Salvador Pons Bordería (ORCID: 0000-0001-5788-5506) y Sara Badia-Climent (ORCID: 0000-0002-7993-5997),
«De la transcripción al análisis: desarrollos técnicos del corpus Val.Es.Co. 3.0».
Biblioteca de Babel: Revista de Filología Hispánica, vol. extra. 2 (2024), pp. 79-113.
DOI: 10.15366/bibliotecababel2024.extra2.003 · **ISSNe**: 2695-6349 · Recibido: 11/09/2024 · Aceptado: 03/11/2024

De la transcripción al análisis: desarrollos técnicos del corpus Val.Es.Co. 3.0[1]

Salvador Pons Bordería
Universitat de València
salvador.pons@uv.es

Sara Badia Climent
Universitat de València
sara.badia@uv.es

Resumen: Este artículo aborda las decisiones teóricas y técnicas adoptadas en la creación del corpus Val.Es.Co. 3.0. El objetivo principal es detallar todos los pasos que se han seguido para lograr crear un corpus oral accesible en formato digital que no solo permita trabajar a los usuarios con el contenido de las transcripciones y el etiquetado de los fenómenos discursivos, sino también con el análisis de su contenido a partir del modelo de unidades del grupo de investigación: subactos, actos, intervenciones, diálogos y discursos. Para ello, el trabajo se ha dividido en tres secciones. En primer lugar, se hace una breve introducción sobre los dos pilares fundamentales que componen el trabajo de creación del corpus Val.Es.Co. 3.0.: la transcripción y su posterior segmentación. En segundo, se describe el proceso de trabajo técnico interno que cada una de las conversaciones ha recibido, desde la transcripción hasta la segmentación de unidades. Por último, la tercera sección expone cómo se visualizan las transcripciones en la web del corpus y detalla las distintas funcionalidades que este pone a disposición de los usuarios.

Palabras clave: corpus oral, lingüística computacional, segmentación discursiva, transcripción, corpus Val.Es.Co., español hablado.

[1] Este trabajo ha sido posible gracias al proyecto CIPROM/2021/038 Hacia la caracterización diacrónica del siglo xx (DIA20), del proyecto PROMETEO de la Generalitat Valenciana, y al proyecto de I+D+I PID2021-125222NB-I00 Aportaciones para una caracterizacion diacronica del siglo xx, financiado por MCIN/AEI/10.13039/501100011033/ y por FEDER Una manera de hacer Europa. Los autores agradecen a los revisores anónimos sus sugerencias y comentarios, que han mejorado notablemente la versión final de este artículo.

From transcription to analysis: technical developments of the Val.Es.Co. 3.0 corpus

Abstract: This article examines the theoretical and technical decisions involved in the elaboration of the Val.Es.Co. 3.0 corpus. Its main goal is to detail the steps taken to develop an accessible digital oral corpus. The Val.Es.Co. 3.0 corpus provides users with spontaneous conversations and a system of discourse-based tags. It also analizes a subset of conversations with the Val.Es.Co. model of discourse units: subacts, acts, interventions, dialogues, and discourses. This article is divided into three sections. Section two outlines the two basic pillars of the creation process for the Val.Es.Co. 3.0 corpus: transcription and its subsequent analysis. Section 3 describes the backend, especially the technical decisions adopted during the processes of transcription and discourse segmentation. Finally, Section 4 explains how the transcriptions are displayed on the website and details the corpus frontend main features.

Keywords: oral corpus, computational linguistics, discourse segmentation, transcription, corpus Val.Es.Co., spoken Spanish.

1. Introducción

El desarrollo de corpus lingüísticos permite acercar la lingüística al lenguaje empleado por los hablantes en contextos reales (Bolaños 2015; Bernal y Hincapie 2018; García-Miguel 2022), ya sea en su forma oral o escrita. Si bien es cierto que los corpus de discursos escritos cuentan con una serie de dificultades relativas, generalmente, al formato en el que se plasma el texto y a los datos que sitúan el discurso (Torruella y Llisterri 1999), los corpus basados en discursos orales se enfrentan a otro tipo de dificultades, puesto que, por su carácter inmediato (Briz 2010), plantean cuestiones específicas que van desde la fase misma de recolección hasta su procesamiento informático (Briz 1996, Pons Bordería 2022).

