INSTITUTO CARO Y CUERVO

Protocolo de transcripción ortográfica CLICC

Línea de Lingüística de corpus
Daniel Bejarano
Andrea Llanos
Ruth Rubio
Johnatan Bonilla
2018

Contenido

Índice de tablas

Índice de figuras

1. Introducción

El protocolo de transcripción ortográfica para el proyecto Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) es una herramienta para normalizar el proceso de transcripción ortográfica de los corpus orales de la institución. Este documento funciona como manual de orientación para transcribir nuevos corpus o normalizar los ya existentes bajo una serie de pautas concretas y homogéneas que faciliten su ingreso, consulta y presentación uniforme en la plataforma CLICC.

El Instituto Caro y Cuervo (ICC) ha desarrollado diversos trabajos investigativos sobre el español y las lenguas de Colombia a través de proyectos como el Atlas Lingüístico Etnográfico de Colombia (ALEC), el Habla Culta de Bogotá (HCB) y el Español Hablado en Bogotá (EHB), entre otros. Durante estas investigaciones se han recolectado datos orales de distinta naturaleza como entrevistas, relatos de tradición oral, conferencias, encuentros y demás. Actualmente, la línea de investigación en lingüística de corpus del Instituto está desarrollando un Sistema Gestor de Contenidos (SGC) para la reunión de estos materiales y su consolidación en una plataforma de administración de corpus lingüísticos en la web. La plataforma, denominada CLICC, permite administrar y almacenar los corpus producto de las investigaciones antiguas y las nuevas y facilita su preservación, divulgación y consulta. De ahí, la necesidad de contar con una serie de pautas para la transcripción ortográfica que permitan adaptar los materiales de audio a formatos homogéneos y tecnologías para la consulta y análisis de corpus.

Este protocolo es producto de la revisión de metodologías de transcripción ortográfica usadas por corpus a nivel nacional e internacional y de la identificación de necesidades de representación de elementos orales en formato escrito. Para este fin, este documento se organiza de la siguiente manera: en el apartado 2 se definen los conceptos básicos; en el apartado 3 se exponen los protocolos de transcripción ortográfica para corpus orales usados en el ámbito internacional, en lengua española y al interior del Instituto Caro y Cuervo; en el apartado 4 se presentan algunos programas de software libre usados para la transcripción y, finalmente, en el apartado 5 se describe la propuesta de transcripción ortográfica para los corpus orales del ICC.

2. Marco teórico

2.1. Corpus

Un corpus es un "conjunto estructurado de materiales lingüísticos en el que se distinguen diversos niveles de representación correspondientes a diferentes grados de elaboración de los datos que lo constituyen" (Llisterri, 1999: 54). Los corpus contienen muestras reales de la lengua en uso recolectadas bajo unos criterios y objetivos determinados que servirán como representación de una parte de la lengua. De acuerdo con Crystal (1991) los corpus "pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificación de hipótesis acerca de una lengua" (p. 32). Debido al abandono del método hipotético-deductivo y al uso de la intuición para el trabajo investigativo, en los años 60 los corpus surgieron como herramientas que privilegiaron el uso de materiales reales y en contexto. De esta manera, la lingüística se constituyó sobre la base del ejercicio empírico.

Por otro lado, es importante mencionar que teniendo en cuenta la información y los datos que maneja cada corpus se pueden llevar a cabo investigaciones, no solo lingüísticas, sino también en otras áreas del conocimiento. Por ejemplo, si las muestras están constituidas por relatos sobre hechos históricos podrían servir para la investigación en este campo.

2.1.1. Corpus orales

Los corpus se pueden categorizar teniendo en cuenta diversos criterios como el número de lenguas que los componen, el medio de producción de las muestras, el grado de tratamiento de los datos, etc. Así, pueden existir corpus monolingües o bilingües; corpus escritos, orales, mixtos y multimodales o corpus anotados, entre otros. Es de nuestro interés describir los corpus orales por el propósito de transcripción al que corresponde el protocolo.

Los corpus orales contienen muestras recolectadas a través del medio oral. Pueden estar compuestos por grabaciones, transcripciones (ortográficas o fonéticas) o las dos. La naturaleza de los corpus orales está directamente relacionada con el propósito mismo que contiene la recolección de la información y los métodos utilizados para lograr las elicitaciones resultantes (Recalde y Rozas, 2009).

Desafortunadamente, la recolección de corpus orales y su tránsito a información para el análisis lingüístico constituye un proceso en el que se perderán elementos comunicativos inevitablemente. Autores como Ochs (1979) o Linell (2005) sugieren que existe una necesaria desventaja al transformar un proceso dinámico como el habla en un producto estático textual, pero concluyen que este ejercicio es el único posible para disponer de los datos de la oralidad en formatos que puedan ser manejados a través de software especializado con propósitos de visualización de la información, etiquetado morfosintáctico, etc.

2.2. Transcripción

Farías y Montero (2005) definen la transcripción como un método para dar cuenta del resultado de algo, en este caso de lo oral. Para Torruella y Llisterri (1999) el proceso de transcripción consiste en el procesamiento de las grabaciones para su posterior uso y se organiza de la siguiente manera: transcripción ortográfica, transcripción fonética o fonológica de acuerdo con los objetivos del corpus y la vinculación del audio con su transcripción que se denomina alineación.

El proceso de transcripción es un momento crítico en el ejercicio de trabajar con corpus orales. Existe el riesgo de caer en una simplificación de la información durante el paso de registros de audio a escritos, pues puede esconderse un gran número de factores que surgen fácilmente en lo oral pero que son especialmente complejos de llevar a otro tipo de formato. Los procesos de transcripción enfrentan constantemente la necesidad de tomar decisiones subjetivas o arbitrarias, dado que debe crearse un sistema lo suficientemente descriptivo de la realidad oral, pero evitando ser demasiado complejo como para dificultar la lectura y el acceso a la información (Recalde y Rozas, 2009). Esta realidad ha llevado a una desvalorización de la labor del transcripción en detrimento de los pasos como la recolección de las grabaciones o el análisis y reporte de los resultados, tal es el caso de Plummer (1989) que recomienda realizar las transcripciones en el plazo de tiempo más breve posible, pues no constituyen el centro de la investigación.

Si bien existe acuerdo en que la transcripción no es el final ni el elemento central de una investigación amparada en corpus orales, también es importante resaltar el valor de una buena transcripción en tanto puede corresponder mejor a la realidad de los datos y facilitar los procesos

posteriores. Es por esto que resalta con ímpetu la definición de unas reglas y criterios claros que permitan un tránsito ágil, sencillo y adecuado de la información oral a escrita; pues la transcripción como herramienta debe respetar al máximo la naturaleza de los datos; aunque debe evitar ser demasiado específico como para solamente aplicarse a un corpus individual.

3. Estado del arte

En este apartado se reseñan algunas propuestas de transcripción desarrolladas a nivel nacional e internacional y que son la base para la creación de una propuesta de transcripción ortográfica propia. Se expone la información bajo las siguientes categorías: *datos generales,* que reúne la forma en la que se presentan los datos de la grabación, de los participantes, el entorno, etc.; *ortografía,* que evidencia el uso de normas utilizadas para la transcripción de los corpus y *elementos extralingüísticos,* en donde se reúnen detalles externos a la producción oral, pero que hacen parte de las circunstancias que suceden durante la recolección de datos. Se toma en cuenta esta estructura, ya que es la organización general que presentan los materiales consultados, sin embargo en algunos casos las convenciones de transcripción se mantuvieron con las categorías propuestas por cada corpus con el fin de revisar las formas de agrupación y sistematización de sus signos y sistemas particulares.

3.1.Propuestas internacionales de estandarización

Las propuestas expuestas a continuación exponen tres modelos que orientan la transcripción ortográfica en lengua inglesa y que son fuente primaria de consulta para la elaboración propuestas de transcripción (Llisterri, 1997).

3.1.1. Text Encoding Initiative (TEI)

La *Text Encoding Initiative* (TEI) es un consorcio que desarrolla colectivamente un estándar para la representación de textos en formato digital. El TEI cuenta con un conjunto de directrices que son usadas para la codificación de textos y su lectura por máquinas y que han sido adaptadas por especialistas en disciplinas como las ciencias humanas, sociales y la lingüística (Brown, 1994). La TEI propone un sistema de etiquetas lo suficientemente completo para construir un lenguaje claro de programación. A continuación se describen sus elementos más destacados.

Los *datos generales* contienen etiquetas como: <date> para la fecha; <time> para la hora; <respStmt> para la existencia de un acuerdo de responsabilidad con el uso de contenido intelectual de un texto, grabación u otro material; <equipment> para dar cuenta de los detalles

técnicos del equipo que se utiliza en la grabación; y <u> para marcar el inicio y finalización de las intervenciones de cada participante.

Desde el punto de vista *ortográfico*, la transcripción de las muestras se ajusta a la norma vigente en la lengua, iniciando los enunciados con letra mayúscula y transcribiendo los números en letras. Es importante mencionar que se prefiere evitar el uso de signos de puntuación, al entender los enunciados como un todo infragmentable por motivos de puntuación.

