recuperación de información (Cf. tb. Briz & Albelda 2009). En conjunto, y a pesar del avance que suponen los proyectos citados, los materiales disponibles son escasos, parciales y de accesibilidad limitada, lo que resulta comprensible dada la gran inversión de tiempo y esfuerzo que supone la elaboración de un corpus oral.

En este contexto surgió la necesidad de elaborar un nuevo corpus con los objetivos generales de

i. incrementar la amplitud y variedad de los materiales disponibles;
ii. contribuir metodológicamente al estudio de los procedimientos de recopilación de registros orales;
iii. desarrollar nuevos recursos de tratamiento y recuperación de los datos del corpus.

1. Características generales del corpus

El corpus ESLORA está formado por entrevistas semidirigidas y conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2014.

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. 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 guión temático-secuencial estricto incluso en su distribución temporal. La contrapartida no deseable 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 para Labov 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 auténtico, 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 (cf. Recalde & Vázquez 2009).

El consentimiento informado de los participantes es un requisito legal para el registro de los datos, requisito 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 ha de mantener la coherencia interna que permita el seguimiento de la continuidad referencial del discurso. En el audio las referencias personales se eliminan y se sustituyen por un ruido, pero dado que la voz es un rasgo identificador del hablante, el acceso a las grabaciones se condiciona al cumplimiento de la finalidad investigadora declarada en los formularios de consentimiento.

2. Metadatos

Un componente fundamental del corpus es el conjunto de metadatos correspondiente a cada grabación. En ESLORA se registra de forma sistemática información sobre

i. la situación del evento grabado: fecha, país, ciudad, localización y circunstancias concretas;
ii. los participantes (edad, sexo, estudios, profesión, lugar de nacimiento), su rol en el intercambio, en el caso de la entrevista semidirigida

[3] http://www.lllf.uam.es/ESP/Corlec.html
[4] http://www.lllf.uam.es:8888/coser/
[5] http://preseea.linguas.net/
[6] http://www.uv.es/corpusvalesco/
[7] http://www.colam.org/om_prosj-espannol.html
[8] http://lablita.dit.unifi.it/corpora/cordial

(entrevistador/a, informante, audiencia) y la relación entre ellos (desconocidos previamente, relación de amistad o parentesco);

iii. el tipo de interacción: entrevista, conversación;

iv. 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 propiedades del evento y 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 datos sociológicos y las declaraciones de los hablantes sobre sus usos y actitudes lingüísticas (cf. http://gramatica.usc.es/proxectos/presegal/att/Cuestionario.pdf).

La muestra de entrevistas incorpora también una prueba de inseguridad lingüística con el objetivo de examinar críticamente este instrumento clásico del método variacionista (cf. Labov 1966: 474-481 y 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 (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. Elaboración del corpus

3.1. 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 (como *Tape-a-Talk*).

La transcripción de las grabaciones se realizó de forma manual con ayuda de los programas Transcriber y 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 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 codificación 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[9]. 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

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

una función interactiva fundamental para el desarrollo normal de las conversaciones, por lo que deben registrarse adecuadamente en texto. Sin embargo, su categorización discreta en el molde escrito oscurece con frecuencia las variaciones de forma y de contenido pragmático que manifiestan en el uso. 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.

3.2. Etiquetación

El sistema de transcripción del corpus ESLORA incluye, además de la representación ortográfica de las unidades verbales, un conjunto limitado de marcas y etiquetas que informan sobre algunas características lingüísticas, paraverbales y contextuales que se consideran comunicativamente relevantes. La selección de las etiquetas del subcorpus de entrevistas está condicionada por los acuerdos adoptados en PRESEEA (vid. nota 9 supra). La lista incluye, además de las marcas de pausa y los signos de interrogación y admiración, etiquetas de ruidos (comunicativos y ambientales), de risas, de forma fónica (alargamiento, palabra cortada, énfasis, vacilación, transcripción dudosa, ininteligible), de léxico (término, siglas), de cita, de lengua (gallego, inglés, portugués, etc.) y una etiqueta de observaciones que recoge información contextual o explicaciones complementarias necesarias para comprender la transcripción y la grabación.

