Corpus de conversaciones coloquiales Ameresco-Tegucigalpa: estado, problemas, soluciones y líneas de trabajo abiertas*

Ameresco-Tegucigalpa colloquial conversations corpus: Status, problems and open lines of work

Danny Fernando Murillo Lanza

Universitat de València

danny.murillo@uv.es

Resumen: *El presente artículo pretende, en primera instancia, presentar el corpus de conversaciones coloquiales Ameresco-Tegucigalpa, el cual se instituye como uno de los primeros corpus orales del español hablado en dicha ciudad y país. En segundo lugar, expone y explica, por un lado, cuáles fueron los problemas metodológicos que se encontraron en el proceso de construcción de la primera parte de dicho corpus y, por otro lado, qué soluciones se emplearon para resolverlos. En tercera instancia, ofrece una serie de sugerencias metodológicas para mejorar el proceso de construcción de un corpus de este tipo. Finalmente, establece, por un lado, cuál es el estado actual del corpus y cuáles son sus líneas de trabajo abiertas.*

Palabras clave*:* conversaciones coloquiales, corpus Ameresco-Tegucigalpa, lingüística de corpus

Abstract: *This article aims, in the first instance, to present the corpus of colloquial conversations Ameresco-Tegucigalpa, which is established as one of the first oral corpus of Spanish spoken in that city and country. Secondly, it exposes and explains, on the one hand, what were the methodological problems that were found in the construction process of the first part of said corpus and, on the other hand, what solutions were used to solve them. Thirdly, it offers a series of methodological suggestions to improve the process of building such a corpus. Finally, it establishes, on the one hand, what is the current state of the corpus and what are its open lines of work.*

Keywords: colloqual conversations, Ameresco-Tegucigalpa corpus, Corpus Linguistics

1. Introducción

1.1. Corpus Ameresco: la idea de un corpus de conversaciones coloquiales del español de América y España

El proyecto Ameresco (América y España español coloquial) nace en el año 2010 como idea original del profesor Antonio Briz, quien fundó y dirige el

* Este artículo ha sido posible gracias al grupo de investigación Val.Es.Co., al proyecto de investigación FFI2016-75249-P, ES.VAG.ATENUACIÓN ("La atenuación pragmática en su variación genérica: géneros discursivos escritos y orales en el español de España y América"), financiado por el MINECO y gracias al apoyo del departamento de Letras y Lenguas de la Universidad Pedagógica Nacional Francisco Morazán de Tegucigalpa, Honduras. El autor del presente trabajo es beneficiario de una ayuda para la Formación del Profesorado Universitario (FPU17/03548), financiada por el Ministerio de Educación, Cultura y Deporte de España.

grupo de investigación Val.Es.Co., grupo que ha creado el corpus Val.Es.Co. El fin último de este proyecto es "profundizar en el estudio de la variedad coloquial del español en geolectos europeos y americanos" (Briz *et al.*, 2020: 3). Una de las principales líneas de trabajo del proyecto –y, quizá, la más importante– es la creación y recopilación del corpus Ameresco (Albelda y Estellés, 2020).

El corpus Ameresco es un macrocorpus que está conformado por un conjunto de corpus de conversaciones coloquiales obtenidas en diferentes ciudades del mundo hispánico –principalmente de América y España–. Actualmente el corpus cuenta con la participación de 14 ciudades de 7 países diferentes (España, Argentina, Chile, Colombia, Cuba, México y Panamá). Tal y como señalan Carcelén y Uclés (2019: 19) se tiene previsto la incorporación de otras ciudades y países, uno de los cuales es Honduras (Tegucigalpa).

Cabe destacar tres cuestiones muy importantes: la primera es que el corpus Ameresco es pionero en el ámbito hispánico ya que aúna, por una parte, conversaciones coloquiales y, por otra, variedades geográficas. La segunda es que el corpus sigue el mismo protocolo de obtención, recogida de datos y transcripción que se estableció para el corpus Val.Es.Co (Briz *et al.*, 2002). Y, la última, es que una de las principales novedades del corpus Ameresco es que ha incluido un etiquetado pragmático-conversacional y alinea la transcripción con el audio, ambas incorporaciones pretenden facilitar la búsqueda automática.

Tal y como se ha comentado anteriormente, el corpus Ameresco (Albelda y Estellés, 2020) sigue el mismo protocolo de obtención, recogida de datos y transcripción del corpus Val.Es.Co (Briz *et al*., 2002). Dado que el proyecto Ameresco trabaja de la mano con una serie de equipos de diferentes ciudades, se ha elaborado un *Protocolo de trabajo de equipos Ameresco* (Briz *et al.*, 2020) en el cual se explica –pormenorizadamente– cuál es la metodología de recolección de datos, transcripción y digitalización del corpus. Por otro lado, Carcelén y Uclés (2019) han redactado un artículo en el que describen con más detalle este protocolo. En consecuencia, el presente no pretende explicar cuáles son los lineamientos metodológicos que se han empleado para grabar y transcribir el corpus Ameresco-Tegucigalpa, ya que pueden consultarse en ambos documentos; no obstante, consideramos necesario plantear de forma sintética los lineamientos metodológicos más importantes del corpus.