Se marcan algunos *elementos extralingüísticos* como: <pause> para pausas; <vocal> para pausas sonoras, voces de duda o pensamiento; <kinesic> para fenómenos de gesticulación; o <incident> para eventos accidentales como ruidos en el ambiente y demás.

3.1.2. Network of European Reference Corpora (NERC)

El proyecto *Network of European Reference Corpora* (NERC) tiene como objetivo crear una representación ortográfica para corpus orales, teniendo en cuenta elementos segmentales y suprasegmentales. NERC propone un sistema para la codificación y la transcripción de corpus basada en el proyecto *Collins-Birmingham University International Database* (COBUILD) el cual desarrolló una base de datos con fines lexicográficos que recogió, inicialmente, un corpus de siete millones de palabras y que actualmente cuenta con más de veinte millones. Uno de sus productos más reconocidos es el diccionario monolingüe "Collins Cobuild English Language Dictionary" (Llisterri, 1997).

El sistema de transcripción propone el uso de *ortografía* convencional para el inglés y las contracciones del *Oxford English Dictionary*, el inicio de enunciados con letra mayúscula y el manejo del punto como único signo de puntuación para la separación de enunciados, las comillas para dar cuenta de citas elicitadas por el informante y el apóstrofe, como es utilizado en inglés, para los posesivos y las contracciones.

3.1.3. Expert Advisory Group on Language Engineering Standards (EAGLES)

El proyecto *Expert Advisory Group on Language Engineering Standards* (EAGLES) propone una transcripción ortográfica de corpus orales agrupando características propias de TEI y NERC y de las tecnologías del habla:

En su mayoría, los *datos generales* en EAGLES como la identidad de los hablantes, los turnos de palabra o los solapamientos siguen los lineamientos propuestos por TEI.

La *ortografía* utilizada es convencional, haciendo caso a contracciones, formas reducidas, dialectales, apóstrofes e interjecciones según la normativa de cada lengua. Para palabras que son elicitadas de forma distinta a la norma, se sugiere transcribirlas según su producción y añadirlas a un banco de datos por cada corpus como casos especiales. Los elementos como números o palabras deletreadas se transcriben de forma completa en letras, respetando la pronunciación de cada hablante.

Se señalan *elementos extralingüísticos* como:<vocal> para vocales semi-léxicos (pausas, dudas); elementos vocales no léxicos (bostezos, risas, estornudos, entre otros); y <event> cuando hay ruidos en el ambiente o producidos por hablantes externos.

3.2. Protocolos de transcripción de corpus orales representativos del español

3.2.1. Corpus Oral y Sonoro del Español Rural (COSER)

El Corpus Oral y Sonoro del Español Rural (COSER) reúne grabaciones obtenidas en áreas rurales de España a partir del año 1990 y están a disposición de los usuarios en línea y de forma gratuita[2]. La directora del proyecto Inés Fernández-Ordóñez afirma que "Por el momento (marzo de 2018), 2.476 informantes están registrados en nuestra base de datos, si bien sólo de algo más de la mitad han sido entrevistados en profundidad" en gran parte mayores, con poca formación escolar y naturales de las zonas a las que se acudió.

Para la trascripción de las grabaciones, COSER define una serie de criterios que dan cuenta de la variedad dialectal de la Península Ibérica. Los puntos más destacados se pueden observar en la tabla 1.

Tabla 1

Convenciones de transcripción del corpus COSER

Datos generales	
E1, E2, E3	Encuestador y el número que le corresponde
[NP:]	Remplaza los nombre propios de los informantes para garantizar la confidencialidad de los datos
[A-Inn]	Ininteligible (calidad de la grabación)
[A-Pln:]	Poco inteligible (calidad de la grabación)
[A-Nul]	Inaudible (calidad de la grabación)
[A-Pau]	Poco audible (calidad de la grabación)
[A-Crt].	Corte de grabación (calidad de la grabación)
Ortografía	

[2] La grabaciones se pueden consultar en la página http://www.corpusrural.es/index.php

Sigue la ortografía normativa de la lengua y los signos de puntuación	
Los números se transcriben en letras, con excepción de los años	
Todos los enunciados e intervenciones inician con mayúscula, excepto aquellas que parten una intervención en un caso de habla solapada.	
Las citas elicitadas por el informante se transcriben entre comillas y precedidas por dos puntos. P.ej. El cura me dijo: "Mañana os caso"	
Si dos vocales iguales se unen por coincidir en posición final e inicial, respectivamente, no se transcriben como un sólo segmento sino cada palabra se anota con ortografía convencional.	
La acentuación de palabras en segmentos en los que no corresponde originalmente se marca con tilde sobre la parte que el informante destaca: p. ej. Áhi, máiz, sabána.	
,	Repetición de las misma palabra previa a una interrupción
-	Cuando el informante abandona una palabra o frase por motivos de autocorrección o modificación de su emisión. Se marca luego de la frase abandonada.
`	Adición o metátesis.
Elementos extralingüísticos	
[HS:]	Habla solapada entre participantes. Después de los dos puntos se específica la persona o personas que participan en segundo plano y su emisión.
[V-Ljn].	Voces lejanas
[RISAS]	Risas
[Rndo:]	Risas durante el habla. Después de los dos puntos la intervención afectada por la risa.
[TOS]	Tos
[CARRASP)	Carraspeo
[ONOMAT]	Sonidos onomatopéyicos
[OTRAS-EM]	Fenómenos de otro tipos
[Asent]	Sonidos de asentimiento: del tipo ajá, uhum, hum, mm, etc.
[PS]	Pausas
[SLNC]	Silencios

Nota. Elaboración propia con base en Ordóñez, I. (dir.) (2005-2018): Corpus Oral y Sonoro del Español Rural . <www.corpusrural.es> [9 de mayo de 2018] ISBN 978-84-616-4937-2

3.2.2. Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA)

El proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA) tiene el propósito de crear un corpus representativo del español de acuerdo con sus variedades geográficas y sociales para su uso con objetivos educativos y tecnológicos. Para la transcripción, el PRESSEA maneja un conjunto de marcas y etiquetas muy diversas que se pueden evidenciar en la tabla 2.

Tabla 2

Convenciones de transcripción del corpus PRESEEA

Ortografía y puntuación	
¡!	Enunciados exclamativos
¿?	Enunciados interrogativos
/	Pausa mínima
//	Pausa

:	Tras código de hablante (I: E: A1:)
Mayúsculas	Inicial de nombres propios y siglas
Elementos cuasi-léxicos funcionales	Interjecciones; apoyos. Escritura ortográfica (ah, ay, aha, mmm, eeh, pff, bah)
Onomatopeyas	Escritura ortográfica (zas, bum, plas)
Etiquetado de ruidos	
<ruido = " "/>	Ruido, con especificación de tipo (p.e. <ruido = "chasquido boca"/>) I
<ruido_fondo></ruido_fondo>	Ruido continuo de fondo AD
<risas = " "/>	Risas, con especificación de emisor/es (p.e. <risas = "E"/>, <risas = "todos"/>) I
<entre_risas></entre_risas>	Risas simultáneas con el habla AD
<registro_defectuoso></registro_defectuoso>	Fragmento de la grabación de mala calidad AD
<interrupción_de_grabación/>	Interrupción de la grabación I
Etiquetado fonico	
<énfasis></énfasis>	Fragmento con pronunciación claramente enfática AD
<alargamiento/>	Alargamiento de sonido D (sin espacios)
<silencio/>	Silencio de un segundo o más I
<palabra_cortada/>	Palabra cortada D
<vacilación/>	Vacilación; titubeo breve I
<sic></sic>	No es descuido de transcripción AD
<ininteligible/>	Fragmento ininteligible I
Etiquetado léxico	
<término></término>	Lexía claramente usada como uso especializado AD
<extranjero></extranjero>	Extranjerismo (excepto usos de la L2 del hablante) AD
<siglas = [ ]></siglas>	Siglas; incluye pronunciación AD
Etiquetado de dinámica discursiva	
<cita></cita>	Cita, estilo directo AD
<simultáneo></simultáneo>	Solapamiento (traslape). También se usa en turnos de apoyo, si fuera necesario AD
Etiquetado de transcripción	
<transcripción_dudosa></transcripción_dudosa>	Transcripción dudosa para transcriptor y revisores AD
<tiempo = " "/>	Anotación de minuto y segundo de grabación. (p.e.<tiempo = "02:45"/>) I
<observación_complementaria = " "/>	Observación complementaria I

Nota. Tomado de PRESEEA (2014-2018): Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá.[http://preseea.linguas.net].Consultado: [9 de mayo de 2018]

3.2.3. Corpus Valencia Español Coloquial (Val.Es.Co)

El corpus Valencia Español Coloquial (Val.Es.Co) es un proyecto desarrollado por el grupo de investigación Val.Es.Co (Valencia Español Coloquial) de la Universitat de València que tiene como objetivo la descripción del español coloquial a través de grabaciones de

conversaciones, llamadas telefónicas, radio, televisión y entrevistas. Este corpus maneja un conjunto de convenciones que se presentan en la tabla 3.

