CUADERNOS DE LA ALFAL
Volumen especial 2, agosto 2023: 228-255
ISSN 2218-0761 en línea
DOI: 10.5935/2218-0761.2023012

CORPUS DEL *HABLA CULTA DE LA CIUDAD DE MÉXICO*: PRESERVACIÓN, SISTEMATIZACIÓN Y REVITALIZACIÓN[1]

HABLA CULTA DE LA CIUDAD DE MÉXICO CORPUS: ADVANCES IN PRESERVATION, SYSTEMATIZATION, AND REVITALIZATION

LAURA CRISTINA VILLALOBOS PEDROZA
Universidad Nacional Autónoma de México
lvillalobos@comunidad.unam.mx

En este artículo se presentan los avances del proyecto de preservación, sistematización y revitalización del *Corpus del habla culta de la Ciudad de México*. Sus materiales se compaginan con los objetivos de los paradigmas actuales de la investigación lingüística, que dan preeminencia a las muestras reales de habla en uso, en contraposición a las investigaciones basadas en intuiciones lingüísticas. Este marco resalta la importancia de revitalizar y difundir los materiales de esta colección como fuente de datos para investigaciones lingüísticas contemporáneas. En cuanto a los avances realizados, se describen las estrategias para preservar los documentos analógicos; se comenta la metodología y los criterios adoptados en la transcripción; se hace hincapié en la relevancia de sistematizar el archivo digital del corpus y se expone el esquema de metadatos que se ha diseñado para su identificación digital. Todas estas acciones se dirigen hacia la revitalización y próxima difusión de las entrevistas en un repositorio. Asimismo, se abordan las dificultades y necesidades futuras en relación con la preservación de los materiales del corpus.

Palabras clave: corpus, lengua en uso, habla culta, preservación

[1] Este trabajo se realizó en el marco del proyecto PAPIIT IN403123 *Corpus Norma Culta de la Ciudad de México: revitalización y sistematización*.

This article presents the progress of the project for the preservation, systematization, and revitalization of the Habla culta de la Ciudad de México Corpus. The corpus materials match the objectives of current paradigms of linguistic research, which encourage the use of actual samples of speech in use as opposed to research based on linguistic intuitions. This framework highlights the significance of revitalizing and disseminating the materials of this corpus as a data source for contemporary linguistic research. The strategies used to preserve the analogical documents are described; the methodology and criteria adopted in the transcription are discussed; the relevance of systematizing the digital archive of the corpus is emphasized; and the metadata scheme designed for its digital identification is presented. All these actions are aimed at revitalizing and disseminating the interviews through a repository. The difficulties and future needs in relation to the preservation of the corpus materials are also addressed.

Keywords: corpus, language in use, high instructed sociocultural level speech, preservation

Recibido: 05 junio 2023 Aceptado: 30 junio 2023

1. INTRODUCCIÓN

El corpus del Habla Culta de la Ciudad de México reúne más de 400 horas de grabaciones de entrevistas llevadas a cabo entre los años 1964 y 1971. Estas entrevistas incluyen muestras de alrededor de 600 hablantes de instrucción culta, de diversas edades y profesiones, y se recopilaron en escenarios cotidianos reales. Este corpus constituye un valioso patrimonio cultural e histórico, expresado a través de conversaciones, testimonios y narraciones que reflejan la vida, las formas de pensar y la praxis social de ese periodo. Por lo tanto, poner a disposición de la comunidad académica el contenido de estas entrevistas puede ser provechoso como fuente para otros trabajos además de la investigación lingüística, como la historia, la sociología, la antropología, entre otras.

El objetivo de este artículo es dar a conocer los avances en el proyecto de preservación, sistematización y revitalización del corpus. Se mencionan las estrategias metodológicas utilizadas para preservar y sistematizar los contenidos, así como el progreso actual en tales estrategias. También se reportan los procesos llevados a cabo en el pasado y los planes a futuro para estos materiales, que son de gran relevancia para la investigación académica.

El artículo se organiza de la siguiente manera: en primer lugar, se aborda la relevancia de los corpus orales con materiales de habla en uso para los estudios lingüísticos (§2). Luego, se contextualiza, brevemente, del surgimiento del corpus dentro del *Proyecto de estudio del habla culta de las principales ciudades de Hispanoamérica*; se resume la distribución de sus materiales y se reflexiona acerca del cambio de paradigma en las investigaciones lingüísticas y la pertinencia del corpus (§3). A continuación, se realiza un recuento de las primeras acciones de preservación,

que comprendieron la digitalización y algunas directrices para la transcripción (§4). En la siguiente sección, se describe la mirada actual hacia la preservación digital y analógica, con acciones concretas para la transcripción anonimizada de los materiales y la sistematización del archivo digital (§5). Por último, se hace un balance del estatus actual de los materiales del corpus mexicano y se plantean las dificultades y necesidades futuras (§6).

2. LOS CORPUS ORALES Y LA LINGÜÍSTICA

En términos muy amplios se puede decir que la ciencia es el estudio sistemático de diferentes aspectos del mundo, ya sean físicos o sociales, a través de la observación y la experimentación. La lingüística se entiende precisamente como el estudio de aquellos aspectos que se resumen bajo el término lenguaje. De manera muy general, tales aspectos abarcan, por un lado, los sistemas lingüísticos, i.e., conjuntos de elementos y reglas combinatorias y sus representaciones mentales; por otro lado, abarcan las expresiones lingüísticas, i.e., enunciados orales y escritos, y los procesos de procesamiento que subyacen a su producción y percepción.

Sin importar el aspecto específico del lenguaje que se estudie ni la perspectiva desde la cual se aborde, es importante que la labor se realice de manera sistemática, lo que implica que la observación y la experimentación desempeñen un papel relevante en el proceso de investigación.

En el campo de la lingüística de corpus, el término *corpus* se refiere a una colección de muestras de uso de la lengua que cumplan con algunas propiedades particulares. Tales propiedades contemplan que los ejemplos de la colección sean casos auténticos de uso real del lenguaje, que la colección sea representativa de la lengua o variedad lingüística que abarca y que la colección sea relativamente amplia (Stefanowitsch 2020).

En este sentido, lingüística de corpus puede entenderse como el estudio de preguntas de investigación lingüística a partir de un análisis sistemático de la distribución de un fenómeno lingüístico en un corpus (Stefanowitsch 2020).

Gracias al cambio de paradigma en las investigaciones lingüísticas, que están cada vez más orientadas hacia los estudios de la lengua en uso, así como a la reducción de los costos tecnológicos para manejar grandes cantidades de información, hay una tendencia creciente a la creación y recuperación de corpus lingüísticos orales. Por mencionar algunos sobre el español de México, se pueden contar el *Corpus sociolingüístico de la Ciudad de México* (Martín Butragueño y Lastra 2011), la base de datos *Etapas tempranas en la adquisición del lenguaje* (ETAL, Rojas Nieto 2007), el *Corpus oral de la Ciudad de México* (Martín Butragueño *et al.* en preparación), varias colecciones orales del repositorio del *Laboratorio Nacional de Materiales Orales* (LANMO). Así también, algunos otros que están en proceso de revitalización, como el *Corpus de español en contacto*, antes llamado Corpus de Español Indígena (Reynoso y Company 2009), el *Corpus del habla popular de la Ciudad de México* (Lope Blanch 1976) y los corpus del *Habla de*

la Ciudad de México y el *Habla popular de la República mexicana* (Pozas Loyo y Martín Butragueño 2014).

Se pueden mencionar también proyectos surgidos en otras latitudes como el *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA 2014), *el Corpus de Buenos Aires* (CORdeBA, Caladiz 2019), *el Corpus de español de migrantes en Argentina* (CORdeMIA, Martínez 2013), los *Corpus lingüísticos del Instituto Caro y Cuervo* (CLICC) (2017) y las partes orales de los corpus de Referencia de la Real Academia Española (CREA y CORPES XXI). Todos estos proyectos sobre el español oral, originados en diversas partes del planeta, dan pista sobre la importancia que ha tomado la oralidad como fuente documental y, además, constituyen pautas de referencia sobre los criterios metodológicos de tratamiento de los corpus orales, que tienen necesidades particulares.

3. EL CORPUS DEL *HABLA CULTA DE LA CIUDAD DE MÉXICO*

Antes de analizar este corpus, es preciso recordar brevemente su germen, para enmarcar la relevancia que tiene en la investigación lingüística volver la mirada hacia la oralidad. El *Corpus del habla culta de la Ciudad de México* nació en 1964, cuando, en el Segundo Simposio del Programa Interamericano de Lingüística y Enseñanza de Idiomas (PILEI), Juan Miguel Lope Blanch –investigador de la Universidad Nacional Autónoma de México y fundador del Centro de Lingüística Hispánica en esa universidad– propuso el *Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades del mundo hispánico*.

