Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU

Francisco Moreno Fernández
(editor)

2

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU

Francisco Moreno Fernández
(editor)

OBSERVATORIO
de la lengua española y las culturas hispánicas en los Estados Unidos

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto *CORPEEU*

Francisco Moreno Fernández

(editor)

Jorge Covarrubias

Domnita Dumitrescu

Andrés Enrique Arias

Andrew Lynch

Francisco Moreno-Fernández

Ricardo Otheguy

María Luisa Parra

Gerardo Piña-Rosales

Carmen Silva-Corvalán

OBSERVATORIO de la lengua española y las culturas hispánicas en los Estados Unidos

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Introducción

El “Corpus del Español en los Estados Unidos” (CORPEEU) es un proyecto iniciado en el “Observatorio de la lengua española y las culturas hispánicas” del Instituto Cervantes en la Universidad de Harvard, con la colaboración de la Academia Norteamericana de la Lengua Española (ANLE). Los trabajos para la construcción del CORPEEU comenzaron en el otoño boreal de 2017, dentro de la Universidad de Harvard, pero previamente se hizo necesaria una detenida reflexión sobre las implicaciones y dificultades, conceptuales y metodológicas, que suponía abordar una tarea tan compleja.

Con el fin de reflexionar sobre el modo de proceder a la construcción de un corpus del español en los Estados Unidos, el Observatorio del Instituto Cervantes en Harvard convocó un debate a puerta cerrada entre nueve especialistas en el español en los Estados Unidos, conocedores de su realidad desde diversas perspectivas: lingüística, sociológica, literaria, periodística, educativa, académica, histórica y política. La reunión de expertos se celebró en la sede del Observatorio el día 22 de septiembre de 2017.

Previamente al encuentro en Harvard, el Observatorio había distribuido una documentación que pretendía servir de base para las discusiones. En ella se incluía un borrador preliminar de proyecto, información sobre la configuración interna del *Corpus del Español del siglo XXI* (CORPES XXI), de la Real Academia Española, el detalle de las fuentes estadounidenses manejadas hasta ese momento para los corpus académicos, información sobre los textos de origen estadounidense incluidos en el *Corpus del Español* de Mark Davies, así

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

como el detalle de todas sus fuentes, e información sobre diversos proyectos de recopilación de materiales del español hablado en los Estados Unidos.

Este documento reproduce la conversación mantenida por los mencionados expertos en el Observatorio del Instituto Cervantes, si bien se han añadido algunas anotaciones y referencias bibliográficas que ayudarán a identificar las citas y alusiones realizadas. El texto respeta, en líneas generales, la literalidad de las intervenciones de cada participante, si bien se han modificado aquellos aspectos propios del coloquio que podrían entorpecer la comprensión de los discursos, se han aclarado enunciados que podrían llevar a interpretaciones inadecuadas, se han corregido algunos errores evidentes y se han suprimido o modificado los elementos conversacionales que resultaban ajenos al principal centro de interés: la construcción de un corpus del español hablado y escrito en los Estados Unidos.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Debate: hacia un corpus del español en los Estados Unidos

Francisco Moreno Fernández [FMF]. Muchas gracias por acudir a esta reunión, en la que tenemos puesta mucha ilusión, mucha esperanza; una reunión organizada por el Observatorio del Instituto Cervantes en la Universidad de Harvard, pero con la implicación directa de la Academia Norteamericana de la Lengua Española. Y que quiere ser un foro de reflexión teórica que lleve a propuestas prácticas, en la medida que sea posible para la creación de un corpus del español de los Estados Unidos, que podría tener relación con la investigación y el conocimiento lingüístico y sociolingüístico generado desde el mundo universitario, así como con la actividad de la Academia Norteamericana y las Academias de la Lengua en general.

Se trata de algo que consideramos, si no una necesidad absoluta, sí algo conveniente en relación con el estudio del español: la elaboración de un corpus, de una colección de materiales que nos permita conocer el español, su estado actual y futuro, incluso pasado, en los Estados Unidos; con el fin de poder abordar distintos tipos de trabajos de investigación, de estudios, de aplicaciones didácticas, de aplicaciones lexicográficas. En fin, ya sabéis que las aplicaciones de los corpus son múltiples y no siempre previsibles.

Por eso decidimos organizar esta reunión, respondiendo también una petición que se nos hizo desde la Real Academia Española, que desde hace dos décadas viene trabajando con corpus del español. La Academia tiene dos grandes corpus de referencia, que intentan recoger materiales a gran escala del español en sus

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

diversos territorios. Uno es el Corpus de Referencia del Español Actual, el CREA, que abarca desde el año 1975 hasta el año 2000; el otro es el CORPES XXI, el Corpus del Español del siglo XXI, que abarca desde el año 2000 hasta la actualidad. Cuando la Academia Española toma decisiones, por ejemplo, sobre qué usos presentar en su gramática, sobre qué entradas incluir en el diccionario o qué acepciones suprimir, siempre se hacen consultas; siempre hay un trabajo lingüístico de base, antes de tomar una decisión, por el que se crea una documentación bastante compleja sobre el estado de los usos lingüísticos que interesan en un momento determinado.

Y desde la Academia Española, con el fin de que los trabajos que se afronten en los distintos países hispanohablantes tenga también un respaldo científico o, al menos, un respaldo documental, testimonial, están animando a crear colecciones de materiales que complementen la gran colección de la Real Academia Española, de tal manera que la información general sobre el uso del español no dependa solo de los materiales reunidos desde Madrid, sino que haya también una aportación directa de cada uno de los países hispanohablantes, dado que la realidad y la visión del español en cada territorio tiene sus particularidades.

Por poner un ejemplo: la Academia Mexicana de la Lengua se está planteando lo mismo para el español mexicano. Intentan crear un corpus del español mexicano, pero visto desde México; es decir, atendiendo a las variedades y a la situación del español en México, lo que incluye el contacto con lenguas indígenas, que es una de sus particularidades, no por las lenguas indígenas en sí mismas, que existen en otros muchos lugares, sino porque el modo en que se manifiestan en México

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

afecta a la particularidad del español de México. Por eso los materiales reunidos deben reflejar esa situación, para un mejor conocimiento general y para una mejor documentación local o nacional de lo que está ocurriendo con el español en México, aunque ya existe un corpus elaborado con fines lexicográficos desde El Colegio de México.

Es decir, en cada país se están tomando decisiones que adecuan o que adaptan a los distintos contextos la necesidad teórica de un corpus. Y, si compleja es la situación de México a la hora de tomar decisiones sobre qué incluir en un corpus mexicano, qué no incluir, cómo hacerlo, qué debe quedar representado, pues, imaginad lo que puede ocurrir con el español en Estados Unidos, donde la complejidad, si cabe, es mucho mayor, por la convivencia con el inglés y por otras claves que conocéis mejor que nadie.

Ese es el sentido de esta reunión. La idea es que podamos sacar de aquí algunas guías, algunas orientaciones, algunas reflexiones que nos permitan abordar trabajos en esa dirección. No sabemos si el producto final será un gran corpus o será otra cosa, porque es difícil crear desde cero un gran corpus, entre otras razones, porque se necesita mucho tiempo y mucho dinero, y no es algo que se pueda conseguir de manera rápida y sin problemas. Pero sí al menos podremos apuntar pistas que nos permitan poco a poco ir acumulando conocimiento y construir una base suficiente para nuestros fines, desde cada uno de nuestros ámbitos. Ese es el sentido de nuestra reunión.

Envié una documentación previa con la idea de que sirviera de base para articular esta reunión. Podemos utilizarla como guía, si os parece bien. Pero el objetivo es

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

que podamos dialogar e intercambiar opiniones de manera mucho más libre que en una mesa redonda, para que el intercambio sea mucho más fructífero y dinámico.

Si fuera necesario, podemos intervenir por turnos, para que nadie pierda la oportunidad de expresar su opinión. En la documentación previa propongo una serie de cuestiones muy generales, como a partir de qué momento puede hablarse de la existencia de un español estadounidense, para incluir una dimensión diacrónica en la colección de materiales, o en qué áreas y con qué límites se distribuye ese español estadounidense, qué rasgos, qué perfil deberían tener los hablantes de Estados Unidos para ser considerados como representativos de un español estadounidense. Y lo mismo en cuanto a los dominios y los contextos de uso. Es decir, se trata de hablar de cuestiones diacrónicas, de cuestiones estilísticas, de cuestiones sociolingüísticas y de cuestiones lingüísticas, para trazar el perfil general de lengua española en los Estados Unidos.

Pero no sé si, antes de entrar en estas cuestiones más concretas, cabría hacer una reflexión primera sobre cómo concebir la necesidad o la conveniencia de abordar un corpus. ¿Se necesita un corpus del español estadounidense? Estoy hablando desde el punto de vista casi de una filosofía de la investigación y de una metodología. ¿Cómo se valora la perspectiva de trabajar en un corpus, al margen de otros aspectos concretos?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Jorge Covarrubias [JC]. Yo quisiera decir, como mi enfoque es el periodístico, que parte de la documentación que podemos buscar sería de la prensa en español en Estados Unidos a partir de 1801, en la primera década del 1800, cuando aparece el primer periódico que creo que se llamaba *El Mississippi*. Está en mi informe. No recuerdo exactamente, pero era en la primera década de 1800. Entonces, creo que el aporte de la prensa hispanounidense sería ese. Empezar a buscar la documentación a partir del primer periódico, de *El Mississippi*, en la primera década del 1800.

Carmen Silva-Corvalán [CSC]. Sería un corpus que recogería realizaciones del español en Estados Unidos. No sería un corpus del español estadounidense.[1] O sea, a mí me parece que es una diferencia más o menos importante. No sé si estamos de acuerdo en que, por ejemplo, yo no represento el español estadounidense; Andrés [Enrique Arias] mucho menos; claro que ha vivido mucho en Estados Unidos, pero no es el español estadounidense. Entonces, lo mismo pasaría con manifestaciones en la prensa. Por dos razones: uno, no sabemos quién estaba escribiendo en la prensa de 1801. Incluso hoy en día no sabemos quiénes son muchas veces los que escriben en la prensa, en revistas. Eso es lo mismo que pasa en cualquier otro país. En Chile también: uno lee muchos artículos en la prensa chilena y resulta que son traducciones enviadas desde otro país.

[1] La contraposición entre un "español estadounidense" y un "español en los Estados Unidos" se trata de forma paralela y similar en el debate editado por Moreno Fernández (2018).

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Entonces, yo no sé de qué manera uno puede llegar a un acuerdo sobre qué es el español estadounidense y cuál es. Es difícil.

Ricardo Otheguy [RO]. Siguiendo exactamente con esa manera de pensar, yo tengo una pregunta, más que un comentario. Y es ¿qué opinión tienen ustedes sobre lo que ya se ha hecho en los otros países y hasta qué punto podemos imitar lo que se ha hecho? Y si estas preguntas ya han tenido respuesta en alguna parte, entonces podemos decir, bueno, vamos a hacer lo que se ha hecho en España, lo que se ha hecho en México. Que, claro, tampoco pueden hablar de un español de España y un español de México por todas esas limitaciones, que aquí son más. Pero yo quisiera saber qué opinión nos merecen esos esfuerzos y hasta qué punto lo que tenemos que hacer es seguir sus pautas o hasta qué punto no. Eso es lo que yo quisiera entender un poco más porque yo todavía eso no lo tengo claro.

Gerardo Piña-Rosales [GPR]. Yo creo que en realidad tampoco nos importa tanto que ese español sea de Estados Unidos. Simplemente es español que se ha hablado aquí, que se ha escrito aquí, en Estados Unidos, independientemente de que sea estadounidense. Lo que sí hay que aclarar precisamente es el criterio que tú decías del periodismo. ¿Cuándo empezar? Y también propongo en este caso, ya que existe un corpus del español de México, que nos aprovechemos de ese corpus porque, al fin y al cabo, el 60% de los hispanos en este país son de origen mexicano. O sea que no partimos de nada; no partimos de cero. Creo que eso es importante.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Pero, ¿qué sería aprovecharnos de ese corpus?

GPR. Bueno, que ellos nos mandaran o nos dieran una serie de pautas. Y que después nosotros simplemente vaciemos lo que nos interese de ese corpus. Nada más, para no partir de cero. Pero creo lo que importante es ver dónde empezar; o sea, cuál es el criterio realmente.

Domnita Dumitrescu [DD]. Yo tengo una pregunta. Esto que usaron ellos [la Real Academia Española], ¿quién es el que dio estos datos [sobre Estados Unidos]? O sea, lo que ya incluyeron en el CREA, en el CORPES y en el corpus complementario de la *Nueva Gramática*. Ustedes vean aquí: hay un montón de trabajos que han sido incluidos.[2]

GPR. Yo la verdad es que le mandé a Guillermo Rojo una bibliografía de estudios, sobre todo de literatura, novelas, cuentos, poesía publicadas aquí en Estados Unidos.

DD. ¿Y está reflejada en este corpus?

GPR. Por lo que he visto, no.

FMF. Bueno, eso tiene que ver con la pregunta que tú hacías [Ricardo Otheguy]. Ahora te doy la palabra, pero quería hablar María Luisa [Parra] primero.

[2] Se trata de un listado de fuentes que incluye 11 textos escritos (2 de ficción y 9 de no ficción) y 343 cabeceras de prensa.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

María Luisa Parra [MLP]. Sí. Era un poco esta idea de que, bueno, hay trabajo similar en otros países. Y creo que sería un trabajo interesantísimo ver si esas mismas pautas, que son más bien un marco teórico, a partir de las cuales se ha definido el español de un cierto lugar, la metodología para definirlo, delimitarlo, si se pueden aplicar acá. Y a mí me parecía importante ver, si no se puede aplicar, por qué; y cuál sería la aportación entonces de un equipo como este para la definición de marcos conceptuales o la redefinición de lo que es un corpus. Creo que sería un trabajo muy interesante.

14 **RO**. No sería solo decidir si se puede aplicar aquí o no. Sería primero valorar lo que han hecho ellos y decidir si nos parece que lo han hecho bien. Porque, aun si lo han hecho bien, puede que no sea aplicable. Pero el paso anterior es: lo que se ha hecho ¿nos merece nuestra aprobación? Yo no sé la respuesta a esa pregunta.

MLP. Porque, si no es aplicable, sería también importante ver por qué no, porque eso nos va a dar muchos datos con respecto a la situación de acá, ¿no?

FMF. Lo que se ha hecho por parte de la Academia Española, respondiendo un poco a tu [RO] pregunta, es reconocido de manera general como válido y como útil en lo que es: es decir, una representación, una muestra de manifestaciones del español, ordenada fundamentalmente por géneros, por tipos de texto. Esa tipología de textos se ha ido afinando desde el CREA hasta la propuesta que se maneja ahora. Y se está revelando, en general, como suficientemente útil. Está siendo útil, aunque se necesitó un corpus complementario para la ejemplificación

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

de la gramática, para la argumentación utilizada en la *Nueva gramática* de la Academia. Está siendo útil para su consulta en infinidad de trabajos, en publicaciones científicas. Es decir, hay una utilidad de fondo que creo manifiesta.

Eso no quiere decir, desde mi punto de vista, que no tenga puntos débiles. Yo creo que uno de los puntos débiles está precisamente en la representatividad geográfica, porque el CREA daba al español de España una proporción de tres a uno frente al español americano. En el CORPES se cambió radicalmente la proporción; no llega a la proporción que maneja Davies en su *Corpus del Español*, donde se le da como un 10% más al español de los países americanos, pero digamos que ahí hay un problema que se ha ido perfilando y que plantea dificultades. Dificultades que, en el caso de los Estados Unidos, son especialmente graves, ya que las fuentes que se han manejado y que se están manejando, para cuestiones gramaticales y lexicográficas, pues son unas fuentes que no se conocen bien. Bueno, se saben cuáles son, pero no se sabe muy bien quién ha decidido que eso esté allí, de dónde ha salido, por qué tipo de decisiones y quién las ha tomado.

Y yo creo que eso se debe no sólo a un mayor o menor conocimiento por parte de los que lo han trabajado en Madrid, que son toda gente experta en la elaboración de corpus, sino a que no se ha consultado, no se ha solicitado información sobre lo que está pasando en los Estados Unidos. Por lo tanto, creo que, en materia de corpus, lo que no hay que hacer es lo que se ha hecho para los Estados Unidos: no

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

hay que tratar como muestras representativas unas muestras que no representan nada. Hay que, digamos, trabajarlo desde el principio, desde la base.

En el caso del corpus de Mark Davies, lo que se ha hecho fundamentalmente, sobre todo en su última versión, ha sido una búsqueda de información en las redes, casi un saqueo, que ha permitido construir un corpus de 2000 millones de palabras, frente a los 300 a 500 millones que tienen los corpus académicos. Davies se centra en un tipo de textos nada más, principalmente para el español contemporáneo. Otra cosa es la parte histórica. La parte histórica tiene sus propias dificultades metodológicas y el procedimiento es distinto. Sí me parece interesante la idea, para eliminar barreras conceptuales, de no intentar definir y perfilar un "español estadounidense", sino recolectar materiales del español que se encuentra en los Estados Unidos ahora mismo, que podrán tener mayor o menor capacidad representativa, pero que son español que se utiliza hoy en los Estados Unidos, de una manera o de otra. Y es bueno testimoniarlo, recogerlo; y saber, tener información de primera mano, datos, de qué está ocurriendo; desde el español "trasplantado" que representamos nosotros –casi todos, menos Andrew Lynch y Domnita Dumitrescu, que usan un español aprendido–, hasta el español más popular.

DD. Quería aclarar cómo surgió esta idea del corpus. Esta idea surgió cuando discutimos el proyecto del diccionario [la 24ª. edición del *Diccionario de la Lengua Española*], el próximo diccionario, que necesita, en cada palabra, especificar la región donde se usa, con un mapa y con ejemplos. Entonces, varios de los

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

participantes en la Comisión Interacadémica dijeron: “Bueno, pero a lo mejor el CORPES y lo que ustedes tienen no refleja bien la realidad de nuestro país”. Entonces, de ahí surgió esta idea de que cada Academia enviara materiales, hiciera un corpus de lo que considera representativo, para incluirlo en este diccionario. Y esto es sumamente importante para el nuevo concepto de “estadounidismos” que empezó a manejarse, que apareció en el último diccionario [23ª. edición del *Diccionario de la Lengua Española*]. Entonces, ustedes dicen que se usa mucho “aplicar”, pero si no lo hemos impactado en el corpus, no podemos incluirlo, porque todo va a estar basado en corpus. Entonces, la idea es hacer un corpus que refleje realmente lo que la gente habla; no textos literarios.