Tabla 3

Convenciones de transcripción del corpus Val.Es.Co

Convención	***Descripción***
:	Cambio de voz
A:	Intervención de un interlocutor identificado como A
?:	Interlocutor no reconocido
§	Sucesión inmediata, sin pausa apreciable, entre dos emisiones de distinto interlocutores
=	Mantenimiento del tuno de un participante en un solapamiento [Lugar donde se inicia un solapamiento o superposición
]	Final del habla simultánea
-	Reinicios y auto interrupciones sin pausa
/	Pausa corta, inferior al medio segundo
//	Pausa entre medio segundo y un segundo
///	Pausa de un segundo o más
(5'')	Silencio (lapso o intervalo de 5 segundos; se indica el número de segundos en las pausas de más de un segundo cuando sea especialmente significativo
↑	Entonación ascendente
↓	Entonación descendente
→	Entonación mantenida o suspendida
Cou	Los nombres propios, apodos, siglas y marcas, excepto las convertidas en «palabras-marca» de uso general, aparecen con la letra inicial en mayúscula.
PESADO	Pronunciación marcada o enfática (dos o más letras mayúsculas).
pe sa do	Pronunciación silabeada.
(( ))	Fragmento indescifrable.
((siempre))	Transcripción dudosa.
((...))	Interrupciones de la grabación o de la transcripción
(en)tonces	Reconstrucción de una unidad léxica que se ha pronunciado incompleta, cuando pueda perturbar la comprensión.
pa'l	Fenómenos de fonética sintáctica entre palabras, especialmente marcados.
°( )°	Fragmento pronunciado con una intensidad baja o próxima al susurro.
H	Aspiración de «s» implosiva.
(RISAS, TOSES, GRITOS...)	Aparecen al margen de los enunciados. En el caso de las risas, si son simultáneas a lo dicho, se transcribe el enunciado y en nota al pie se indica «entre risas».
Aa	Alargamientos vocálicos.
Nn	Alargamientos consonánticos.
¿¡ !?	Interrogaciones exclamativas.
¿?	Interrogaciones. También para los apéndices del tipo «¿no?, ¿eh?, ¿sabes?»
¡!	Exclamaciones
Letra cursiva	Reproducción e imitación de emisiones. Estilo directo, característico de los denominados relatos conversacionales
Notas a pie de página	Anotaciones pragmáticas que ofrecen información sobre las circunstancias de la enunciación. Rasgos complementarios del canal verbal. Añaden informaciones necesarias para la correcta interpretación de determinadas palabras (la correspondencia extranjera de la palabra transcrita en el texto de acuerdo con la pronunciación real, siglas, marcas, etc.), enunciados o secuencias del

	texto (p.e., los irónicos), de algunas onomatopeyas: del comienzo de las escisiones conversacionales, etc.

Nota. Tomado de Cabedo, Adrián y Pons, Salvador (eds.): Corpus Val.Es.Co 2.0. Consultado online en http://www.valesco.es

3.2.4. Corpus para el estudio del español oral (ESLORA)

El corpus para el estudio del español oral (ESLORA), desarrollado por el Grupo de Gramática del Español de la Universidad de Santiago de Compostela, está conformado por 60 horas de entrevistas y 20 horas de conversaciones naturales de Galicia obtenidas entre 2007 y 2015. El corpus posee un sistema de consultas simples y combinadas que tiene en cuenta variables sociales (edad, nivel de estudios, sexo) y categorías lingüísticas (clases de palabras, lemas, categorías morfológicas); además, permite recuperar los fragmentos de audio que corresponden a las consultas hechas. El sistema de etiquetas usado en la transcripción de las grabaciones es similar al corpus PRESEEA, puesto que el subcorpus de entrevistas se incluye dentro de este. Sus marcas más representativas se muestran en la tabla 4.

Tabla 4

Convenciones de transcripción del corpus ESLORA

Convención	***Descripción***
alargamiento	Aumento de cantidad que afecta a algún sonido de la palabra marcada.
Cita	El fragmento resaltado reproduce estilo directo.
énfasis	Señala casos de pronunciación especialmente acentuada.
ficticio	El nombre ha sido cambiado para preservar el anonimato de los hablantes.
lengua[nombre=xx]	El fragmento está en una lengua diferente al español (xx puede ser gl: gallego, en: inglés, pt: portugués, it: italiano, fr: francés, el: griego)
palabra_cortada	El segmento marcado representa un fragmento de una palabra.
Risa	Marca un segmento en que se ríe un hablante.
Sic	Señala algunos errores de dicción para que no se interpreten como errores de transcripción.
Sigla	Indica que una cierta forma es una sigla.

Nota. Tomado de Grupo de Gramática del Español. Corpus para el estudio del español oral (ESLORA). Universidad Santiago de Compostela. Recuperado de: http://eslora.usc.es/ [Mayo 15 de 2018]

3.2.5. Corpus Oral del Español como Lengua Extranjera (CORELE)

El Corpus Oral del Español como Lengua Extranjera (CORELE) recoge 13 horas y 36 minutos (15 minutos por entrevista, aproximadamente) de grabaciones producidas por estudiantes de español como lengua extranjera y busca ser una herramienta para el análisis de errores en la oralidad. Se propone como un elemento enriquecedor de los procesos de enseñanza/aprendizaje de lenguas, permitiendo conocer particularidades como la influencia de la lengua materna o la

frecuencia de cada error, entre otras. El CORELE planteó sus convenciones de transcripción con base en los *Codes for the Human Analysis of Transcricpts* (CHAT) y las convenciones usadas por el Spanish Learner Language Oral Corpora (SPLLOC). En la tabla 5 se observan los símbolos usados por el CORELE.

Tabla 5

Convenciones de transcripción del corpus CORELE

Marcas de información prosódica	
Símbolo	**Descripción**
/	Pausa no final y no autónoma.
//	Pausa no final pero autónoma. La usamos en estos casos: oraciones yuxtapuestas de sentido completo (a veces son subordinadas); oraciones en estilo directo o citas de palabras textuales dentro delenunciado; incisos o aclaraciones sobre la idea principal del enunciado, u oraciones intercaladas del tipode *¿cómo se dice?, no sé*, etc.; y la parte final de un enunciado, que sigue a una pausa sin serindependiente.
Marcas de final de enunciado y continuación de turno	
///	Pausa final de enunciado.
¿?	Pausa final de enunciado interrogativo. Separamos de la palabra contigua los signos con un espacio
¡!	Pausa final de enunciado exclamativo. Separamos de la palabra contigua los signos con un espacio, salvo las interjecciones.
…	Pausa final de enunciado sin terminar o con entonación suspendida.
+	Interrupción por otro hablante.
¬	Continuación de turno tras la intervención del otro hablante.
Fenómenos de la oralidad espontánea	
[/]	Reformulación, reinicio o repetición involuntaria de palabra o sintagma.
[///]	Reformulación sintáctica o reinicio de oración. Lo empleamos cuando el hablante reestructura el enunciado pero continúa la idea que expresaba.
=	Autointerrupción o abandono del enunciado (el hablante no continúa con la idea anterior).
→	Alargamiento vocálico oconsonántico a final depalabra.
<>	Solapamiento. El enunciado del segundo hablante que se solapa aparece precedido del signo [<].
&	Ante palabra incompleta.
&eh &ah &mm	Apoyos vocálicos.
#	Pausa no prosódica sin intención expresiva (Ej. cuando hablante está pensando unapalabra).
Xxx	Fragmentos ininteligibles.
{%alt: }	Fallos de producción o elisiones típicas del habla espontánea. Se transcribe la palabra correcta y se translitera el sonido entre llaves.
{%com: }	Comentarios. Indicamos con esta marca puntualizaciones sobre la situación comunicativa o el discurso (p. ej. si la persona habla susurrando o silabea una palabra, si disminuye el volumen o pone énfasis en un fragmento, etc.).
hhh {%act: }	Signos paralingüísticos: assent[asentimiento]; blow[soplido]; click[chasquido]; cough[tos]; doubt[duda]; laugh[risa]; onomatopoeia [onomatopeya];[pregunta]; y sigh[suspiro]
Fenómenos de interés para el análisis de la interlengua	

@ + tres letras del código de lengua	Se añade al final de una palabra extranjera (generalmente del inglés o de la L1 del informante)
@c	Se añade al final de una forma verbal malconjugada.
{%err: }	Error de forma léxica. Se transcribe correctamente y se translitera la forma errónea empleada por el hablante.
@g	Se añade a continuación de una palabra para indicar que el estudiante extranjero ha imitado el discurso del oyente. Las palabras imitadas en otros enunciados posteriores a aquel en que aparece la primera imitación no las marcamos.
@n	Se añade al final de una creación léxica (normalmente por analogía con la forma de una palabra de la L1 usando una raíz o mecanismos de inflexión de la L2). Lo empleamos cuando nos resulta muy difícil o imposible recuperar la palabra que el hablante busca.
{%pho: }	Pronunciación errónea (o no propia del español estándar). A continuación del término (o el grupo de palabras) transcrito ortográficamente, incluimos la transcripción fonética con los símbolos del Alfabeto Fonético Intnal. (AFI) entre corchetes.