Este proyecto, tal como lo presentó Lope Blanch en 1964, surgió de su preocupación sobre el conocimiento, para entonces incompleto y con grandes lagunas, sobre el español americano. El objetivo era, pues, contribuir a describir las hablas de Hispanoamérica, que habían sido caracterizadas, a lo más, como rústicas, arcaizantes e innovadoras (Lope Blanch 1967, 1993). El proyecto proporcionaría un estudio riguroso y completo del habla normal urbana de los principales centros de población iberoamericanos (Lope Blanch 1986). En la primera reunión de la Comisión del proyecto, en 1966, se decidió estudiar las hablas de las ciudades de México, Bogotá, Lima, Santiago de Chile, Buenos Aires, La Habana, Madrid y Montevideo. Luego, se sumaron San Juan de Puerto Rico, Caracas, Sevilla, La Paz, San José de Costa Rica, Granada, Las Palmas de Gran Canaria, La Habana y Córdoba (Argentina), según indica Valencia (2014: 5). En este proyecto, *norma* no significa prescripción, sino que "está concebida en los términos en que la definió Coseriu (1962). Es un concepto de norma que asume, de forma inmanente, el carácter variable del hecho lingüístico"[2] (Caravedo 2003: 15).

[2] Es preciso mencionar que el objetivo de estudiar la norma culta de las grandes ciudades incluye aspectos como el poder irradiador de las ciudades, la expansión de sus usos hacia las ciudades del interior y la fijación de formas de

Entre 1964 y 1971 se recopilaron las 518 entrevistas que conforman el acervo sonoro del corpus del Habla de la Ciudad de México, que se recogieron en 412 cintas magnéticas de carrete abierto, y fueron realizadas con el apoyo de jóvenes investigadores en formación, con una metodología minuciosa, guiada por los objetivos de investigación del proyecto[3].

Una vez que concluyó el proceso de levantamiento de las entrevistas, se seleccionó una muestra representativa de los materiales, la que se transcribió y posteriormente se publicó en un volumen editado por Lope Blanch (1971).

Cuatro tipos de entrevista se contemplaron en la recopilación de los datos: conversaciones dirigidas, conversaciones libres, elocuciones formales en actos públicos –como conferencias y clases– y grabaciones secretas. Al realizar estas últimas no se notificaba a los entrevistados durante la grabación, sino al finalizarla, cuando se pedía su autorización para utilizar los materiales (Lope Blanch 1971: 6). El deseo de observar distintos tipos de entrevista deja entrever la preocupación por los tipos textuales, en la que subyace la idea de que el contexto impacta la selección de las formas lingüísticas en todo sentido, desde lo prosódico hasta lo sintáctico.

Por diversas razones, durante el curso de las décadas se han perdido tres de las cintas del corpus. En la Tabla 1, se muestra la distribución de las entrevistas que conserva el Centro de Lingüística Hispánica. El corpus se compone de 514 entrevistas contenidas en 409 horas. La mayor parte de ellas son de diálogo dirigido y de diálogo libre[4].

Tipo de entrevista	**Número de entrevistas**	**Número de horas**
Diálogo dirigido	332	254
Diálogo libre	115	99.5
Elocuciones formales	39	33

habla oficiales. El estudio de la norma culta planteaba algunos beneficios en el pasado, como la enseñanza escolar, la enseñanza del español a hablantes de otros idiomas o la castellanización de las comunidades indígenas en Hispanoamérica. Aunque algunos de estos objetivos no estén alineados con los propósitos socioculturales y educativos actuales, es importante valorar y revitalizar los materiales surgidos de este proyecto para utilizarlos con los propósitos de investigación en nuestra época.

[3] Para conocer con más profundidad los detalles del proyecto del estudio del habla culta de las principales ciudades de Hispanoamérica, ver Lope Blanch (1986).

[4] La idea del diseño del corpus preveía la misma cantidad de entrevistas para diálogo dirigido y diálogo libre, 40% para cada tipo. Sin embargo, el corpus presenta un sesgo hacia el diálogo dirigido y es que en realidad la información sobre este campo en las papeletas que acompañan las cintas es sumamente heterogénea. Por ejemplo, algunas entrevistas se marcan simplemente como “diálogo”, “diálogo entre varios”, otras como “conversación”, por lo que esta distribución es aún aproximada. Hemos tratado de interpretar las etiquetas de las papeletas considerando cuestiones como el número de participantes. Por las razones anteriores, la distribución final podría cambiar una vez que se termine de revisar y transcribir el corpus.

Grabaciones secretas	28	22.5
Total	514	409

Tabla 1. Distribución de los tipos de entrevista en el *Corpus del habla culta de la Ciudad de México*

En la Tabla 2, se muestran los datos de las personas entrevistadas, agrupadas por edad: jóvenes, de 25 a 35 años, fueron 234 hablantes; adultos, de 36 a 54 años, 242 hablantes; y adultos mayores, de 55 años o más, 129. En total, participaron 605 hablantes, de los cuales 304 eran mujeres y 301, hombres.

Tipo de entrevista	**Informantes**	**N/Edad**		**Mujeres**	**Hombres**
Diálogo dirigido	359	142	(25-35)	77	65
		137	(36-54)	71	66
		80	(>54)	43	37
Diálogo libre	175	67	(25-35)	34	33
		74	(36-54)	33	41
		34	(>54)	20	14
Elocución formal	42	12	(25-35)	4	8
		23	(36-54)	6	17
		7	(>54)	0	7
Grabación secreta	29	13	(25-35)	9	4
		8	(36-54)	3	5
		8	(>54)	4	4
Total	605	234 242 129	(25-35) (36-54) (>54)	304	301

Tabla 2. Datos de los informantes del *Corpus del habla culta de la Ciudad de México* por tipo de entrevista, edad y género.

El *Corpus del habla culta de la Ciudad de México* contempló las tres propiedades requeridas por los corpus señaladas por Stefanowitsch (2020). En efecto, dicho corpus (i) documentó entrevistas donde se observa "la realización viva de un sistema, de una lengua" (Lope Blanch 1967: 258); (ii) las entrevistas son representativas de los hablantes de español de instrucción alta en la Ciudad de México en ese periodo y, (iii) la colección abarca más de 400 horas de entrevistas, una amplitud considerable.

Podemos decir, sin lugar a dudas, que en el proyecto se originaron los primeros trabajos de lingüística de corpus sobre el español de América, investigaciones que en su momento buscaban proporcionar conocimiento completo del habla media urbana de los centros de población iberoamericanos. visionario fue Lope Blanch al proponer estudiar la lengua en uso, con lo que fundó toda una tradición de donde surgirían los primeros estudios que contemplaban la lengua viva, no la lengua escrita, como fuente valiosa para describir las propiedades lingüísticas de las variedades hispanoamericanas del español.

Este interés por comenzar a documentar el español de México a través de usos lingüísticos reales y no formas lingüísticas plásticas, surgidas de la intuición del analista, se adelantó al menos algunos lustros a las perspectivas que dejan atrás las visiones del lenguaje como un sistema aislado, desprovisto de su uso en la interacción y cognición humana. Tales perspectivas proponen que el conocimiento lingüístico se plasma en representaciones mentales sensibles al contexto y a las probabilidades estadísticas, que son, a la vez, cognitivamente plausibles y suficientemente poderosas para dar cuenta de la complejidad del uso real de la lengua. Estos modelos se centran en el peso de los hechos distributivos del uso real de la lengua para formar las representaciones mentales de los hablantes sobre las estructuras lingüísticas a todos los niveles, desde la fonética y la fonología hasta la morfología, la sintaxis y la pragmática (Bybee y Hopper 2001; Boyland 2009).

Con todo lo anterior, se deja en claro que los materiales de las hablas cultas son de suma relevancia para, a partir de datos reales, comprender mejor los mecanismos lingüísticos que subyacen a nuestros intercambios comunicativos. Además, constituyen parte del patrimonio sociocultural sonoro de la época. Por lo tanto, resulta de gran interés revitalizar los materiales surgidos del proyecto del Habla Culta y, en nuestro caso particular, los del *Corpus del habla culta de la Ciudad de México*.