GPR. ¿Por qué no?

DD. Bueno, también. Pero en el caso de los autores, lo que dijo Guillermo Rojo es: “si ustedes usan literatura, el problema viene cuando un autor peruano ha vivido varios años en México y luego se ha mudado a otro país, Argentina, y luego se fue a Estados Unidos”. Va a ser muy difícil saber qué es lo que refleja realmente el español de Estados Unidos. Mientras que la idea principal era recoger un corpus hablado, de interacción en internet, o sea, cosas del habla diaria.

GPR. Yo no puedo estar de acuerdo con eso. Es decir, entiendo estos casos, ¿no? Pero la gran mayoría de lo que tenemos en literatura es muy claro. Aparte de las referencias que haga al país donde vive el autor. Es parte del español de Estados Unidos. Eso no se puede quedar fuera jamás.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. Por eso debería ser un corpus de las manifestaciones del español en Estados Unidos. Bueno, y podría entrar incluso hasta nuestro texto [Silva-Corvalán y Enrique Arias 2017], aunque está escrito en un español de España, porque Andrés me lo lee y me lo relee para que yo no ponga ahí ningún chilenismo.

CSC. Y por eso tengo esta pregunta. ¿qué es representativo para el español en Estados Unidos?

RO. Precisamente Paco [FMF] ha dado una respuesta a esa pregunta, que si estamos hablando del español de un cubano en Estados Unidos o de un chileno; estas son barreras conceptuales que podemos eliminar simplemente diciendo que vamos a hacer un récord de lo que se dice en Estados Unidos, en lo que llamamos "español", dígalo quien lo diga. Y no podemos borrar esos problemas porque precisamente eso…

CSC. Entonces, hay que definirlo. No se puede poner estadounidismo porque no importa quién lo dice.

FMF. El problema de fondo no es un problema de Estados Unidos ni nuestro. Es un problema general. Es que vamos hacia un mundo globalizado donde el concepto y las barreras nacionales están saltando por los aires. La movilidad, la súperdiversidad de Blommaert y de otros, está haciendo saltar las barreras nacionales de una manera radical. Por lo tanto, cada vez va a tener menos sentido intentar estabular el origen de cada rasgo; simplemente no va a funcionar.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

DD. Lo que quieren es saber en qué país se usa, no necesariamente que un uso sea privativo de Estados Unidos. Como se ha hecho en el *Diccionario de Americanismos* [ASALE, 2010], donde dice: "Esto se usa en tal, tal, tal país y en Estados Unidos" porque ha sido encontrado en materiales de Estados Unidos. Entonces, si hablamos de "estadounidismo", porque lo dice Gerardo [GPR] o porque lo digo yo, yo estaría de acuerdo.

GPR. "En": la preposición lo resuelve todo.

GPR. "En" Estados Unidos. Ya está.

JC. Ahora yo tengo una pregunta. Para la elaboración de un corpus en Estados Unidos, al ver esto recién de las proporciones. Nosotros sabemos que dos tercios provienen de México. ¿Tiene sentido o es ridículo tratar de hacer un corpus buscando que dos tercios sean, este, basados en los mexicanismos, un 9% en los puertorriqueñismos un porcentaje un poquito menor de dominicanos y cubanos? Porque si la proporción es tan grande que un poquito más de dos tercios son de origen mexicano, entonces lógicamente hay un número mayor proporcionalmente de términos mexicanos. No sé si tendría sentido buscar esas proporciones.

FMF. Bueno, hay que tener en cuenta que un corpus general incluiría lengua escrita y lengua hablada. Dentro de la lengua escrita, podrían entrar textos literarios, creados aquí, escritos desde Estados Unidos y en Estados Unidos, o textos más cotidianos de lengua tecleada, como se dice ahora, donde sería muy difícil aplicar proporciones migratorias. En todo caso, de aplicarse proporciones,

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

sería en la parte oral del corpus. Desde Estados Unidos se puede hacer lo que se considere más apropiado para los Estados Unidos. En esa parte del corpus oral, es donde sí se podrían, en todo caso, manejar esas proporciones. La cuestión es si merece la pena o no; o si son más importantes otro tipo de factores, de variables sociales, como el nivel educativo, la generación, el tiempo de residencia. Si esto es mucho más importante que ser de origen mexicano, puertorriqueño o ser de…

JC. Claro, no. Me basaba un poquito en la idea del español del Suroeste, que era en esa proporción, mucho mayor.

20

RO. El problema es que el corpus, según yo lo entiendo, no es una lista de palabras, sino una lista de textos. Entonces, la pregunta que tú harías sería: cuando vamos a seleccionar textos escritos, tecleados y hablados, ¿de dónde se recoge la muestra? Esa es la pregunta que tú estás haciendo. Y la muestra, pues, habría que ver exactamente cuáles son los factores. El origen del hablante, el origen nacional del hablante, sería uno de muchos factores para crear la muestra.

CSC. Yo creo que eso sería importante porque se vería claramente que hay autores aquí que son básicamente ecuatorianos. Había un escritor ecuatoriano que vivió toda su vida fuera de Estados Unidos y que últimamente ha publicado una novela en español, creo, traducida al español y está incorporado en el corpus de CORPES.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

DD. ¿Cómo decidieron eso? ¿Esta selección? Otra cosa que quería recordar es que se habló en la reunión[3] de incluir traducciones, porque los traductores a veces tienen que hacer una selección pensando en lo más general. Entonces, por ejemplo, van a incluir traducciones al español de obras escritas en otros idiomas porque piensan que el traductor es un creador también. Entonces, eso es muy importante para los documentos legales.

CSC. ¿Cómo se sabe quién es el traductor? ¿No podía haber sido traducido en, no sé, en Colombia o en Argentina?

DD. Bueno, hay traductores aquí oficiales. O sea, que habría que identificar a los traductores también. Sí, pero la idea es que el traductor tiene que escoger un término entre varios pensando en que este término va a ser entendido por la mayoría de los hispanos.

CSC. ¿En Estados Unidos?

DD. En Estados Unidos. Y eso es importante para los documentos reales. Sí, bueno, en relación con todas las variables, desde mi punto de vista, efectivamente se puede tener en cuenta el origen, como un factor más. Además, se puede tener en cuenta como factor principal para seleccionar el hablante o el texto, o como factor subsidiario; es decir, como información complementaria que permita luego recuperar la información del origen de cada uno de los hablantes. Así se le podía

[3] DD se refiere a una reunión de la Comisión Interacadémica para la elaboración de la 24ª ed. del Diccionario de la Lengua Española.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

dar más importancia al hecho de si el hablante procede del suroeste o si procede de Florida o de la costa Este, y luego aportar la información del origen familiar o generacional.

MLP. Solamente para complementar lo que estaba diciendo sobre el problema de limitarlo al tema del origen del hablante.

CSC. No, no lo estamos limitando.

MLP. Bueno, o que quede como algo subsidiario.

CSC. Es un informe subsidiario.

MLP. Porque además los hablantes cambiamos; elegimos palabras dependiendo de con quién hablamos: si yo hablo con mi amiga puertorriqueña, voy a usar el *revolú* y, cuando le quiero mandar un texto a mis amigas en México, digo, me sale *revolú*, pero digo: "¿Eso se dice en México? Creo que no. Ya no se dice en México. ¿Quién lo dice?" Entonces, también uno, como hablante acá, va adquiriendo otros términos. Lo usas en un contexto y ya no lo usas en otro. Y la generación. Mi hijo, que usa *parquear*, es de origen mexicano, pero está eligiendo usar un término que es más frecuente en otros grupos que en el mexicano, ¿no?

DD. Yo solo quiero completar esta idea de la traducción. *Billón* y *trillón* entraron en el diccionario [DLE 2014] a petición de los traductores, que decían que es un lío traducir al español de otras formas porque, si ponían billón en el sentido de España, se estropearía todo el sentido de la traducción.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. Esas son traducciones hechas para Estados Unidos.

FMF. De todas maneras, *billón* y *trillón* también se utilizan en otros países, con un valor estadounidense. No es una cosa exclusiva de aquí.

RO. Al igual que *parquear* se utiliza en la lengua nativa de los cubanos.

FMF. En otros corpus se demuestra que tiene un uso mucho más amplio.

Andrew Lynch [AL]. Yo creo que, si hay algo que realmente se pueda llamar estadounidismo, es algo que se va a manifestar en el plano oral más que nada. Entonces, la pregunta que me interesa, o la cuestión que me interesa, es qué proporción o qué porcentaje de este corpus en particular, a diferencia de los otros corpus de países donde la lengua mayoritaria es el español, que aquí no es el caso, qué porcentaje de este corpus debe ser oral, para que sea, entre comillas, representativo de las manifestaciones del español en Estados Unidos. Porque, por ejemplo, todo el mundo en Miami te dice "te llamo para atrás". Bueno, casi todo el mundo, pero nadie escribiría eso. "Llamar para atrás" es, para mí, un estadounidismo en cierto sentido. Todo el mundo lo usa. Todo el mundo lo entiende, etcétera. Yo creo que se usa también en Puerto Rico, hasta cierto punto. Pero eso no se va a dar en ningún texto literario ni de las noticias ni de… Y como ese lexema o ese ítem, hay toda una serie que ustedes ya conocen muy bien.

RO. En textos tipo *Facebook* y eso, sí que se puede encontrar; tipo *Twitter*, *Facebook*…

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

AL. Supongo, pero lo más obvio y lo más fácil sería un corpus oral. Entonces, la pregunta es: ¿qué porcentaje –ustedes seguramente hayan comentado este tema anteriormente– se propone para este corpus que sea oral y no escrito? Creo que lo oral, vuelvo a insistir, sería más importante para este corpus en particular que en el caso de los otros corpus.

GPR. Pero, perdón, esa oralidad se refleja también en la literatura. Es decir, un monólogo en primera persona, de acuerdo. Pero si un escritor quiere reflejar realmente la realidad en la que vive…

AL. Además, el español en Estados Unidos es un fenómeno más oral…

GPR. No digo que no, pero digo que en la literatura se refleja también la oralidad. ¿Y qué se ha hecho en estos corpus en cuanto a esa pregunta?

FMF. Pues normalmente la parte oral de un corpus general ocupa no más del 10%.

CSC. Muy poco representativo del español de Estados Unidos.

FMF. Pero estamos hablando de corpus de referencia, generales, donde se ha adjudicado esa proporción. No sé si nuestro destino final, desde el punto de vista de la Academia, es construir un gran corpus de referencia o si es simplemente reunir unas colecciones sólidas de materiales, suficientemente amplias, bien organizadas internamente, al margen de las proporciones. Para mí como lingüista, es más importante tener un buen corpus oral, que el hecho de que esos materiales supongan un 10% o un 15% de un corpus general. Si el corpus oral es bueno,

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

suficientemente amplio, lo voy a utilizar al margen de la proporción que suponga en el total. Pero bueno, efectivamente, suponiendo que vamos hacia un posible corpus general del español en los Estados Unidos, probablemente haya que contar con una proporción mayor de lengua oral de la que se le da en los corpus generales que se manejan en la Academia Española. Yo estoy de acuerdo.

MLP. Sí, yo tengo una pregunta en términos de este *continuum* oralidad y escritura. Porque hay oralidades que tienden hacia las normas de la lengua escrita y hay escrituras que reflejan la oralidad. Entonces, ahí es un *continuum*. Hay intersecciones. Estoy pensando un poco, por ejemplo, en el lenguaje que usa Junot Díaz, que escribe en inglés, pero incorpora esa oralidad en español. Y, si no la tiene, pues entonces pierde el chiste el texto, ¿no? Pero es un texto literario. Entonces, una pregunta sería: ¿cómo se trabaja esa diversidad de géneros y esos continuos? ¿Los marcas? ¿Se seleccionan o se definen?

FMF. Hay distintas soluciones. En el caso del corpus académico general, se trabaja con una separación más clara entre lengua escrita y lengua hablada, aunque la lengua escrita incluye también manifestaciones procedentes de internet, que es lengua tecleada. Sin embargo, en la clasificación que la Academia Mexicana está proponiendo para el español de México, se están manejando conceptos de semioralidad; es decir, estadios intermedios entre lo escrito y lo oral, que efectivamente reflejarían al menos una escala de cuatro grados. Habría una lengua escrita con más cuidado y otra lengua escrita más espontánea (son distintos grados de espontaneidad). Y habría una lengua hablada más elaborada,

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

más cuidada, y una lengua hablada más espontánea. Se puede también jugar con eso en la tipología de textos que se decida a manejar. Hay que tomar la decisión adecuada, aunque no sea lo tradicional.

RO. Yo quisiera preguntar si hay consenso o no lo hay en cuanto a eliminar la preocupación por la representatividad. Porque, no sé, en la conversación surgió sobre las manifestaciones y la representatividad. Me parece a mí que o tenemos como meta la representatividad o tenemos como meta las manifestaciones y … reconocer el caos.

CSC. Así lo habíamos acordado.

RO. Entonces no hay que preocuparse de si eso es representativo. ¿Estamos de acuerdo en eso?

CSC. Ya Paco [FMF] nos había dicho que lo vamos a ver como el español *en* Estados Unidos. Y entonces entrarían todas las manifestaciones del español en Estados Unidos.

FMF. Hay además otra cuestión técnica, pero con mucho fondo, que tiene que ver con los corpus y que Andrés [Enrique Arias] conoce muy bien. Cuando se trabaja con un corpus, conviene contar con una buena diversidad de muestras, porque, cuando uno va a consultar un corpus, lo que quiere es que salgan las cosas que se dicen. Entonces, como trabajes con una tipología restringida o solo con manifestaciones que consideres representativas, emblemáticas, te puedas perder

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

una buena parte de la realidad. O sea, hay que buscar la forma de que haya una suficiente diversidad, variedad tipológica, de textos, con manifestaciones distintas, de tal manera que, cuando uno vaya a buscar, encuentre el "pa' trás" o "te llamo pa' trás", encuentre las "aplicaciones" y encuentre lo que tenga que encontrar, en las proporciones que sea, pero que lo encuentre. O sea, que se sepa si eso existe o no y si aparece en una proporción enorme o si estamos hablando de algo esporádico.

CSC. Bueno, si es una proporción enorme, eso ya no lo podríamos decir si no es un corpus proporcional, como sugería alguien.

RO. ¿Proporcional? Pero Carmen [CSC], "proporcional" ya introduce el concepto de representatividad.

CSC. Por eso te digo. Si Paco [FMF] dice que la persona que consulta un corpus querría saber con cuánta frecuencia, entonces ya estaríamos introduciendo proporción ahí. Pero tengo una pregunta: es decir, un corpus puede ser suficientemente honesto como para plantear que esto no es representativo de nada, sino que representa el español usado en Estados Unidos y quién usa eso. De alguna manera, tú lo has dicho. En otras aportaciones, en otros países, por ejemplo, en Chile, me decía Abelardo San Martín, que cuando ellos han recogido el corpus de Santiago, ponen como condición que la persona haya o nacido en Santiago o haya llegado a Santiago a una cierta edad y luego lo jerarquizan de acuerdo con el nivel de educación, etcétera. Es decir, son corpus muy acotados. Y

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

el nuestro no sé cuán acotado podría ser. ¿Quién puede aparecer en un corpus masivo como el que queremos producir para Estados Unidos con diferentes manifestaciones?

Andrés Enrique Arias [AEA]. Yo solo iba a comentar dos cosas. Sobre la cuestión de la representatividad, obviamente la cuestión es compleja. Pero yo pienso que, para que aflore un amplio rango de fenómenos, vamos a necesitar, pues, que haya representación de muchos géneros, es decir, de textos orales, escritos, de, como dices tú [FMF], tecleados. Pero lo que quería decir en realidad es otra cosa. Se me está ocurriendo que, como usuario de corpus, a mí me parece que, en el fondo, un corpus se puede diseñar de tantas maneras, que todo lo que se está diciendo aquí puede ser válido en español en Estados Unidos o restringirlo. Pero a mí me parece que lo esencial y lo que realmente quiere el usuario es que los metadatos, es decir, la información sobre los textos, sea muy clara; de tal manera que, si yo encuentro algo y me dicen "Bueno, autor nacido en Perú" o lo que sea, yo ya, si quiero, puedo desechar eso dependiendo de lo que yo esté buscando; o según la generación o tal.

AEA. Es decir, que el corpus yo creo que, cuanto más inclusivo sea, mejor, porque va a ayudar a que el que quiere buscar… Claro, uno no puede imaginar quién lo va a usar. Y lo puede usar gente a la que posiblemente en verdad no le interesa un vernáculo estadounidense o algo así. Igual lo que quiere incluso es analizar la literatura o lo que sea o… Es decir, que, si uno tiene un corpus amplio, permite que lo use, pues, más gente, que puede venir con cualquier tipo de intención, que ni

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

siquiera podemos imaginar realmente a quién se le puede ocurrir entrar ahí y para qué. Lo esencial es la claridad. A mí lo que me frustra muchísimo de algunos corpus es, cuando encuentro un ejemplo interesante, pero no me dan la oportunidad de contrastar de dónde viene, de dónde ha salido. Entonces ya no sé qué hacer con ese ejemplo. No sé qué significa. Quizá la gran restricción es decir: "Vamos a restringir los materiales a aquellos de los que sepamos bien la procedencia; o sea, la biografía del hablante o todo lo que se pueda saber sobre de dónde viene".

RO. Me parece magnífico lo que dices. Eso quiere decir que, si vamos a saber siempre qué pata puso ese huevo, vamos a incluir todos los huevos. Quiere decir que, si es un letrero que ponen en una tienda, que está escrito macarrónicamente, lleno de disparates, para utilizar todo ese lenguaje, eso no es motivo de exclusión. Que diga, eso que tú me acabas de enseñar, todas las cosas que une leen, que cualquier persona que tenga interés en la normatividad dice: "Eso es un disparate"; eso entra también, mientras el usuario pueda luego llegar a ver que se trata de un letrero que pusieron en una bodega del Bronx, que lo escribió el bodeguero que nació aquí, que sus padres también crecieron aquí. ¿Esa es la idea?

AEA. Bueno, pero eso es difícil.

MLP. Sería muy bonita esa idea, pero si el letrero que vamos a enseñar era de Dallas, del aeropuerto, vamos a empezar a conjeturar: "¿Pero esto lo habrá escrito

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

alguien que es de segunda generación o lo ha escrito alguien que es anglo y no sabe español?"

RO. El usuario sabrá que lo que encontramos en el aeropuerto de Dallas.