En primer lugar, respecto a la selección de hablantes que forman parte de la muestra de cada corpus, se han seguido parámetros referidos a sexo (mujer y varón), edad (18-25, 26-55, ≥56) y nivel sociocultural (bajo, medio y alto) (Labov, 1972). Idealmente se espera que cada equipo recoja un mínimo de 27 conversaciones y un total de 72 hablantes.

En segundo lugar, en relación con las grabaciones, estas deben ser secretas[1] pero deben contar con la autorización de los hablantes. Los participantes de la conversación deben guardar una relación vivencial de proximidad y el marco interaccional de la misma debe ser familiar, por lo que se recomienda grabar en espacios interiores en los cuales se pueda obtener la mejor calidad de audio. Por otro lado, se recomienda que el número de interlocutores de las conversaciones sea orientativamente de 2 a 4 y que el tiempo de grabación ocupe aproximadamente de 20 a 60 minutos. En última instancia, es importante señalar que cada grabación se debe acompañar de las autorizaciones de los hablantes y de una ficha técnica que contiene información general sobre la grabación.

Finalmente, respecto al proceso de transcripción y el alineado cabe destacar dos cuestiones: la primera es que todas las grabaciones deben estar transcritas partiendo de la base de las convenciones del sistema de transcripción del Grupo Val.Es.Co. (Briz *et al*., 2002), las cuales fueron ligeramente adaptadas a las características del corpus Ameresco. La segunda es que, en el proceso de transcripción de las conversaciones, se distinguen dos fases y se aplica un estándar doble de transcripción: "por una parte, una transcripción ancha en Word, transcrita en el sistema clásico adaptado y llevada a cabo por los equipos locales; por otra, una revisión, un alineado y un etiquetado por parte del equipo central" (Carcelén y Uclés, 2019: 25).

1.2. Corpus Ameresco-Tegucigalpa: la idea del primer corpus de conversaciones coloquiales de Honduras

Tal y como se señaló en apartados anteriores, el proyecto Ameresco nació con el objetivo de profundizar el estudio de la variedad coloquial del español en geolectos europeos y americanos. La idea es intentar recopilar muestras reales del habla oral –en concreto, de conversaciones coloquiales– de la mayoría de países y ciudades hispanohablantes. Cabría destacar que algunos países, por ejemplo, España, México, Venezuela, Colombia o Chile ya han desarrollado o participado en proyectos[2] en los que se han recolectado muestras orales de algunas de sus ciudades –sobre todo entrevistas y, en menor medida, conversaciones coloquiales–. Por el contrario, hay otras ciudades y países del

[1] Primero, se pide permiso a los hablantes para grabarlos en algún momento futuro, no especificado. Segundo, una vez realizada la grabación y habiendo sido informado el hablante de que acaba de ser grabado, puede escuchar la conversación y si está de acuerdo, firma la autorización. Tercero, los hablantes deben firmar la sección para el tratamiento de datos personales, de acuerdo con la normativa vigente, y aceptar los términos. Si no se tiene dicha autorización, el archivo no podrá utilizarse y deberá ser destruido (Carcelén y Uclés, 2019).

[2] Algunos de los proyectos más importantes –aunque no los únicos -son el corpus PRESEEA (Moreno 2005; PRESEEA, 2014-), el corpus Val.Es.Co (Briz *et al.*, 2002), o , incluso, el corpus Ameresco (Albelda y Estellés, 2020). Véase, por ejemplo, los trabajos de Albelda y Briz (2009) o Briz y Carcelén (2019) para conocer más corpus orales del mundo hispánico.

mundo hispánico como Honduras, por ejemplo, que no cuentan o no han construido corpus orales como los mencionados anteriormente.

Como resultado de la deficiencia o la falta de corpus orales, se tiene muy poca constancia de estudios lingüísticos del español hablado en países como Honduras o en algunas de sus ciudades más importantes como Tegucigalpa, San Pedro Sula, La Ceiba o Choluteca. En consecuencia, se podría afirmar que el español hablado en Honduras es una variedad –desde una perspectiva investigadora– muy poco explorada o desconocida en algunos niveles lingüísticos, así lo afirmaba Herranz (1990): "El español de Honduras y Nicaragua siguen siendo las hablas que cuentan con menos estudio" o conocen un notorio retraso.

La mayoría de estudios lingüísticos sobre el español hablado en Honduras, tal y como se puede consultar en Herranz (1990), se han desarrollado o bien desde una perspectiva dialectológica (Herranz, 2001) o bien sociolingüística (Hernández, 2014). La mayoría de estos trabajos abordan los diferentes niveles lingüísticos: fonético-fonológico (Lipski, 1987; Herranz, 2001; Hernández, 2006, 2013a), léxico (Membreño, 1982; Izaguirre, 1955; Schwimmer, 2001; Ventura, 2013) y morfológico y sintáctico (Wijk, 1969; Scavnicky, 1974; Castro, 2001; Hernández, 2013b; Flores y Reyes, 2019). Por otra parte, habría que señalar que la mayoría de estos estudios se han obtenido a partir del análisis de cuestionarios, encuestas o entrevistas lingüísticas que no pueden consultarse.