Nota. Tomado de Corpus Oral del Español como Lengua Extranjera (CORELE). Recuperado de: http://cartago.lllf.uam.es/corele/home_es.html [Mayo 15 de 2018]

3.2.6. Corpus Sociolingüístico de Mérida – Venezuela (CSMV)

El Corpus Sociolingüístico del habla de Mérida-Venezuela (CSMV) es un proyecto realizado por el Grupo de Lingüística Hispánica de la Universidad de Los Andes que nació en 1990. El CSMV recoge muestras extensas de habla espontánea para facilitar la investigación lingüística y cuenta con 40 horas de grabación de entrevistas realizadas a 80 informantes de Mérida.

Los criterios de transcripción del corpus CSMV son los siguientes: con respecto a los *datos generales*, marcar la letra inicial del nombre del informante para mantener la confidencialidad de los hablantes. En O*rtografía*, guardar la mayor fidelidad posible con las grabaciones; transcribir duda o falsos inicios de elicitación a través de puntos suspensivos; copiar las repeticiones del hablante de seguido, sin ninguna marca especial; transcribir con ortografía convencional conjunciones u onomatopeyas; usar los signos de puntuación convencionalmente; y mantener la morfofonología de las palabras que forman parte de expresiones o usos dialectales por parte del informante con la ortografía convencional. Sobre los *elementos extralingüísticos* se recomienda utilizar corchetes dobles [[ ]] para indicar sonidos como risas, ruidos, tos e interrupciones de la grabación y hacer uso de paréntesis para marcar segmentos ininteligibles de la manera: (no se entiende).

3.2.7. Corpus de Referencia del Español Actual (CREA)

El Corpus de Referencia del Español Actual (CREA) está compuesto por un estimado de 160 millones de palabras, provenientes de fuentes escritas y orales reunidas entre los años 1975 y 2004 que buscan facilitar el estudio lingüístico de la lengua según sus variedades, su norma o distintos intereses investigativos.

Dentro de su extensión, el CREA reúne los corpus: Análisis de la conversación de la Universidad de Alcalá de Henares (ACUAH), Macrocorpus de la Norma Lingüística Culta de las Principales Ciudades del Mundo Hispánico (ALFAL), Estudio sociolingüístico de Caracas, 1977 (Caracas-77), Estudio sociolingüístico de Caracas, 1987 (Caracas-87), Corpus de Encuestas en Asunción de Paraguay (CEAP), Corpus Oral de la Variedad Juvenil Universitaria del Español Hablado en Alicante (COVJA), Corpus para el Estudio del Español Hablado en Santiago de Compostela (CSC), Corpus Sociolingüístico de la Ciudad de Mérida (CSMV) y Corpus Oral de Referencia del Español Contemporáneo (UAM).

De acuerdo con Pino (1988) los criterios para la transcripción del CREA están basados en los estándares de la TEI, el EAGLES y el NERC. De forma general, la *ortografía* sugiere adaptarse al uso normativo del español; iniciar enunciados con mayúscula; integrar abreviaturas y acrónimos de acuerdo a la pronunciación del hablante y deletreo en palabras, separando cada contenido deletreado con un guión; marcar interjecciones entre corchetes; transcribir palabras focalizadas con mayúsculas en su totalidad; usar comillas para marcar citas textuales y transcribir errores de producción entre asteriscos. La puntuación rescata el uso de punto al final del enunciado y usos de signos de interrogación, admiración, comas y dos puntos según normas de la lengua. Los *elementos extralingüísticos* como correcciones del discurso sobre la marcha o pausas considerables se marcan con tres puntos suspensivos, los enunciados ininteligibles o poco claros con uso de tres signos de cierre de interrogación y las palabras titubeantes con dos asteriscos posteriores al fragmento.

3.2.8. Corpus Oral de la Lengua Española en Montreal (COLEM)

El Corpus Oral de la Lengua Española en Montreal (COLEM) está compuesto por entrevistas semidirigidas con una duración aproximada de una hora cada una, recogidas en entornos naturales. El COLEM se caracteriza por su trabajo con el habla espontánea propia de

ámbitos familiares y estilos conversacionales. Para la recolección del corpus se tienen en cuenta datos sociolingüísticos como la edad, el origen, el sexo y el nivel de educación; y datos relacionados con la situación de contacto lingüístico y migración.

El director del corpus, Enrique Pato (2018), comenta que la transcripción ortográfica se rige bajo los criterios del NERC y del EAGLES. Además, tiene en cuenta aspectos propios de la metodología de corpus como COSER y PRESEEA. El COLEM estudia el habla de comunidades hispanas de Montreal teniendo en cuenta tres factores: el primero de ellos "la continuidad estructural de la variedad propia del español de origen", el segundo se enfoca en "el contacto lingüístico con el francés y el inglés"; finalmente, hay un tercero relacionado con "la nivelación dialectal, fruto del contacto con otras variedades del español en Montreal" (Pato, 2018).

3.3.Protocolos de transcripción del ICC

3.3.1. Corpus del Español Hablado en Bogotá (EHB)

El Corpus del Español Hablado en Bogotá (EHB) recoge 234 grabaciones de narraciones semilibres de una hora aproximadamente y 242 grabaciones fonéticas de 10 minutos aproximados.

Las grabaciones de EHB fueron transcritas y analizadas con un programa de computador desarrollado como tesis de grado por dos estudiantes de sistemas, Hugo Suárez y Jorge Molina. La aplicación se llamaba SATES y comenzó a funcionar en 1994. Como resultado de los procesos de transcripción y análisis se publicaron dos trabajos, el primero de ellos una muestra impresa con la trascripción de 30 grabaciones: El español hablado en Bogotá. Relatos semilibres de informantes pertenecientes a tres estratos sociales (Montes, 1997) y un análisis de los resultados realizado a partir del software SATES: El español hablado en Bogotá. Análisis previo de su estratificación social (Montes, 1998). Estas transcripciones tienen algunas marcas de fenómenos lingüísticos específicos como la aspiración de la s (loj mijmoj, curaj, ej), la elisión de consonantes (universidá, onde, salú) o el yeismo (yego). Las convenciones usadas para la transcripción se observan en la tabla 6.

Tabla 6

Convenciones de transcripción del corpus EHB

Convención	***Descripción***
ENC.–	Encuestador

INF. −	Informante
...	Vacilación o suspensión
..	Después de palabra incompleta o trunca
‘	Separa repeticiones
“ ”	Citas textuales
J	Aspiración de s
[--]	Lo que no se entendió. Lo que parece que es o que resulta ininteligible
iii	Repetición de fonemas finales
( )	Cuando el informante tose o le da risa
(?)	Se cree que el informante dijo algo pero no se está seguro
Mj., o Mm.	Sonido que se produce frecuentemente en los relatos

Nota. Tomado de Montes (1997). El español hablado en Bogotá. Relatos semilibres de informantes pertenecientes a tres estratos sociales. Bogotá: Instituto Caro y Cuervo.

3.3.2. Corpus del Habla Culta de Bogotá (HCB)

El Corpus del Habla Culta de Bogotá (HCB) está compuesto por aproximadamente 600 encuestas de una duración promedio de 30 minutos. Las grabaciones se recopilaron entre 1972 y 1984 en el marco del proyecto “Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica”.

La transcripción ortográfica de las encuestas del HCB se realizó durante la investigación en formato papel y con máquinas de escribir. Actualmente se están realizando procesos de digitalización por OCR, digitación y conversión a los distintos formatos para consulta y análisis. De acuerdo con Otálora y Fernández (1986) y después de una revisión general de las transcripciones, los criterios y convenciones usados para su realización se presentan en la tabla 7.

Tabla 7

Convenciones de transcripción del corpus HCB

Datos generales	
Enc. —	Encuestador
Inf. —	Informantes
Inf. A —Inf. B —	Si hay dos o más hablantes
X	Para remplazar los nombres de informantes y de otras personas en la grabación
[...]	Omisión de pasajes del texto por dificultades de audición.
[sic]	Indica que una parte de la transcripción no es error de imprenta sino uno realización apartada del uso general.
Ortografía	
Se siguieron la normas de ortografía convencionales de la época, solamente se registraron las diversas variantes de la palabra “entonces” (ento’es, ´tonces, ´to’es, ‘to’ces, ent’os, ‘to’s) como ejemplo ilustrativo de la adopción de distintas formas lingüísticas para una misma unidad, incluso en el mismo informante.	
Se procuró mantener la prosodia de los hablantes, sin embargo en varias ocasiones fue necesario adoptar criterios de puntuación para facilitar la comprensión del texto	

Las transcripciones se revisaron varias veces para asegurar la fidelidad del texto con la grabación	
Elementos extralingüísticos	
...	Señalar los fenómenos de sintaxis propios del habla como vacilación, pausas prolongadas, interferencias entres los hablantes, repeticiones, frases incompletas, entre otros.
[Risas]	Risas

Nota. Elaboración propia con base en Otálora y Fernández (1986). El habla de la ciudad de Bogotá: Relatos para su estudio. Bogotá: Instituto Caro y Cuero.