FMF. El espíritu que quieren darle al corpus de México es precisamente el de Archivo del Español en México; es decir, de lo que se manifiesta ahora. E incluye también letreros; es decir, incluye el paisaje lingüístico, como se dice ahora en sociolingüística, porque es testimonio de un estado de lengua que ahora mismo puede ser considerado como barbaridad, pero que puede ser antecedente de un cambio lingüístico que se va a producir en no sé cuántas generaciones.

RO. Dentro de nada, ¿estarán todos llamando para atrás?

FMF. Seguramente.

CSC. Cuando yo venía de Chile, nadie decía "habían": "Habían unos niños ahí afuera". La lengua va cambiando. Hay que incluirlo todo. Mientras se pueda hacer lo que dice Andrés [AEA], ser honestos en cuanto al origen de lo que se está incluyendo.

FMF. Eso técnicamente está resuelto porque hay un sistema de cabeceras de textos donde se pueden incluir las etiquetas que se quiera. Solo hay que tener un buen sistema clasificatorio, para ir identificando, con todos los rasgos que se consideren pertinentes, cada uno de los textos. Lo que ocurre es que algunos corpus lo incluyen, pero no permiten recuperar la información, porque están

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

blindados. Solo se sacan concordancias y no se accede a toda la información de la cabecera. Pero yo creo que, efectivamente, en un trabajo así hay que contar con una cabecera detallada. Como la tenemos en PRESEEA [Proyecto para el Estudio Sociolingüístico del Español de España y de América]. PRESEEA tiene una cabecera sociolingüística donde se habla de ese itinerario lingüístico de cada uno de los informantes, incluso del desarrollo de la interacción: con quién hablas, si habla con uno solo, si habla con dos, y el origen de cada uno de los interlocutores, para que eso se pueda interpretar.

RO. Porque me pareció que Andrés [AEA] estaba un poco escéptico ante el letrero en el aeropuerto de Dallas o en la bodega en el Bronx. Muchas veces no vamos a saber quién es el autor. ¿Y estaríamos dispuestos a decir que, mientras haya un dato metatextual que diga de dónde sacamos el dato, se incluye? ¿No importa si se sabe quién lo escribió o no se sabe?

CSC. Mientras se diga: "Autor anónimo". Ya sabrá el usuario si lo usa o no lo usa. Si estos datos los voy a usar con otros anónimos o no.

RO. ¿Y ese ha sido el criterio en México?

FMF. Bueno, ese es el espíritu que quieren darle, un espíritu de archivo, con testimonios reales de distintas fuentes, que incluyan desde usos bilingües de indígenas hasta paisaje lingüístico. Pero claro: eso no es para el corpus que luego

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

se aportaría a las Academias en general. Eso es para el uso de la Academia Mexicana, con fines lingüísticos que interesan a México por su peculiar historia.

JC. Yo creo que sería interesante aplicar a la prensa el concepto de semioralidad del que hablaba María Luisa [MLP] porque muchos de estos pasquincitos que empiezan desde 1800 se hacían más para vender que para informar. Y entonces, muchos de ellos, la mayoría, en vez de tener un lenguaje formal, simplemente reflejaban el habla del público al que querían venderle algo. Entonces creo que ese criterio se aplicaría porque está reflejando el habla de ese momento, con la ventaja diacrónica de que, en *El Mississippi*, podemos ver qué términos se estaban utilizando hace 200 años.

CSC. Bueno, yo tengo una pregunta en relación con esto. Los *corpora* que se han producido en otros países, ¿qué profundidad histórica tienen?

FMF. Bueno, hay un corpus diacrónico del español que incluye desde el origen de la lengua.

CSC. El de México.

FMF. México también tiene un corpus histórico.

CSC. En el presente porque, no sé, cuando pienso yo en el año 1800, en el Suroeste, que es hoy en día prácticamente una extensión de México, en 1900 había solamente 100.000 hispanos en todo el Suroeste. Es decir, era una población relativamente muy pequeña.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. En el caso de México, tienen un corpus histórico. El corpus de español contemporáneo aún no lo tienen. Están comenzando ahora, discutiendo precisamente dónde ponen el origen; sabiendo que, si vamos más allá de 50 años hacia atrás, ya estaríamos ante una diacronía distinta. Por lo tanto, si queremos que sea un corpus contemporáneo habría que definir si es a partir del año 2000, del año 90…

CSC. ¿Y qué tiene en mente la Academia aquí, digamos, que es lo que necesita y está estimulando este proyecto?

DD. No pusieron ningún tipo de límites. Dijeron, cada Academia decide lo que quiere hacer.

CSC. El año en que empieza.

RO. Pero, para fines académicos, se puede tener lo que yo pienso que sería el criterio: o sea, una falta de criterio. ¿A la Academia le interesaría saber lo que dice el letrero en la bodega del Bronx?

FMF. A la Academia Norteamericana sí le interesa saber que eso existe y luego enviar a la Academia Española o a la Asociación de Academias –no es la Academia Española sola, sino la Asociación de Academias– la información que se solicite del tipo que se solicite, si es que disponemos de ella. Lo que no podemos permitir o aceptar desde aquí, desde la Academia Norteamericana es que, arbitrariamente, en un diccionario general del español, se incluyan cinco palabras procedentes de

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Estados Unidos de las cuales tres son dudosas porque no tienen el respaldo de ningún tipo de documentación. O que se incluya una definición de *espanglish* que no ha sido consensuada desde Estados Unidos o que se tomen decisiones sobre qué textos literarios incluir en un corpus sin que haya una decisión suficientemente respaldada. En fin, eso es lo que no queremos que ocurra; que se tenga una imagen deformada de lo que ocurre en Estados Unidos, que bastante complicada es la situación como para que se deforme más y se aleje más de la realidad.

CSC. Yo la pregunta la hice porque Jorge tiene interés en lo que pasaba en el año 1801. Y yo digo, bueno, si nos vamos a preocupar de lo que era el español, digamos, tantos años atrás, ahí sí que la tarea va a ser demasiado monumental; que quizá deberíamos poner, por lo menos, para nosotros un límite temporal; digamos, como decías tú [FMF], un máximo de 50 años o algo así.

MLP. Cuando pensamos en levantar todos estos datos, ¿va a ser en todo el país? Es que estoy pensando, por ejemplo, en regiones como Nuevo México, donde la historia permea. Donde todavía se usan palabras que se usaron hace muchísimo tiempo y que ahora son parte del habla de las comunidades de ahí; que pueden ser estos… -no sé si tal vez usar la palabra– estratos, pero que reflejan una historia.

CSC. Bueno, están reflejadas en la historia del español de Nuevo México, pero son datos recogidos ahora.

CSC. También hay un corpus recogido por [Garland] Bills, por ejemplo, que empezó ya hace unos 40 años.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

MLP. Claro. Y no sé si, como parte de esos datos de interés para el usuario del corpus, también se puede incluir: "Bueno, esta es una palabra que puede tener sus orígenes en el español de 1801".

CSC. La etimología de la palabra.

RO. Todas las palabras que se usan en Nuevo México se han usado durante mucho tiempo en la historia.

RO. *Mesa, agua, cara*. Todas tienen su historia.

GPR. Pero, si no se usan en otros estados, a lo mejor convendría indicar en qué estado predomina esa palabra.

RO. Esa es la idea de Andrés [AEA], que hay que poner, para que los usuarios sepan de donde se sacó.

CSC. De dónde se sacó. Eso es fundamental.

JC. Bueno, de estado o región, ¿no?

FMF. Propongo lo siguiente. De lo que hemos ido hablando hasta ahora, hay, desde mi punto de vista, tres argumentos, tres aspectos, que pueden generar consenso. Uno de ellos es dar prioridad, no tanto a un concepto unitario o uniforme de español estadounidense, sino a una colección de manifestaciones del español en los Estados Unidos, y buscar qué manifestaciones y cómo, pero hablando siempre de manifestaciones del español en los Estados Unidos. Una segunda idea general

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

es que el criterio sea inclusivo; es decir, que se acepte una diversidad de textos. También se decidirá dónde comienza y dónde termina esa diversidad, si entran carteles o no, si entra un tipo de literatura o no, pero que haya diversidad, para que de la consulta del corpus se pueda extraer información muy variada. Y, en tercer lugar, que la información vaya perfectamente etiquetada, con detalle suficiente, a través de los metadatos.

CSC. Eso es fundamental.

JC. ¿Cuál es la tercera, perdón?

FMF. Que haya un sistema de etiquetado con metadatos detallados, que, en el caso de Estados Unidos, es muy importante. Bueno, en todos los casos es muy importante, pero en nuestro caso casi más, por la propia diversidad del origen de las manifestaciones.

RO. Pero, el segundo criterio, me parece que, si lo dejas así, diciendo "diversidad", lo que tú estás diciendo y lo que va o oír tu público o la audiencia o el lector van a ser dos cosas distintas. Porque la diversidad siempre va amparada bajo un concepto de lo que es y no es español. Y, entonces, si verdaderamente ese archivo va a incluir lo que una persona con un poco de escolaridad llama "disparates" y "espanglish" y cosas que nadie dice, eso hay que decirlo explícitamente, hay que decir si opinamos así; que lo que llamamos "diversidad" incluye lo que se considera disparatado, escrito por uno que no sabe español o no. Pero tenemos que decidirlo y dar la cara; y decir: "Hemos tomado esta decisión". Si los disparates

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

que se ponen en los letreros no son español y, por lo tanto, no lo registramos, o si queremos registrarlo. Me parece que tenemos que decirlo con todas las letras porque, si no, cuando tú dices "diversidad", el que te oye no reconoce la gravedad de lo que tú estás diciendo.

CSC. Lo que pasa es que, como hemos decidido, parece que van a ser manifestaciones del español en Estados Unidos. Si yo veo en el aeropuerto de Dallas o en el Bronx un letrero que me parece un… ¿cómo era la palabra? un disparate, si yo lo reconozco como español, es español. Puede no haber concordancia entre el sujeto y el verbo, pero lo mismo pasa en Chile: no hay concordancia entre el sujeto y el verbo en numerosos tipos de textos. Pero yo lo reconozco como español; será con errores o no, pero es español, mientras digamos cuál es el origen. Mientras la metadata sea explícita, detallada, yo no veo ningún problema con incorporarlo.

RO. Pero quiere decir que, entonces, si tenemos manifestaciones del español que carecen de concordancia, que la fraseología está calcada del inglés, como dicen (yo lo digo de otra manera, pero bueno); si la selección léxica es mayormente de préstamos del inglés o de léxico español con el sentido del cognado en… Todas esas tragedias que tanto preocupan a los que se preocupan, a los que aman el español, ¿verdad?, si no vamos a hacerles caso a los amantes del español que quieren protegerlo y cuidarlo, tenemos que decirlo porque, si no, es imposible que estas personas tan educadas en esta mesa, vayan a registrar esas cosas. Por favor. O sea, hay que decirlo o no decirlo. Es lo que les quiero decir.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. Se entiende.

FMF. Cuando hablo de diversidad y variedad, me refiero sobre todo a variedad tipológica; es decir, que tengamos textos procedentes de literatura escrita de una serie de géneros, que haya una lengua tecleaba procedente de blogs, que haya grabaciones sociolingüísticas de estratos más cultos, de estratos menos cultos; es decir, que haya una tipología suficientemente amplia, que no lo limitemos solo a lengua culta o a literatura escrita publicada en medios nacionales, sino que sea un corpus diverso en ese sentido.

CSC. Porque hay una cantidad de léxico chileno que yo no usaría jamás, construcciones sintácticas que también se escapan de mi variedad de español chileno. Y, sin embargo, forman parte del español de Chile y están marcadas como nivel socioeconómico cultural más bajo, en los cuatro niveles que reconoce Abelardo [San Martín], por ejemplo. Eso es lo mismo que querríamos nosotros para el español de acá. Yo sé que tú sabes mucho de corpus también, pero Andrés [AEA], que es un experto, tendrá que ayudar con esto de los metadatos.

Lengua escrita

FMF. Muy bien. Entonces, si esos tres puntos están consensuados, más o menos, podríamos seguir adelante con esta discusión. Para no andar de un lado para otro, podríamos hacer algunos comentarios o reflexiones sobre los tipos de textos que se podrían incluir. Por ejemplo, a propósito de la lengua escrita, ¿qué habría que

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

recoger o que no habría que recoger en un corpus? ¿Vale la tipología de textos que utiliza la Academia Española?

GPR. Según; no todos. Ahí habría que hacer una selección, ¿sabes? Y ampliarlo además. Ahí falta muchísimo, claro.

FMF. La idea es hablar ahora de lengua escrita. Luego podemos hablar de internet y luego de lengua hablada, para ver cómo proceder en cada uno de los casos. Ya veremos qué es lo que podemos hacer en cada uno de los campos. Una cosa es la teoría y otra es cómo proceder en concreto con cada una de esas manifestaciones.

MLP. Sí, nada más una pregunta porque me perdí un poquito al final. Entonces, si vamos a hacer esta selección de textos, los anuncios ¿los vamos a incluir? Y, si encontramos un anuncio de seguros que dice, "As Fast as a Flying Chancla", ¿ese texto se va a incluir? ¿Ese tipo de textos, como decía Ricardo con mezclas o con errores? Es lengua escrita. Está en un anuncio, si decidimos incluir anuncios. ¿Esos textos se incluirían?

FMF. Bueno, podrían incluirse si, como digo, nuestra idea es recoger manifestaciones de distinto tipo.

MLP. Sí, lo que pasa es que dentro de cada manifestación también va a haber un espectro de expresiones y de usos. Y entonces hay unos que van a estar más apegados a la norma y otros que van a ser más innovadores, ¿entonces vamos a incluir todos?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Claro, podría ser. Lo que ocurre es que no tendría por qué atenderse a todo desde el principio. Habría que ordenar a qué se le da preferencia. Lo que propongo ahora es hablar principalmente de la lengua escrita en cuanto a creación literaria.

AEA. De todas maneras, yo creo que nadie, o igual lo he entendido mal, pero nadie tiene una intención de que sea un corpus que gravita hacia lo normativo o hacia lo convencional o hacia lo estándar.

CSC. Eso sí. Eso es importante.

AEA. Porque me parece que lo que estabas diciendo antes [RO] era como respuesta a algo que no ha dicho nadie. ¿Vamos a incluir disparates? Obviamente sí. Yo no veo por qué no.

RO. No, tú tienes razón. No es que lo haya dicho nadie hoy aquí, pero yo creo que, si no explicitamos esa postura, va a haber muchos otros públicos que sí nos van a malentender. No conozco lo suficiente el CREA y el CORPES para saber, por ejemplo, si ellos tomaron esta actitud tan promiscua ante la selección de los textos. Quizás es un prejuicio mío. Yo me temo que no.

AEA. Creo que tienen también materiales de periódicos, revistas y algo de oral, ese 10%. Pero, bueno, los corpus tienen que ser sucios, digamos, ¿no?

RO. Pero lo que yo no sé es hasta dónde llega la suciedad. Y creo que tenemos que decidirlo nosotros. Yo estoy proponiendo una suciedad máxima, pero creo que a lo mejor haya personas aquí que piensen que eso no debe ser así, que eso sería desvirtuar lo que es el español de Estados Unidos. Yo creo que, si estas cosas no se hablan, no sé... Aunque no estoy dirigiéndome a nadie que lo haya dicho aquí,

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

tú tienes razón. Dice Andrés que los corpus o los *corpora* suelen ser sucios. Muy bien, pero lo que tenemos que decidir aquí es hasta dónde llega la suciedad, hasta dónde llega el churro. Y yo estoy proponiendo que, si queremos reflejar el español de los Estados Unidos de alguna manera, esa recogida de manifestaciones que no quiere ser representativa, sino que quiere ser lo que haya, tiene que incluir muchísimas cosas que van a hacer, desde el punto de vista nuestro, disparatadas.

CSC. No sería un corpus del español de Estados Unidos, sino que está "en" Estados Unidos. Por eso, en un principio pensaba que iba a ser español de Estados Unidos, porque había que marcar estadounidismos, pero veo que no, que es manifestaciones del español en Estados Unidos.

RO. Claro, la preposición resuelve nada más que parte del problema de la suciedad porque, si uno quiere verdaderamente que sea un saco muy grande, la palabra "de" limita mucho. Bueno, ponemos "en." Ahora nos entendemos. Pero la palabra "español" también limita mucho, siempre y cuando haya públicos aquí en esta mesa, públicos que digan, pero eso no es español. Por lo tanto…

RO. Eso es espanglish; eso es lo que sea; lo que quieras llamarlo.

CSC. Entonces tendríamos que llamarlo del español y del espanglish en Estados Unidos.

RO. Lo que hay que decir es lo que entendemos por ese nombre.

FMF. No, pero en tu línea de pensamiento, ese mismo problema se plantea para el español de o en cualquier país. Mira, Moreno de Alba llama su libro *El español en América* y hay una larga discusión sobre si debe hablarse de español de

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

América o español en América. Y también se habla sobre si son español o no las distintas manifestaciones lingüísticas de los indígenas. Exactamente el mismo problema, calcado, se da en la frontera con el catalán, ¿verdad?, donde no sabes si se habla español o qué se habla.

CSC. Y los libros de Moreno de Alba lo que corrigen es la manera de hablar en México.

FMF. Es exactamente el mismo problema.

RO. Por eso digo. Yo creo que, si nosotros no explicitamos, nos van a malentender. Y, si hay posiciones consensuadas, deberíamos explicitarlas.

JC. En respuesta a lo que dice Ricardo, yo sería partidario de la máxima promiscuidad.

CSC. De acuerdo.

AEA. Digamos que esto se suele establecer en unos porcentajes: tanto de esta fuente, tanto de esta…

FMF. Sí, la idea era, como decía antes, hablar de tipos de textos. Porque hay cuestiones a propósito de la literatura que hay que abordar: ¿qué tipo de autores? ¿qué tipo de perfil? Al margen de si es novela, si es ensayo, si es tal…

GPR. Bueno, eso también es importante.

FMF. También es importante, pero es más fácil de resolver. Bueno, si me oyera la gente de teoría literaria, me diría: "¿cómo que el género es algo fácil de resolver? Claro que no".