Todo lo anterior permitiría afirmar dos cuestiones: la primera es que para llevar a cabo dichos estudios no se han construido corpus orales similares a los referidos aquí; y, la segunda, es que estos estos trabajos no han profundizado en el estudio de fenómenos relacionados con otros niveles como el pragmático, por ejemplo; tampoco se han centrado –en sentido estricto– en el análisis de registros como el formal o el coloquial –salvo a nivel léxico–, ni han estudiado géneros discursivos como la conversación coloquial.

Si se toma como punto de partida el panorama que se ha descrito brevemente hasta aquí, se podría afirmar que el nacimiento del corpus Ameresco-Tegucigalpa está más que justificado. Un corpus oral que pretende recoger conversaciones coloquiales obtenidas en la ciudad capital de Honduras, Tegucigalpa, y en sus zonas aledañas.

El corpus Ameresco-Tegucigalpa tuvo sus genes a finales de 2018 cuando las doctoras Marta Albelda y Maria Estellés presentaron la idea al autor de este artículo quien, posteriormente, contactó a la doctora Rosario Buezo Velásquez, profesora del departamento de Letras y Lenguas de la Universidad Pedagógica Nacional Francisco Morazán (UPNFM) de Tegucigalpa, Honduras; con el objetivo de que conociera el proyecto. Tras algunas reuniones entre miembros del corpus Ameresco como las profesoras Marta Albelda, Maria Estellés y el profesor Antonio Briz; la doctora Buezo y Danny Murillo, autor de este artículo, se decide construir el corpus Ameresco-Tegucigalpa.

La primera parte del corpus se recopiló y transcribió durante los meses de julio

a septiembre de 2019 gracias al apoyo de la profesora Sandra Liz Irías, quien era la profesora responsable de la asignatura Seminario del Español de Honduras del grado del Profesorado en la Enseñanza del Español de la sede central de la UPNFM. En esta asignatura se planificó y asignó un puntaje de la evaluación total a la recolección y transcripción que llevarían a cabo las estudiantes. En total, participaron 17 estudiantes (todas mujeres). Durante este proceso se contó con las orientaciones del autor de este artículo y la coordinadora técnica del corpus.

2. Metodología

En este apartado no se pretende explicar cuál ha sido la metodología empleada para la recolección y transcripción del corpus, ya que esta se encuentra descrita detalladamente en Briz *et al.* (2020) y Carcelén y Uclés (2019), sino que aspira a explicar cuáles han sido las fases de trabajo que se han seguido para la construcción de la primera parte del corpus Ameresco-Tegucigalpa.

La construcción del corpus Ameresco-Tegucigalpa se ha estructurado en las siguientes fases de trabajo: la primera consistió en la explicación a las estudiantes del protocolo de trabajo y las directrices para el proceso de grabación; en la segunda, se dio paso al proceso de grabación de las conversaciones; en la tercera, se revisaron y aceptaron las grabaciones por parte del equipo local; en la cuarta, se explicó cómo debían transcribirse las conversaciones; en la quinta, se realizaron prácticas de transcripción; en la sexta, se dio pasó a la transcripción ancha de las conversaciones grabadas siguiendo el sistema de transcripción; en la séptima, se revisaron, mejoraron y aceptaron las transcripciones. En una octava fase –posterior a la recolección y transcripción– se entregaron todos los archivos correspondientes a la coordinación técnica del corpus. En una última etapa, la coordinación técnica del corpus revisó, aceptó o rechazó las conversaciones grabadas y transcritas. Las siguientes fases se explicarán en el apartado 4.

3. Resultados y análisis

3.1. Proceso de grabación

3.1.1. Grabaciones recolectadas

Las conversaciones grabadas fueron 17, de las cuales 10 fueron aceptadas por la coordinación técnica del corpus Ameresco. Las otras 7 han sido descartadas, pero se incluirán dentro de un repositorio[3] que recopilará el material lingüístico que no cumple con alguno de los parámetros de calidad establecidos. En este caso, estas 7 grabaciones no han sido aceptadas por problemas relacionados con la baja calidad del audio, los espacios seleccionados

[3] Disponible en <http://esvaratenuacion.es/archivos>.

para grabar o la falta la alternancia de turnos no predeterminada de la conversación (Briz, 2010: 5).

Tal y como se detalló en el apartado 1.1., todos los equipos de trabajo deben recoger una muestra que cumpla con los parámetros referidos a sexo, edad y nivel sociocultural. Por ello, a continuación, en la Figura 1 se da detalle de los datos sociolingüísticos de los hablantes que participan en las conversaciones aceptadas.