3.4.Generalidades

Luego de revisar los protocolos de transcripción de varios corpus orales del español, encontramos distintos elementos comunes y que consideramos importantes para la construcción de una propuesta de transcripción propia. A continuación, mencionaremos varios de estos puntos y evaluaremos la conveniencia e importancia para definir las convenciones de transcripción.

En primer lugar, sobre los datos generales, se suele guardar la confidencialidad de la información personal de los informantes buscando remplazarla con una letra o número. Para marcar los encuestadores o informantes se usan símbolos como "I." "E. 1" "INF." Etc. En algunos corpus se suele marcar la calidad de la grabación, pero no hay convenciones que sean similares. Asimismo, se encuentran sistemas como la TEI, el NERC, o el EAGLES que manejan varias etiquetas para la información vinculada con los datos de recolección de la grabación. Sin embargo, en CLICC estos datos suelen estar registrados en la ficha de metadatos y directamente en el sistema para facilitar la realización de las búsquedas.

Con respecto a las pautas ortográficas, los trabajos de transcripción suelen seguir la norma de lengua que se encuentre vigente y se toman ciertas decisiones arbitrarias para casos particulares como elicitaciones alejadas de esta norma, fenómenos propios de la producción oral como alargamientos, adición o elisión de sonidos y demás casos. En cuanto a los signos de puntuación, no hay un criterio uniforme, pues algunos corpus justifican su inclusión, mientras que otros prefieren evitarlos o manejar solo unos pocos como los puntos o (en casos mucho más esporádicos) las comas. Por otro lado, son pocos los sistemas de transcripción que sugieren llevar a las versiones normativas algunas elicitaciones propias de producción individual de los hablantes o formas dialectales; pues, en general, se prefiere la conservación del material tal como es producido, valiéndose de marcas o etiquetas para indicar los fenómenos que puedan tener lugar.

Finalmente, hay una amplia gama de etiquetas para dar lugar a *elementos extralingüísticos*. Los más tenidos en cuenta son la risa, los murmullos o sonidos propios de la

duda o el pensamiento, los ruidos en el ambiente de grabación o interrupciones fuera del control de los participantes, las pausas y la gesticulación.

Para el caso de los Corpus Lingüísticos del Instituto Caro y Cuervo, se sugiere realizar transcripciones que sigan la norma de lengua vigente para el español, ya que los fenómenos propios de los hablantes en cuanto a fonética o sintaxis harían parte de procesos de transcripción y anotación posteriores para la realización de investigaciones. Adicionalmente, es recomendable no emplear signos de puntuación dentro de los enunciados y valerse únicamente de símbolos para marcar las pausas y silencios, ya que introducir marcas adicionales puede resultar arbitrario al intentar privilegiar la lectura de los usuarios de la plataforma frente a la producción de los hablantes. También, es importante manejar la información de los hablantes de forma confidencial, sin revelar datos como sus nombres y demás de carácter personal[3].

Por otro lado, deben incluirse marcas que den cuenta de eventos fuera de la información lingüística que hacen parte de la grabación, del ambiente en el que se recogen los datos, y de los fenómenos propios del habla oral. Finalmente, es de suma importancia lograr consolidar un sistema de transcripción ortográfica que permita una lectura sencilla de las grabaciones, que no sea demasiado complejo y que no sature los contenidos o a los usuarios con etiquetas que deban ser entendidas e interpretadas para el correcto acceso a los materiales. El ideal es obtener transcripciones sencillas, claras y amigables con su consulta.

3.5. Programas usados para la transcripción

En este apartado expondremos algunos programas para la transcripción de audio disponibles en línea de forma gratuita.

3.5.1. EasyTranscript

EasyTranscript es un programa que soporta distintos formatos de audio y vídeo y que maneja una interfaz sencilla para el usuario (ver, figura 1). Además, emplea herramientas de edición de texto parecidas a las que se pueden encontrar en word (negrita, cursiva, selección de fuentes, entre otras). El software incluye el tiempo de trabajo de transcripción y conmutadores de

[3] Las únicas transcripciones que incluyen este tipo de datos son las de los corpus del ALEC, pues las grabaciones tomadas en su momento los incluyen por contar con la autorización de los informantes.

pedal para manejar la reproducción del audio con los pies mientras se realiza la transcripción con las manos (Institute for Social Innovation [ISI], s.f.).

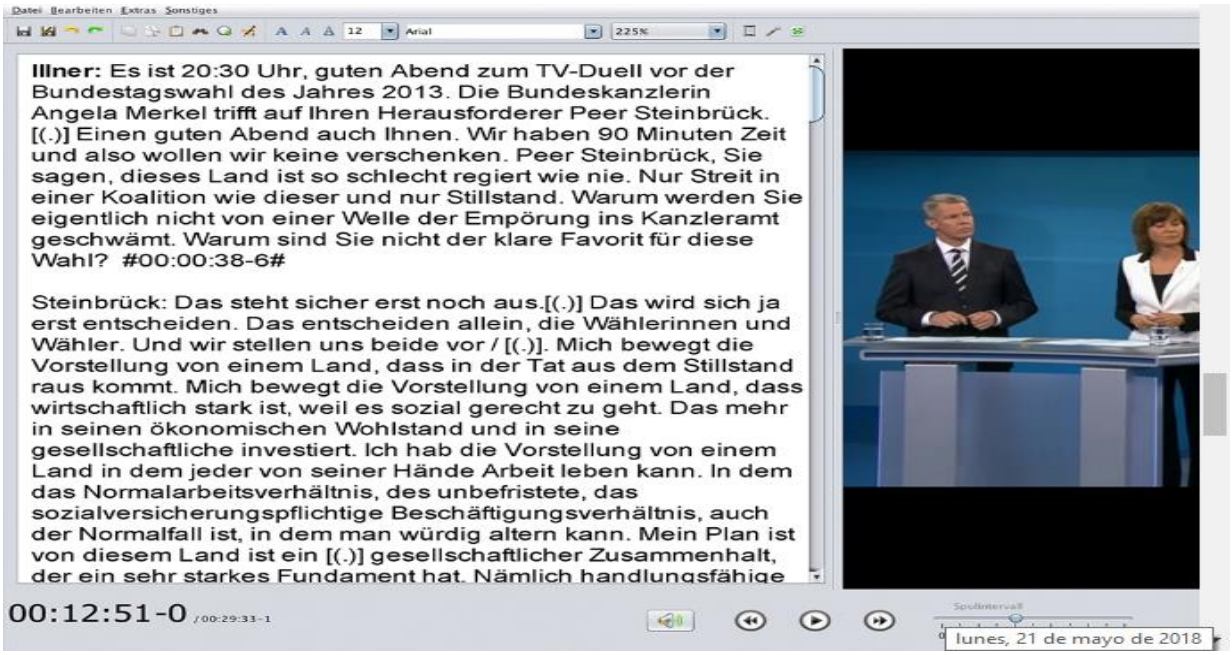

Figura 1. Interfaz de usuario de EasyTranscript. Recuperado de: https://e-werkzeug.eu/index.php/de/produkte/easytranscript

3.5.2. TranscriberAG

TranscriberAG es una herramienta para la segmentación, nivelación y transcripción de señales de audio. Esta aplicación cuenta con herramientas para la edición y almacenamiento del audio y de la transcripción; para deshacer cambios, cortar, pegar, importar clips, etc.; para la búsqueda de fragmentos de audio; para la anotación de ruidos, turnos, comentarios, pronunciación, entre otros.; para la reproducción, pausa y manejo del audio; y para el manejo de las herramientas del sistema (TLA-team, s.f.).

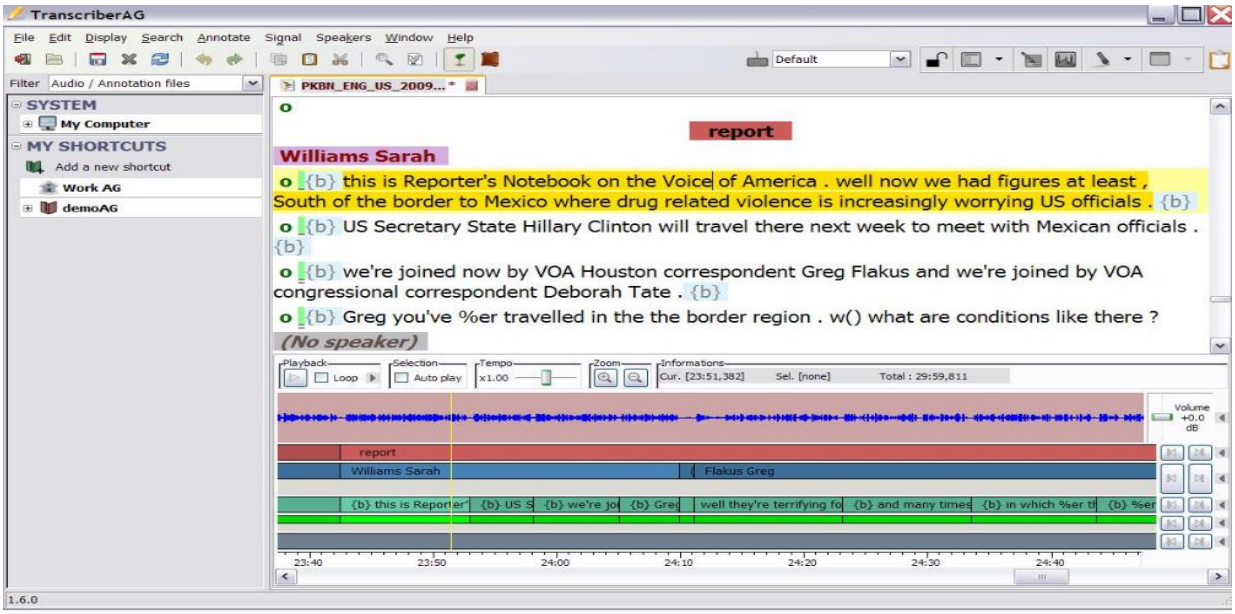

Figura 2. Interfaz de TranscriberAG. Recuperado de: https://e-werkzeug.eu/index.php/de/produkte/easytranscript