4. PRIMERAS ACCIONES DE PRESERVACIÓN: DIGITALIZACIÓN Y AVANCE EN LA TRANSCRIPCIÓN

Las primeras acciones de rescate del corpus mexicano tuvieron lugar en la primera década del siglo XXI. Entre 2007 y 2010, Julio Serrano comenzó algo que el mismo Lope Blanch ya había proyectado, el traslado de las entrevistas del formato analógico al digital (Báez 2004; Vigueras 2004). Este paso fue crucial para garantizar la manejabilidad de los documentos sonoros,

puesto que los aparatos, sujetos a la obsolescencia tecnológica, a los estragos del tiempo y circunscritos a un espacio físico, dificultaban la ~~escucha~~ audición de las cintas. Este primer paso, llevado a cabo de forma meticulosa y documentado en Serrano (2009), constituyó la entrada de este corpus al mundo digital.

Con ello, se facilitaron las posibilidades de consulta a través de cualquier computadora. Una vez que todos los materiales del corpus se tuvieron disponibles en formato digital, fue posible implementar un proyecto de transcripción. Para ello, Serrano adaptó y utilizó los criterios de transcripción del *Corpus sociolingüístico de la ciudad de México* (Martín Butragueño y Lastra 2011), basados en gran parte en las normas del *Proyecto para el estudio sociolingüístico del español de España y América*, *PRESEEA* (Moreno Fernández 2003, citado en Martín Butragueño y Lastra 2011). Conforme a tales criterios y formatos, se transcribieron veinticuatro entrevistas, que se pusieron a disposición en formato PDF en el sitio web *El habla de la ciudad de México* (Serrano 2011), acompañadas por muestras de audio de cinco minutos. En ese sitio aún están disponibles tales materiales, así como también otras 32 entrevistas digitalizadas y sus muestras de audio, que corresponden a las primeras entrevistas que se transcribieron y que habían sido publicadas por Lope Blanch (1971).

5. PRESERVACIÓN DIGITAL Y ANALÓGICA: SISTEMATIZACIÓN Y PREPARACIÓN PARA CONSULTA

Durante las últimas décadas, se han incrementado los esfuerzos hacia la preservación documental, orientada, en el caso de los documentos sonoros, a salvaguardar los contenidos y asegurar su accesibilidad permanente. Un paso importantísimo en la preservación consiste, precisamente, en la migración de los contenidos analógicos a formatos digitales. La preservación entendida de forma holística, debe seguir una serie de procesos encaminados a "conservar, administrar, gestionar y proporcionar acceso, difusión y aprovechamiento permanente –por siempre –del audio digital" (Rodríguez Reséndiz 2020: 58).

El Centro de Lingüística Hispánica del Instituto de Investigaciones Filológicas de la UNAM ha mostrado gran interés por difundir y maximizar el impacto de la documentación y la creación de diferentes materiales lingüísticos realizadas por sus académicos durante más de cincuenta años. Por esta razón, en el 2018 comencé la coordinación de un proyecto, en el que se busca implementar el *Repositorio Digital de Corpora y Bases de datos del CLH 'Juan M. Lope Blanch'*. El repositorio está aún en proceso de implementación, pero se tienen ya avances sustantivos en el diseño estructural, técnico y tecnológico, y en su plataforma es donde se albergarán los elementos del corpus que aquí nos concierne.

Actualmente, las acciones relativas al corpus se orientan hacia tres objetivos. En primer lugar, se procuró asegurar la preservación física de los soportes analógicos. En paralelo, las acciones

presentes se enfocan hacia la preparación de los materiales para la difusión, lo que incluye la transcripción, descripción y selección de fragmentos de audio de las entrevistas. Por último, los esfuerzos se han encaminado hacia la preservación digital integral a largo plazo del acervo sonoro, considerando el ciclo de vida del objeto digital (Rodríguez Reséndiz 2020) y corresponden con el diseño y sistematización del archivo digital y un esquema de metadatos. A continuación se abordan las estrategias que se han implementado para avanzar en el cumplimiento de las tres metas señaladas.

5.1. Conservación y preservación de los documentos sonoros

El primer proceso atendido fue la conservación y restauración de los soportes físicos, que no estuvo prevista en la etapa de digitalización. Las cintas se hallaban resguardadas en el Laboratorio de Lingüística del Centro de Lingüística Hispánica en gabinetes de madera incrustados en la pared, con algunos riesgos de exposición a la humedad y a temperaturas altas en algunas épocas del año. Ahí pasaron cerca de medio siglo, hasta que en 2022, gracias a un convenio de comodato establecido con la Fonoteca Nacional de México, fueron trasladadas a bóvedas especializadas,[5]. Posteriormente, a través del sistema NOA se procedió a la catalogación, digitalización y descripción.[6].

Vale mencionar que, si bien ya se contaba con versiones digitales de audio, se determinó redigitalizar la colección completa para cumplir los estándares del proceso de preservación seguido por la Fonoteca Nacional. Gracias a tales estándares será posible generar versiones mejoradas de algunas entrevistas que inicialmente no contaban con una buena calidad, pues la Fonoteca utiliza técnicas especializadas, como el uso de hornos especiales, donde la cinta recupera durante algunas horas las propiedades perdidas por el deterioro, y en ese lapso es posible realizar la digitalización con resultados óptimos. Gracias a estas acciones, se ha avanzado en la preservación de estos documentos sonoros, muy valiosos para la documentación de lengua en uso real de hace más de cinco décadas y con especial valor histórico para el Centro de Lingüística Hispánica (Báez 2004).

5.2. Transcripción de las entrevistas

Además de la salvaguarda del contenido de las cintas, se ha seguido con otro proceso que se había previsto desde el inicio: la transcripción de las entrevistas. Este proceso es necesario porque los audios contienen al inicio una sección con 32 rubros de información sobre la entrevista, los entrevistadores y los datos sociolingüísticos de los entrevistados. Estas cabeceras incluyen

[5] Convenio con número de registro DGAJ-DPI-42-120522-176-R.

[6] El sistema NOA es una plataforma tecnológica que permite el flujo de trabajo para la digitalización, ingesta y catalogación de materiales analógicos. NOA Archive Transfer Technology: https://www.noa-archive.com/company-profile/aboutnoa/

información sensible sujeta a protección de datos personales, que incluye el nombre de los entrevistados, su edad, su lugar de residencia, su nivel de instrucción, su lugar de nacimiento y otros datos familiares. Además, la traslación del audio a texto facilita el acceso al contenido de las entrevistas. Generar versiones textuales del contenido de las cintas permite conocer el contenido de las cintas sin necesidad de escuchar las entrevistas completas. Con ello se generan nuevas fuentes primarias, que pueden ser localizadas por palabras clave ~~en~~ a través de buscadores y repositorios.

Desde las primeras transcripciones, las entrevistas pasaron por procesos de anonimización, principalmente a través de la eliminación de nombres propios y algunos lugares. En el primer periodo, se realizó el "cambio en los nombres personales y la sustitución de los apellidos por iniciales" (Lope Blanch 1971: 7). En el segundo periodo, a cargo de Julio Serrano, se adoptaron, con ciertas adaptaciones, los procesos de anonimización del *Corpus Sociolingüístico de la Ciudad de México* (Martín Butragueño y Lastra 2011).

Asimismo, con la finalidad de generar transcripciones fieles al contenido y de capturar la realidad oral que se manifestaba en ellas, se ideó un método de transcripciones y revisiones. En las primeras transcripciones "cada investigador transcribía una encuesta, y esa transcripción era después revisada –mediante una nueva audición de la cinta magnetofónica– por otro investigador diferente" (Lope Blanch 1971: 7), mientras que la revisión final quedaba a cargo del coordinador del proyecto. El mismo sistema se utilizó también en el segundo periodo de transcripciones, pero se añadieron dos revisiones más.

Si bien existen herramientas tecnológicas disponibles que podrían agilizar los procesos de transcripción, en la etapa actual, se han mantenido la transcripción y la revisión manual, porque se busca que las versiones de las entrevistas ~~que~~ reflejen lo más fielmente posible la oralidad. Tal fidelidad se ha logrado a través de un sistema de transcripción con matices sobre actitudes e intenciones lingüísticas que, aun con los avances tecnológicos actuales en inteligencia artificial, una computadora no podría identificar adecuadamente.[7]

Es necesario señalar que la etapa actual de sistematización y revitalización del corpus ha sido auspiciada por la UNAM, en el marco de dos proyectos PAPIIT: el primero de 2019 a 2021, coordinado por María Ángeles Soler Arechalde y de 2023 a 2026, por Alejandra Vigueras Ávila, con un equipo de académicos y estudiantes colaboradores de proyectos.

