Desafíos y reflexiones sobre utilización de corpus libres

El caso de PRESEEA en el estudio de la reformulación

Luisa Fernanda Acosta Cordoba
UMR 5191 ICAR, 3LA Université Lyon 2

Our main purpose is to present our experience working on a free spoken-language corpus in Spanish. As part of my doctoral thesis project, concerned with rephrasing used in spoken interaction, I decided to study this phenomenon by analyzing the data available in the PRESEEA corpus. In order to annotate a part of this corpus, we created a TextGrid file on PRAAT associated with each audio file. Then, we imported those files onto ELAN to start the annotation process, with categories related to five language levels relevant for our research: prosodic, syntactic, macrosyntactic, interactional and rephrasing phenomena. In addition to sharing our methodological work experience, we would like to emphasize the importance of free spoken corpora for the development of the research in linguistics.

Keywords: spoken language, free corpora, corpus segmentation, rephrasing.

1. Introduccion: el desarrollo y la utilizacion de corpus para el estudio de la lengua hablada

Sin duda esenciales para las ciencias del lenguaje, el desarrollo y la utilización de corpus supone imperativos metodológicos particulares cuando se trata del estudio de la lengua hablada (Traverso, 2008). Así, si en el caso de la lengua escrita existen numerosos recursos a partir de producciones literarias, periodísticas o académicas, para el estudio de la lengua oral los recursos son considerablemente limitados, debido tanto a las dificultades en la constitución de corpus, como a ciertas políticas de investigación que restringen el acceso libre a los datos. Esta situación no es exclusiva a la lengua española, sin embargo, respecto a otras lenguas europeas, hemos podido constatar una mayor dificultad en el acceso a cierto tipo de corpus.

CHIMERA. Romance Corpora and Linguistic Studies 5.2 (2018), 267-273. ISSN 2386-2629
© 2018 The Author; distributed under the Creative Commons Attribution License

El objetivo de este artículo es presentar brevemente nuestra experiencia de investigación a partir de un corpus ya constituido. Este artículo se inscribe tanto en el trabajo de tesis de la autora[1], como en los estudios adelantados en el proyecto ANR SegCor[2]. Así, queremos hacer énfasis en el interés que supone la adaptación de un corpus constituido a un nuevo proyecto, cuyas perspectivas de investigación transforman la orientación y la interpretación de los datos originales. De igual manera, buscamos resaltar la importancia y la inmensa utilidad de poner a disposición del público académico corpus libres, condición que consideramos primordial para el avance de la investigación lingüística española.

El atributo principal de la constitución integral de un corpus oral es la posibilidad de adaptar, según los intereses de estudio del investigador, los diferentes factores ligados al trabajo de campo. Así, las variedades lingüísticas representadas (la población de estudio: edad, género, sociolecto, etc.), la gestión de la interacción (entrevista dirigida, conversación informal, soportes visuales, etc.) y las condiciones técnicas de grabación (cabina insonorizada, lugar público, etc.) corresponden a expectativas específicas y pueden ser orientadas en las etapas tempranas de la investigación. El trabajo de campo supone, sin embargo, una movilización significativa de los medios de trabajo y una inversión importante de tiempo, que pocos jóvenes investigadores pueden permitirse. Asimismo, el estudio de incontables fenómenos lingüísticos, como la reformulación, puede realizarse sin que sea necesario partir de un corpus diseñado exclusivamente para tal efecto.

Partir de un corpus ya constituido en la primera etapa de nuestra investigación representa, en primer lugar, una optimización importante del tiempo y los recursos a nuestra disposición. En segundo lugar, esta decisión nos permite conducir una reflexión sobre el objeto dinámico en que puede convertirse un corpus pues, como lo indica Véronique Traverso (2008), más que un conjunto estático de grabaciones y otros materiales, es decir de *données primaires*, un corpus cuenta también con *données secondaires*: "*le corpus est aussi le résultat d'une série d'interventions du chercheur, depuis le terrain jusqu'à la réalisation des représentations des*

[1] Tesis en curso en la Université Lyon 2, financiada por la Escuela doctoral 3LA (http://3la.univ-lyon2.fr/).