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

GPR. Pero, bueno, está el problema, digamos, de lo temporal, la temporalidad; es decir: cuándo empezamos a hablar de literatura en lengua española en Estados Unidos. ¿Vamos a empezar por Farfán de los Godos? Obviamente, no. Entonces, podemos tener en cuenta, quizás, desde que Estados Unidos se constituye como nación, de 1789 en adelante. Eso también se podría hacer, pero tenemos que ponernos de acuerdo. Después, en los textos mismos, sobre todo en cuestión de la novelística, pues yo me limitaría, sobre todo, a aquellas novelas, aquellos textos que hacen referencia directamente al país donde han sido escritos. Yo creo que eso es importante. Y ahí se refleja además la lengua, sobre todo en una narrativa, digamos, de tipo realista, entre comillas, claro está, donde lo que se refleja es la realidad. Entonces ahí obviamente la gente va a decir "ir para atrás" o "para adelante" o como tú quieras porque refleja, en esos diálogos, realmente cómo habla la gente, que es algo que es exactamente así. Creo que esos son los criterios importantes a la hora de elegir estos textos, los que representan un poco el español en Estados Unidos. La cuestión del tiempo es decidir cuándo comenzamos.

FMF. Bueno, ¿qué más comentarios hay sobre cuestiones literarias?

CSC. Pues, podemos primero ponernos de acuerdo desde cuándo.

FMF. Pues yo creo que nuestro interés abarca lo histórico, evidentemente, pero eso implica una metodología distinta para elaborar el corpus en algunos aspectos, técnicamente. Yo creo que habría que dar preferencia a lo contemporáneo.

GPR. Sin duda.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Entonces, habría que tomar la decisión sobre cuándo empieza lo contemporáneo, ¿no? ¿Ponemos el año 2000 o es demasiado tarde? ¿El año 80, cuando el censo ya incorpora al factor hispano?

CSC. 1980 fue crucial porque hubo una inmigración enorme.

GPR. Pero también los años setenta, ¿verdad?, con todos los movimientos del *Latino Power*. Sobre todo las novelas chicanas, algunas en español. De la gente de la generación anterior, a Tino Villanueva, por ejemplo. Esos no se pueden quedar fuera.

RO. Yo creo que Gerardo [GPR] tiene mucha razón. Vamos a defender a los viejos. Toda esa literatura de los años de la liberación del *Young Lords*, todo lo que se escribió en español en ese tiempo parecería ser muy interesante.

CSC. 1960.

GPR. Sí, tampoco hace falta una fecha exacta. Pero más o menos, ¿no?

CSC. Después de la Segunda Guerra Mundial.

MLP. Esa literatura está relacionada con los temas de identidad acá.

RO. Y, además, hablando del otro lado, de la parte de más valor literario, el español en Estados Unidos empezó a mejorar muchísimo a partir del año sesenta, con la llegada de los cubanos.

AEA. Yo quisiera comentar una cosa. Si el programa va a ser un corpus histórico, realmente tiene otra metodología. Y tiene complicaciones muy variadas, ¿no? En el caso del corpus que he estado compilando yo en Mallorca, la idea era: ¿desde

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

cuándo se puede decir que hay un español con unas características definidas que lo caracterizan como el castellano de Baleares? ¿Cuándo empieza eso? Y, claro, el problema es decir, cuando has comentado de documentos o periódicos de 1801, ¿hay una continuidad desde entonces hasta ahora de rasgos que puedan ser característicos del español estadounidense? Claro, eso es complicado. Y lo que se está comentando aquí, cuando se está hablando de años sesenta para aquí, creo que lo que hay un poco implícito, en lo que se está diciendo, es eso, que desde los sesenta hasta aquí, vemos ya un movimiento en el que hay cierta continuidad o en el que pueda haber, no sé si ideologías o identidades o rasgos lingüísticos, que continuarán hasta hoy. Y yo creo que remontarse más es más complicado.

JC. No, lo que decía de posguerra también podría ser para fijar una fecha históricamente decisiva.

RO. Andrés, ¿qué se hace en estos corpora académicos? ¿Cuándo empiezan?

AEA. Bueno, tú lo has dicho [FMF]. El CORPES es desde el 2000. Bueno, lo hicieron así porque el CREA iba hasta el 2000.

RO. ¿Y el CREA cuándo empieza?

FMF. En el 75. Todo lo anterior al 75 va al CORDE, el *Corpus Diacrónico del Español*. Y ahora hay un corpus para el nuevo diccionario histórico, que amplía el CORDE desde la Edad Media.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Pero me parece un poco raro porque que algo que se escribió en los años sesenta, o sea un estado diacrónico anterior al actual…

AEA. Casi te metes en el terreno de lo histórico, ¿no? Alguien que escribía en los años sesenta, con madurez, no sé si está vivo ahora. Sí, hace 57 años. No hace tanto, pero, si escribió a los 20 años, tiene 77 ahora. Quiero decir que, si la idea es representar el español actual, el de hace 57 años, no sé si sea español actual. ¿Es eso español actual?

RO. Pero quizá lo más sencillo sería, dado que ha habido pautas ya creadas, que hay esos tres momentos que los académicos o las personas que van pensando sobre esto han decidido hacer, yo creo que podríamos hacer lo mismo. Una manera de hacerlo sería, vamos a hacer los mismos cortes. Vamos a crear tres corpora con ese criterio. Yo creo que eso sería lo más fácil.

CSC. Y empezar por el contemporáneo.

RO. Y empezar por el contemporáneo, exacto. A lo mejor hay diferentes grupos de trabajo que pueden empezar con los tres.

CSC. Eso es una posibilidad. La otra es lo que dice Paco [FMF]: o sea, a partir del censo oficial, que fue en 1980.

FMF. No creo que sea un problema demasiado grave porque podemos decidir, pensando en un corpus estadounidense, que la parte literaria quede representada desde los años sesenta por todo ese movimiento chicano, que efectivamente tiene su prolongación en la literatura actual y que sirve de inspiración para los que están escribiendo ahora mismo; es decir, que hay una continuidad. Y, a la hora de enviar

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

datos para un corpus general o si se hace una consulta, se pueden enviar solo datos del corte sincrónico que se quiera. Pero, para nuestros fines e intereses, yo creo que sí es bueno reflejar la realidad de los Estados Unidos, porque los criterios que utilizaron en el CREA para fijar el año 75 fueron un poco circunstanciales.

RO. Bueno, claro, porque para España esos fueron cortes importantes en la vida española.

FMF. Sí, mediados de los años setenta. Y también tuvo que ver con la tecnología. Se quiso hacer un corpus actual, pero se comenzó a concebir en los años ochenta. Fue algo circunstancial. No fue fruto de una reflexión teórica.

RO. Y la realidad es que todos estos problemas se resuelven, me parece a mí, también con lo que ha dicho Andrés: mientras sea fácil recuperar, entonces no importa tanto cómo se haga.

CSC. Todo está en la información que se refleje.

FMF. Y ahora, desde el punto de vista literario, yo creo que las tipologías que se suelen manejar, de narrativa o ensayo, poesía, teatro, todo eso, se pueden adoptar como se han utilizado en otros corpus sin ningún tipo de problemas. En el caso de los Estados Unidos, tendríamos, por lo menos, que reflexionar sobre esa literatura que originalmente no está escrita en español. O sea, las traducciones de Achy Obejas al español de la literatura de Junot Díaz, ¿eso entraría o no entraría?

GPR. Yo creo que no, pero bueno, pertenece a la literatura norteamericana. Lo de Junot Díaz es inglés básicamente.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Sí, sí, por eso hablo de la traducción de Achy Obejas; es decir, una escritora de origen cubano que ha traducido al español estadounidense, en Estados Unidos.

GPR. Yo dejaría aparte en ese caso las traducciones.

CSC. Pero ¿quién hizo la traducción?

FMF. Achy Obejas. Bueno, de una de las obras, la hizo Achy Obejas, que es una periodista que ha escrito para el *Chicago Tribune*, que es de origen cubano y que es muy muy conocida. Ella misma es autora literaria. Era por poner un ejemplo.

GPR. De Junot Díaz lo ha hecho también, de la primera novela, Lago, Eduardo Lago.

FMF. Todas están en inglés originalmente. Lo que pasa es que han sido traducidas al español. También es un proceso creativo, porque yo creo que lo que hace Achy Obejas es casi creación literaria porque selecciona mucho qué es lo que pasa al español y qué deja en inglés, supongo que de acuerdo con el autor. Pero, vamos, a eso me refería, que hay un ámbito más difuso sobre lo que convendría hacer una reflexión general.

JC. Además, yo creo que otra fuente, una pequeña fuente, que podríamos tener en cuenta es la publicidad porque en algunos de estos periódicos, periodiquitos, pasquines, hay anuncios publicitarios que son interesantísimos y expresivos. Entonces, creo que esa podría ser una fuente interesante, en torno a los anuncios publicitarios en la prensa.

GPR. Pero ahora estamos con la literatura.

JC. Pero digo como otra de las fuentes para el corpus.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. Las traducciones son problemáticas, pero, de nuevo, si se ponen traducciones, sí es fácil incorporar una traducción.

GPR. Sí es fácil. ¿Por qué no? Ahora ¿cómo se hace eso? En mi ignorancia, no sé decirte. ¿Hay que pedirles permiso a los autores?

FMF. Ah, eso es una cuestión muy peliaguda porque afecta a los derechos de autor. En general, hay dos soluciones. Hay una negociación con las editoriales y con los autores para dirimir los derechos, si hay derechos o no hay derechos, si se exige pago o no se exige pago. Pero hay dos formas de orientarla: o bien se pide el uso del texto con estos fines científicos, garantizando que solo se hace pública la concordancia; es decir, que no se ofrece el texto completo, sino simplemente muestras en concordancias; o bien negocias los derechos para que te dejen mostrar fragmentos amplios; si no la obra completa, sí fragmentos amplios de la obra. Pero es una negociación, efectivamente, que implica cuestiones jurídicas.

CSC. ¿Pero no hay un límite de palabras debajo del cual se puede publicar?

FMF. Sí, bueno, depende de la legislación del país. En España, sí. Con fines científicos y pedagógicos o educativos, hay un límite de 250 o 350 palabras, o no sé cuántas, que te permiten reproducir sin permiso de ningún tipo. Pero, en este caso, como se trata de introducir las obras completas, los corpus solo ofrecen la concordancia, sin darte más información de ningún tipo, porque entonces ya implicaría una negociación de derechos de autor.

GPR. Perdona mi ignorancia, pero eso textos, esas novelas que se van a utilizar, etcétera, ¿están digitalizadas?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Sí o las proporcionarían digitalizadas las editoriales, cosa que ocurre cada vez más frecuentemente, por razones obvias, o bien se han digitalizado, como en el caso de la Academia, la Academia Española.

RO. Pero hay cosas muy básicas que yo no entiendo. Una preguntita rápida. Cuando se dice que hay equis millones de palabras en un corpus, ¿eso quiere decir que se cuenta la obra entera que sea? Por ejemplo, un corpus sobre el español latinoamericano tiene *Cien años de soledad*, sin duda. ¿Tiene la novela entera en ese corpus?

50 **FMF.** Sí.

RO. Sí. O sea que, si uno busca un rasgo, si ese rasgo aparece en *Cien años de soledad*, ¿aparece?

FMF. Sí.

RO. Pero eso no quiere decir que uno entra en la página del corpus y vea una novela íntegra.

FMF. No. Eso es.

RO. ¿Qué ve uno?

FMF. Normalmente, la concordancia. O sea, si buscas una palabra concreta, te da la concordancia. Y se puede decidir cuál es el límite de número de palabras por delante y por detrás.

MLP. La concordancia es el contexto…

FMF. Realmente es el contexto, digamos, inmediato.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

MLP. Para que uno pueda entender el sentido.

FMF. Eso es. Podría identificarse el significado, para los casos de homonimia y todo eso. No sé si hay alguna consideración más. En algún momento, habría que abordar una tarea como la de elaborar una lista, un listado; porque si tienes que crear un corpus, hay que decidir qué textos entran y qué textos no entran. ¿Tenemos ya algún listado? ¿Se puede tener un listado?

GPR. Hay bibliografía de literatura en lengua española, ¿no? Pero lo que yo pregunto es, ¿quién va a decidir? O sea, ¿quién va a conocer todo ese material? Obviamente tiene que tener acceso al material, ver lo que hay, leerlo, ¿no? y decidir si esa novela va o no va. ¿Quién lo va a hacer?

DD. Una pregunta, ¿tú [FMF] dices además de lo que ellos [RAE] han puesto o utilizando lo que ellos han puesto?

FMF. ¿La lista que ha manejado la Academia? Pues no sé. Yo creo que, como es una lista académica, lo debería revisar lo debería revisar la Academia Norteamericana con criterios estadounidenses. Esa lista, para los que son expertos en literatura, ¿esa lista de la RAE es representativa?

GPR. No, qué va.

CSC. Acuérdate que hemos eliminado de nuestro vocabulario la palabra "representativo".

FMF. Sí. Me estoy reprimiendo.

GPR. No, además, falta muchísimo.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. No me quedó claro lo de la traducción.

FMF. Sí. Y además ahora viene muy a cuento, porque lo literario también implica la posibilidad de la traducción.

RO. O sea, si hay una novela escrita en inglés que se tradujo en Estados Unidos al español, ¿qué hemos decidido? No me quedó claro en la conversación.

FMF. Se hablaba de las traducciones. Que hay una recomendación de incluir también las traducciones, puesto que la traducción tiene un elemento creativo importante. Y además con autoría.

DD. Pero yo pensaba más que nada en documentos.

RO. Y cuando lleguemos a los siguientes géneros, va a ser importantísimo. Todas las instrucciones médicas, por ejemplo, que se traducen del inglés al español. Creo que va a ser importante pensar en eso. Pero ahora que estamos en lo literario, no entiendo en qué hemos quedado, si es que hemos quedado en algo.

CSC. Yo creía que habíamos dicho que tenían que ser traducciones hechas por traductores que son de Estados Unidos.

GPR. En algún caso, el autor se autotraduce. También ocurre. Es decir, él tiene la versión en inglés, la ha escrito en inglés, y el mismo autor se autotraduce al español. Si acaso, yo creo que sí, que se podría aceptar. Porque es que, si no, es amplísima la cantidad de traducciones que hay, ¿no?, de obras del inglés al español, del español al inglés. Eso es un galimatías.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Lo que dice Carmen es acotar un poco a traducciones hechas en los Estados Unidos.

DD. Sí.

CSC. Con individuos que viven en Estados Unidos.

RO. Exacto. ¿Y eso también sería demasiado amplio?

GPR. Yo creo que sería muchísimo eso. Demasiado.

RO. ¿Las casas editoriales hacen las traducciones al español aquí o van a México y encargan a un traductor en México que las haga?

GPR. Las hacen aquí, en México y en España y en Argentina…

RO. Entonces, parecería que las que se hacen aquí, bien podrían ser parte de esto, ¿no?

GPR. Sí, pero creo que eso es difícil. Es complicado.

CSC. Yo creo que tendrían que ser traducciones de autores que consideramos americanos.

GPR. ¿Americanos?

CSC. Como tú, por ejemplo. Supongamos que tú [GPR] escribes en inglés y te traducen lo que has escrito en inglés. Bueno, has vivido en Estados Unidos tantos años que ya eres norteamericano.

GPR. Ya. Como quieras.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

AEA. Sí, bueno. Una cosa que, hay que tener en cuenta siempre es el diseño general del asunto, incluso cuando uno está diciendo, bueno, ¿qué tipo de obras literarias? Un problema que estoy viendo ya, con esta fecha que hemos dicho del año 80… ¿Era el 80 al final? Tentativamente… El problema que yo veo es que uno idealmente lo que quiere es que todos los géneros estén más o menos equilibrados o representados de una manera. Es decir, que no haya un –yo qué sé— 90% de literatura y un 10%… ¿no? ¿Y entonces qué va a pasar con esto? Si vamos a utilizar algunos géneros que son como más recientes, tipo blogs o cosas de esas tecleadas, en el año 80, obviamente, no hay. Entonces… Y también quiero decir que, si va a haber literatura del 80 hasta aquí, pues habrá que distribuirla de manera equilibrada. Es decir, tanto número de palabras de los años 80, tanto de los 90, tanto de los 2000, etcétera. Entonces, ¿hay en los otros géneros que queremos incluir una representatividad de esas décadas también? Al final va a quedar descompensado. Alguien que haga una búsqueda del año 60, si solo va a ver novelas y nada más, eso va a ser un problema, ¿no?

CSC. No, pero es un problema para él o para ella, que tendrá que decir si el trabajo que está haciendo corresponderá a lo que se produjo en los años 60.

AEA. Es decir, que los metadatos sean los que sirvan para que cada uno escoja lo que quiera. Y lo otro que iba comentar. Sobre traducciones: realmente, si hay suficiente material original literario, yo creo que las traducciones serían como decir: “Bueno, es que no hay material, necesitamos traducciones”. Pero si hay material…

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

GPR. A menos que haya algún caso en que la novela, después de haber sido traducida al español, haya adquirido una gran importancia. Y yo me pregunto, por ejemplo, qué ocurre aquí en el CORPES con novelas que se han escrito en catalán, por ejemplo, y que después se han traducido al español y son muy conocidas y las enseñamos en las universidades. Es el mismo caso.

JC. Yo estoy de acuerdo con Andrés. Yo creo que, sacando del terreno literario, las traducciones prácticamente no aportarían mucho. Yo traduje dos libros, 1.400 palabras, y creo que no habría ningún material para incluir entre los estadounidismos. O sea, no eran traducciones literarias, eran traducciones, qué se yo, el bebé, la embarazada, qué esperar cuando está esperando, qué espera del primer año de vida. Y yo creo que no había ningún elemento útil como para nuestro corpus en ese tipo de traducciones, no literarias.

CSC. Quizás habría que tomar una decisión en estos momentos, un tanto drástica: decir que no vamos a incluir traducciones.

CSC. Por ahora. Porque realmente ya se ve que hay muchos problemas.

FMF. ¿Y cómo se podría abordar la elaboración de un listado de autores o de obras?

GPR. Pero ahora sí. Veo que eso sería bueno crear una Comisión en la Academia, porque en la Academia hay muchos escritores de los Estados Unidos, y compilar esa bibliografía entre todos, ¿no? Lo más exhaustiva posible.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Claro. Alguien que elabore una primera propuesta que luego se vaya depurando o comentando, ¿no? [...] No sé si damos por cerrada esta propuesta de que en algún momento se elabore un listado o que se revisen los listados que ya está manejando la Real Academia Española, para completarlos y adecuarlos. Yo creo que ese trabajo es importante y, cuanto antes se haga, antes se podrá proponer una forma de completar el corpus. Pasamos entonces a hablar de lengua hablada. O sea, al margen de la proporción que tenga, de cómo sería la configuración interna de esa parte del corpus dedicada a la lengua hablada. ¿Quién se anima a hacer algún comentario?