Edad	**Sexo**	**Nivel sociocultural**			**Total**
		alto	medio	bajo	
18-25	Varón (4-4-4)	4	1	0	**5**
	Mujer (4-4-4)	11	0	0	**11**
26-55	Varón (4-4-4)	2	0	1	**3**
	Mujer (4-4-4)	4	1	2	**7**
≥ 56	Varón (4-4-4)	0	1	0	**1**
	Mujer (4-4-4)	0	0	2	**2**
Total		**21**	**3**	**5**	**29**

Fig. 1: Tabla de los datos sociolingüísticos de los hablantes que han sido grabados en la primera parte del corpus Ameresco-Tegucigalpa

A partir de la tabla anterior se pueden destacar algunos datos importantes: en primer término, se puede ver que el número de hablantes que participan en las diez conversaciones son 29. Según la variable edad: 16 pertenecen al grupo de edad de 18 a 25 años, 10 al grupo de 26 a 55 años y 3 al de 56 o mayores de esa edad; según la variable sexo participan 9 varones y 20 mujeres; y, según la variable nivel sociocultural hay 21 hablantes con nivel alto, 3 con nivel medio y 5 con nivel bajo. Todos estos datos permiten afirmar que la mayoría de los hablantes que participan en las conversaciones recolectadas tienen entre 18 a 55 años de edad, son mujeres y su nivel sociocultural es alto. Esto significa que no es un corpus homogéneo, por el momento, aunque dicha homogeneidad se logrará después

En segundo término, se puede visualizar, por una parte, cómo hay 3 estratos sociolingüísticos que están completos: hablantes mujeres y varones que tiene entre 18 a 25 años y que cuentan con un nivel de instrucción alto, y hablantes mujeres que tienen entre 26 a 55 años de edad y cuyo nivel de instrucción también es alto. Esto tiene una explicación, puesto que la mayoría de las estudiantes que recolectaron dichas conversaciones eran mujeres, tenían entre 18 a 55 años de edad y su nivel de instrucción era alto, de ahí que sus amigos o familiares también tuvieran estas mismas características.

En tercer término, se puede ver, por el contrario, cómo hay otros estratos sociolingüísticos de los cuales no hay ninguna representación: en el primer rango de edad, de 18 a 25 años, no hay hablantes mujeres de nivel medio y bajo, ni hombres de nivel sociocultural bajo. En el segundo rango de edad, de 26 a 55 años, no hay ningún hablante varón con nivel de instrucción medio. Y,

finalmente, en el tercer rango de edad, de 56 o mayores de esa edad, es el grupo que menos representación tiene ya que no hay hablantes con nivel de instrucción alta, tampoco mujeres con nivel medio, ni hombres con nivel sociocultural bajo.

En último punto, cabe destacar que, aunque hay estratos sociolingüísticos que sí que cuentan con alguna representación, estos no están completos totalmente, por ejemplo: el grupo de varones de 26 a 55 años, con nivel de instrucción bajo solamente tiene un hablante.

3.1.2. Problemas encontrados, soluciones adoptadas y sugerencias para futuros procesos

Este subapartado se concentrará en explicar, por una parte, cuáles fueron los problemas metodológicos que se encontraron durante el proceso de grabación y, por otro lado, se detallarán cuáles fueron las medidas o soluciones adoptadas para resolver dichos problemas, así como sugerencias para futuros procesos.

3.1.2.1. Origen o procedencia de los hablantes

El protocolo de trabajo del corpus Ameresco ha establecido que los hablantes que participen en las grabaciones deben ser originarios y vivir en la ciudad en la cual se está grabando las conversaciones o, al menos, haber vivido ahí desde hace diez años.

Cumplir este parámetro fue difícil para las personas responsables de grabar las conversaciones del corpus Ameresco-Tegucigalpa, ya que la mayoría de ellas no eran originarias de esta ciudad, sino que se habían traslado a vivir ahí para poder estudiar su grado universitario. Este hecho hacía, por una parte, que las estudiantes no pudieran participar como hablantes activas en la conversación, y, por otro lado, que no conocieran a muchas personas (amigos o familiares a los cuales pudieran grabar) que fuesen originarias de dicha ciudad.

Para superar este inconveniente se tomaron dos medidas: por una parte, si las investigadoras no eran originarias de Tegucigalpa ni habían vivido ahí, al menos, desde hace diez años, estas podían participar en la conversación de forma pasiva (más como oyentes y no como hablantes activas), pero en dicha conversación debería haber, al menos, un hablante que fuese originario de Tegucigalpa o que hubiese vivido en la ciudad desde hacía, al menos, diez años. De esta forma, la variedad diatópica que se pretendía recoger se veía representada por al menos un hablante o más. Por otra parte, si para una investigadora era imposible encontrar hablantes originarios de Tegucigalpa, esta podía solicitar a algún amigo, familiar o conocido o a alguna compañera –que sí conociera a este tipo de hablantes– que le grabara una conversación siguiendo todos los parámetros establecidos.

A partir de dicha experiencia, se puede sugerir, en primer lugar, al equipo de trabajo local del corpus Ameresco-Tegucigalpa o a cualquier grupo

de investigación que recoja sus corpus orales mediante la ayuda de estudiantes de sus respectivos grados o másteres, que se diagnostique y seleccione previamente aquellas asignaturas en las cuales la mayoría de estudiantes son originarios de la ciudad de la cual se quiere recolectar la muestra; en este caso, sería fundamental que el equipo de trabajo del corpus Ameresco-Tegucigalpa estudie previamente en qué asignatura puede encontrar una mayor cantidad de estudiantes capitalinos. En segundo lugar, en caso de que lo primero no sea posible, se recomienda seguir las medidas adoptadas anteriormente explicadas, esto es: en toda conversación que se grabe podría haber 1 hablante pasivo (más como oyente) no originario de la ciudad + 1 o 2 hablantes activos que sean originarios de la ciudad.