[2] El objetivo del proyecto SegCor (*Segmentation de corpus oraux*, ANR-15-FRAL-0004) es la anotación de múltiples niveles (chunk, sintaxis, macrosintaxis e interacción) de un corpus oral francés (corpus Clapi, ICAR UMR 5191, http://clapi.icar.cnrs.fr/ y ESLO, LLL UMR 7270, http://eslo.huma-num.fr/) y alemán (corpus FOLK, Institut für Deutsche Sprache, http://agd.ids-mannheim.de/folk.shtml). Una parte de los objetivos y de la metodología de trabajo del proyecto SegCor ha sido aplicada a mi trabajo de tesis. No obstante, el corpus del presente estudio no hace parte del corpus de trabajo del proyecto SegCor.

données sous forme de transcriptions" (Traverso, 2008: 315). En resumen, defendemos la noción del corpus como un objeto de estudio cambiante gracias a las perspectivas de estudio que diferentes investigadores aportan en las etapas de tratamiento y análisis.

2. Corpus PRESEEA

Entre los corpus libres y en línea disponibles en español, hemos escogido PRESEEA (*Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*) por dos razones centrales: la posibilidad de descargar los archivos (audio en formato Waveform Audio Format y transcripción en formato texto simple) y la diversidad de variedades dialectales presentes en el corpus. Pese a que otros corpus españoles presentan situaciones interaccionales más simétricas y, en consecuencia, más pertinentes con nuestro enfoque conversacional, es un gran impedimento no poder contar con los archivos para su tratamiento informático. Asimismo, en nuestro proyecto de tesis deseamos explorar diversas variedades dialectales del mundo hispánico y, desafortunadamente, una buena parte de los otros corpus presentan pocas variedades.

El proyecto PRESEEA se inspira en una concepción colaborativa de corpus de lengua hablada, con un énfasis particular en las variables sociolingüísticas. Así, los locutores son seleccionados y clasificados según sus características sociales: género, edad, nivel de estudios y origen geográfico. Las entrevistas son orientadas en función de temas generales que buscan producir secuencias tanto narrativas, descriptivas como argumentativas. Los equipos de las universidades que componen el proyecto tiene un margen de decisión respecto al material utilizado en la captura del audio y en la gestión de la entrevista, margen que hemos podido constatar al consultar los audios disponibles. En efecto, la calidad de los audios y la espontaneidad del registro lingüístico son disimiles entre los equipos. Dado que queremos dar cuenta, entre otros, de los niveles prosódico e interaccional, escogimos los puntos cuyos audios fueran de mejor calidad y cuyas interacciones fueran lo más espontáneas posible. Así, dos ciudades nos resultaron satisfactorias para el estudio piloto: México D.F. y La Habana.

3. Tratamiento informático y análisis de datos en PRAAT y ELAN

El objetivo del tratamiento informático que proponemos es poder contar con una herramienta de segmentación, anotación y consulta de corpus capaz de contener diferentes caracterizaciones de una misma producción lingüística. En otros

términos, buscamos, por ejemplo, poder asociar a una unidad sintáctica, un tipo de contorno enunciativo o un tipo de acción conversacional. El interés principal es lograr identificar fenómenos concurrentes o divergentes e, igualmente, caracterizar ciertas unidades mayores, como el turno de conversación, en función de otras, como las unidades sintácticas. Se trata entonces de una concepción multinivel del corpus, que nos permitirá analizar la interrelación de diferentes variables que intervienen en la aparición y realización de la reformulación.