DD. Carmen tiene un corpus de lengua hablada que podríamos incorporar.

CSC. Sí. Yo no tendría absolutamente ningún problema. El único problema es que yo le he dado los derechos de autor a USC [*University of Southern California*]. O sea que cualquier persona puede acceder al corpus. Esta puesto en internet. Tiene dos problemas. Uno, que las transcripciones fueron hechas por alumnos míos a mano, en una época en que todavía, en los años 90, no teníamos computadoras para todos los chicos y están recién empezando a digitalizarse. O sea que no es un corpus realmente al que se pueda acceder fácilmente, porque es en estos momentos es solamente oral; o sea, audio.

AEA. ¿Pero están digitalizando, o sea escaneando, las transcripciones manuscritas o lo están tecleando?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. No. Están tecleándolas. Y, al mismo tiempo que las teclean, van borrando los nombres que identifican a los individuos que hablan, porque la ley prohíbe la identificación de los individuos. Así, la parte oral, audio, está disponible para cualquier persona, pero solamente para propósitos científicos. Entonces, por ejemplo, la parte de los niños la tomó la Universidad de Carnegie, que pidió permiso para acceder y se pasaron todos los datos de los niños al corpus que ellos tienen en *Carnegie University*. Los adultos, todavía nadie ha pedido nada, pero es cuestión de ponerse en contacto con la Biblioteca y, si queremos pasar todos esos datos a nuestro corpus, la biblioteca tendrá mucho gusto en cederlos. Yo se los cedí a USC.

RO. ¿Cuántas entrevistas son, Carmen?

CSC. Son casi 200 horas o un poco más. En realidad, los metadatos los tengo que seguir corrigiendo; he corregido algunos, porque, cuando se digitalizaron mis datos, cuando la biblioteca digital los pasó ya a sus computadoras, yo estaba con mucho trabajo poniendo al día el manual nuestro [CSC y AEA] y también terminando mi libro sobre adquisición bilingüe. Entonces, han puesto los metadatos que aparecían en cada cassette, que pueden no estar completos ni ser exactamente correctos. Así es que está un poquito en pañales. Sin las transcripciones, es problemático porque requiere escuchar por mucho tiempo todas estas grabaciones, para poder ver lo que hay allí. Pero, te digo, están abiertas al público y cualquier persona puede usarlas. Yo te puedo dar el enlace.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. ¿Y la calidad del sonido es buena, suficientemente buena?

CSC. Yo diría que hay muchas grabaciones con buena calidad y otras... Hay una variedad. Cuando yo hice mi estudio, escogí las mejores grabaciones y esas transcripciones yo las edité. Habían sido hechas por mis estudiantes, algunos argentinos, que habían trabajado ya mucho conmigo. Así, algunas traducciones están mejores que otras y las que he editado yo están también en papel. Las correcciones las hacían con un lápiz rojo para que se viera qué correcciones había entrado yo. Pero yo estaría absolutamente encantada de que ustedes pudieran hacer uso de eso.

FMF. Apuntas entonces una posibilidad de abordar el trabajo sobre la lengua hablada que sería contar con tu corpus, pero probablemente también con otros corpus de transcripciones sociolingüísticas que existan en los Estados Unidos. Esa podría ser una forma de comenzar: es decir, contar con lo que ya existe, si es que se quiere aportar o ceder para un uso de este tipo. ¿Creen que habría algún problema en eso? ¿En Miami qué es lo que hay; corpus que ya existan?

AL. Yo no conozco ningún corpus en Miami. Tengo grabaciones que he hecho yo a lo largo de los años, pero nada está digitalizado. Yo de hecho estoy en proceso de proponer un grupo para recoger un corpus de español en Miami, más o menos al estilo que ha seguido Ricardo en Nueva York. No sé si será tan extenso, pero sí un corpus que sea "representativo", perdonen, de los distintos grupos en Miami, distintas generaciones, etcétera. Y, si eso se concreta, pues claro, yo igual que

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Carmen con mucho gusto se puede entrar aquí. Yo, de hecho, esto lo había comentado contigo [FMF] hace un año. Yo creo que sería genial que tuviéramos una colección de entrevistas de estilo sociolingüístico, controlando una serie de variables en distintas zonas de Estados Unidos, pero siguiendo siempre una misma metodología; tener una metodología uniforme, como ha hecho Ricardo en Nueva York, pero que incluya Miami, Chicago, Los Ángeles, etcétera etcétera. Supongo que las grabaciones que tendrá Carmen…

CSC. Son solamente mexicoamericanos; inmigrantes recientes que no habían estado en Estados Unidos por más… O sea que habían llegado a Estados Unidos después de los 10 años o de los 14 años, creo, con la excepción de los jóvenes, que podrían haber llegado de los 10 años en adelante; segunda generación y tercera generación. Y solamente mexicoamericanos, son más o menos 60 personas, o más.

AEA. ¿Y las grabaciones son de los años 80, 90?

CSC. Todas de los 80. Y después está el otro corpus del Este de Los Ángeles; perdón, del Valle de San Fernando, que es todavía más pequeño: son 20 hablantes, todos adolescentes, todos mexicoamericanos de segunda generación. Con mucho *Spanglish*, con mucha alternancia entre el inglés y el español. Tú [DD] usaste parte de esos datos también. Pero te digo, todo eso, lo que está transcrito está a mano y hay que corregirlo mucho, desafortunadamente.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Sí. Ese problema lo tenemos todos. El corpus que yo hice en Nueva York, para empezar, no lo hice yo, lo hicimos Ana Celia Zentella y yo, así que cualquier cosa que se fuera a hacer habría que consultar con Ana Celia.

FMF. Sí, por supuesto.

RO. Tendríamos que hablar los dos. Está todo en audio y todas las transcripciones son digitales; ese corpus se ha usado muchísimo. Ha habido artículos de muchísima gente utilizando ese corpus. Yo soy un poco renuente a hacerlo público, porque nunca hemos tenido dinero para corregir nada más que dos tipos de rasgos. Uno, el uso de los pronombres y la parte léxica, porque va a salir un libro ahora, basado en ese corpus, de Rachel Varra, que ella estudió bien. Pero yo pongo la mano en el fuego que lo que está transcrito en cuanto a los pronombres y a las palabras prestadas del inglés, eso es exactamente lo que está en el audio. Pero yo no estoy seguro de que, sin corregirlo, uno pudiera decirle al mundo: "Esto que está aquí escrito es exactamente ese audio", porque no se hizo con ese fin todavía. Entonces, harían falta recursos para que alguien se sentara y volviera a oír, para que ese corpus en su totalidad fuera fiable. Así que esa parte a mí siempre me ha tenido un poco preocupado y hemos tratado de ver si alguien se anima a hacer esa revisión, porque es un trabajo de romanos.[4] Tenemos, también, como 200 horas y hay que revisarlo. Lo que sí quizá se podría hacer, si Ana Celia estuviera

[4] La revisión de una treintena de entrevistas del corpus Otheguy-Zentella se realizó finalmente en el Observatorio del Instituto Cervantes en la Universidad de Harvard a lo largo de 2018.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

dispuesta es, hacerlo a cuentagotas. Es decir, yo tengo entrevistas con personas de seis nacionalidades y dos generaciones. Yo no hice lo de la estratificación generacional tan bien como lo hizo Carmen. Yo no tengo nada más que dos. Tú [CSC] tenías más grupos, ¿verdad? Tú tenías como tres grupos generacionales.

CSC. Sí. Primera, segunda y tercera. Exacto. Porque muchos de tercera generación te hablan español con mucha fluidez y otros, por supuesto, con poca fluidez también.

RO. Yo hice dos nada más. Pero esos datos están digitalizados. Si se pudieran corregir, para que uno pudiera confiar completamente, sea cual sea el rasgo que se quiera estudiar… Eso es lo que yo todavía no podría decir. Pero, con todas esas salvedades, pues, yo también. Y también con la salvedad de que la universidad… Yo estoy ahora en mi primer mes de jubilación y la jefa de cátedra de mi Departamento ya ha dicho lo que dice Carmen: "Un momentito, tú hiciste esto aquí con fondos de *National Science Foundation*, pero eran fondos de aquí y queremos que tu corpus sea de aquí". Entonces, también hay que hablar con la Biblioteca y todo eso. Pero yo acabo de jubilarme, así que nada de eso se ha hecho todavía. En eso estamos. Pero, con todo, teniendo en cuenta todas esas salvedades, pues sí, sería muy lindo poder contribuir con el corpus Otheguy-Zentella.

FMF. Imagino que estáis hablando de horas de grabación. En número de palabras o de ítems, como se maneja también en la terminología de los corpus, ¿de cuántos estamos hablando? ¿Tenéis una idea o no?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Ni idea. Nunca lo he hecho. Será cuestión de entrar y mirar. Para nuestro libro [Otheguy y Zentella 2012] utilizamos 140 entrevistas y, luego, los artículos que se han hecho han utilizado subconjuntos de eso. Hay quien ha utilizado 10; hay quien ha hecho 20; hay quien ha hecho… Y, lo mismo, la calidad del audio varía mucho. Por ejemplo, cuando Danny Erker hizo su tesis, que la hizo con ese corpus, quería saber de /s/ final y entonces dijo: "No, esta cinta no nos sirve; este fichero no; este sí." Y así fueron haciéndolo.

FMF. Nosotros, en el corpus PRESEEA, para cada comunidad, cada ciudad investigada, que como máximo tienen 108 informantes –dependiendo del tamaño puede haber desde 54 a 108– el número de palabras viene a ser entre 300.000 y 500.000. 500.000 si son unas 100 entrevistas.

CSC. ¿De cuánta duración, cada entrevista?

FMF. Varía también un poco, pero suelen ser 45 minutos; algunas tienen una hora u hora y cuarto. En algunos lugares han hecho media hora; por eso varía entre 300.000 y 500.000.

RO. ¿Y cuánto es una hora? ¿Cuántas palabras hay en una hora más o menos? Así *grosso modo*.

FMF. No sé. No he hecho el cálculo. Contamos con que 100 informantes suponen unas 500.000 palabras. Más o menos, ¿no?

CSC. 100 horas, 500.000 palabras.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

AEA. Me suena poco, porque en el PRESEEA de Palma, con 54 entrevistas, te sale medio millón.

RO. ¿Cuantas palabras tú crees que haya, entonces, en las 200 entrevistas con Carmen?

CSC. Son de una hora más o menos.

FMF. Estaríamos hablando, entonces, de un máximo de un millón.

CSC. Porque, yo, esos cálculos los hice con tres o cuatro ciudades al principio, para tener una idea del volumen total que podría efectivamente alcanzar. 200 horas. No 200 entrevistas; 200 horas.

RO. ¿Y cuántas personas fueron? ¿Recuerdas o no?

CSC. Las que yo estudié, 50. Pero tengo más entrevistas que no estudié. Todas están ahí.

GPR. ¿Por qué no son fiables esos textos que han sido corregidos? [a RO]

RO. Bueno, porque eso lo hace un lingüista muchas veces novicio, una persona que estaba oyendo y transcribiendo. Y, cuando tú las revisas, hay equivocaciones. La gente se equivoca.

CSC. Problemas de ortografía, muchos.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Pero, además de eso, lo que tú oyes, lo que el segundo y el tercer par de orejas oyen, no es lo mismo que el primero.

FMF. A veces se oyen cosas increíbles, disparatadas. ¿Y cómo es posible que alguien haya entendido esto que claramente es así? ¿Verdad?

RO. Te cambian la preposición; te cambian el tiempo verbal; ponen un lexema que no era. Hay que oírlas una segunda o una tercera vez para poderle decir al público, en general: "Esta transcripción refleja este audio". Bueno, por otra parte podríamos ceder los audios, sin transcripción, y entonces ahí sí que no hay problema.

CSC. Por eso cualquiera tiene acceso a los míos. Están ya puestos.

FMF. Por eso preguntaba también por la calidad. Ahora hay sistemas de transcripción automática, que se utilizan para subtitular cualquier tipo de vídeo, y lo hacen casi en tiempo real. Te cobran un buen dinero por hacerlo, pero…

MLP. Pero este es el mismo, creo que es el mismo problema, que el que transcribe tiene que revisar, porque yo he pasado videos en mis clases, donde están hablando en español, y empiezan a salir unas transcripciones que no tienen nada que ver con lo que están diciendo.

JC. Mira, lo automático no es fiable. Complicadísimo. Porque tú lo ves en la televisión; yo le pongo automático *Closed Caption*, creo que es– y entonces yo escucho y las transcripciones a veces tienen tremendas equivocaciones. No es

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

muy confiable. Es como la traducción automática: de pronto uno ve cosas que no tienen sentido. Y hay muchas muchas equivocaciones. Yo lo veo constantemente y lo controlo con la televisión.

FMF. Sí. Bueno, hay distintos sistemas. Ahora estamos probando un sistema de subtitulado de todos nuestros vídeos y nos hemos sorprendido de lo bien que lo hacen. Pero hay que revisarlo, porque no lo hacen perfecto. Siempre hay fallos, malas interpretaciones o un mal sonido en un momento determinado. Pero, bueno, quita mucho trabajo.

MLP. Cuando yo transcribí en su momento habla infantil, no es nada más que uno oiga bien; hay que entrenar para que la otra persona sepa cómo se transcribe, las implicaciones de corregir, de incluir, de no incluir, de normalizar el habla o no hacerlo.

JC. Pero, a veces, incluso aunque seas un lingüista que conozca el léxico del que se está hablando, hay curvas de entonación que bajan al final de la oración y uno no entiende cuál es la última palabra que se dijo. Muchas veces ocurre.

AEA. Lo mejor es que el que acaba de hacer la entrevista, mientras se acuerda de todo eso, se ponga a transcribir. Pero lo difícil es cuando transcribes una entrevista que hizo otro y te imaginas lo que ha dicho.

FMF. ¿Qué más corpus creéis que podríamos incluir? ¿Con qué corpus podríamos negociar o hablar?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Casi todo el mundo tiene; todo el que ha publicado trabajos de sociolingüística. Danny [Erker] tiene un corpus de hablantes de Boston. Todo el mundo tiene mucha gente. El problema va a ser siempre el mismo: si esa persona puede donarlo. Yo sé que Danny, por ejemplo, en lo de Boston, que aprendió hacerlo conmigo, lo ha hecho mucho mejor. Quizás lo de él esté más listo para compartir, no sé.

JC. Ahora. Yo creo que una fuente rica, pero también, claro, exige mucho trabajo, es la grabación de las radios hispanas.

FMF. Sí. Eso es otra una cuestión importante. Porque aquí estamos hablando de entrevistas sociolingüísticas, que suponen un tipo de lenguaje y un tipo de interacción, pero no es toda la lengua hablada, evidentemente. Probablemente convenga incorporar al menos otros dos tipos de lengua hablada: uno sería lengua hablada en medios de comunicación, aunque tiene el problema de que hay mucha lengua escrita para ser leída y a veces es difícil…

JC. Pero en la radio, hay chicas, hay mucha gente que habla espontáneamente. Y es muy interesante, muy variado y muy espontáneo.

FMF. …y el otro tipo sería el de las conversaciones espontáneas. Conversaciones más familiares o coloquiales, en otro tipo de contextos, que no sea la entrevista sociolingüística. Son como tres tipos de textos hablados, suficientemente variados, ¿no? para representar la lengua hablada en general.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

AEA. He mirado algunas entrevistas del PRESEEA. Son de 45 a 50 minutos y tienen unas 10.000 palabras de media; entre 9.000 y pico 11.000. Es decir que, de 54 entrevistas, me salen casi 600.000 palabras.

FMF. Entonces, las comunidades de 100 horas supondrían 1.200.000 palabras, o algo así. También varía según la duración y según la velocidad de habla y cosas así, ¿verdad? Colegas, estamos hablando de muchos millones para hacer un corpus representativo.

CSC. No es representativo. Acuérdate.

FMF. Bueno, vamos a quitarle un poco de hierro a lo de "representativo". Hay una representatividad estadística, que no buscamos. Evidentemente, hay una representatividad cualitativa: todo lo que salga de un lugar en cierta manera lo representa. Y luego está la representatividad vista desde afuera: un material que presentemos desde los Estados Unidos, si los demás lo reconocen como de Estados Unidos, nos está representando, frente a materiales de otro origen. Es decir que no estamos hablando de una representatividad estricta, cuantitativa, estadística y sublimada, sino de algo mucho más sencillo.

AEA. Sí. Pero en lo que yo no había caído es en que hacer algo sobre Estados Unidos tiene el problema de lo legalista que es todo en este país. Porque en otro sitio, bueno, tú cuelgas cualquier cosa en la red, a todo el mundo le da igual, nadie va detrás de ti. Pero aquí veo que está todo sujeto; todo tiene dueño; todo tiene derechos.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

CSC. Bueno, yo tuve que firmar unos papeles. Yo no quería que se perdieran mis grabaciones. Así que fui a hablar a la Biblioteca y me dijeron. "Estaríamos encantados". Entonces ellos pusieron mucho dinero para poner en un *site* todos esos datos. Cuesta. Y yo estoy segura de que ellos no pondrían absolutamente ningún pero. Si es sin fines de lucro, es solamente para..., normalmente los suelen ceder. Es cuestión de decirles, me han dicho a mí: "Mira si hay gente que...", Y se lo dijeron a *Carnegie Mellon University*: "Todo lo que ustedes tienen que hacer es poner ahí, cuando se usen, que vienen de la Biblioteca Digital de USC".

68

FMF. Sí. Yo creo que eso se puede negociar desde una institución, puede ser desde el Observatorio en Harvard, desde la Academia. En fin, se explican los fines, se negocia, se acuerda, se firma y se presenta en la forma necesaria.

RO. Pero qué interesante que tú lograste que ellos te dieran fondos para digitar lo que estaba nada más que en audio.

CSC. Lo hicieron ellos. Yo fui solamente a revisar. Entregué todo, les organicé todo y colaboré un poquito con los metadatos, la "metadata", como dicen ellos. Pero no está completa todavía. Y están las entrevistas con niños. Siéntese en el suelo y juegue con ellos. Es decir: uno aplica toda esta metodología. Hay que tenerlo en cuenta. Son entrevistas muy espontáneas, están hechas con la metodología laboviana, son bastante buenas, pero no son completas. Pero creo yo que sí reflejan un español oral de estas diferentes localidades. Y los datos de Rena [Torres Cacoullos] son excelentes, los que hay de producción en Nuevo Mexico.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Mejores que los míos porque las entrevistas, muchas, fueron hechas por sus estudiantes novomexicanos.