3.1.2.2. Calidad del audio

Otro de los problemas que se encontraron durante el proceso de grabación está relacionado directamente con la calidad del audio de las conversaciones grabadas.

Los corpus orales –como es el caso del corpus Ameresco– deben aspirar a tener la mejor calidad de audio posible por diversas razones: en primer lugar, porque esto facilita de forma abismal el proceso de transcripción; en segundo término, porque de esta manera se puede ofrecer a los usuarios una mayor cantidad de material lingüístico, esto es: si se transcribe la mayor parte de la conversación, se puede disponer de más material lingüístico para ser analizado; y, finalmente, porque solo así se podrían efectuar, por ejemplo, estudios prosódicos que necesiten la mayor calidad de audio. Por ello –tal y como se indicaba– la calidad del audio de muchas de las conversaciones que fueron entregadas presentaba este problema.

Si un corpus oral puede toparse con esta dificultad, lo es más si el tipo de género discursivo que el corpus oral quiere recolectar –de forma secreta– es la conversación coloquial. Esto último quedó evidenciado al momento de revisar muchas de las conversaciones grabadas. La baja calidad de audio que estas presentaban se traducía en ruidos fuertes que opacaban las intervenciones de los hablantes tales como el sonido de una televisión encendida, la música alta de un dispositivo, el sonido de una máquina de coser, el murmullo de otros hablantes que no participan en la conversación, el ruido de los carros o las motocicletas que pasan por la calle, el sonido que se produce cuando se toca o friega un plato o cuando se está cocinado, los ladridos de un perro, entre muchos otros. Todo este tipo de interferencias acústicas –que reducen en gran medida la calidad del audio de las grabaciones– no son en absoluto raras si se piensa en el tipo de género discursivo que se está recolectando: la conversación coloquial; de hecho, lo normal es encontrarnos con este tipo de interferencias, ya que las conversaciones coloquiales prototípicas surgen en el seno de un marco interaccional muy familiar (Briz, 2010: 11), marco en el cual son comunes este tipo de sonidos.

Es importante resaltar que, precisamente, dado que en este tipo de contextos o espacios como el comedor o el salón de una casa puede producirse una conversación coloquial y también pueden aparecer sonidos como el ladrido de un perro, no se puede aspirar siempre a que el audio de la conversación sea perfecto o sin ruidos, de hecho, eso podría hacernos pensar que la conversación no es coloquial, natural, ni espontánea; ahora bien, hay que intentar que los ruidos no se prolonguen de forma excesiva en la grabación, ni que se escuchen fuertemente. Sin embargo, muchas de las conversaciones entregadas tuvieron que ser rechazadas tanto previamente por la revisión local como posteriormente por la coordinación técnica del corpus por alguna de las dos restricciones señaladas anteriormente.

En cuanto a la revisión local, para solucionar dicho problema se pedía a las investigadoras lo siguiente: en primer lugar, si la conversación tenía una mala calidad de audio, se solicitaba que se repitiera el proceso de grabación. En segundo lugar, se orientaba nuevamente a la investigadora y se explicaba qué debía tener en cuenta para que su grabación tuviera una mejor calidad de audio: grabar en espacios interiores y no exteriores o evitar grabar cuando la televisión estuviera encendida, por poner algunos ejemplos. Finalmente, se mostraban a la investigadora grabaciones de conversaciones con muy buena calidad de audio y grabaciones en las que había mucho ruido. La idea era que pudieran ver las diferencias de una y otra grabación y que fuesen conscientes de cuándo era ideal grabar y en qué espacios.

Para futuros procesos de grabación, se recomienda lo siguiente: en primer punto, es fundamental e imperativo que las personas que van a grabar las conversaciones sepan con claridad y sin vacilo cuáles son los espacios o lugares ideales para grabar una conversación coloquial y cuáles no. Solamente así podrán evitar que aparezcan ruidos o interferencias que entorpezcan la grabación hasta el punto de que no se escuchen las intervenciones de los hablantes y, en consecuencia, se tenga que rechazar la grabación. En segundo, tal y como se llevó a cabo en este proceso de grabación, puede ser ilustrativo que las personas que van a grabar escuchen, por una parte, grabaciones de conversaciones con una buena calidad de audio y con pocos ruidos; y, por otra parte, grabaciones con ruidos prolongados y fuertes, es decir, conversaciones con una mala calidad de audio. Seguidamente, en último punto, se podría pedir, a las personas que grabarán, que transcriban un minuto de dos conversaciones: una con poco ruido y otra con muchas interferencias. Esto serviría para reflexionar sobre las dificultades que generó la calidad del audio al momento de transcribir una conversación y otra.

3.1.2.3. Conversación dirigida por algún hablante

Como señala Briz (2010: 5): "Lo verdaderamente definidor del género conversacional […] es que la alternancia de turno no está predeterminada y es libre". Por lo tanto, una conversación coloquial no debe estar dirigida por el

investigador, ya que no se trata ni de un debate ni de una entrevista en la que el entrevistador pregunta al entrevistado.