Para el tratamiento del corpus partimos de algunos materiales disponibles en línea. El sitio de internet del proyecto PRESEEA permite descargar diez minutos de audio de las entrevistas, cuya duración total es de una hora aproximadamente. Además de los criterios ya evocados, en la selección de las entrevistas ningún otro fue empleado, ni de orden temático ni de orden sociolingüístico. Para la fase piloto, seleccionamos tres entrevistas de México D.F. y dos de La Habana[3].

3.1 Alineación entre la transcripción y el audio en PRAAT

La alineación entre el audio y la transcripción es un procedimiento indispensable, pues permite crear anotaciones y segmentaciones en un mismo software. En esta etapa utilizamos el software PRAAT (Boersma y Weenink, versión 6.0.37) para la creación, a partir de la transcripción en texto simple, de un archivo tipo TextGrid alineado con cada audio.

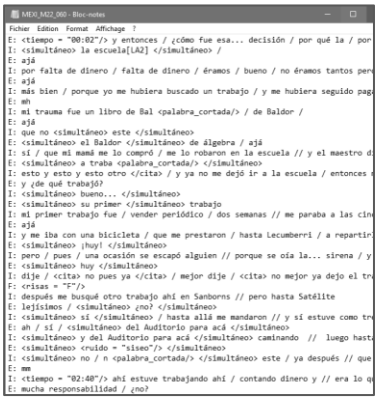

Figura 1. Transcripción texto simple

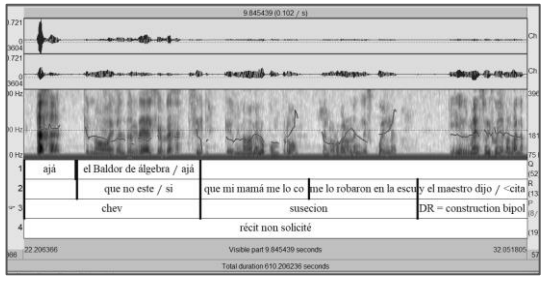

Figura 2. Captura de pantalla de TextGrid en PRAAT

La figura 1 muestra la trascripción en su estado bruto, sin ningún vínculo automático con el audio. En la figura 2, tenemos la misma transcripción, esta vez

[3] MEXI_M22_060, MEXI_H22_054, MEXI_H11_078, LHAB_H22_049 y LHAB_M22_055.

alineada en PRAAT. Las dos primeras líneas corresponden a las dos locutoras presentes en la interacción. En las dos últimas, anotamos diferentes fenómenos, a partir de los cuales creamos una lista fija de categorías pertinentes para el estudio de este corpus. Notemos que una de las ventajas del archivo TextGrid es la posibilidad de indicar con precisión los solapamientos de los turnos de conversación, anotados como 'chev' en la imagen dos. Precisión que es imposible de obtener en un archivo de texto tradicional. En resumen, el paso de la transcripción de texto simple a TextGrid nos permite recuperar información ligada a la progresión lineal del discurso.

3.2 Segmentación y anotación en ELAN

Una vez creado el TextGrid, procedimos a su importación a ELAN (Max-Planck-Institute for Psycholinguistics, versión 5.2). Este software permite asociar diferentes categorías a un mismo segmento discursivo, gracias a un sistema de múltiples líneas, que observamos en el costado izquierdo de la Figura 3. Las líneas