3.5.3. Listen N Write

Listen N Write es un programa muy liviano, de manejo muy sencillo y con posibilidades de interacción con el audio que permiten fijar puntos a lo largo de la reproducción para volver a ellos en cualquier momento (Torroba, 2015)

Figura 3. Interfaz de Listen N Write. Recuperado de: https://e-werkzeug.eu/index.php/de/produkte/easytranscript

3.5.4. Transcription Aid

Transcription Aid ofrece una interfaz sencilla para el usuario, opción de guardado automático y facilidad para auto-completar palabras recurrentes que ya se han guardado en el transcriptor.

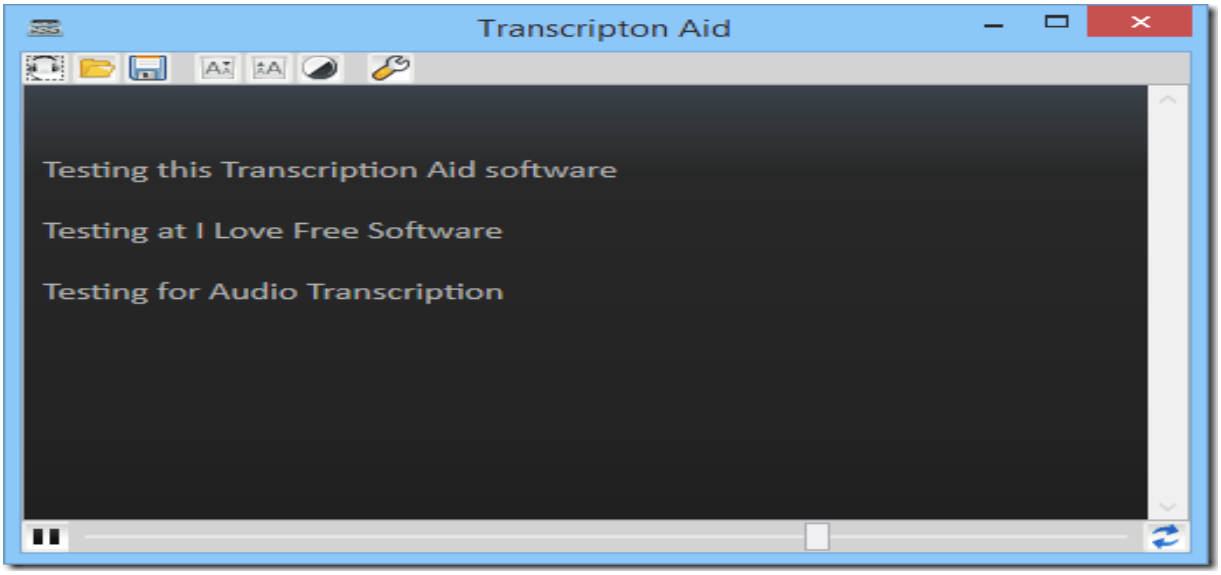

Figura 4. Interfaz de Transcription Aid. Recuperado de: https://e-werkzeug.eu/index.php/de/produkte/easytranscript

3.5.5. ELAN

ELAN es una herramienta profesional para la anotación de audio y vídeo. Permite la adición de varias capas para agregar texto, comentarios, traducción o cualquier característica que se quiera marcar sobre el material multimedia. Las capas o niveles de anotación pueden estar relacionadas jerárquicamente y se pueden alinear con la señal de audio o referirse a otras anotaciones existentes. Además, la transcripción se almacena en formato XML y siempre está en Unicode (Max Planet Institute for Psycholinguistics, s.f.).

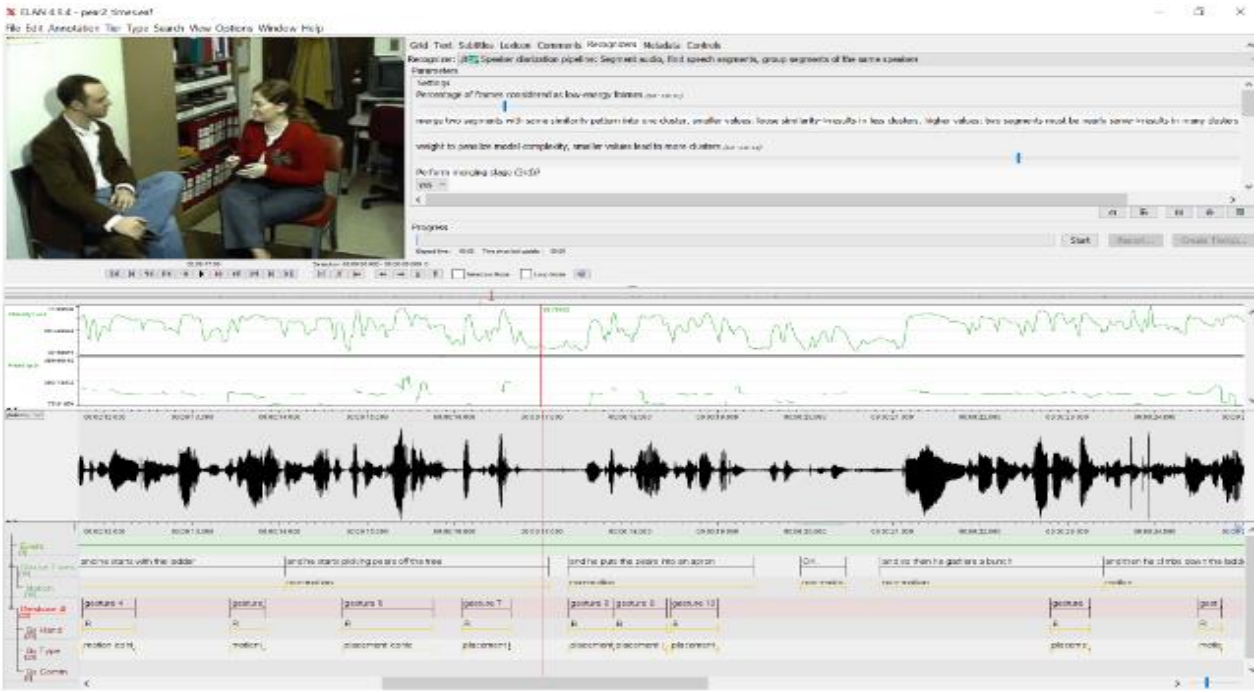

Figura 5. Interfaz de ELAN. Recuperado de: https://tla.mpi.nl/tools/tla-tools/elan/screenshots/

3.5.6. PRAAT

Praat es un software para el análisis fonético del habla que permite la transcripción de las señales de voz en varios niveles, la visualización del espectrograma, el oscilograma y el análisis de las propiedades acústicas de la emisión como la frecuencia fundamental (F0), la intensidad, la duración, la curva melódica y los formantes, entre otros. Praat es uno de los programas más populares en el análisis lingüístico del habla y contiene varios recursos para graficar, programar y generar estadísticas. Adicionalmente, se trata de software de uso libre.