Esto –precisamente– es lo que hacían algunos hablantes de las conversaciones que se recolectaron en esta primera parte del corpus Ameresco-Tegucigalpa: dirigir la conversación, forzarla, preguntar a los hablantes; en suma, no dejar que la misma fuese espontánea. Cabe destacar que este problema fue detectado en la mayoría de los casos por parte de la coordinación técnica del corpus. En la revisión local fue difícil saber hasta qué punto un hablante estaba dirigiendo o no la conversación. Asimismo, es pertinente señalar que la mayoría de las grabaciones en las que un hablante/investigador dirigía la conversación estaban conformadas por dos interlocutores: la persona que grababa (que era quien dirigía la conversación) y el participante que no sabía que estaba siendo grabado.

Dado que este problema –en la mayoría de casos– fue detectado a posteriori por la coordinación técnica y no cuando se recogía el corpus, por ello, la sugerencia para futuros procesos de grabación es que en las conversaciones participen, al menos, tres hablantes (el que graba y otros dos que no saben que están siendo grabados), de esta forma, se garantiza que el primero no dirija la conversación ya que su rol sería de hablante pasivo.

3.2. Proceso de transcripción

3.2.1. Conversaciones transcritas

Una de las directrices que todos los equipos de trabajo deben seguir es que cada conversación grabada que haya sido aceptada debe tener su correspondiente transcripción. En este caso, tal y como se ha explicado, las conversaciones recolectadas se transcribieron en Word siguiendo las convenciones del Grupo Val.Es.Co. (Briz *et al*., 2002) y utilizando una selección de signos imprescindibles para una transcripción ancha (Briz *et al*., 2020: 8).

En esta primera parte del corpus Ameresco-Tegucigalpa se intentó que las 17 conversaciones aceptadas a nivel local tuvieran sus respectivas transcripciones en Word. Del total correspondiente, se entregaron 15 transcripciones completas de 20 minutos de conversación, 1 se transcribió en un 40% y 1 no se transcribió.

3.2.2. Problemas encontrados, soluciones adoptadas y sugerencias para futuros procesos de transcripción

Los diversos problemas o dificultades que se hallaron en el proceso de transcripción de las conversaciones recolectadas están relacionadas –como es lógico– con el uso y la aplicación de los diferentes símbolos del sistema de transcripción. Es muy importante explicar que esta era la primera vez que las estudiantes se enfrentaban a un proceso de transcripción de la lengua oral. Esto es, de lejos, una cuestión significativa, ya que, si transcribir la oralidad de los corpus es "sin duda […] la tarea más costosa y complicada, pero también una

tarea necesaria" (Briz, 2012: 128), lo es más, desde nuestra perspectiva, si es la primera vez que dicha labor se efectúa.

Con el objetivo de conocer cuál fue la experiencia que tuvieron las estudiantes en esta fase, se elaboró y aplicó un cuestionario en el que –entre otras cuestiones– se les preguntaba: ¿cuáles habían sido las principales dificultades que encontraron?, ¿qué símbolos del sistema le habían generado duda o no entendía cómo se debían usar?, ¿qué medio o reproductor facilitaba la transcripción?, etc. En términos generales, las investigadoras señalaron –como era de esperar– que el proceso de transcripción de una conversación coloquial puede ser tedioso, cansado o complicado.

Las principal complicación que encontraron fue el uso de algunos símbolos de transcripción: en mayor medida, símbolos como el inicio y el final del habla simultánea ([ ]) o los reinicios y autointerrupciones sin pausas (–), y, en menor medida, las pausas (/), el estilo directo (*cursiva*) o las notas a pie de página, tal y como se puede ver en la Figura 2:

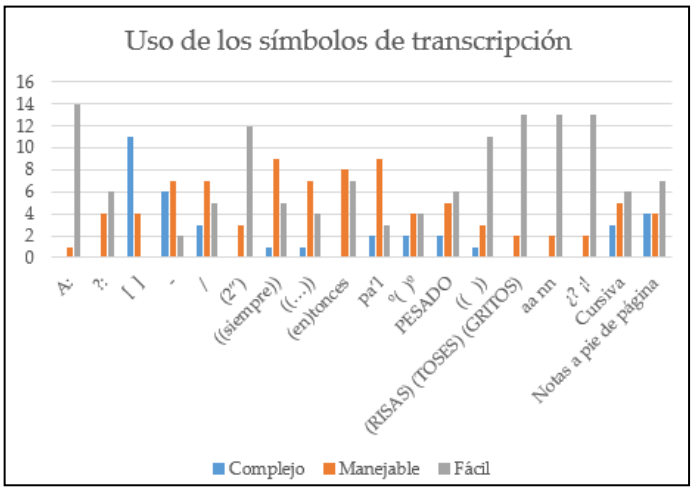

Fig. 2: Gráfico del nivel de dificultad que –según las transcriptoras– tenían los símbolos de transcripción empleados

Aunque la única manera para que el uso de símbolos de transcripción no resulte difícil es la práctica, no obstante, se tomaron las siguientes medidas: en primer lugar, para el habla simultánea y las pausas, se sugirió a las transcriptoras que usaran el programa ELAN[4], un programa para la anotación lingüística que permite asociar pequeños fragmentos transcritos con un código de tiempo, no para transcribir ahí, sino como un reproductor de audio. En este se pueden ver con mayor claridad las pausas o los solapamientos y, además, permite una mejor reproducción de las grabaciones ya que se puede repetir un fragmento cuantas veces sea necesario sin necesidad de mover el cursor. En segundo lugar, para el estilo directo, se mostraron ejemplos reales de este tipo de discurso para que pudieran distinguirlo sin problemas. En tercer término,

[4] Disponible en <https://archive.mpi.nl/tla/elan>.

para las notas a pie de página se mostraron tutoriales sobre cómo insertarlas en Word.