Estos problemas aumentan cuando el corpus incluye, además de una transcripción, un análisis del material lingüístico, y no solo por las cuestiones teóricas relativas al tipo de interpretación, sino por las dificultades técnicas que plantea incorporar una carga de información adicional al texto transcrito. Dichas dificultades afectan tanto al diseño del *frontend* (interfaz, disposición de las ventanas de resultados, botones de búsqueda, etc.) como, sobre todo, al *backend* (secuenciación de la información, diseño de la hoja de ELAN, vocabulario de metadatos,

estructura del corpus, etc.). Este trabajo pretende abordar dichas cuestiones a partir del trabajo realizado en la elaboración de la versión 3.0 del corpus Val.Es.Co. (Valencia Español Coloquial), lo que puede ser de interés para la lingüística de corpus en general y para la del español hablado en particular.

Este artículo se organiza en tres partes: primero, se revisarán las características relevantes para el diseño de un corpus en el ámbito del español hablado. En segundo lugar, se mostrarán las decisiones informáticas tomadas en el corpus Val.Es.Co. 3.0. y los pasos seguidos para lograr su adecuada visualización en la página web y su correspondiente exportación. Por último, se mostrará cómo se visualizan las conversaciones en la página web y las distintas funcionalidades accesibles para los usuarios.

2. Estado de la cuestión: el trabajo de creación de un corpus

El interés por los corpus orales en la lingüística cobra especial relevancia con el desarrollo del Análisis de la Conversación (Sacks, Schegloff y Jefferson 1974)[2] y, en el ámbito hispánico, con los trabajos pioneros de Criado de Val (1964) y el proyecto PILEI dirigido por Lope Blanch (Lope Blanch 1971, 1976, 1986). Este interés teórico supuso, además, el desarrollo de una metodología para el reflejo del habla , ya que los sistemas ortográficos se revelaron insuficientes para tal fin. Gail Jefferson fue la primera investigadora en abordar estas cuestiones y en desarrollar un sistema de transcripción propio de las conversaciones, denominado hoy en día *jeffersoniano* (Jefferson y Sacks 2000; Jefferson 2004; Margret *et al*. 2009; Bassi 2015). Esta metodología implica aspectos tales como la transcripción de las conversaciones y su etiquetado (§ 2.1) y, en un nivel diferente de estudio, su análisis (§ 2.2.).

2.1. La transcripción de la conversación

Transcribir la oralidad conversacional implica lograr un equilibrio entre fidelidad y utilidad: por un lado, la transcripción debe ser lo suficientemente estrecha como para reflejar los fenómenos que cada corpus pretende reflejar; por otro, debe ser adecuada a los intereses del investigador, es decir, no contar con una sobrecarga de información que no permita o dificulte el estudio de su objeto (Pons Bordería 2022).

[2] La etnometodología del habla (Garfinkel 1967) ya había abordado la recogida y transcripción de material lingüístico oral; no obstante, sus intereses se centraban, esencialmente, en el análisis de las características sociales de los individuos y su reflejo en el habla (Zimmerman 1978; O'Keefe 1979), a diferencia de las disciplinas lingüísticas, que buscan analizar el lenguaje en sí mismo o en relación con rasgos de los individuos, del contexto o la sociedad.

Esta cuestión más general se puede concretar con la pregunta de cuánta información debe contener la transcripción: tratándose un corpus de lenguaje oral, una primera respuesta podría ser «todo aquello que se dice». Pero comunicarse, especialmente cuando se trata de discursos cara a cara, va más allá de las palabras: la entonación, el paralenguaje, los gestos e, incluso, las acciones extralingüísticas cobran un papel tan fundamental que, en ocasiones, dotan de sentido a determinados enunciados que, sin dicha información, serían incomprensibles (Cestero 2014; Cabanes 2023). En este punto el investigador debe buscar un equilibrio entre su objeto de estudio, que puede ser variable, y la cantidad de información que decida incluir en su corpus.