[7] Las labores de transcripción en la etapa en curso son realizadas por colaboradores del Laboratorio de Lingüística del Centro de Lingüística Hispánica, becarios y prestadores de servicio social en el marco del proyecto PAPIIT IN403123. El proyecto de sistematización y revitalización del corpus ha contado con el trabajo de los siguientes transcriptores: Adriana López, Alexander Sanabria, Centli Torres, Karen Benítez, Luis Fernando Aguilar, María Fernanda Vázquez, Samantha Michaus y Sarahí San Juan. Agradecemos a todos ellos y a la DGAPA, pues sin el apoyo de los participantes beneficiados por las becas del proyecto, su desarrollo no habría alcanzado el nivel actual.

5.2.1. Etapas y procedimiento

En esta sección, se describirá esquemáticamente el procedimiento que se ha seguido en las transcripciones. Las entrevistas se transcriben siguiendo un protocolo riguroso, que asegura que las versiones en texto reflejen de modo fidedigno el contenido de los audios, cuidando no exponer datos personales de los participantes ni datos que puedan revelar su identidad. Se ha mantenido, con algunas adaptaciones, la metodología de transcripción seguida por Julio Serrano, que hemos referido en la sección previa, con algunas adaptaciones.

La entrevista es transcrita por una persona, a partir de los criterios que se anotan en 5.2.2. En el siguiente paso, otra persona coteja el texto con el audio de la entrevista, corrige errores u omisiones y guarda la nueva versión en un documento nuevo, que constituye la primera revisión. Enseguida, una tercera persona, distinta al transcriptor y al primer revisor, hace el mismo proceso a partir de la primera revisión y genera el documento de la segunda revisión. El mismo proceso es llevado a cabo por el tercer revisor, una persona distinta a los colaboradores previos, y se crea la tercera revisión. Los documentos se generan en el formato de texto editable DOCX.

Cada documento se guarda de manera independiente y se genera tomando en cuenta los criterios de contenido, organización y formato, que se abordan en la sección siguiente. Para evitar errores potenciales en la identificación de las entrevistas, se mantiene el nombre de las personas entrevistadas en el encabezado de las transcripciones.

Finalmente, se designa una persona encargada de generar las versiones públicas de las entrevistas. En este proceso, se sigue el protocolo de las revisiones previas, pero además se eliminan los nombres de los entrevistados y entrevistadores. El revisor final exporta la transcripción al formato no editable PDF/A, que según el Archivo General de la Nación (2015), es uno de mejores formatos para preservar documentos electrónicos y asegurar su pervivencia.

Al finalizar el proceso de transcripción y revisión, cada entrevista habrá generado, al menos, cinco documentos diferentes, cuatro de trabajo y uno definitivo. Los documentos se nombran con su identificador en el corpus, siguiendo una nomenclatura definida (ver 5.5.2.1.) y un sufijo identificador: *_trans*, *_rev1*, *_rev2*, *_rev3* y *_revfin*, según corresponda[8].

5.2.2. Consideraciones y criterios

Las convenciones actuales del corpus se basan en un sistema de transcripción similar a TEI (Text Encoding Iniciative) (TEI Consortium 2021), que incluye una serie de normas de codificación estandarizadas, que permiten el tratamiento semiautomatizado de los datos. Como señalan Hidalgo Navarro y Sanmartín Sáez (2005), este sistema contempla "una serie de etiquetas que plasman tanto los datos externos a la muestra (fecha, texto oral o escrito, transcriptor, etc.), como internos (énfasis, tono, risa, ruido, etc.)".

[8] Las entrevistas grabadas por ambos lados de la cinta generan el doble de documentos. Para obtener un archivo digital idéntico al físico, se ha optado por generar un documento textual por cada lado de la cinta.

En las transcripciones del corpus, los datos externos se señalan en un encabezado, que incluye datos necesarios para identificar la entrevista, y sigue criterios estrictos en orden, formato y organización. Se han incorporado algunos campos y se han reorganizado los preexistentes[9].

Los campos del encabezado se organizan en cuatro rubros. En primer lugar, se identifican los creadores y coordinadores del corpus; en segundo lugar, los campos relacionados con la creación del contenido del audio, es decir, la entrevista; en tercer lugar, las iniciales que permitirán identificar a los participantes de la entrevista en la transcripción. Por último, los créditos a los transcriptores y revisores, así como las observaciones que estos puedan aportar. (*cf.* Figura 1).

<creator= Universidad Nacional Autónoma de México>
<creador= Instituto de Investigaciones Filológicas>
<creator= Centro de Lingüística Hispánica 'Juan M. Lope Blanch'>
<collection= Repositorio digital de corpora y bases de datos del CLH 'Juan M. Lope Blanch' (RepositorioCLH)>
<corpus= Norma Culta de la Ciudad de México (Mx)>
<coordinadores= Lope Blanch, Juan M.; Luna Traill, Elizabeth Guadalupe; Rojas Nieto, Cecilia; Soler Arechalde, María Ángeles; Vigueras Ávila, María del Carmen Alejandra >
<cita= Universidad Nacional Autónoma de México: Repositorio digital de corpora y bases de datos del CLH Juan M. Lope Blanch. *Norma Culta de la Ciudad de México*, <https://ru.filologicas.unam.mx/> [Fecha de consulta]>

<ID= Mx-CCLXXXVIII-B-347>
<cinta= Mx-CCLXXXVIII lado B>
<máster= cinta analógica de carrete abierto ¼ pulg >
<señal= mono>
<duración de la cinta= 50 minutos>
<digitalización= PCM, 44.1 khz, 16 bits>
<idioma= español>
<texto= oral>
<número de encuesta= Mx-347>
<tipo de encuesta= Conferencia>
<ciudad= Ciudad de México>
<fecha de grabación= 25-02-1970>
<lugar de encuesta= México D.F.>
<tema= diálogo>
<observaciones= NR>

[9] En la segunda etapa de transcripciones del corpus ya se incluía un encabezado. En la Figura 1 se puede observar en gris oscuro los campos consignados en esa etapa y en negro los campos nuevos o modificados.

<informante1= I=>
<código del informante1= ME-392-Mx-HA-70>
<origen= I= Veracruz, México>
<informante 2= X>
<código del informante 2= ME-393-Mx-MA-70>
<origen= X= México, D.F.>
<entrevistador=E= >
<codificación de informantes= "País - No.informante - Corpus - Sexo(H/M)Edad(J/A/M) - Año de entrevista; J= <34; A=35-54, M=>55">

<fecha de transcripción= 24-08-2020>
<procesador= PC, Word >
<transcripción= Ochoa Zenil, Antonio>
<revisión1= Castillo Bautista, Laura Marlene, 17-10-2020>
<revisión2= Hinojosa Romero, Fátima, 28-10-2020>
<revisión3= Guillén Rodríguez, Lydia Citlalli, 10-11-2020>
<revisiónfinal= >
<observaciones en general= >

Figura 1. Campos de identificación en el encabezado en las transcripciones.

El sistema de transcripción basado en TEI permite identificar fenómenos discursivos en el texto a través de una serie de etiquetas, lo cual simplifica la localización y recuento de casos de fenómenos etiquetados. Además, se trata de un sistema difundido globalmente, lo cual favorece la interoperabilidad y la interpretación de las distintas marcas. La descripción detallada de estas convenciones se entrega en el Anexo, al final de este artículo.

5.3. Descripción de las entrevistas

En la etapa actual se implementó además la descripción de las entrevistas. La descripción tiene como objetivo en unas pocas líneas dar cuenta del contenido temático y el esquema conversacional de la entrevista. Los criterios de descripción de las entrevistas están basados en las reglas de catalogación contenidas en la Norma Mexicana NMX-R-002-SCFI-2011 (2012), dedicada a la catalogación de documentos fonográficos. Los rubros que se incluyen en la descripción son: título, participantes, producción, serie o proyecto, resumen, contenido, duración de la grabación, idioma, palabras clave, condiciones de acceso y créditos. La descripción se realiza por una sola persona, generalmente el transcriptor, y se guarda en un documento en formato DOCX, nombrado de acuerdo con los criterios del corpus (ver 5.5.2.1) y el sufijo identificador *_descr*. En la Figura 2 se muestra un ejemplo de descripción de una entrevista.