FMF. María Luisa, ¿querías comentar algo?

MLP. Estaba pensando en cómo se resuelve el problema de los contextos de uso de la lengua oral. Estaba pensando en el español médico. Pero, bueno, con eso no se puede hacer nada porque es una cuestión confidencial, ¿no? No tenemos acceso a eso. Nada más estaba pensando. Y que hay varios contextos donde seguramente hay muchísimas palabras, muchísimas expresiones. Sería muy rico poder obtener información de esos contextos, pero no creo que se pueda, ¿no?

AEA. Pero Ricardo hablaba más de documentos públicos. En los hospitales suele haber folletos y mucho material.

GPR. Y la televisión también, supongo, ¿no? Entrevistas.

FMF. Sí. Radio y televisión.

GPR. Tener acceso a las entrevistas, que han hecho 40, 50 entrevistas, en HITN, "educa y entretiene".

CSC. Creo que ellos las tienen transcritas. No solamente la parte audio o oral, sino que también la transcripción, de acceso inmediato. Y en *Youtube*.

DD. Yo Iba a preguntar, ¿cómo hacemos con las conversaciones orales? ¿Necesitamos hacer grabaciones nosotros?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Otro género hablado que sería interesante es, todo lo que es el deporte, las narraciones en español de los *Mets* o de los *Yankees*. Todo eso puede ser muy interesante como datos del español en Estados Unidos.

FMF. Sí. Dentro de los medios de comunicación, entrevistas, en radio y en televisión, narraciones deportivas, coloquios, por ejemplo, deportivos también, ¿no? Hay muchos debates o coloquios.

JC. Sí, en fútbol, en boxeo, en básket, en béisbol; que yo no sé nada de béisbol, pero...

CSC. En *soccer*. Los programas de la mañana, a las ocho o nueve de la mañana que son...

FMF. Evidentemente, hay otra vía, a propósito de los materiales verbales, que es emprender la recogida y la elaboración de estudios nuevos, para los que se podrían utilizar métodos en la línea que tú proponías, estudios homogéneos en distintos lugares. A lo mejor la metodología de PRESEEA podría servir de referencia para todo eso.

AL. ¿No hay PRESEEA de Estados Unidos? Perdona.

FMF. No, no. Intentamos crear un grupo de Miami, cuando estaba Humberto López Morales en plena actividad y visitaba Miami constantemente, pero finalmente no lo logró hacer.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Nos vamos a tropezar con lo que dice Andrés [AEA]: el leguleyismo de la ciudad, de la sociedad norteamericana en general. Porque, en las entrevistas mías, por lo menos cada entrevistado firmó que se le permitía entrevistarlo. Lo cual es siempre interesante, porque, si usted me está hablando, parecería que me ha dado permiso.

FMF. Bueno, eso lo hicimos en Alcalá de Henares y en otras ciudades. Tenemos los recibos de todos los informantes, con su aceptación. Y, además, al comenzar a hablar en una entrevista, se pedía una confirmación de la aceptación, pero bueno… Lo que ocurre es que emprender una línea de PRESEEA en los Estados Unidos, pues lleva tiempo. PRESEEA no es algo que se consiga de un día para otro. En Baleares han tardado un poquito en conseguir PRESEEA-Palma de Mallorca, ¿no? Y lleva tiempo, pero bueno, es una línea que no hay que desechar.

RO. Volviendo otra vez a estos tres córpora que ya se han hecho, ¿tienen lengua hablada también o no? Estos tres: el CREA…

FMF. Sí, tienen lengua hablada. Lo que hicieron para el CREA, que es el que tiene el subcorpus de lengua hablada más grande, fue, en general, reunir materiales que ya estaban recogidos para otros proyectos. Desde Alcalá, aportamos un corpus de conversaciones que recogió Ana María Cestero. Todo el corpus de la norma culta, las transcripciones, se incorporaron directamente. De hecho, firmamos un acuerdo entre PRESEEA y la Academia Española para pasarle las entrevistas de PRESEEA y que pudieran integrarse en el corpus de español hablado. O sea que sí

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

tienen. Lo que ocurre es que la proporción que le dan, dentro del corpus general, es de un 10%.

RO. Lo que estoy pensando es que, si se pudiera crear con los córpora que existen un corpus de PRESEEA de Estados Unidos, ¿tendría sentido eso? ¿O habría un PRESEEA de Miami, un PRESEEA de Los Ángeles…?

FMF. Los PRESEEA por definición son de ciudades. Lo que ocurre es que sí se puede plantear un proyecto colectivo, más amplio, nacional. En Guatemala hicieron un PRESEEA-Guatemala general y aplicaron el sistema de encuestas en Guatemala y en otra media docena de ciudades. Lo llamaban PRESEEA-Guatemala, pero en realidad incluía una serie de ciudades.

RO. Pero, por ejemplo, las entrevistas de Carmen en Los Ángeles, las mías en Nueva York, las de Andrew [Lynch] en Miami, a las de Danny [Erker] aquí, ¿podrían, de alguna manera, reformarse para que fueran PRESEEA?

CSC. No. Las mías no.

FMF. Las de PRESEEA son semidirigidas, siguen un guión de módulos, con una metodología para obtener temas y estilos discursivos diferentes. Y supongo que en las entrevistas que se han hecho en los Estados Unidos se han seguido criterios muy diferentes. Bueno, pasamos entonces a la cuestión de la lengua escrita.

RO. Me perdona. ¿Ese es un interés que tienen en el Observatorio?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Sí, bueno. El Observatorio lo que pretende es, de este abanico de posibilidades que estamos comentando, ver qué puede hacer, qué es factible en un año. Nosotros estamos dispuestos, a lo largo de todo este año, a trabajar en la elaboración de una parte de esto que estamos hablando. Dependiendo de los objetivos que decidamos, vamos a trabajar en recoger materiales de literatura o de medios de comunicación; o vamos a trabajar con algún corpus de lengua hablada que podamos obtener a corto plazo, para ver la manera de crear una cabecera y un método de transcripción que sea válido…

RO. ¿Pero tú tienes, el Observatorio tiene recursos como para poner, por ejemplo, a alguien a oír un audio y una transcripción y chequearla?

FMF. Ahora mismo, sí. A lo largo de este año lo tenemos garantizado. Pero esa tarea, digamos, es un grano de arena para lo que se necesita. Intentaremos, dentro de todas estas posibilidades, obtener el máximo rendimiento con el esfuerzo que podamos hacer.

RO. Entonces, concretamente en el caso mío y de Ana Celia, si Ana Celia estuviera de acuerdo, y yo te traspaso 20 transcripciones (son seis países) o que fueran 30 para que fueran 5 informantes de cada país de origen… Están hechas, pero con este temor que tengo yo de que no se han verificado lo suficiente. ¿Tú tienes recursos para que alguien las vuelva a oír?

FMF. Sí.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Bueno, pues entonces eso sería una manera de hacerlo para mí.

FMF. Por algún lugar hay que empezar. Este podría ser el comienzo, pero, como vemos, hay muchas posibilidades. Si esto se concreta en un plazo corto, sí podría ser.

RO. Yo estoy muy dispuesto y creo que Ana Celia estaría también dispuesta. Claro, con todo lo que ya sabemos que no son nada más que para fines científicos, que no puede haber fines de lucro.

FMF. Por supuesto. Y eso habría que ponerlo en un papel y firmarlo.

RO. Bueno, pues, yo lo consultaré con Ana Celia [Zentella].[5]

FMF. Y, luego, el sistema de cabeceras y de etiquetados sería el mismo que se utiliza en el corpus del siglo XXI, que además es muy simple: ya incorporó toda la experiencia del CREA anterior, que era muy ambicioso en cuanto a etiquetas, y se simplificó al máximo. Y se podría utilizar también como complemento del sistema de etiquetado de PRESEEA, que también está muy simplificado y que etiqueta los rasgos más objetivos; no los rasgos "opinables", sino los rasgos que son claramente percibidos y conocidos.

DD. ¿Y el corpus de Rafael [Orozco] es diferente al tuyo [RO]?

---

[5] Finalmente, en 2018 se firmó un acuerdo para incorporar al CORPEEU una treintena de entrevistas del corpus Otheguy-Zentella.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Totalmente. No solo diferente, sino probablemente mucho mejor. Así que yo creo que sí. Lo de Rafael seguro que está muy bien hecho; trabaja muy bien. Creo que es nada más que de colombianos. La única ventaja que tiene el corpus de Ana Celia y mío es que tenemos seis nacionalidades y entonces es más completo. Pero lo de Rafael seguro que está magnífico. Estoy seguro.

Lengua escrita en intermet

FMF. Bueno, muy bien, pues es una posibilidad para empezar. Pasamos entonces a cuestiones de lengua escrita y a cuestiones de internet y de redes sociales, que incorporan lengua tecleada y otras manifestaciones. ¿Cómo podríamos orientar ese corpus? Para empezar, hay un hecho que creo importante; y es que existe un mega corpus elaborado por [Mark] Davies a partir de páginas electrónicas, de periódicos, de blogs… Hay un listado que he incluido en la documentación y que supone muchísimos millones y millones de palabras para los Estados Unidos. Ahí tienen los millones para Estados Unidos.[6]

AEA. Este es el *Corpus del Español*, ¿no?

FMF. Tiene 180 millones de palabras solo de Estados Unidos. Es decir, que hay ahí ya un material, un poco misceláneo, porque tiene noticias, tiene blogs, tiene páginas de contenidos diversos, de entretenimiento, de deportes, de cosas

[6] Puede consultarse el listado a través de este enlace: https://www.corpusdelespanol.org/x.asp?c=3

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

extraídas directamente de la red. Si eso ya está ahí, con ese volumen, a lo mejor se podría plantear una negociación con Davies.

JC. Perdón, ¿cuál es ese corpus?

FMF. Se llama *Corpus del Español*, de Mark Davies, de Brigham Young. Está en la documentación que os pasé, con la tipología e incluso las fuentes que manejó, para que valoremos si son fuentes adecuadas.

RO. Yo he tenido estudiantes que han mirado a Davies, pero yo nunca lo he mirado.

FMF. Ese es el listado que el propio Davies da de las páginas electrónicas de las cuales ha extraído los textos en español.

JC. Ahora, mira. Este… Yo hice un trabajo para la *Enciclopedia del español en Estados Unidos* sobre las jergas juveniles y en ese trabajo utilicé redes sociales de contacto. Al final opté por una de ellas, que era la que me permitía distinguir por nacionalidad. De hecho, busqué *hispanics,* latinos, hispánicos, divididos por edad, por sexo y por procedencia. Y entonces ahí encontré una fuente riquísima de material que utilicé para eso, para ver las características de las jergas juveniles. En internet encontré 40 rasgos e hice dos ejemplos de cada uno. Redes de contacto social. Los chicos que están hablando entre sí, reciben mensajes, envían mensajes.

RO. ¿Qué red en concreto era?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

JC. Estos eran *Hi5*, que era una de esas redes … Esa fue la que me permitió distinguir por rasgo, nacionalidad, estado, procedencia. Y, entonces, eran todos chicos hablando entre sí. Eran, no sé, hasta 19 o 20 años. Fue una fuente riquísima de contacto, ¿no? Porque allí los chicos estaban hablando de la manera más espontánea posible. También están los foros de chat, donde hay conversaciones que son así, vertiginosas y creativas, tecleadas.

FMF. Tecleadas, sí. Nosotros hemos estado haciendo últimamente unas búsquedas en el Observatorio. Uno de los proyectos en los que hemos venido trabajando ha consistido en la detección de anglicismos en el español de los Estados Unidos. Entonces, hemos acudido a redes sociales para extraer textos en bruto y, a partir de ahí, hacer el análisis y ver cuáles son anglicismos, sobretodo, innovadores, neologismos; anglicismos del español de Estados Unidos. Bueno, lo hemos hecho con *Twitter*, porque en *Twitter* se pueden comprar millones de mensajes y luego tú buscas la manera de filtrarlos y limpiarlos, de analizarlos. Ahora estamos intentando trabajar con blogs y con redes. Hemos visto también las redes de contacto, pero la mayor parte de esos contactos son contactos sexuales. El problema que tienen esos contactos, esas conversaciones, esos chats, es que son conversaciones muy fragmentadas. Son palabras sueltas; por supuesto, mezcladas con emojis o emoticonos y con todo. Con lo cual lo que se pueda aprovechar del discurso es mínimo; te obliga a procesar millones de elementos, de palabras, para que el fruto final, desde un punto de vista lingüístico, sea muy poco, en esos chats, por lo que hemos visto. Pero sí, hay blogs efectivamente; de hecho

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

algunos de esos blogs, con los que ha trabajado Davies, sí aportan bastante información. Pero planteaba, si eso se considera adecuado, hablar con Davies para una posible cesión; aunque no sé si Davies lo cedería o lo vendería.

AEA. Lo vende seguro.

FMF. SÍ, sí, bueno. Además, es que no te habla de precios, pero dice que te deja consultar hasta cierto punto. Si quieres más, te lo vende.

AEA. Lo que te va a vender es listas de palabras, frecuencias de palabras, de letras. No te puede vender los textos en sí porque no son suyos. Lo que el explota es que te vende listas de palabras por frecuencia y cosas de ese tipo.

FMF. Sí, como lo que se sacan con NGram, ¿no? Ese sistema es de *Google*, de búsqueda en *Google Books* para cualquier secuencia de palabras o palabras sueltas; y te da su frecuencia a lo largo de los años.

RO. Pero, a ver si te entiendo, Andrés [AEA]. Cuando se terminara este proyecto y hubiera un gran corpus del español en Estados Unidos, tampoco se le daría a nadie ese corpus, ¿verdad?

AEA. Opcionalmente. Pero, claro, como los textos tampoco serían propiedad en su gran mayoría de los que gestionan esto, entonces no podrían darlo. Cuando se termine este proyecto y exista ese corpus, la persona que entra en la página dónde esté ese corpus puede hacer búsquedas de si se dice "llamar para atrás" o si no se dice, o lo que sea; se le dan las ocurrencias y las concordancias, pero nunca se

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

le va a dar el texto. ¿No? Me imagino que lo que habrá es una concordancia, que es el contexto precedente y posterior; no sé cuántas palabras, 10 para un lado y 10 para el otro. Normalmente, en casi todos los corpus, tipo CORDE, pinchas en la palabra y te dan un contexto un poco más amplio, 250 o 299 palabras.

CSC. Claro, por lo menos. Si no, no te sirve a nada, pues.

RO. Por eso no entiendo lo que nos has advertido sobre Davies. Davies entonces es igual.

AEA. Davies lo que hace es vender. O sea, tú puedes hacer consultas. Lo malo es que, cuando haces más de un número de consultas, te empiezan a aparecer ventanas: "Regístrese porque usted ha hecho ya más de no sé cuántas consultas" Y luego te empiezan a aparecer ventanas: "¿Ha usted considerado donar dinero?" O también aparecen ventanas: "Le damos los 99 primeros resultados. Si quiere más, tiene que pagar tal".

FMF. Entonces, una posibilidad es utilizar ese material como un corpus general. Es decir: si estamos interesados en comprobar la frecuencia de uso de una determinada forma, porque nos han hecho una consulta, vamos a poner, de la Asociación de Academias, podríamos decirle a Davies: "¿Cuánto me cobras por darme las ocurrencias y las concordancias de tales palabras? Y dice: "Son $300".

AEA. Podría ser algo así. De todas maneras, veo que él da un listado de fuentes, ¿no? Si interesa crear un corpus propio, esto puede servir de inspiración. Uno ve

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

aquí, seguramente, algunos blogs de esos. Lo que pasa es que aquí no ponen la referencia, como en nuestra bibliografía, donde dice: "Consultado tal día", para saber exactamente, pero todas tienen una fecha detrás. Alguna sí la tiene explícita y otras no.[7]

RO. Pero entonces, la misma pregunta. Carmen ha dicho que ella con mucho gusto cede los datos y yo te he dicho que voy a investigar con Ana Celia a ver cómo podemos hacerlo también. Si existiera un banco de, qué se yo, 20-30 audios y transcripciones que se han cedido al Observatorio o al Cervantes o a la Academia –todavía no sé exactamente a quién–, los que buscaran en ese corpus ¿tendrían, entonces, acceso a toda la entrevista o, igual, nada más que a ítems y concordancias? Eso a mí no me queda claro.

FMF. El acceso para el público sería solo a las ocurrencias; a los ejemplos con las concordancias en la medida que se decida.

CSC. Una concordancia amplia o, si no, no son muy útiles.

FMF. Sí, bueno, la concordancia que digamos; la que legalmente se nos permita para fines científicos. Pueden ser, eso, 200 palabras o 300 palabras. Lo que no impide que, internamente, si lo tiene la Academia, esta pueda hacer sus consultas.

[7] Realizado un posterior análisis de la procedencia de los materiales de los Estados Unidos en el corpus de Mark Davies, comprobamos que en su inmensa mayoría procedían de páginas alojadas en servidores de los Estados Unidos, pero con contenidos de escritores de fuera y destinados a un público externo a este país.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Pero no para hacerlo público, sino para trabajar con ello. Igual que cuando un investigador entra en un corpus para hacer un recuento de /s/ o de un determinado léxico. Pero eso no sería público.

CSC. ¿Y eso? Porque la transcripción es una transcripción sin ningún código, o sea de palabras, de estructuras.

FMF. Habría marcado y etiquetado.

CSC. ¿Quién lo hace y cómo se hace?

FMF. Podríamos intentar hacerlo nosotros, si es que las transcripciones no vienen con ningún tipo de etiquetas.

CSC. El mío no tiene etiquetas. La transcripción es una transcripción.

FMF. Los textos de PRESEEA sí tienen. El sistema de etiquetas PRESEEA es simplificado, bastante objetivo, porque no implica decisiones subjetivas por parte del etiquetador o del transcriptor. Eso se puede abordar. Tenemos experiencia en hacerlo.

CSC. Pero no es automático.

FMF. No. Hay que ir transcripción por transcripción. Sería relativamente fácil: mientras vas oyéndolo, vas poniendo la etiqueta correspondiente.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Y en la sección oral del CREA y del CORDE XXI, la sección oral, ¿tiene etiquetado?

AEA. ¿Pero estáis hablando de etiquetado del tipo de "risas", "interrupciones" ...?

RO. No. Estoy hablando de etiquetado gramatical. Yo pensaba que estábamos hablando de etiquetado gramatical.