Por otro lado, es imperativo señalar que –desde el punto vista del revisor de las transcripciones– se identificaron otras dificultades o errores frecuentes cometidos por las transcriptoras. Estos errores están relacionados directamente con los símbolos de transcripción, la anonimización de nombres y sobrenombres, las pérdidas vocálicas o consonánticas o la confusión entre el estilo directo e indirecto. Asimismo, algunos otros fallos tienen que ver con intentar transcribir fonética y no ortográficamente y cometer fallos ortográficos en las transcripciones. Muchos de estos no pudieron ser solucionados en su totalidad, pero otros, como la anonimización, sí. Por todo lo descrito anteriormente, para próximos procesos de transcripción se sugiere lo siguiente:

En primer lugar, antes de que se proceda a transcribir, se tienen que realizar la mayor cantidad de sesiones o capacitaciones posibles en las que se pueda conocer detalladamente el sistema de transcripción que se empleará. Para ello, es fundamental que cada símbolo del sistema se acompañe de muestras o ejemplos en los que sea necesario usarlos (cuantos más ejemplos se muestren, mejor será). Otro fin de dichas sesiones es que se puedan realizar una serie de prácticas de transcripción. Estas serán esenciales para que el transcriptor se percate de las dificultades que puede encontrar durante la transcripción de una conversación coloquial y, además, pueda llegar a discutir sobre las soluciones que ellos darían a tales problemas. En segundo lugar, se sugiere –solo en caso de ser necesario– que se use el programa ELAN para reproducir el audio –y no para transcribir– ya que en este se pueden visualizar con mayor claridad las pausas o el habla simultánea. En tercer lugar, para facilitar la anonimización se pueden usar recursos en línea como *palabrasque.com*[5] para la búsqueda, por ejemplo, de los nombres ficticios. Por último, se recomienda transcribir de forma procesual: primero transcribir 5 minutos, luego, revisar estos minutos, corregirlos y continuar con los próximos 5 minutos, revisar y corregir. Así hasta acabar con toda la conversación. Esto permitirá solucionar los problemas en el proceso y no al final.

4. Estado actual del corpus y líneas de trabajo abiertas

El corpus actualmente cuenta con 10 conversaciones coloquiales, tal y como se detalló en el apartado 3.1.1., no obstante, este número debe ser ampliado, pues todavía hay estratos sociolingüísticos que no han sido representados. De ahí la utilidad del presente artículo, dado que se puede tomar como punto de referencia para obtener mejores resultados en próximos procesos de grabación y transcripción.

Por otro lado, actualmente la primera parte del corpus se encuentra en una nueva fase que consiste en un etiquetado pragmático conversacional, un

[5] Disponible en <https://www.palabrasque.com/>.

alineado de transcripción-audio y una anonimización de los audios de las 10 conversaciones aceptadas. Posteriormente, la coordinación técnica revisará el etiquetado, el alineado y la anonimización para, finalmente, publicar las conversaciones en la página web del corpus[6].

5. Consideraciones finales

La construcción del corpus de conversaciones coloquiales Ameresco-Tegucigalpa aportará beneficios, sin lugar a dudas, al estudio de la lingüística hispánica, en general, y a la hondureña, en concreto. A partir de este se podrán desarrollar nuevos trabajos sobre el español hablado en Tegucigalpa.

Si cierto es que este corpus es un recurso de gran valor, también lo es que su proceso de construcción es complejo, tal y como señalan Carcelén y Uclés (2019:18). Así se ha podido notar en este artículo con la exposición de una serie de problemas metodológicos que se han encontrado durante la recolección y la transcripción de las primeras conversaciones que forman parte del mismo. De ahí la importancia de reflexionar sobre estas fases y proponer algunas soluciones como las que aquí se han sugerido, sobre todo, si se piensa iniciar o continuar la construcción de un corpus –como es el caso del corpus Ameresco-Tegucigalpa–.

[6] Disponible en <http://esvaratenuacion.es/>.

Bibliografía

ALBELDA, Marta y Maria ESTELLÉS (2020): *Corpus Ameresco*. [en línea], disponible en <www.corpusameresco.com> [consultado en abril de 2020].

ALBELDA MARCO, Marta y Antonio BRIZ GÓMEZ (2009): "Estado actual de los corpus de lengua española hablada y escrita: I+D", en Instituto Cervantes: *El español en el mundo: Anuario del Instituto Cervantes*. Madrid: Instituto Cervantes, 65-226.

BRIZ GÓMEZ, A. y Grupo Val.Es.Co. (2002): *Corpus de conversaciones coloquiales*. Anejo 1 Oralia, Madrid: Arco Libros.