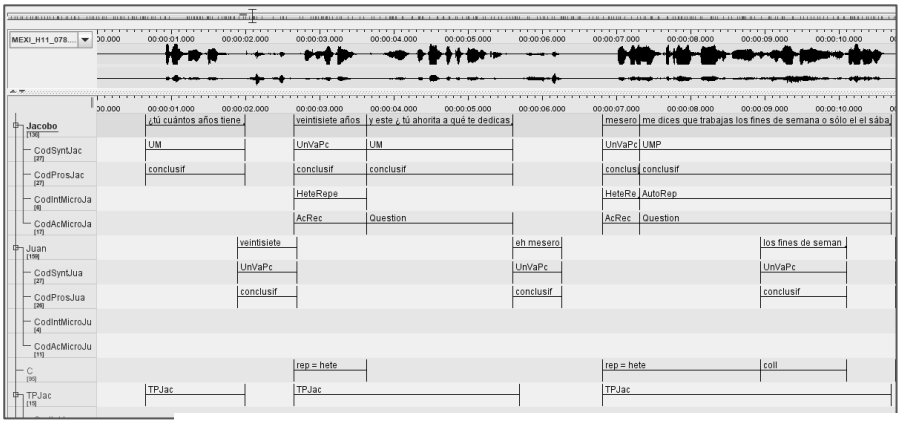

Figura 3. Captura de pantalla de anotación en ELAN

'Jacobo' y 'Juan', contienen la transcripción hecha en PRAAT, las otras corresponden a un nivel de anotación y contienen categorías de codificación fijas. Así, por ejemplo, a la línea 'CodAcMicroJac' corresponde etiquetas ligadas a las acciones conversacionales (Selting, 2000): pregunta, respuesta, marca de recepción, etc.

De igual manera, hemos establecido etiquetas para otros niveles: la sintaxis (unidad completa, no verbal, abandonada, etc.), la prosodia (contorno conclusivo, suspensivo, etc.), la macrosintaxis (Blanche-Benveniste, 2010), los fenómenos

interacccionales (reparación, construcción incremental, etc. Traverso, 2007), los fenómenos sintácticos (Feuillard, 1989) y la reformulación. Finalmente, ELAN también permite crear una macro unidad, en este caso 'TP', turno de conversación, la cual puede incluir varios segmentos y su propia codificación.

4. Perspectivas de estudio

El proceso de anotación es un trabajo laborioso que exige una aguda observación de los datos, especialmente para establecer un sistema de etiquetas fijas y funcionales. El interés de la sistematización de dicho proceso es poder utilizar las herramientas de búsqueda de ELAN para recuperar todas las ocurrencias de una etiqueta y, de igual manera, las concurrencias entre varias categorías. Una tarea que manualmente sería dispendiosa y poco precisa.

Nuestro estudio no solo ofrece una nueva perspectiva metodológica sobre el corpus, sino también una pista de análisis pues, por ejemplo, en el trabajo con la entrevista MEXI_H11_078 hemos hallado recursos particulares en la gestión interaccional de la entrevista. Estos aspectos, que serán el objeto de futuros trabajos, no han sido analizados, según los trabajos que hemos podido consultar, por el equipo PRESEEA de México.

Agradecimientos

Agradezco al proyecto PRESEEA y a sus políticas de contenidos, particularmente a los equipos del Colegio de México y de la Universidad de La Habana. Agradezco igualmente el apoyo de mis directoras y a mi asesora de tesis, las profesoras Nathalie Rossi-Gensane, Véronique Traverso e Isabel Colón de Carvajal. Asimismo, quiero agradecer la ayuda de mis colegas Elizaveta Chernyshova, Biagio Ursi y Carole Etienne.

Referencias bibliográficas

Blanche-Benveniste, C. 2010. *Approches de la langue parlée en français*. Paris: Ophrys.

Colón de Carvajal, I. 2013. *Guide practique pour annoter sur ELAN*. Publicado por la Cellule de Corpus Complexes del Laboratorio ICAR UMR 5191. http://www.icar.cnrs.fr/ccc/ressources/ (consultado en junio, 2018).

Feuillard, C. 1989. La Syntaxe fonctionnelle dans le cadre des théories linguistiques contemporaines. Thèse d'État, Université Paris V.

PRESEEA : *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*. Alcalá de Henares: Universidad de Alcalá. http://preseea.linguas.net/ (consultado abril, 2018).

Selting, M. 2000. The construction of units in conversational talk. *Language in Society* 29: 477-517.

Traverso, V. 2008. Analyser un corpus de langue parlée en interaction : questions méthodologiques. *Verbum* XXX(4), 313-328.

Traverso, V. 2007. *Analyser les conversations*. Paris: Armand Colin.