FMF. Eso no lo incluye, pero hay algunos corpus que sí lo tienen. Lo que tiene Guillermo Rojo para el español de Galicia sí está etiquetado gramaticalmente. En PRESEEA, no. En PRESEEA solo son las etiquetas, las marcas de solapamiento, risas... Por eso decía que son rasgos objetivos que no suponen mayor complicación y que son relativamente fáciles de transcribir. Pero no está etiquetado gramaticalmente, aunque hay sistemas automáticos para ello.

CSC. Nosotros no pusimos nada de risas. A veces, en la transcripción, pero en general no hay nada. Porque no me interesaba hacer un estudio...

RO. Claro, eso es lo que pasa, que uno hace estas cosas pensando en un rasgo y entonces, cuando uno lo publica, el consumidor piensa que el que lo hizo lo hizo fidedigno totalmente al audio y es ahí donde está la preocupación.

FMF. Bien. Retomando el hilo que llevábamos respecto a la lengua tecleada, podríamos ver, por un lado, la posibilidad de recuperar o conseguir textos nuevos a partir de la "inspiración" que nos ofrece este listado, lo que implicaría un proyecto, claro, un trabajo técnico, que podríamos hacer también desde el

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Observatorio. Eso parece algo factible en un plazo corto. Sería una buena manera de entrar en este ámbito de la lengua tecleada. ¿Sí? ¿Os parece?

RO. ¿Te puedo hacer una pregunta que quizás esté atrasada con respecto a tu temario? Si alguien te dice: "Profesor Moreno Fernández, ¿usted para qué quiere hacer esto o para qué el Observatorio quiere hacer esto, el Cervantes o la Academia, si ya lo hizo Davies? Ya Davies tiene un corpus. ¿Por qué ustedes están reinventando la rueda? ¿Cuál es la respuesta a esas preguntas?

FMF. La respuesta es que el acceso a los materiales de Davies es limitado y nosotros queremos tener acceso a una información lingüística completa.

RO. Y si Davies fuera menos empresarial y no hubiera esas ventanas que se abren y piden dinero, ¿haría falta hacer esto?

FMF. Si lo de Davies fuera completamente gratis, completamente abierto y nos dejara acceder a toda la información lingüística, no sólo a las unidades que se han buscado específicamente… Queremos también acceder a textos para proceder a otro tipo de análisis, de fraseología, más complejos, de cuestiones de sintaxis o de otro tipo.

AEA. Yo creo que lo interesante, comparándonos con los materiales de Davies, sería hacer algo más limpio, más pequeño obviamente. Claro, es que estas cosas de Davies son a lo bestia, ¿no? 2 billones de palabras porque es todo sacado de la web. Pero, claro, el problema que tiene eso es saber dónde te metes ahí. Yo creo

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

que lo interesante sería hacer un corpus más confiable en el sentido de que se sepa de dónde salen los datos, de tal entrevista, de tal autor. Esto de Davies es casi como meterte en *Google*.

CSC. No hay metadata.

FMF. Esa es la clave. Si lo hacemos desde aquí, si lo hacemos nosotros, sí podríamos controlar qué tipo de textos se incorporan. Y podríamos además acceder a las cabeceras con todos los metadatos que permitan saber exactamente de qué se trata en cada momento, con criterios cualitativos.

RO. ¿Y con Davies no se sabe eso?

AEA. No.

RO. Entonces, hay dos problemas: no se sabe de dónde viene y cobra por todo eso. Esos son problemas muy serios.

FMF. Se sabe de dónde viene, porque que se da esto [listado de fuentes]. ¿No? Pero lo que hay dentro no lo puedes controlar porque no tienes acceso a todo el texto y no puedes hacer valoraciones lingüísticas. No hay esos metadatos sobre el origen del texto.

CSC. ¿Y qué pasa después que tú te vayas? ¿Se acaba el proyecto?

FMF. No, bueno. Seguro que el Observatorio asumirá los proyectos que están en marcha y continuará con otros proyectos que irán surgiendo. Probablemente, la

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

garantía de continuidad de este proyecto está en su vinculación con la Academia Norteamericana.

RO. ¿Y hay proyectos paralelos? ¿Hay grupos como este que están tratando de hacer algo para la Academia en México, en Colombia y Argentina? Esa parte no la entiendo. Esto sería un corpus que se aportaría, de alguna manera, que la Academia Norteamericana de la Lengua lo va a aportar a la Academia en Madrid. No sé. Esa parte no la entiendo.

FMF. Bueno, el Cervantes también está implicado. La Academia Norteamericana está asociada con las demás Academias, que pueden solicitar información de cosas que están ocurriendo en Estados Unidos; o la propia Academia Norteamericana puede estar interesada en aportar información para que sea incorporada al diccionario general o a otras obras, en los diccionarios del estudiante o los que sea. Entonces, se aporta desde aquí. La propiedad sería del Cervantes y de la Academia Norteamericana. Esto no se daría a la Academia Española ni a la Asociación de Academias, porque no todo les va a interesar.

GPR. Si interesa, que paguen.

JC. También pondrías ventanitas.

CSC. ¿Entonces, la ANLE es la que tiene que contactar, por ejemplo, con [la Biblioteca de USC]?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Si se garantiza una continuidad del Observatorio, sería un proyecto de Instituto Cervantes y la ANLE.

CSC. Eso sería mucho mejor: Cervantes-ANLE, yo creo. Y otra pregunta que tenía es, supongamos que te mandemos unas 10 excelentes grabaciones, con 10 transcripciones que ustedes podrían revisar y todo eso, ¿ustedes luego podrían cederle esas transcripciones a nuestra Biblioteca Digital?

FMF. Por supuesto.

CSC. Ya. Porque a ellos les gustaría mucho.

FMF. En eso consistirían los términos de la negociación. Vamos a hacerlo con fines científicos. Nos cedéis unos materiales generosamente y nosotros los devolvemos, los ponemos a vuestra disposición.

CSC. Una maravilla. Van a estar dichosos.

AEA. ¿Tú sabes si hay posibilidades de escanear la manuscritura? Porque de esto se ha intentado hacer muchos desarrollos informáticos. Lo digo por las transcripciones que están hechas a mano.

CSC. Tú deberías verlas, eso sí. Están hechas en papel amarillo, que no copia bien y muchas veces por los dos lados de la hoja, porque estábamos tratando de ahorrar dinero. Teníamos una beca de la *National Science Foundation* con cierto

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

límite de dinero. Pues, no, creo que por eso las estamos digitalizando. Hay una chica, supuestamente de México, de Monterrey que está…

FMF. …taipeando.

CSC. …tecleando.

JC. Ahora hay unas fuentes de financiación que permiten solicitar al público. Yo conozco a una persona que escribió un libro magnífico sobre la guerra de las Malvinas, con fotos y entrevistas a gente de los dos lados, excombatientes y todo. Y entonces, bueno, yo le hice la edición. Y entonces solicitó por medio de…, este, no me acuerdo del nombre; es para solicitar fondos públicos, y consiguió $50.000. La cuestión era formular el proyecto, entonces esa empresa se encarga de difundirlo. Cada uno lo difunde a sus conocidos: por ejemplo, nosotros podemos difundirlo a 50 numerarios, no sé: 100 correspondientes, 100 colaboradores, no sé; y esta persona consiguió $50.000. Uno establece un proyecto y fija un plazo. Entonces, yo, por ejemplo, ofrecí $50 con una tarjeta de crédito. Si para esa fecha se logra el objetivo, entonces se cobra todo lo que prometieron. Si no se llega a ese objetivo, no cobra nadie. Y creo que la empresa cobraba, no sé, un 3%, una cosa mínima, y esta persona consiguió $50.000 para ese libro. A lo mejor nosotros podemos decir, bueno, necesitamos $100.000 con fecha, no sé, 10 de septiembre del 2018, el proyecto es esto y después difundirlo a los cuatro vientos.

AEA. Todo el mundo anda pidiendo.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

JC. Si se consigue, se consigue y, si no, no se pierde nada, porque no hay que pagarle a esa empresa ni nada. No se consiguió. Pero esta persona para un libro consiguió $50.000. Podríamos solicitar eso, definir un proyecto específico y después difundirlo a todo el mundo. Todo fue con pequeñas donaciones. Ahora, conozco otra persona también, que fue demasiado ambiciosa para financiar un documental y que no consiguió el dinero. Lo que pasa es que el plazo era demasiado corto y el objetivo era demasiado ambicioso. Pero, si uno no es ambicioso y lo difunde a todo el mundo que pueda estar interesado, no sé, no se pierde nada con probar.

CSC. Pon el nombre de Jorge como iniciador de ese plan.

JC. No. Yo puedo preguntarle a esa persona a ver cómo es, cómo lo hizo, cuál es el nombre y cuál es el procedimiento.

GPR. James Fernández también lo hizo así con el proyecto de inmigrantes españoles a los Estados Unidos.

FMF. Hemos hablado del uso o del tratamiento que se le podría dar a toda esta parte de lenguaje tecleado. Al margen de estas páginas informativas de blogs, ¿habría otro tipo de textos procedentes de internet que convendría incluir?

CSC. Como lo decía Jorge, los *webseminars* que se hacen en español. Yo nunca he participado en un *webseminar* en español, pero puede haberlos ¿no?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Hay otro ámbito de la lengua de internet que es el de las informaciones, textos oficiales o textos informativos, que se difunden fundamentalmente desde páginas de instituciones administrativas o de empresas, que se hacen en español.

AEA. ¿Dices la página de la Seguridad Social y cosas así?

DD. Sí, es buena idea.

MLP. Y perdón, ahí sí, Por ejemplo, yo no sé si hay hospitales que tengan sus páginas traducidas en diferentes idiomas.

DD. Sí, este es el tipo de documento de traducción que yo pensaba.

FMF. Pueden ser páginas oficiales pero también páginas de la administración, de los hospitales.

DD. Por esto pensaba yo que las páginas oficiales son traducidas por traductores.

GPR. Bueno, es lo que lleva Laura Godfrey,[8] ¿no? para el gobierno. Todo eso sigue. Eso no lo han quitado.

FMF. ¿Qué más?

MLP. El *Linguistic Landscape* que mencionaste.

[8] Manager de GobiernoUSA.gov desde 2005.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Bueno, eso no entraría en internet. El paisaje sería otro tipo de lengua escrita. Bueno, hemos hablado, en cuanto a lengua escrita, sobre todo de literatura, pero está esa otra lengua escrita que tú comentabas, también, la publicidad, textos publicitarios, paisaje lingüístico, que podría incorporarse y que para el caso de los Estados Unidos tiene un valor también muy significativo.

JC. Sí, publicidad impresa, pero también en internet, digital.

FMF. Los datos de *Twitter* son muy complicados de manejar. Nosotros hemos controlado el origen y hemos hecho una búsqueda de unidades léxicas, pero para sintaxis... A veces es ilegible, entre la combinación de emoticonos, de abreviaturas, de faltas de ortografía, faltas de tecleado..., nunca sabes por dónde va a venir; es muy difícilmente manejable como texto.

JC. Claro, abreviaturas de este lenguaje telegráfico.

FMF. Sí, pero, además, muchas veces son abreviaturas que no responden a una convención fija, sino que se escriben sobre la marcha; que se entienden, pero que no son esperables.

AL. Esta es una pregunta de pura ignorancia, pero Andrés [AEA], ¿este tipo de cosa es muy difícil de hacer? Porque, como has dicho tú, esto [listado de URL en español que se estaba comentando] parece salir de una búsqueda en *Google*. Porque no sé cuáles habrán sido los criterios para elegir algunos de estos textos: marcianosmx.com. Hay de todo: prensa y deportes...: *Gossip Center*, *El Hispano*

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

News; y luego ves cosas de *FEMA*, de *Federal Emergency Management Agency*; o sea hay de todo: de *West Lake University*, estudiar la Biblia, etc. Aparte de la cuestión de cuáles han sido los criterios para recopilar una cosa así, para esto, me siento y digo, este, este y lo otro; de este genero un poco, de este genero, otro poco, etc. Es cuestión de un día, ¿no? Pero más allá de eso, ¿qué tan fastidioso es? ¿Qué tan laborioso es hacer los metadatos? El etiqueteo y tal.

AEA. Yo creo que voy a meterme uno de estos días para investigar un poco más a ver cómo está hecho, pero da la impresión de que está hecho con algún tipo de búsqueda, ¿no?, en *Google* con unos parámetros para que sean páginas de los Estados Unidos. Y luego, ya, pues a recopilar. Prácticamente. No sé. Aquí no creo que haya mucho filtrado de ningún tipo. Sencillamente parece que son páginas de los Estados Unidos en español y ya está. El problema que tiene esto es que nunca sabes quién lo ha hecho.

AL. O sea, ¿será una cosa semejante para los propósitos de ustedes? Esto, al menos, no parece tan difícil. Lo que más difícil me parece es, o más laborioso, ¿se dice? *laborious*, sería la recopilación de un corpus oral, de entrevistas o de metadatos, etcétera, ¿no? Habría que invertir o pensar en más en la cuestión de recursos. Pero esto, tan difícil no me parece. Como dicen, puede ser *famous last words*, ¿no? porque no sé si es muy complicado hacer los metadatos.

FMF. Tiene razón. Por lo que he visto, con los sistemas de búsquedas que manejan *Twitter* y *Facebook*, su búsqueda es automática, detectan la lengua

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

automáticamente, porque, cuando no está etiquetada la lengua, hay unos algoritmos que te permiten saber, tras leer una secuencia de caracteres, en qué lengua está. Entonces, una vez que has detectado la lengua dentro de una web determinada, se cuentan cuántas palabras hay, para saber si es un texto aceptable para tus fines o no, si es demasiado corto o si es demasiado largo, y se extrae directamente. Además, hay programas que trabajan los textos directamente en código fuente y los tienes en TXT en cuestión de minutos.

RO. Pero yo quería preguntar, si pensamos en la primera parte de la conversación, en donde dijimos que incluiríamos paisaje lingüístico, avisos, anuncios, independientemente de su corrección, porque vamos a poder decir, esto proviene de un aviso que apareció en una guagua en Nueva York… Sea como sea, ya el usuario sabe de de dónde sacó ese dato. Exactamente, cuál es la diferencia entre esa actitud y la de Davies. Vuelvo a la pregunta de si estamos haciendo algo que ya Davies ha hecho. Uno podría decir: "Bueno, aparece este dato y salió de un blog que encontró Davies".

AEA. Lo de los letreros, yo no tengo ningún inconveniente en lo de los letreros, siempre que puedas decir esto es del aeropuerto de Houston, una fotografía de tal año o lo que sea. Claro, yo el problema que veo con ponerse a navegar por la red es que muchas veces no se sabe realmente lo que es, quién lo ha hecho y de dónde sale. Esa es mi única cautela, pero quizás sí se pueda. Yo es que realmente no tengo experiencia en compilar corpus de la web. Ahora se hace mucho, ¿no?

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

RO. Pero, antes explícanos, ¿qué es lo que no se sabe en la web que sí se sabe en el caso del anuncio que apareció en el aeropuerto de Houston en el año 2014? ¿Ahí no sale que salió del blog tal y eso no es una información equivalente?

FMF. Cuando hicimos el trabajo con *Twitter*, por ejemplo, miramos miles de mensajes con el único criterio de que estuvieran en español y geolocalizados en Estados Unidos. Pero había muchos que, obviamente… había algunos en catalán, por ejemplo. Obviamente, ese es uno de Barcelona que está aquí de vacaciones y envió un Twitter, ¿no? A eso me refiero, que es una cosa tan inmensa la web, que hay pocas garantías de saber si realmente se trata de alguien de los Estados Unidos. Exige un trabajo cualitativo.

RO. Ya veo. No se sabe que sea de Estados Unidos.

AEA. A eso me refiero.

FMF. Nosotros lo que hicimos fue revisar 80.000 tuits, para hacer una limpia cualitativa que impidiera contar como un anglicismo usado en el español de los Estados Unidos el que utilizó casualmente un catalán que fue a ver a su amigo en Miami y puso un tuit en su red. Pero eso solo se detecta si vas analizando uno a uno. Entonces, lo que propone Andrés es que se revisen cualitativamente también esas páginas. Se puede hacer una revisión primero para evitar que te metan una bola intragable.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

MLP. Una pregunta. Y, por ejemplo, cuando hay versión impresa y en internet de un periódico, ¿la consideramos lengua escrita, impresa o digital?

FMF. Normalmente, dentro de la lengua escrita hay "Prensa". Yo creo que iría a la categoría de lengua escrita, a prensa. Para la parte de internet, estamos hablando de otro tipo de textos; más blogs o páginas oficiales, que pueden ser administrativas, de hospitales, de centros de enseñanza, en fin.

JC. En la prensa también hay muchos textos que son indistintos. Hay otros que tienen una divergencia y otros que ofrecen más material en un medio que en el otro, en un soporte que en el otro.

FMF. Sí, pero para una tipología de este tipo, es lengua escrita de todas maneras.

CSC. Perdonen, voy a volver medio minuto a lo de atrás. ¿Saben qué no recogen los corpora que ya existen de la lengua hablada? Reuniones en comunidades latinas, hispanas. Hay tantas reuniones de centros de vecinos, durante las cuales hay una interacción con el representante del grupo de vecinos y los vecinos que han ido a esa reunión. Yo fui a dos durante la época que estuve haciendo grabaciones y no grabé nada. Fui a unas, que ellos llamaban las tardeadas; no sé si en México se hacen también; es una reunión que hacen el día domingo por la tarde y a veces entonces uno habla y dice algo y después de comer y de hablar y de todo esto, empiezan todos a bailar, incluso. Es casi como una fiesta y la hacen los días domingo. Y la otra fue una reunión en Pico Rivera, de comunitarios, o sea de la comunidad. Y también fue muy interesante porque ahí sí que es muy

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

espontáneo. Pero eso yo no lo he grabado nada. Y no sé si en algún momento se podrá lograr, guardar ese tipo de lengua de la misma manera.

FMF. Creo que ese tipo de lengua tiene un interés lingüístico, pragmático, enorme. Pero creo que hay también un problema técnico, cuando se trata de reuniones de mucha gente: en las reuniones no hay micrófonos para que se oiga bien a todo el mundo, el traslape de turnos es continuo…

CSC. Por eso yo no grabé nada. Fui a dos y dije: "No, aquí yo no". Es una pena porque eso nos dice que, aunque uno vaya, y vaya a tres o cuatro, y escuche mucho y tome nota, no te lo aceptarían, ¿no? Porque tiene que estar de alguna manera grabado.

DD. Pueden ser solamente fichas.

CSC. Tomar notas.