BRIZ, Antonio (2010): "El registro como centro de la variedad situacional. Esbozo de la propuesta del grupo Val.Es.Co. sobre las variedades diafásicas", en Irene Fonte y Lidia Rodríguez Alfano (eds.): *Perspectivas dialógicas en estudios del lenguaje*. Nuevo León: Universidad Autónoma de Nuevo León, 21-56.

BRIZ, Antonio (2012): "Los déficits de los corpus orales del español (y de algunos análisis)", en Juliá Jiménez, Tomas, Belén López Meirama, Victoria Vázquez Rozas y Alexandre Veiga (eds.): *Estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Departamento de Lingua Española. Servizo de Publicacións e Intercambio Científico da Universidade de Santiago de Compostela, 115-137.

BRIZ GÓMEZ Antonio y Andrea CARCELÉN GUERRERO (2019): "El futuro iberoamericano del español. La investigación del español oral y en español", en Instituto Cervantes: *El español en el mundo: Anuario del Instituto Cervantes*. Madrid: Bala Perdida Editorial.

BRIZ, Antonio *et al*. (2020): *Protocolo de trabajo equipos Ameresco.* [en línea], disponible en <http://esvaratenuacion.es/material/protocolo-de-trabajo/>. [consultado en abril de 2020].

CASTRO MITCHEL, Amanda Lizet (2001): *Los pronombres de tratamiento en el español de Honduras*. München: Lincom Europa.

CARCELÉN, Andrea y Gloria UCLÉS (2019): "Diseño y construcción de un corpus oral multidialectal. El corpus Ameresco". *Normas*, 9, 17-36.

FLORES MEJÍA, Francis Paola y Allan José, REYES MURILLO (2019): *Formas de tratamiento «vos» «tú» y «usted» en los estudiantes de la asignatura Español General de primer ingreso año 2019 UNAH*. León: Universidad Autónoma de Nicaragua UNAN-León.

HERNÁNDEZ, Hilcia (2014): "Actitudes lingüísticas en Honduras. Un estudio sociolingüístico sobre el español de Honduras frente al de otros países de habla hispana", En Ana Beatriz Chiquito y Miguel Ángel Quesada Pacheco (eds.): *Actitudes lingüísticas de los hispanohablantes hacia el idioma español y sus variantes*. Bergen, Bergen Language and Linguistic Studies (BeLLS), 715-792.

HERNÁNDEZ TORRES, Ramón Augusto (2006): "La /s/ ápico-alveolar de Olancho, Honduras: un estudio dialectológico". *Revista de la Academia Hondureña de la Lengua Española*, 15, 93-116.

HERNÁNDEZ TORRES, Ramón Augusto (2013a): *Atlas lingüístico pluridimensional de Honduras. Nivel Fonético*. Tegucigalpa: Editorial Universitaria.

HERNÁNDEZ TORRES, Ramón Augusto (2013b): *Atlas lingüístico pluridimensional de Honduras. Nivel Morfosintáctico*. Tegucigalpa: Editorial Universitaria.

HERRANZ, Atanasio (1990): *El español hablado en Honduras*. Tegucigalpa: Guaymuras.

HERRANZ, Atanasio (2001): *Formación histórica y zonas dialectales del español de Honduras, II Congreso de la Lengua Española Valladolid, España*. [en línea], disponible en <https://n9.cl/9lkh> [consultado en abril de 2020].

IZAGUIRRE, Carlos (1955): "Hondureñismos: vocablos, giros y locuciones más corrientes usadas en Honduras". *Boletín de la Academia Hondureña de la Lengua*, 1 (1), 59-123.

LABOV, William (1972): *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

LIPSKI, John M, (1987): *Fonética y fonología del español de Honduras*. Tegucigalpa: Guaymuras.

MEMBREÑO MÁRQUEZ, Alberto (1982): *Hondureñismos, 5ª edición*. Tegucigalpa: Guaymuras.

MORENO FERNÁNDEZ, Francisco (2005): "Corpus para el estudio del español en su variación geográfica y social: el corpus PRESEEA". *Oralia: Análisis del discurso oral*, 8, 123-140.

PRESEEA (2014-): *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá*. [en línea], disponible en <http://preseea.linguas.net> [consultado en abril de 2020].

SCAVNICKY, Gary E.A. (1974): "Los «sufijos» no españoles y las innovaciones sufijales en el español centroamericano", *Thesaurus, 29*, 68-117.

SCHWIMMER, Eric (2001): *Dictionary of Honduran Colloquialisms, Idioms and Slang Revised and expanded. A Spanish-English dictionary of words and expressions used in written and spoken Spanish in Honduras, plus a list of acronyms, words that express local origin and shortened forms of proper names*. Tegucigalpa: Litografía López.

VENTURA, Julio (2013): *Atlas lingüístico-etnográfico de Honduras. Nivel Léxico*. Tegucigalpa: Editorial Universitaria.

WIJK, Henri van (1969): "Algunos aspectos morfológicos y sintácticos del habla hondureña". *Boletín de Filología de la Universidad de Chile*, 20, 13-16.

Fecha de recepción: 15/04/2020
Fecha de aceptación: 06/07/2020