Título: Mx-CCCII-B-363
Participantes:
<Informante 1> ME-416-Mx-HA-70
Producción: Ciudad de México: Centro de Lingüística Hispánica (CLH) "Juan M. Lope Blanch", IIFL, UNAM, 15 de marzo de 1970.
Serie o proyecto: Corpus Norma Culta de la Ciudad de México.
Resumen: El informante comparte sus opiniones respecto al movimiento estudiantil de 1968 en México. Habla de la educación en las universidades públicas y los subsidios que éstas reciben. Asimismo, trata la importancia de la Constitución mexicana y la necesidad de enseñársela a los mexicanos desde la infancia.
Contenido:
01:00-08:13 El informante 1 comparte su opinión en torno al movimiento estudiantil de 1968 en México.
08:14-17:15 Habla sobre enseñar a los niños la importancia de la Constitución mexicana y el proceso que implicó su creación. También expresa algunas opiniones sobre la formación que deberían tener los jóvenes para el progreso de México.
17:16-32:12 Comparte su opinión respecto a los subsidios del estado para las universidades. Asimismo, habla tanto de la situación educativa como laboral de los estudiantes y los profesores de la universidad pública.
Duración de la grabación: 00:32:12.
Idioma: español.
Palabras clave: movimiento estudiantil, universidad, constitución, jóvenes, profesores, maestros, colegiaturas, México.
Condiciones de acceso: es necesaria la autorización del Centro de Lingüística Hispánica para su copia, difusión o retransmisión parcial o total. Se requiere carta responsiva acerca del uso que se dará al material.
Créditos: Descripción, Adriana López Guevara.

Figura 2. Ejemplo de descripción de una entrevista del *Corpus del habla culta de la Ciudad de México*.
Nombre del archivo digital: Mx-CCCII-B-363_descr.docx

5.4. Avances en las transcripciones y descripciones

Luego de la descripción, transcripción y sucesivas revisiones, cada entrevista tendrá al menos seis documentos de texto generados a partir de ella: la descripción, la transcripción, tres revisiones intermedias y la revisión final. Actualmente, las labores de transcripción continúan y, hasta el momento, como muestra en la Tabla 3, se han transcrito 222 horas de grabaciones, lo que representa más del 50% de las entrevistas del corpus. Más de 260 entrevistas han pasado por los procesos de descripción y transcripción. Como se puede observar, se priorizó la transcripción de tres tipos de entrevista: (i) el diálogo libre, por ser manifestaciones de conversaciones en interacción natural, no en un formato de pregunta-respuesta; (ii) las grabaciones secretas, por ser las más espontáneas y no tener el efecto del observador, y (iii) las elocuciones formales. Se ha

procedido así, porque se publicarán en el repositorio, en conjunto con los audios completos y sin la sección inicial donde se mencionan los datos personales.

Tipo de entrevista	Número de entrevistas	Descripción y transcripción
Diálogo dirigido	332 entrevistas (245 horas)	28% 93 entrevistas / 73 horas
Diálogo libre	115 entrevistas (99.5 horas)	97% 112 entrevistas / 97 horas
Elocuciones formales	39 entrevistas (33 horas)	97% 38 entrevistas / 32.5 horas
Grabaciones secretas	28 entrevistas (22.5 horas)	89% 25 entrevistas / 19.5 horas
Total	514 entrevistas (409 horas)	52% 268 entrevistas / 222 horas

Tabla 3. Resumen de avances de descripción y transcripción del *Corpus del habla culta de la Ciudad de México*

5.5. Sistematización y preservación del archivo digital

Siguiendo en la línea de la preservación digital, se comenzó el diseño e implementación de un archivo digital que cumpla con las recomendaciones de preservación de la IASA, las cuales tienen como finalidad "asegurar el acceso a la información a lo largo del tiempo" (IASA 2020:6). El objetivo principal es conformar un archivo digital del corpus, basado en OAIS (Open Archival Information System), un modelo para la gestión, archivo y preservación a largo plazo, siguiendo el protocolo de integración y publicación de colecciones de la DGRU de la UNAM (Pérez Ortiz y Giménez Héau 2017).

Para acercarnos al objetivo, se realizó una serie de acciones, entre otras, la restructuración de la base de datos del corpus, el diseño de una plantilla de metadatos y algunas medidas, como mejoras en el almacenamiento y planes para el acceso y la difusión de las entrevistas. A continuación se describen los procesos de cada una.

5.5.1. Restructuración de la base de datos

El primer paso para lograr un archivo digital ordenado a partir del archivo físico, fue reestructurar la base de datos, de modo que esta reflejara el acervo sonoro del Habla. Una de las

dificultades a las que nos enfrentamos en ese proceso fue la existencia de una tabla con la información del corpus en formato Excel.

Esa primera tabla se estructuró a partir de la información de los entrevistados, por lo que cada entrada de la base de datos remitía a una persona. El razonamiento de una organización con estas características responde a las necesidades de la investigación en sociolingüística, pues la información de los participantes resulta de gran importancia, por lo que debe estar disponible en todo momento. Sin embargo, desde el punto de vista archivístico y de preservación, este diseño representó un problema porque el catálogo no reflejaba el archivo físico de los materiales analógicos, pues las entrevistas con más de un participante estaban duplicadas en la tabla de Excel.

Por tal razón, y tomando en consideración las recomendaciones y diagnóstico del acervo digital sonoro del CLH realizado por Nydia León[10], así como la propuesta del equipo de la DGRU de la UNAM, se optó por diseñar una base de datos organizada a partir de los documentos sonoros digitales. Para ello, se migró la información de los informantes a una base de datos. La tabla principal de la base de datos se organizó, entonces, a partir de las cintas y se generaron tablas secundarias sobre los entrevistados.

Con esta nueva base de datos, los colaboradores de la DGRU, bajo la coordinación de Rubén Sáenz, diseñaron un sistema de ingesta y organización para el archivo digital del corpus, que hasta el momento nos ha permitido organizar y preservar los documentos digitales del corpus.

5.5.2. Metadatos

Para consolidar esta base de datos fue necesario determinar, en primer lugar, el esquema de metadatos de los documentos sonoros digitales. Los metadatos son "datos sobre los datos" y son el complemento de los datos digitales, para que puedan ser reconocidos, procesados y distribuidos en ambientes digitales. El conjunto de los datos y los metadatos constituye lo que en el modelo OAIS se llama *paquete de información*, que engloba el contenido del objeto digital y las directrices que aportan los metadatos para contextualizar y procesar tal objeto en todo sentido (Lavoie 2014).

La plantilla de metadatos del corpus está basada en el esquema estandarizado DublinCore (2020), diseñada en apego a los esquemas de preservación PREMIS-METS (2015; 2017). Dicha plantilla se corresponde con la tabla principal de la base de datos y está compuesta por 31 campos. Dicha plantilla tuvo dos fases: la primera comprende 19 campos y se muestra en la Tabla 4, donde se incluye el número de campo, el nombre del campo en el *Repositorio digital de corpora y bases de datos del Centro de Lingüística Hispánica 'Juan M. Lope Blanch'*, la descripción y, en la última columna, la correspondencia con campos DublinCore[11].

[10] Nydia León, especialista en bibliotecología, colaboró en el proyecto en 2019.

[11] Los campos sombreados en gris aparecerán en la plantilla extendida, mientras que los campos sin sombreado serán los que aparecerán en el registro de la plantilla general.

	Repositorio CLH	Descripción	OAI-DC
1	ID	Identificador alfanumérico único, con criterios específicos: Mx-Cinta-lado-entrevista	(dc: identifier. id)
2	Título	Nombre del recurso que se utilizará para encabezar el registro. Incluye el tipo de entrevista, el género y edad de los entrevistados. Se omiten los nombres propios	(dc:title)
3	Corpus	Nombre del corpus	(dc:collection)
4	Lugar	Lugar de la entrevista (tan específico como sea posible)	(dc:coverage.spatial)
5	Fecha	Fecha de la entrevista	(dc:date)
6	Tema	Constituido por palabras clave. Si hay un lexicón disponible, se utiliza. Si se ha recabado en el corpus, respetar los términos originales.	(dc:subject)
7	Resumen	Breve descripción del contenido de la entrevista (se utiliza la descripción).	(dc:description.abstract)
8	No. informante	Número de control se dio a cada entrevistado.	(dc:source)
9	Idioma	Idioma(s) de la entrevista	(dc:language)
10	Duración	Duración de la entrevista	(dc:format:extent)
11	Institución	Para este corpus, la institución creadora será la UNAM, pero se puede duplicar el campo para incluir otras instituciones participantes cuando sea pertinente.	(dc:creator) (dcterms:provenance)
12	Entidad	Para este corpus, la entidad creadora será el IIFL, pero se puede duplicar el campo para incluir otras entidades participantes cuando sea pertinente.	(dc:creator) (dcterms:provenance)
13	Centro	Para este corpus, el centro creador será el CLH, pero se puede duplicar el campo para incluir otros centros participantes cuando sea pertinente.	(dc:creator) (dcterms:provenance)

14	Responsable del corpus	Se refiere al coordinador o coordinadores del corpus	(dc:contributor.cre)
15	Acceso	Nivel de acceso que tiene el recurso descrito (p.e., acceso restringido, acceso abierto).	(dcterms:accessRights)
16	Formato	Formato del archivo digital, p.e., WAV, MP4 PDF, DOCX, TXT, etc.	(dc:format)
17	Tipo de contenido	Tipo de recurso que contiene el archivo digital. Se indica el tipo de entrevista: conversación dirigida, conversación libre, elocución formal, grabación secreta	(dc:type)
18	Medio físico	Tipo de formato del documento físico, si es que existe, p.e., “Cinta de carrete abierto de 1/4"	(dcterms:MediaType)
19	Formato de cita	Forma en que se debe citar el recurso	(dcterms.bibliographicCitation)

Tabla 4. Esquema de metadatos para el *Corpus del habla culta de la Ciudad de México*

Este conjunto de metadatos es suficiente para describir semánticamente el contenido y contexto de creación de las entrevistas. Y dado que los metadatos se incrustarán en los objetos digitales y constituirán el registro de cada elemento del corpus en el repositorio, decidimos que los campos sociolingüísticos de los entrevistados no formarían parte de la plantilla de metadatos. Por supuesto, la información sociolingüística está resguardada en la base de datos interna del corpus (*vid*. 5.5.1.), en las tablas secundarias y tablas de relaciones.