GPR. Y centros religiosos. Porque en Estados Unidos funcionan. Desde luego, los centros religiosos tienen aquí mucho predicamento, nunca mejor dicho.

JC. Pero hay alguno de estos falsos profetas que utilizan un lenguaje así muy directo, muy vulgar, muy espontáneo.

CSC. Sí, muy genuino hay que decir. Muy genuino.

JC. Y muy interesado generalmente.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

FMF. Y, además, lo religioso puede venir también a través de medios de comunicación, porque muchos de ellos emiten por televisión. Bueno, como se nos van agotando las baterías poco a poco, si os parece, hago un resumen muy rápido de lo que hemos visto y luego haré un comentario final, si lo veis conveniente. ¿De acuerdo?

Como resumen, comienzo recordando el asunto de las manifestaciones del español en los Estados Unidos, huyendo de esa pretensión de representatividad absoluta, cualitativa, conceptual, del español estadounidense, que nos complica enormemente. Junto a eso, es importante manejar el criterio de la diversidad de muestras, de la variedad tipológica y del origen de las muestras, para conseguir una mayor eficacia del corpus a la hora de encontrar elementos de distintas características. Y eso implica un uso escrupuloso, en la medida de lo posible, de metadatos, para identificar correctamente e interpretar correctamente cualquier tipo de dato de los manejados.

Tomando eso como pauta general, en lo que se refiere a materiales de lengua escrita, hablamos de aceptar la tipología de géneros que se manejan en otro corpus, que está bastante estandarizada, y hablamos de manejar para los Estados Unidos una producción literaria que tuviera su origen en los años 60, para poder recoger toda esa literatura chicana que sigue teniendo influencia hasta la actualidad, sin perjuicio de que, a la hora de poder compartir datos con otros, con otras asociaciones o universidades, se les ofrezca información desde el año 80 o

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

desde el año 75 o desde el año 2000, según la necesiten.

Para poder avanzar en el ámbito de la lengua literaria, necesitaríamos contar con un listado de obras que nos permita conocer y también explicar a los de fuera de qué hablamos cuando hablamos de literatura estadounidense en español. Dentro de la lengua escrita, aparte de la literatura, habría que tener en cuenta también elementos como el paisaje lingüístico, con sus debidas etiquetas, y la publicidad o los avisos que aparezcan en distintos lugares.

En el apartado de la lengua hablada, hemos comentado la posibilidad de hablar con universidades e investigadores que ya cuentan con corpus, para poder llegar a acuerdos sobre cómo compartir esos datos, a veces como datos que van y que vuelven, trabajados por un lado o por otro en las condiciones que se decidan. Eso implicaría, en fin, entrar en contacto con los responsables de distintos corpus de lengua hablada, actuales o los que puedan llegar en el futuro, con la posibilidad de trabajar de manera casi inmediata, si se llega a un acuerdo razonable, con el corpus de Nueva York, que es un corpus muy variado en cuanto al origen de sus voces y que tiene un número suficiente de encuestas y con suficiente calidad. Tarea que podríamos plantear para el Observatorio, también, de una manera inmediata, en cuanto llegáramos al acuerdo.

Aparte de las entrevistas de los corpus que ya existen, hemos hablado de la posibilidad de incorporar materiales de lengua hablada en medios de comunicación, de radio y televisión, incluyendo entrevistas, narraciones deportivas, coloquios o debates de distintos temas, políticos, deportivos, temas

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

sociales. Veremos la manera de incorporar coloquios de distinto tipo, grabaciones de coloquios o de conversaciones, que podrían ser de las comunidades de vecinos o de otro tipo; no sé, habría que ver cuáles son los ámbitos que lo ponen más fácil. Incluimos, también, discursos religiosos que se transmitan a través de los medios de comunicación social o grabaciones que se puedan hacer en directo; con la puerta abierta para organizar en el futuro un PRESEEA-Estados Unidos como macroproyecto que permita trabajar en distintas comunidades. Ya veremos si hay posibilidades y quién se anima a hacerlo.

Y en lo que se refiere al espacio de la lengua tecleada, del mundo de internet, hemos hablado de hacer una recopilación similar a la que ha hecho Davies, pero con un control, una revisión cualitativa, lingüística, que permita tener más seguridad sobre el tipo de material que tenemos entre manos; al que someteríamos a un sistema de etiquetados y de marcas, de metadatos, para identificar, en la medida de lo posible, cada uno de los textos. Hemos hablado también de incorporar páginas oficiales o institucionales de universidades, de hospitales, de entidades administrativas, de centros educativos o de otras fuentes.

Para el trabajo en internet, podemos ver la forma de realizar una extracción de datos en un tiempo razonable, que nos permita obtener algo a lo largo de este curso académico. Y luego intentar otras fuentes, ¿no? como los webinarios. Habría que conseguir también los permisos y ver la calidad técnica que tienen. Si alguien se anima a hacer un webinario, pues que lo diga y lo grabamos, para poder incorporarlo directamente. Estas son las notas mas importantes que he tomado.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

No sé si se me ha olvidado algo importante o no. Y, si no hay nada importante que se me haya olvidado, ¿hay alguna valoración, algún comentario final, alguna prevención, alguna recomendación?

JC. En cuanto a las fuentes de financiación, yo me comprometo a buscar toda la información para ver cómo podemos solicitar esa financiación en internet. Claro, habría que establecer primero un proyecto fijo, una fecha fija, una cifra, pero por lo menos puedo buscar cómo es la mecánica para hacer esa solicitud, que creo que no es muy complicada.

FMF. Lo que podemos hacer, para darle forma a ese proyecto que nos permita abordar la búsqueda de financiación, es pasar estas notas a un formato próximo a un proyecto. Haría lo posible para que se sepa por dónde queremos caminar, quién está interviniendo y cómo estamos participando. Y, a partir de ahí, ver lo que se necesita, precisar más, para poder pedir dinero aquí o allá.

AEA. Yo lo que veo aquí es un desafío, bueno, en el sentido de que no sé si la idea es partir de una planificación en que uno dice, bueno, queremos un corpus que tenga, por ejemplo, esos tres elementos, literatura, o sea materiales impresos, materiales orales, materiales tecleados, en qué proporción… Regionalmente, cómo se va a distribuir, también de que época, del 60 hasta aquí. Eso parece que está muy bien. Pero luego el problema está realmente al final: qué es lo que puedes hacer, no lo que quieres hacer; es decir, qué es lo que te encuentras. Materiales orales: igual vas a encontrar que te los regalan de un sitio, pero de otro sitio no encuentras; o los hay, pero no te los quieren prestar. Quiero decir: ¿cómo

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

se pueden llenar luego los huecos de lo quede sin rellenar? O hay autores de cierta región, pero no de otra. O hay materiales de una región o no de otra. O los materiales de la web: muchas veces no se va a poder saber de qué región son o de que época, bueno, de época sí. Pero quiere decir que todo eso, bueno, hay que tenerlo en cuenta.

FMF. Sí, los problemas van a ser muchos.

CSC. Muchísimos.

100 **FMF.** Bueno, lo comentado, entonces, es que el paraguas en el que se incluiría este proyecto sería el de la Academia Norteamericana vinculada al Instituto de Cervantes. Porque el instituto está aquí, pero también está en Nueva York; y hay otros lugares desde de los que se puede apoyar el proyecto, de distintas formas.

CSC. Pero el proyecto lo sigues dirigiendo tú, claro.

FMF. Sí, sí. Yo podría hacerlo. Aunque no estuviera en los Estados Unidos, yo me comprometería a seguir trabajando en ello, porque forma parte de mi vida de investigador. Y creo que es de lo más interesante que se pueda hacer ahora en el mundo hispanohablante.

RO. Yo creo que lo que faltaría saber, quizás, es, cuando hablamos de la representatividad o de la no representatividad, hasta qué punto estamos diciendo que esto es un problema especial para los investigadores que quieran hacer esto en los Estados Unidos. O estamos diciendo que la representatividad ha sido

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

siempre una quimera y que no lo vamos a pensar así aquí porque no lo puede pensar nadie. Esas dos alternativas quedan abiertas en la forma que tú lo has dicho y no sé si queremos explicitar eso un poco más.

FMF. Bueno, yo creo que lo que se dijo aquí fue que todo ese asunto de representatividad, que nos parece tan difícil para los Estados Unidos, si se pensara en relación con otros países, sería igualmente difícil.

AEA. Es un ideal, la representatividad.

RO. Yo creo, entonces, que podríamos decirlo con todas las letras, que no queremos que esa sea una de nuestras metas, porque nos parece que no es una meta factible, ni aquí ni en ningún otro sitio. O sea no estamos diciendo eso como algo especial de los Estados Unidos, aunque quizás sea un problema más grave en Estados Unidos. Creo que esas cosas hay que decirlas.

CSC. Tendría que tener una introducción.

RO. Y, claro, lo otro que ya sabes que voy a decir es que, para el paisaje lingüístico, podríamos aclarar que no tenemos un criterio de exclusión; que lo que veamos, por mucho que ofenda, pues se incluye.

CSC. Esos son manifestaciones del español en Estados Unidos.

FMF. Que aparecerán mezcladas como están mezcladas en todas partes, ¿no? El penúltimo libro de Blommaert sobre la sociedad global está hecho sobre paisajes

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

lingüísticos con mensaje en lenguas mezcladas, que es a lo que nos lleva a la globalidad, ¿no?

CSC. Yo les mostré a los colegas un fichero, un anuncio en Barcelona, que estaba la mitad en español y la mitad en inglés.

RO. Pero la diferencia es que Blommaert está trabajando en un ambiente totalmente académico, en el sentido universitario, por decirlo así, en que todo eso se ve, sin duda, de esa manera. A nadie se le ocurre. Pero como esto es una cosa de la Academia Norteamericana de la Lengua y del Cervantes, va a haber muchos públicos en nuestro mundo que van pensar: "No, esa gente lo que está haciendo es pulir y dar esplendor". Todavía están en eso. Y para que se sepa que no estamos en eso, quizás haya que decirlo con todas las letras, mientras que Blommaert no lo tiene que decir. Pero, un proyecto de Francisco Moreno Fernández, con el Cervantes y con la Academia, hay público que va a decir: "Eso me huele a prescriptivismo y a corrección y a exclusión". Y, si no lo dices, se pierde credibilidad.

GPR. Tienes razón.

FMF. Por parte de la Academia habría que explicar que el espíritu de un corpus no es prescriptivo, sino que es más bien descriptivo. Y que, además de descriptivo, es inclusivo y testimonial; forma parte de la historia del español de los Estados Unidos, que ahora tiene un valor y que en el futuro tendrá otro. Es decir, presentarlo de esa manera, ¿no?: no como un proyecto para fijar normas, porque

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

en este caso no se trata de eso, sino de utilizar herramientas que sirvan para diferentes fines.

CSC. El que quiera fijar normas a partir de eso, que lo haga si acaso quiere.

FMF. Se puede hacer lo que sea: el que quiera describir, que solo que describa; el que quiera utilizarlo para explicar cómo evoluciona la lengua, que lo haga. Creo que ese es el valor, el gran valor, que se aporta.

AEA. Muy bien. Pero una cosa que no se me había ocurrido cuando hemos hablado de materiales. Yo no sé si hay posibilidad de acceder a composiciones escritas por estudiantes de herencia o si eso es material que…

MLP. Yo lo pensé, y estaría feliz de donar lo que yo tenga, pero tendría que verificar si necesito permiso del IRB o de los estudiantes.

DD. Yo tengo las de mi salón de CAL State de LA del año 94 o por ahí me firmaron un permiso.

FMF. Bueno, vamos a cerrar oficialmente esta conversación. Gracias, muy sinceramente, por todo este tiempo, todas estas aportaciones y esta sabiduría compartida. Muchas gracias.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Referencias bibliográficas

Asociación de Academias de la Lengua Española (2010). *Diccionario de americanismos*. Madrid: Santillana.

Bills, Garland D. y Vigil, Neddy A. (2008). *The Spanish Language of New Mexico and Southern Colorado: A Linguistic Atlas*. Albuquerque: UNM Press.

Blommaert, Jan (2010). *Language and Superdiversity*. London: Routledge.

Blommaert, Jan (2013). *Ethnography, Superdiversity and Linguistic Landscapes: Chronicles of Complexity*. Bristol: Multilingual Matters.

Cestero Mancera, Ana (1996). *Análisis de la conversación de la Universidad de Alcalá*. Alcalá de Henares: Universidad de Alcalá. Corpus inédito

Covarrubias, Jorge (2016). "El periodismo en español en los Estados Unidos". *Informes del Observatorio / Observatorio Reports*. 019-03_2018SP.

Davies, Mark (2016). *El corpus del español*. Brigham Young University. En línea. http://www.corpusdelespanol.org [Consultado el 27-12-2018]

Díaz, Junot (1996) [1997]. *Drown / Negocios*. Trad. de Eduardo Lago. New York: Vintage.

Díaz, Junot (2007) [2008]. *The Brief Wondrous Life of Oscar Wao / La maravillosa vida breve de Óscar Wao*. Trad. de Achy Obejas. Barcelona: Mondadori.

Dumitrescu, Domnita (2011). *Aspects of Spanish Pragmatics*. New York: Peter Lang.

Enrique Arias, Andrés (2014). "Efectos del contacto de lenguas en el castellano de Mallorca: una perspectiva histórica". En A. Enrique-Arias, M. Gutiérrez, A. Landa y 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. 271-297.

Erker, Daniel (2010). "A subsegmental approach to coda /s/ weakening in Dominican Spanish". *International Journal of the Sociology of Language*, 203: 9-26.

Erker, Daniel (2014). *Spanish in Boston Project*. Boston University. http://blogs.bu.edu/danerker/research/ [Consultado el 27-12-2018].

Fernández, James y Luis Argeo (dirs.) (2018). *Invisible Immigrants (Spaniards in the US 1868-1945)*. Proyecto en línea: https://www.kickstarter.com/projects/538868554/invisible-immigrants-spaniards-in-the-us-1868-1945?lang=es

Lara, Luis Fernando (dir.) (2019). *Diccionario del español de México*. México: El Colegio de México.

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Lopez Morales, Humberto (coord.) (2008). *Enciclopedia del español en los Estados Unidos*. Madrid: Instituto Cervantes- Santillana.

Lynch, Andrew (2009): "A Sociolinguistic Analysis of Final /s/ in Miami Cuban Spanish". *Language Sciences,* 31 (2009): 767-790.

Moreno de Alba, José G. (1988). *El español en América*. México: Fondo de Cultura Económica.

Moreno de Alba, José G. (1992). *Minucias del lenguaje*. México: Fondo de Cultura Económica.

Moreno Fernández, Francisco (2006). "Información básica sobre el 'Proyecto para el Estudios Sociolingüístico del Español de España y de América' – PRESEEA (1996-2010). *Revista Española de Lingüística*, 336: 385-392.

Moreno Fernández, Francisco (coord.) (2013). *Proyecto para el Estudios Sociolingüístico del Español de España y de América' – PRESEEA*. En línea [Consultado el 27-12-2018].

Moreno-Fernández, Francisco (2018). "El español estadounidense a debate". *Informes del Observatorio / Observatorio Reports*. 043-09_2018SP.

Moreno Fernández, Francisco y Antonio Moreno Sandoval (2018). "Configuración lingüística de anglicismos procedentes de *Twitter* en el español estadounidense". *Revista signos,* 51-98: 382-409.En línea: http://dx.doi.org/10.4067/S0718-09342018000300382.

Orozco, Rafael (2018). *Spanish in Colombia and New York City*. Amsterdam: Joh Benjamins.

Otheguy, Ricardo y Ana Celia Zentella (2012). *Spanish in New York. Language contact, dialectal leveling, and structural continuity*. Oxford: Oxford University Press.

Real Academia Española y Asociación de Academias de la Lengua Española (2014). *Diccionario de la lengua española*. 23ª ed. Madrid: Espasa.

Real Academia Española: Banco de datos (CORDE) [en línea]. *Corpus diacrónico del español.* <http://www.rae.es> [Consultado el 27-12-2018]

Real Academia Española: Banco de datos (CORPES XXI) [en línea]. *Corpus del español del siglo XXI.* <http://www.rae.es> [Consultado el 27-12-2018]

Real Academia Española: Banco de datos (CREA) [en línea]. *Corpus de referencia del español actual.* <http://www.rae.es> [Consultado el 27-12-2018]

Rojo, Guillermo (2008). *Proyecto para el estudio sociolingüístico del español de Galicia* (PRESEGAL). En línea: https://gramatica.usc.es/proxectos/presegal/?lang=es

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP

Samper, José Antonio, Clara Eugenia Henríquez y Magnolia Troya (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Las Palmas: Universidad de Las Palmas de Gran Canaria.

San Martín, Abelardo y Silvana Guerrero (2016). "Marcadores de reformulación en el corpus PRESEEA de Santiago de Chile". *Forma y función*, 29: 15-38.

Silva-Corvalán, Carmen (1994). *Language contact and change. Spanish in Los Angeles*. Oxford: Oxford University Press.

Silva-Corvalán, C. (2014). *Bilingual* language *acquisition*: *Spanish* and English in the first six years. Cambridge: Cambridge University Press.

Silva-Corvalán, Carmen y Andrés Enrique Arias (2017). *Sociolingüística y pragmática del español*. 2ª. ed. Washington, DC: Georgetown University Press.

Torres Cacoullos, Rena (2011). "El estudio de la variación morfosintáctica: volver a la «complementariedad débil» por los canales de gramaticalización". En P. Martín Butragueño (ed.), *Realismo en al análisis de corpus orales (primer coloquio de cambio y variación lingüística*. México: El Colegio de México, pp. 391-410.

Varra, Rachel (2018). *Lexical borrowing and de borrowing in Spanish in New York City*. London: Routledge.

Villanueva, Tino (1979). *Hay otra voz Poems*. New York: Mensaje.

Villanueva, Tino (ed.) (1980). *Chicanos: Antología Histórica y Literaria*. México: Fondo de Cultura Económica.

Wanzer, Darrel Enck (2010). *The Young Lords: A Reader*. New York: New York University Press.

Francisco Moreno Fernández (editor)

Instituto Cervantes at Harvard University

Instituto Franklin – Universidad de Alcalá

Academia Norteamericana de la Lengua Española

OBSERVATORIO de la lengua española y las culturas hispánicas en los Estados Unidos

Hacia un corpus del español en los Estados Unidos. Debate para la génesis del proyecto CORPEEU
Informes del Observatorio / Observatorio *Reports*. 049-03/2019SP
ISBN-13: 978-0-578-43060-7 (online) doi: 10.15427/OR049-03/2019SP