Transcriber proporciona un sistema cómodo para la introducción de etiquetas compatibles con PRESEEA, bien modificando o sustituyendo las que vienen por defecto, bien creando directamente otras nuevas. Además, el programa agiliza el proceso de transcripción con el uso de combinaciones de teclas para añadir las etiquetas, 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[10].

La interfaz de trabajo de Transcriber y las facilidades que ofrece para la introducción de etiquetas se adecua bien a un sistema de codificación diferencial con anotaciones verbales. La figura 1 muestra una captura parcial de la interfaz del programa mientras que en la figura 2 se ve el mismo fragmento en formato XML:

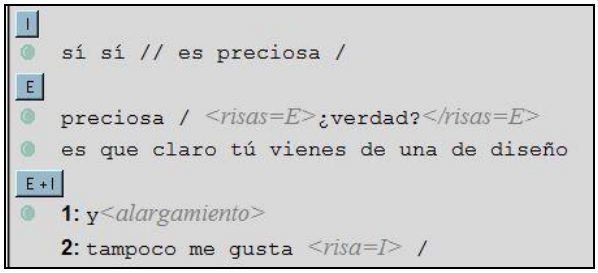

Figura 1. Fragmento de la interfaz de Transcriber

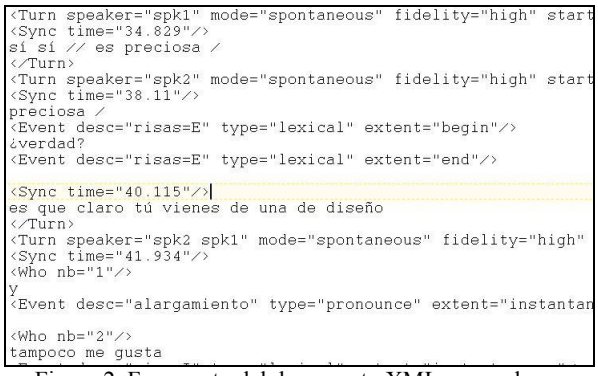

Figura 2. Fragmento del documento XML generado en Transcriber

En las figuras 1 y 2 puede observarse el carácter verbal y diferencial de las etiquetas "risas" (de apertura y de cierre), "risa" y "alargamiento" (estas últimas instantáneas), pero también se ve el uso de indicaciones de tipo simbólico integradas en la transcripción, como las barras inclinadas "/" y "//" que marcan pausas.

Las transcripciones realizadas con ELAN incluyen prácticamente el mismo conjunto de etiquetas que las de Transcriber, pero dadas las características del programa y el sistema previsto de explotación (cf. infra apdo. 4), para la fase de transcripción se optó por un sistema de representación más simbólico que verbal. Aunque ELAN es una herramienta que prevé la utilización de líneas o niveles dependientes de anotación para cada línea principal, por el momento en ESLORA las intervenciones de cada hablante se transcriben en la línea principal correspondiente y en ella se incluyen también las indicaciones lingüísticas y paralingüísticas relevantes. Solo reciben codificación en línea independiente las pausas que superan las dos décimas de segundo y las observaciones contextuales aclaratorias necesarias para interpretar la conversación. La figura 3 recoge una parte de la pantalla de ELAN con las líneas principales de tres hablantes (H1, H2 y H3), las pausas y la línea de observaciones. Nótese que los signos <@> y </@> funcionan como las etiquetas de apertura y cierre de "risas", que en Transcriber (XML) se representaban, respectivamente, como <Event desc="risas=X" type="lexical" extent="begin"/> y <Event desc="risas=X" type="lexical" extent="end"/> (donde

[10] No obstante, la transferencia de los archivos de Transcriber al formato requerido por el proyecto PRESEEA no es directa. De hecho, no se usan los archivos TRS sino la opción de exportación a documento TXT que ofrece el programa, y a ese TXT se aplican unas rutinas de reconversión y control que afectan tanto a las cabeceras de metadatos como a las etiquetas.

X representa la inicial del hablante). Por su parte, el signo "=" representa la etiqueta XML de Transcriber <Event desc="alargamiento" type="pronounce" extent="instantaneous"/>.