No obstante, al plasmar únicamente esos 19 campos en la descripción de los objetos digitales, capturaríamos solo una parte del espíritu del diseño metodológico que planteó Lope Blanch; dejando de lado el género y los grupos etarios de las personas entrevistadas.

En este punto apareció otro problema: algunas entrevistas tienen hasta cuatro participantes, por lo que incluir un campo de metadatos para el género y edad de los participantes no sería suficiente. La solución fue añadir once campos más, correspondientes a la información de género y edad de cuatro informantes. Esto permitirá filtrados y búsquedas a partir de los grupos etarios y del género de los entrevistados.

En cuanto al método de relación entre los metadatos y los objetos digitales, se optará hacerlo a través *DSpace* (Lyrasis 2022), la herramienta que es utilizada para la gestión de los repositorios digitales de la UNAM. Está previsto que la relación de los metadatos y la ingesta al repositorio se lleve a cabo por medio de la carga masiva de paquetes, que incluirá los objetos digitales y los metadatos asociados. La arquitectura y la gestión de los datos del repositorio aún están en la fase de diseño. La reestructuración de la base de datos y el uso de un esquema de metadatos simples

con estándares internacionales permitió crear un archivo digital estructurado, con una identificación del origen y rastreo de relaciones de cada documento digital. De este modo, se permite la interoperabilidad con otros sistemas para dar difusión a los materiales del corpus y asegura su preservación digital.

5.5.2.1. Nomenclatura de los objetos en el archivo digital

Para garantizar una buena gestión de los archivos digitales y la accesibilidad a todos los contenidos relacionados, se nominan con criterios explícitos y rigurosos. La nomenclatura del objeto digital remite al corpus fuente y a las claves de control internas. Además, los nombres digitales facilitarán la carga masiva de paquetes al repositorio. Los criterios seleccionados para este corpus son los siguientes:

- *Siglas del corpus*. Se toman dos o tres letras que representen al corpus indicado. El *Corpus del habla culta de la Ciudad de México* tiene las siglas *Mx*.
- *Número de cinta*. Si el corpus se registró en soportes analógicos y, en la organización del archivo físico se asigna un número a cada soporte, se consigna tal cual. En el corpus las cintas están en romanos.
- *Lado de la cinta*. Al igual que el rubro previo, si el corpus se registró en soportes analógicos y estos tenían un identificador adicional relacionado al soporte analógico, como el lado de la cinta o del disco, se consignan tal como se hizo en el archivo físico. Si la entrevista fue grabada en el lado A, se escribe "A"; si fue grabada en ambos lados, se escribe "AB".
- *Número de entrevista*. Corresponde al consignado en el corpus.

Al nombrar el archivo, los rubros deben ordenarse de izquierda a derecha y deben dividirse por un guión corto, así: Mx-XXXIX-AB-053.

Con estos criterios cada archivo digital es identificable. En el ejemplo, el nombre del archivo permite saber que el audio pertenece al *Corpus del habla culta de la Ciudad de México*, a la cinta XXXIX por ambos lados y a la entrevista 53. De este modo, el archivo digital refleja la estructura del archivo físico de soportes analógicos: se organiza por cintas y no por entrevistados.

La nomenclatura corresponde al campo ID referido en la Tabla 4.

5.5.3. Otras acciones de preservación digital

Se han llevado a cabo otras acciones encaminadas a la preservación de las entrevistas del corpus: resguardo de los materiales en distintos soportes, físicos y digitales, generación de versiones comprimidas de audio, selección y recorte de muestras para difusión.

Las digitalizaciones actuales de los audios se guardan en ~~el~~ formato de audio no comprimido WAV. Se han generado también versiones comprimidas en MP3 para que los datos sean más manejables. Se cuenta con respaldos en distintos soportes físicos: dos copias completas en DVD, cinco copias en discos duros, resguardados en distintas ubicaciones de la Ciudad de México. Así también, buscando proteger los datos ante amenazas de desastres naturales, se ha comenzado a depositar los audios digitales en formato WAV en un servidor de la UNAM, a través de la plataforma de ingesta diseñada por la DGRU.

En cuanto al archivo digital de las fuentes textuales, las descripciones, transcripciones y revisiones, se almacenan en tres computadoras locales del Laboratorio en Lingüística y se sincronizan en la nube para facilitar el trabajo de los colaboradores. Se realiza un respaldo semanal en dos discos duros.

Además, se ha comenzado a seleccionar y recortar muestras de las entrevistas destinadas a ~~su~~ difusión en el repositorio. Se trata de entre cinco y siete minutos, en formato comprimido, en los que no se mencionen datos personales.

Finalmente, antes de que la colección de las cintas fuera trasladada a la Fonoteca Nacional de México, se fotografiaron algunos soportes originales, las guardas y las papeletas que los acompañaban, como se muestra en la Figura 3. Estos recursos digitales relacionados se resguardan en el archivo digital con la nomenclatura descrita en la sección precedente.

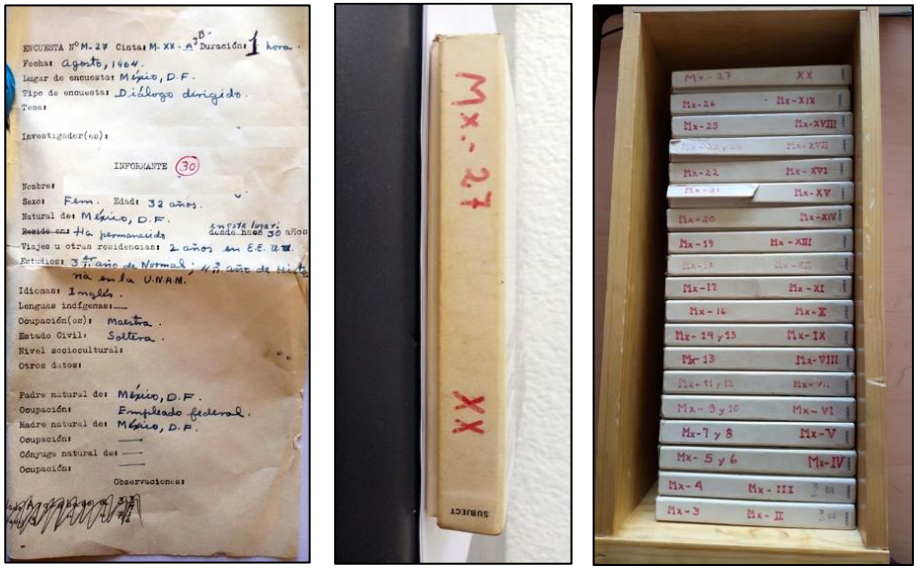

Figura 3. Ejemplos de otros recursos digitales relacionados con las entrevistas, de izquierda a derecha: Mx-XX-AB-27_papeleta.jpeg, Mx-XX-AB-27_guarda-canto.jpeg y Mx-archivero01_frente.jpeg. Autora: Nydia León, 2019.

6. Apuntes finales, dificultades y acciones futuras

Hasta el momento, se han tomado acciones encaminadas a la preservación, la sistematización y la difusión de las entrevistas. En primer lugar, los soportes analógicos se han trasladado a bóvedas especializadas, con lo que se posibilita alargar la vida de los materiales originales del corpus y obtener nuevas versiones digitales de los contenidos que, por diversas cuestiones en la primera etapa de digitalización, presentaron defectos. En segundo lugar, se ha descrito la metodología adoptada para crear nuevas fuentes primarias textuales de las entrevistas del corpus, y se ha dado a conocer el avance actual de las transcripciones que, al momento, supera el 50%. En tercer lugar, se han explicado los criterios para la restructuración de la base de datos del corpus, así como estándares internacionales de metadatos; con ello se ha organizado de forma óptima el archivo digital de los audios y sus derivados. Todas estas labores apuntan a proporcionar –en el más breve plazo– el acceso en línea de las entrevistas a través del repositorio.