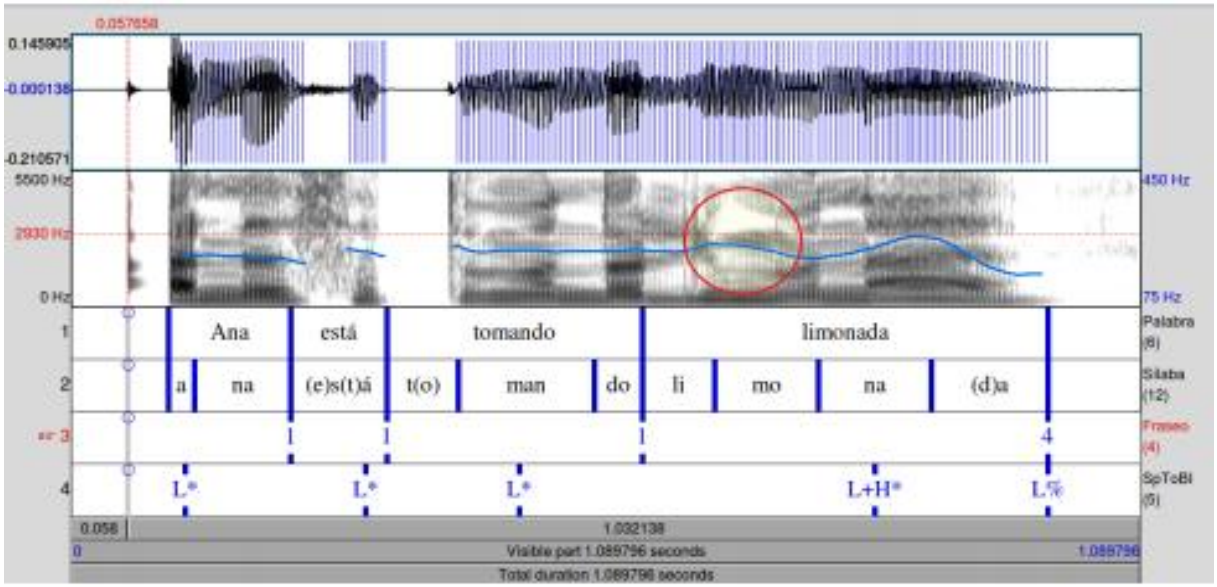

Figura 6. Interfaz de transcripción de PRAAT. Recuperado de: Correa (2014). Manual de análisis acústico del habla con PRAAT. Bogotá: Instituto Caro y Cuervo. Imprenta Patriótica, 130 pp. (Series Minor; 49).

3.5.7. EMU Speech Database Managment System (EMU-SDMS)

EMU Speech Database Managmente System (EMU-SDMS) es una colección de herramientas de software para el manejo y análisis de bases de datos de audio diseñadas para funcionar en el lenguaje de programación R (Winkelmann et al., 2017). EMU-SDMS cuenta con un etiquetador interactivo que posibilita la visualización de espectrogramas y otras formas de onda de la señal de habla, y que permite la creación de etiquetas jerárquicas, así como secuenciales, para cada enunciado (Winkelmann, 2016).

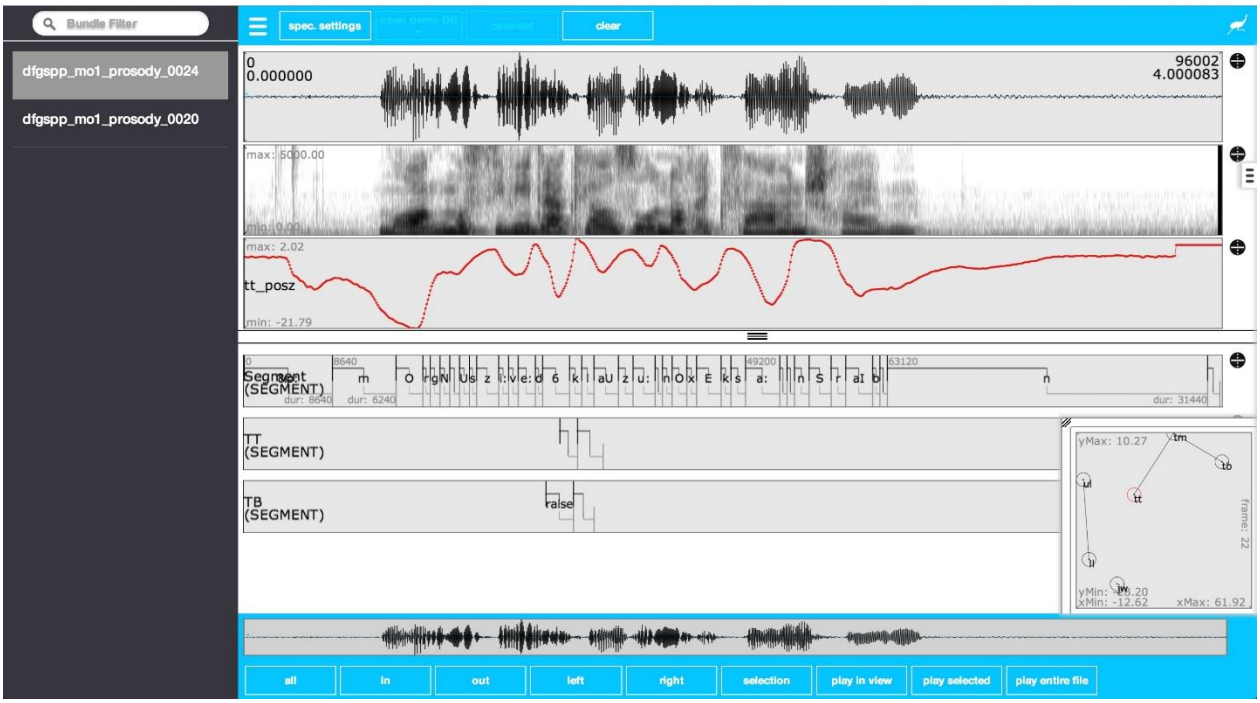

Figura 7. Screenshot of EMU-webApp. Recuperado de https://ips-lmu.github.io/EMU.html

3.5.8. Observaciones sobre los programas usados en CLICC

La plataforma CLICC busca integrar herramientas de análisis lingüístico que permitan a investigadores y usuarios manejar y analizar los datos fácilmente y almacenar información nueva que pueda alimentar los corpus de manera ágil. Hasta ahora se integró la herramienta ELAN para la visualización de la alineación de algunos audios y su presentación en la interfaz del usuario. Posteriormente, se buscaría integrar herramientas como PRAAT y, más adelante, EMU-SDMS, por lo cual se recomienda a los investigadores trabajar con estas tres herramientas para la transcripción. Al mismo tiempo, resaltamos que hay algunos grupos de investigación que han usado ELAN y PRAAT para la realización de la transcripción de sus corpus y esta ha sido una de las razones por las que son herramientas de prioridad de integración en CLICC. Por otro lado, es importante destacar que estas últimas aplicaciones son de uso y código libre, cuentan con varias herramientas de análisis y sus formatos y archivos de resultado son compatibles entre sí.

ELAN permite interactuar con el material multimedia a través de una interfaz sencilla que facilita la realización de comentarios o transcripción de fragmentos alineados temporalmente con

su elicitación. Además, permite crear distintas líneas de anotación para organizar la información según las etiquetas o el orden de jerarquías que el investigador decida.

Por su parte, PRAAT presenta la misma lógica de transcripción alineada temporalmente que posee ELAN, pero genera archivos del tipo *.Textgrid* de fácil lectura por parte de sistemas que soportan la visualización de la transcripción a la par de la emisión del contenido sonoro o del mismo ELAN. Adicionalmente, facilita el análisis de las características acústicas de los sonidos, de manera que se constituye en una herramienta de utilidad para el especialista o el interesado en el análisis fonético.

Por último, es importante comentar que el uso de software para la transcripción asistida debe ser revisado y complementado por la mano del usuario. Los resultados obtenidos automáticamente por medio de una herramienta tecnológica pueden omitir o no dar fe de la información lingüística contenida en las grabaciones. En ese sentido, siempre será indispensable la participación del investigador para la consecución de una transcripción final.

4. Propuesta de transcripción ortográfica para los Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC)

Tras la revisión de varios protocolos de transcripción y la definición de elementos que consideramos fundamentales para la representación escrita de materiales producidos en contextos de comunicación oral. En la tabla 8 presentamos la propuesta de transcripción ortográfica para los corpus orales del ICC. Al respecto es importante mencionar que los corpus y materiales que ya tienen transcripción mantendrán los criterios definidos durante la investigación. Con respecto a los nuevos corpus esperamos que sigan esta guía que permitirá la normalización de la información en CLICC.

Tabla 8

Convenciones de transcripción ortográfica de CLICC

Símbolo	Descripción	Ejemplo
Marcas de participantes y grabación		
ENC: ENC1: ENC2:	Señala la participación del encuestador. En caso de existir más de un encuestador se agrega el número del encuestador.	Encuestador ENC: Encuestador 1 ENC1: Encuestador 2 ENC2:
INF: INF 1: INF 2:	Indica la participación de un informante. En caso de existir más de un informante se agrega el número del informante.	Informante INF: Informante 1 INF1: Informante 2 INF2:
INF:(GRUPO)	Indica la participación simultánea de varios informantes. Se debe utilizar cuando el número de personas no se puede determinar con claridad (i.e. un coro cantando al unísono).	INF:(GRUPO) Oh señor ten piedad.
***	Marca el lugar de un nombre propio y se utiliza para guardar la confidencialidad de los informantes. Los nombres propios que no comprometen la privacidad de los informantes no se eliminan.	ENC: Cuénteme Don *** ¿Cuántos años tiene?
[X:]	Interlocutor no reconocido. Su intervención se debe incluir entre corchetes.	ENC: Cuénteme Don *** ¿Cuántos años tiene? [X: ya toca comer]
<INF1::> <<INF2:>>	Indica solapamiento entre los participantes. El enunciado del segundo hablante que se solapa se marca con <<>>	<INF1: ¿su casa es grande o pequeña?> <<INF2: pequeñita>>