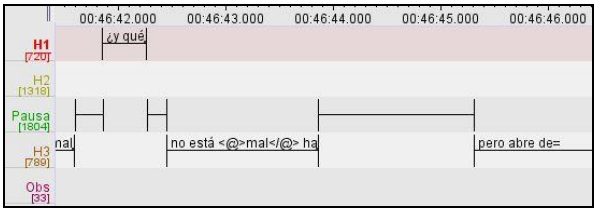

Figura 3. Fragmento de la pantalla de ELAN en modo de "Anotación"

4. Recuperación de información

La transcripción manual de las entrevistas y conversaciones que forman ESLORA incorpora, pues, además de las marcas de sincronización texto-voz, una serie de indicaciones sobre la interacción y sus actores (metadatos, identificación de hablantes, segmentos, solapamientos) y sobre ciertas características lingüísticas, paraverbales y contextuales de los registros de habla. Las marcas y etiquetas introducidas junto con la codificación estándar ortográfica de los textos permiten recuperar información textual y sonora de forma automática formulando búsquedas simples o combinadas de cadenas textuales, marcas y etiquetas. No obstante, se ha considerado necesario enriquecer el texto para facilitar su explotación en aplicaciones lingüísticas y computacionales. Para ello se están llevando a cabo tareas de procesamiento de los materiales que incluyen, en primer lugar, la conversión de las marcas y etiquetas de entrevistas y conversaciones a un formato de representación XML unificado y, seguidamente, la lematización y etiquetación morfosintáctica de los textos. En tanto no está disponible un etiquetador de elaboración propia adaptado a textos orales, se ha utilizado provisionalmente el analizador FreeLing (Padró *et al*. 2010). En el momento de redactar esta presentación, ESLORA dispone ya de una versión en pruebas de la aplicación de búsqueda que comprende 250.278 unidades gramaticales del subcorpus de entrevistas (informantes con educación universitaria), accesible en <http://galvan.usc.es/eslora>.

5. Recapitulación y perspectivas

ESLORA es un corpus que documenta el uso de una variedad poco estudiada -y hasta cierto punto marginada- del español: el español hablado en Galicia. Su registro en un corpus y las posibilidades de análisis que se abren con él contribuyen a legitimar esta variedad como objeto de estudio lingüístico, a conocerla y a reconocerla al lado de otros dialectos del español. Aparte de este reconocimiento, las principales ventajas y posibilidades de explotación del corpus ESLORA derivan del detalle, la sistematicidad y la fiabilidad de los datos que aporta. Por una parte, incluye la necesaria información contextual en forma de metadatos y de anotaciones textuales. Por otra parte, permite el acceso directo a los registros de audio, con la consiguiente posibilidad de contrastación empírica de la transcripción y etiquetación que se ofrece. Además, el corpus se acompaña de herramientas de búsqueda y recuperación de información que permiten explotar los materiales codificados en diferentes áreas lingüísticas, tanto de orientación descriptiva como aplicada. Por último, el corpus ofrece información relevante sobre técnicas de recopilación de datos lingüísticos y sociológicos (entrevistas, conversación espontánea, cuestionarios, sistemas de transcripción) y por tanto constituye una contribución metodológica al desarrollo y mejora de los sistemas de recogida y procesamiento del habla.

Referencias

Briz Gómez, A. & M. Albelda. 2009. Estado actual de los corpus de lengua española hablada y escrita: I+D". En *El español en el mundo. Anuario del Instituto Cervantes 2009*. <http://cvc.cervantes.es/lengua/anuario/anuario_09/briz_albeida/p01.htm >

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington, D.C.: Center of Applied Linguistics. 2ª ed. Cambridge: CUP, 2006.

Padró, L., M. Collado, S. Reese, M. Lloberes & I. Castellón. 2010. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA* La Valletta, Malta. May, 2010. <http://nlp.lsi.upc.edu/freeling>

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, M. & V. Vázquez Rozas. 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. <http://www.um.es/lacell/aelinco/contenido/titulos.html>

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