En lo referente a las dificultades, una de las principales que hemos enfrentado –debido a que el corpus posee cierta profundidad histórica–, es que ha habido distintos procesos de sistematización, diferencias en la catalogación de los datos de las entrevistas, a lo largo del tiempo. Esto ha ocasionado la existencia de varias versiones de la base de datos, con información contrastante, por lo que ha sido necesario cotejar con los audios. Las inconsistencias se han ido documentando en una bitácora y conforme se confrontan con el contenido de las entrevistas, se modifican en la base de datos actual.

Otra dificultad tiene que ver con la ausencia de autorizaciones por escrito de los entrevistados para el uso y difusión de los materiales. Lope Blanch menciona que los participantes de las entrevistas autorizaron el uso de ellas, incluso en los casos de las grabaciones secretas, pero las disposiciones legales de la época no exigían una autorización expresa. Por esta razón, los audios –con excepción de las elocuciones públicas–, no se alojarán de manera abierta en el repositorio y sus transcripciones seguirán un riguroso proceso de anonimización. Asimismo, puesto que el contenido del corpus se utiliza con fines de investigación académica, siempre que un audio se consulta, se pide al consultante firmar un acuerdo de confidencialidad y uso responsable de la información.

Otra limitación significativa ha sido la dilación en la implementación de una plataforma con una interfaz de consulta. Aunque no se ha mencionado aquí, durante estos años se han contemplado diferentes posibilidades, como la integración del corpus a proyectos como *Telemeta*[12] o el desarrollo de un sistema propio a través de herramientas de uso libre, pero hemos carecido de personal para materializar alguna de las dos alternativas. No obstante, gracias al apoyo de la DGRU, el corpus tendrá no solo una interfaz de consulta, sino un sistema de organización digital óptimo para la preservación.

[12] *Telemeta* es un sistema colaborativo gratuito y de código abierto de gestión de colecciones multimedia, dedicado a proyectos colaborativos de archivo multimedia, laboratorios de investigación y humanidades digitales: http://telemeta.org/

Por último, uno de los objetivos a corto plazo, es la publicación del corpus en el *Repositorio de corpora y bases de datos del Centro de Lingüística Hispánica 'Juan M. Lope Blanch*'.

Faltan aún muchos esfuerzos para concretar cabalmente los objetivos de la preservación digital de los materiales sonoros de este corpus: el principal es tener una infraestructura que permita la comprobación automatizada de la integridad de los datos de los archivos digitales, así como la migración de medios y formatos para monitorear la obsolescencia y garantizar el acceso perenne a todos los documentos del corpus.

REFERENCIAS BIBLIOGRÁFICAS

Archivo General de la Nación. 2015. *Recomendaciones para proyectos de digitalización de documentos*, México, Secretaría de Gobernación/AGN. [en línea]. Disponible en: https://www.gob.mx/cms/uploads/attachment/file/146401/Recomendaciones_para_proyectos_de_digitalizacion_de_documentos.pdf

Asociación Internacional de Archivos Sonoros y Audiovisuales (IASA). 2020. *La salvaguarda del patrimonio audiovisual: Ética, principios y estrategia de preservación IASA-TC 03,* 2º ed., (Tr. de Asociación Española de Documentación Musical), IASA. [en línea]. Disponible en: https://www.iasa-web.org/tc03-es/etica-principios-estrategia-preservacion

Báez, Gloria. 2004. Juan M. Lope Blanch y el centro de lingüística hispánica, en Gloria Báez y Elizabeth Luna Trail (eds.), *Disquisiciones sobre filología hispánica. In memoriam Juan M. Lope Blanch*, México, UNAM: 47-53.

Boyland, Joyce Tang. 2009. Usage-based models of language, en David Eddington (ed.), *Experimental and quantitative linguistics*, Munich, Lincom: 351-419.

Bybee, Joan y Paul J. Hopper. 2001. *Frequency and the emergence of linguistic structure*, Amsterdam, John Benjamins.

Caladiz, Adriana (coord.). 2019. *Corpus de español de Buenos Aires (CORdeBA)*. [en línea]. Disponible en: http://arcas.fahce.unlp.edu.ar/arcas/portada/colecciones/cordeba

Caravedo, Rocío (dir. y ed.). 2003. Introducción al *Léxico del habla culta de Lima*. Lima, Pontificia Universidad Católica del Perú, Fondo Editorial.

DublinCoreTM. 2020. *DCMI metadata terms.* [en línea]. Disponible en: https://www.dublincore.-org/specifications/dublin-core/dcmi-terms/

Hidalgo Navarro, Antonio y Julia Sanmartín Sáez. 2005. Los sistemas de transcripción de la lengua hablada, en *Oralia*, 8: 13-36.

Instituto Caro y Cuervo. 2017. *Corpus lingüísticos del Instituto Caro y Cuervo (CLICC)*. [en línea]. Disponible en: https://clicc.caroycuervo.gov.co/

Laboratorio Nacional de Materiales Orales. ENES Morelia UNAM. [en línea]. Disponible en: https://lanmo.unam.mx/repositorionacional/

Lavoie, Brian. 2014. *The open archival information system (OAIS) reference model: Introductory guide*, Great Britain, Digital Preservation Coalition.

Lope Blanch, Juan M. 1986. *El estudio del español hablado culto. Historia de un proyecto*, México, UNAM.

Lope Blanch, Juan M. 1967. Proyecto del estudio del habla culta de las principales ciudades de Hispanoamérica, en *El simposio de Bloomington. Agosto de 1964. Actas, informes y comunicaciones*, Bogotá, Instituto Caro y Cuervo: 255-267.

Lope Blanch, Juan. M.1971. *El habla de la ciudad de México: Materiales para su estudio* México, UNAM.

Lope Blanch, Juan M. 1976. *El habla popular de la ciudad de México: Materiales para su estudio*, México, UNAM.

Lope Blanch, Juan M. 1993. *Ensayos sobre el español de América*, México, UNAM.

Lyrasis. 2022. *DSpace source code BSD license*. Disponible en https://dspace.lyrasis.org/

Martín Butragueño, Pedro y Yolanda Lastra (eds.). 2011. *Corpus sociolingüístico de la Ciudad de México*, Vol. 1: Hablantes de instrucción alta, México, El Colegio de México.

Martín Butragueño, Pedro; Érika Mendoza y Leonor Orozco (coords.). en preparación. *Corpus oral del español de México (COEM)*, Ciudad de México, El Colegio de México.

Martínez, Angelita (coord.). 2013. *Corpus de español de migrantes en Argentina (CORDEMIA)*. [en línea]. Disponible en: http://arcas.fahce.unlp.edu.ar/arcas/portada/colecciones/cordemia

Norma Mexicana NMX-R-002-SCFI-2011. 2012. *Documentos fonográfico-lineamientos para su catalogación*, México, Secretaría de Economía. [en línea]. Disponible en: https://rva.fonotecanacional.gob.mx/cartelera/normas_fonograficas.html

Pérez Ortiz, Tila María y Joaquín Giménez Héau (eds.). 2017. *Manual de datos abiertos de colecciones universitarias digitales*, México, UNAM.

Pintzuk, Susan. 2014. Corpus linguistics, en Patrick Colm Hogan (ed.), *The Cambridge encyclopedia of the language sciences*, Cambridge, Cambridge University Press: 231-232.

Pozas Loyo, Julia y Pedro Martín Butragueño (coords.). 2014. *Proyecto para la preservación y estudio del corpus lingüístico oral «Juan M. Lope Blanch»*, Ciudad de México, El Colegio de México.

PREMIS Editorial Committee y METS Editorial Board. 2017. *Guidelines for using premis with mets for exchange.* [en línea]. Disponible en: https://www.loc.gov/standards/premis/guidelines2017-premismets.pdf

PREMIS Editorial Committee. 2015. *Premis data dictionary for preservation metadata version 3.0*. [en línea]. Disponible en: https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

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.uah.es

Real Academia Española: Banco de datos. *Corpus de referencia del español actual (CREA)*. [en línea]. Disponible en: http://www.rae.es

Real Academia Española: Banco de datos. *Corpus del español del siglo XXI (CORPES)*. [en línea]. Disponible en: http://www.rae.es

Reynoso, Jeanett y Concepción Company. 2009. Criterios de edición de un corpus oral: El español indígena de México, en Belem Clark, Concepción Company, Laurette Godinas y Alejandro Higashi (eds.), *Crítica textual. Un enfoque multidisciplinario para la edición de textos*, México, El Colegio de México: 309-321. https://doi.org/10.2307/j.ctv6mtc1g.26

Rodríguez Reséndiz, Perla Olivia. 2020. *El archivo digital sonoro*, México, UNAM.