[Ruido de fondo:]	Ruido de fondo. La marca de [Ruido de fondo:] indica los ruidos de fondo la grabación. Los dos puntos aclaran la naturaleza del ruido producido. Si existen varios ruidos se separan por ;	INF: iba caminando y de repente se cayó [Ruido de fondo: voces] una casa [Ruido de fondo: puerta] [Ruido de fondo: viento] [Ruido de fondo: pitos] [Ruido de fondo: buses; carros]
[Ruido de fondo continuo:]	Ruido de fondo continuo.	INF: [Ruido de fondo continuo: voces] iba caminando de repente se cayó una casa
[Ininteligible]	Señala la imposibilidad de transcripción por condiciones del audio o por la realización del enunciado	INF: antes uno podía salir a cazar pero ahora eso se volvió un complique porque [Ininteligible] y cómo sale uno.
(( ))	Transcripción dudosa	INF: Yo pensaba que estaba bien ((decir lo de la finca))
[Corte]	Indica el corte abrupto de la grabación por motivos relacionados con las condiciones de la grabadora (terminación del carrete, límite del audio, etc.)	INF1: luego se pone en el macerador y [Corte]
Marcas de enunciado y turnos de habla		
/	Pausa mínima	INF1: luego se pone en el macerador / y
//	Pausa	INF1: luego se pone en el macerador y se macera//
[Silencio]	Silencio (lapso o intervalo de 3 segundos en adelante).	ENC: ¿Le gusta su barrio? INF: [Silencio] Pues sí/aunque es un poco inseguro//
+	Interrupción por otro participante	INF1: iba caminando y de repente+ INF2: ¿se cayó?
¿?	Enunciados interrogativos	ENC: ¿Le gusta su barrio?
¡!	Enunciados exclamativos	INF1: iba caminando y se dio un ¡totazo!
...	Vacilación, suspensión o entonación suspendida. Después de una oración incompleta.	INF: voy a comer y...
..	Palabra incompleta o trunca	INF: Me voy para Maniza..
<INF1::> <<INF2:>>	Indica solapamiento entre los participantes. El enunciado del segundo hablante que se solapa se marca con <<>>	<INF1: ¿su casa es grande o pequeña?> <<INF2: pequeñita>>
Fenómenos de registro oral		

→	Alargamiento vocálico o consonántico a final de palabra.	INF1: y → / hay pan //
[AV]	Apoyo vocálico: AVmm, AVeh, AVah Indica la emisión de marcadores discursivos para momentos de asentimiento, aprobación, duda o negación.	ENC1: ¿usted quiere decir que no lo conoce? INF1: AVmm/ sí/ tiene razón//
=	Señala el abandono de una palabra por motivos de autocorrección o modificación del mensaje.	INF: Mi señora tiene un perro = un gato pequeño.
“ ”	Encierran la voz de la tercera persona en una cita producida por el informante. Discurso directo.	INF1: Entonces el señor dijo: “Tuvo que ser usted”.
,	Repetición	INF: Cuando estaba en la, en la, en la misa.
[SS:]	Signos suprasegmentales [SS:onomatopeya] [SS: soplido] [SS: suspiro] [SS: tos] [SS: risa] [SS: chasquido] [SS: tarareo]	ENC1: el sonido se parece al de un helicóptero toco toco toco [SS:onomatopeya]. INF1: Tiroriroriro [SS: tarareo]
[ALT:]	Alteraciones o elisiones propias del registro oral. Se transcribe la palabra correcta y al lado la transliteración.	INF1: en → / todos [ALT: to] los lugares del pueblo / sí //

Nota. Elaboración propia

Para la transcripción ortográfica se recomienda seguir las normas ortográficas vigentes y respetar al máximo el habla hallada en las grabaciones. Las convenciones se organizan en tres categorías: marcas de participantes y grabación para marcar datos sobre los informantes, encuestadores y sobre la calidad de la grabación; marcas de enunciado y turnos de habla para representar las pausas, turnos, cortes entre otros; y fenómenos de registro oral para marcar los elementos extralingüísticos propios del registro oral.

Los símbolos de puntuación no se usarán, ya que la marca de inicio o final de enunciado podrían variar de acuerdo con la percepción del investigador, de esta manera se manejarán los signos presentados en la categoría 2. Los números se transcriben en letras, con excepción de las fechas (p. ej. treinta animales - 20 de agosto de 1918). No hay categorías para la marca de fenómenos fonéticos, léxicos, sintácticos, etc., ya que consideramos que estos se deben marcar en los procesos de transcripción fonética o de anotación morfosintáctica que se podrían llevar a

cabo en fases posteriores de alimentación del corpus y de acuerdo con las necesidades específicas de cada investigación.

5. Referencias

Brown, M. B. (1994). WHAT IS THE TEI. *Information Technology and Libraries, 13*(1), 8-8.

Cabedo, Adrián y Pons, Salvador (eds.): Corpus Val.Es.Co 2.0. Recuperado de: http://www.valesco.es

Corpus Oral del Español como Lengua Extranjera (CORELE). Recuperado de: http://cartago.lllf.uam.es/corele/home_es.html Consultado: [Mayo 15 de 2018].

Correa, A. (2014). Manual de análisis acústico del habla con PRAAT. Bogotá: Instituto Caro y Cuervo. Imprenta Patriótica, 130 pp. (Series Minor; 49).

Crystal, D. (1991). The Cambridge Encyclopedia of Language. Cambridge: Cambridge University Press.

ELAN (Version 5.0.0-beta) [Computer software]. (2017, Abril 18). Nijmegen: Max Planck Institute for Psycholinguistics. Recuperado de: https://tla.mpi.nl/tools/tla-tools/elan/

Fernández-Ordóñez, I. (2018). COSER Corpus Oral y Sonoro del Español Rural. Recuperado de http://www.corpusrural.es

Farías, L., & Montero, M. (2005). De la transcripción y otros aspectos artesanales de la investigación cualitativa. International Journal of Qualitative Methods, 4(1), 1-14.

Grupo de Gramática del Español. Corpus para el estudio del español oral (ESLORA). Universidad Santiago de Compostela. Recuperado de: http://eslora.usc.es/. Consultado: [Mayo 15 de 2018]

Linell, P. (2005). The written language bias in linguistics. Its nature, origins and transformations. Londres: Routledge.

Llisterri, J. (1996). EAGLES Preliminary recommendations on Spoken Texts EAG--TCWG--SPT/P. Barcelona, España. Recuperado de http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html

Llisterri, J. (1997). Transcripción, etiquetado y codificación de corpus orales. Seminario de industrias de la Lengua, Curso "Etiquetación y extracción de información de grandes corpus textuales", Fundación Duques de Soria, Soria. Recuperado de: http://liceu.uab.es/~joaquim/publicacions/FDS97.html

Malinowski, J. 1935. An Ethnographic Theory of the Magical Word. Coral Gardens and Their Magic, vol. II. Londres: Allen & Urwin.

Montes (1997). El español hablado en Bogotá. Relatos semilibres de informantes pertenecientes a tres estratos sociales. Bogotá: Instituto Caro y Cuervo.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. Schieffelin (Eds.), Developmental Pragmatics. New York: Academic Press, 43-72.

Pato, E. (2014). COLEM, Corpus Oral de la Lengua Española en Montreal. Montreal: Universidad de Montreal. Recuperado de: https://esp-montreal.jimdo.com/colem/. Consultado: [10 de mayo de 2018]

Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. RLA. Revista de lingüística teórica y aplicada, 46(1), 93-119.

Pino, M. (1998). Transcripción, codificación y almacenamiento de los textos orales del corpus CREA. Versión 2.0. Instituto de Lexicografía, Real Academia Española. 29/07/1997. En J. A. Samper, C. E. Hernández Cabrera y M. Troya (Eds.), *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (MC-NLCH)*. [CD-ROM] Las Palmas de Gran Canaria: Servicio de Publicaciones de la Universidad de Las Palmas de Gran Canaria.

Plummer, K. (1989). *Los documentos personales* (Introducción a los problemas y la bibliografía del método humanista). Madrid: Siglo veintiuno.

PRESEEA (2014-2018): Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá. Recuperado de: http://preseea.linguas.net. Consultado: [9 de mayo de 2018]

Otálora y Fernández (1986). El habla de la ciudad de Bogotá: Relatos para su estudio. Bogotá: Instituto Caro y Cuero.

Recalde, M., & Rozas, V. V. (2009). Problemas metodológicos en la formación de corpus orales. A Survey on Corpus-based Research/Panorama de investigaciones basadas en corpus, 37-49.

Torroba, L. (2015). Softzone: Listen N Write, transcribir archivos de audio nunca fue tan fácil. Recuperado de https://www.softzone.es/2015/10/10/listen-n-write-transcribir-archivos-de-audio-nunca-fue-tan-facil/. Consultado: [19 de mayo de 2018].

Torruella, J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. Filología e informática. *Nuevas tecnologías en los estudios filológicos, 45-77.*

Winkelmann, R. (2016). The EMU Speech Database Management System (EMU-SDMS) [En línea]. Recuperado de: https://ips-lmu.github.io/EMU.html

Winkelmann, R., Harrington, J. y Jänsah, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language* 45, p. 292-410.