Rojas Nieto, Cecilia. 2007. La base de datos ETAL (Etapas tempranas en la adquisición del lenguaje), en Alejandra Vigueras (ed.), *Jornadas Filológicas 2005: Memoria*, México, UNAM-IIFL: 575-599.

Serrano, Julio. 2009. Rescate de los archivos sonoros del centro de lingüística hispánica Juan M. Lope Blanch", en Laura Elena Sotelo Santos (ed.), *Jornadas Filológicas 2007: Memoria*, México, IIFL-UNAM: 297-306.

Serrano, Julio. 2011. *El habla de la Ciudad de México.* [en línea]. Disponible en: https://www.iifilologicas.unam.mx/elhablamexico/

Stefanowitsch, Anatol. 2020. *Corpus linguistics. A guide to the methodology*, Berlin, Language Science Press.

TEI Consortium. 2021. *TEI P5: Guidelines for electronic text encoding and interchange*, [versión 4.3.0], TEI Consortium. [en línea]. Disponible en: https://tei-c.org/Vault/P5/4.3.0/doc/tei-p5-doc/en/html/

Valencia, Alba. 2014. Introducción a Marcadores discursivos en la norma culta hispánica 1964-2014, en *Cuadernos de la ALFAL* 5: 4-12.

Vigueras, Alejandra. 2004. Juan M. Lope Blanch y el proyecto de la norma culta, en Gloria Báez y Elizabeth Luna Trail (eds.), *Disquisiciones sobre filología hispánica. In memoriam Juan M. Lope Blanch*, México, UNAM: 221-224.

ANEXO

CONVENCIONES DE TRANSCRIPCIÓN DEL *CORPUS DEL HABLA CULTA DE LA CIUDAD DE MÉXICO*

Las marcas que hacen referencia a los datos internos de la muestra se marcan en el cuerpo del texto y se agrupan en ocho rubros, que se resumen en la siguiente Tabla. El primer rubro se refiere a las unidades discursivas, que incluyen la forma de marcar los turnos, los traslapes de turnos y las emisiones truncas. El siguiente conjunto de marcas aborda la forma de señalar las referencias a las personas, tanto las que participan en la entrevista como las que no, y también los nombres de lugares y personajes públicos. El tercer y cuarto apartados se dedican a marcas de la oralidad; el tercero aborda la prosodia, es decir, cuestiones suprasegmentales, que remiten al énfasis y las cesuras en el habla; en el cuarto se señalan cuáles aspectos vinculados con la fonética deben señalarse y cuáles no, por ejemplo, las modificaciones segmentales, las lexicalizaciones, los préstamos y extranjerismos, las siglas y las onomatopeyas. En el quinto rubro se concentran las marcas paralingüísticas que ocurren en concomitancia con el texto, como la risa al hablar, la voz baja, la imitación o el tono maternés. En el sexto se incluyen notaciones especiales como las interjecciones y las citas directas. En el séptimo se indica la forma en que los transcriptores y revisores pueden hacer observaciones y el último rubro se dedica a otras cuestiones, como la marcación del tiempo y los cambios de código.

	Descripción	**Marca**	**Notas**
Unidades discursivas	Turno	X:	(seguido por un tabulador, sin espacio blanco). El turno empieza siempre con minúscula, y al final no se pone nada en especial.
	Traslapes	[ ]	(sin espacio). Se pone entre corchetes aquello que se enunció con traslape en ambos turnos.

	Palabra trunca	t-	(sin espacio)
Referencias a personas[13]	Nombres propios (participantes de la entrevista)	Inicial	Se sustituyen por iniciales
	Nombre propios de familiares o amigos	Inicial	(ídem)
	Nombres de instituciones, personajes públicos o históricos, calle, obras artísticas.		Se transcriben completos a menos de que permitan identificar al informante.
	Entrevistador principal	E:	
	Primer informante registrado en la ficha de la entrevista	I:	
	Más participantes en la entrevista (en orden de aparición)	X:, Y:, Z:, A:, B:, C:, R:, S:, T:	
	Hipocorísticos, diminutivos o apodos	(hipoc.) (dim.) (apod.)	Se anota al lado de la inicial (con un espacio).
	Identidad dudosa	(identidad de Z dudosa en este turno)	Si la voz de la persona no se identifica claramente.
Prosodia	Énfasis fuerte	¡¡ !!	(sin espacio)
	Énfasis moderado	¡ !	(sin espacio)
	Interrogación	¿ ?	(sin espacio)
	Suspensión voluntaria	...	(sin espacio en la palabra a que se adjuntan).
	Alargamiento	a: m::	(sin espacio) Los dos puntos van a la derecha del sonido alargado.
	Pausa breve	/	(sin espacio a la izquierda, un espacio a la derecha de la barra)
	Pausa media	//	(ídem)

[13] Es muy importante que en el encabezado se marque la función de los participantes y su correspondiente letra inicial.

	Pausa larga	(pausa) (lapso)	La primera para intervalos de 1.2 a 2 segundos, el segundo para para intervalos de más de 2 segundos.
	Pausa a final de turno		Al final de turno no se marca pausa, salvo que exista una buena razón para marcarla.
Fonética	Ortografía ordinaria		El texto se transcribe en ortografía ordinaria.
	Pronunciación	pues<~pos>	(un espacio después del ángulo de cierre).
	Fonetización en presencia de otras marcas	¿verdad <~verdá>?	Los ángulos de fonetización van en adyacencia al texto fonetizado, por lo que quedan dentro de otras marcas, como las de admiración o interrogación
	Lexicalizaciones	nomás, quesque, dizque, ei	Se transcriben sin fonetización si están consignadas en el *Diccionario del Español de México* (DEM), a menos que tengan una pronunciación especial.
	Préstamos y extranjerismos	hippie <~jipi>	Se transcriben tal cual si están consignados en el DEM. De lo contrario, se transcriben como en la lengua de origen. Si tienen una pronunciación particular se incluye una fonetización.
	Siglas	XEW<~equis e doble ú>	En mayúsculas con fonetización
	Onomatopeyas	zas, bum, plas	Escritura ortográfica
	Onomatopeya de la risa	ja/ ja /ja	Se distingue de la risa espontánea, se utiliza la sílaba *ja* ó *je*.
Marcas paralingüísticas	Risa al hablar	<rie> TEXTO </rie>	
	Voz baja	<vb> TEXTO </vb>	La persona baja a propósito el volumen de la voz.

	Imitación de la voz de otro hablante	<imt> TEXTO </imt>	
	Voz de maternés	<mts> TEXTO </mts>	La persona habla en un tono más alto del habitual al dirigirse a un infante o mascota.
Notaciones especiales	Interjecciones: Marcador fático Afirmación Negación	mh mhm ó ajá m'm	
	¡ah! o *¡ih!* suspirado	ah (susp.)	Casos en que la corriente de aire es ingresiva, indican sorpresa o desconcierto.
	Silbidos	(silb.)	
	Chasquidos o clics	(chasq.)	Clic (bilabial o dental) con corriente de aire ingresiva. Suele indicar incredulidad.
	Citas y estilo directo	" "	(sin espacio)
Observaciones	Comentarios del analista	( )	Cuando aparezcan en el texto; comentarios más detallados, preferiblemente que vayan en notas. Ejemplos de comentarios: (vacilación) (corrección) (risas de todos)
	Interrupciones en la grabación	(señal deficiente 2 min)	Se anota la duración aproximada
	Fragmento ininteligible	<…>	(sin espacios dentro)
	Texto inseguro	<voy a estar allá>	(sin espacio)
	Ruidos	(risa) (ruido) (carraspeo)	
Otras cuestiones	Marcas de tiempo	[05:00]	Se anotan cada cinco minutos
	Habla con una persona ajena a la entrevista	<3p> TEXTO </3p>	La persona interrumpe su discurso para hablar con alguien (una "tercera persona") que interrumpe la conversación
	Cambio de lengua	<lengua= "inglés">	Cuando el hablante cambia de

		TEXTO </lengua>	código lingüístico, introduciendo frases completas en otra lengua. No aplica para préstamos léxicos ni extranjerismos.