Se sigue de esto que no existe una única forma de transcribir, sino varias, en función del carácter más o menos amplio de la investigación. En este sentido, Pons Bordería (2022: 43) distingue hasta cinco niveles de detalle en la transcripción, que dependen del grado de complejidad y de la cantidad de fenómenos que incluya:

Nivel I: codificación ortográfica.

Nivel II: codificación según los principios del Análisis de la Conversación.

Nivel III: codificación de nivel II con el añadido de la información prosódica.

Nivel IV: codificación de nivel III con el añadido de información kinésica y paralingüística.

Nivel V: codificación de nivel IV con el añadido de vídeo(s) de los participantes.

Asimismo, dentro del nivel de información kinésica y paralingüística (niveles III y IV) pueden codificarse distintos tipos de información (Poyatos 1994, 2018), desde la gesticulación voluntaria o involuntaria del cuerpo hasta las actividades realizadas por los participantes, pasando por el registro de aquellos elementos contextuales o situacionales que puedan influenciar la comunicación (Poyatos 2018: 23-24; Cabanes 2023).

Muchos de los corpus de español hablado actuales se sitúan entre los niveles II y III, como ocurre con el corpus del Proyecto para el estudio sociolingüístico del español de España y de América (PRESEEA), el Macrocorpus de la norma lingüística culta de las principales ciudades de España y América (MC-NC), el Corpus Oral del Lenguaje Adolescente (COLA) o el corpus AMERESCO, entre otros (Rojo 2016; Briz y Carcelén 2019). Si bien aprovechan la ortografía del

español, reflejan, al mismo tiempo, ciertos fenómenos de la oralidad. Por ejemplo, la elisión de la *d* intervocálica en los participios como en *llega(d)o* o los alargamientos vocálicos mediante la repetición de la letra del sonido correspondiente como en *uun*. Suele ser frecuente, en el caso de los corpus de conversaciones, señalar además los solapamientos ([ ]), las intervenciones inmediatas (§) o las pausas en los enunciados (/). Por último, los corpus orales también suelen dan cuenta de ciertos fenómenos paralingüísticos, como las risas (RISAS), o ciertas acciones necesarias para entender la conversación.

Los corpus se consultan en abierto a través de la red mediante un sistema de búsqueda. Este debe permitir la identificación no solo de las palabras, sino también de la información de la que el grupo de trabajo en cuestión haya decidido dar cuenta. Para ello, el estándar que se ha impuesto desde hace años se basa en el lenguaje de etiquetado XML (Santamaría 1999; Brun 2005) y, en el ámbito de la lingüística, frecuentemente en el sistema TEI (Alcaraz y Vázquez 2016; Del Rio y Allés-Torrent 2023). No obstante, este sistema se ha ido modificando y adaptando de acuerdo con los objetivos e intereses de cada grupo de investigación.

Así, en el texto, las etiquetas ofrecen información sobre fenómenos conversacionales (solapamientos, énfasis, habla simultánea), prosodia, elisiones o acortamientos y observaciones necesarias para entender la transcripción. Gracias a este sistema el usuario de los corpus orales puede consultar transcripciones que, además de dar cuenta de las palabras de las grabaciones, identifican también estructuras lingüísticas o fenómenos discursivos, como se puede ver en los siguientes ejemplos:

(1) **PRESEEA – BARR_H22_037**
I: <tiempo = "02:33"/> yo de Vill <palabra_cortada/> de Villanueva<alargamiento/> añoro<alargamiento/> // principalmente mis padres // mis viejos // y añoro la paz / que<alargamiento/> // que <vacilación/> se perdió<alargamiento/> // o sea se ha ido perdiendo // eeh // esa tranquilidad<alargamiento/> // esa gente sana // de la cual yo yo <vacilación/> dejé // o sea sí<alargamiento/> había / problemas / normales / como en todos las regiones yo pienso que // del mundo pero // pero<alargamiento/>…

En (1), correspondiente al proyecto PRESEEA, se puede observar el uso de distintas etiquetas, como la localización temporal del fragmento (<tiempo = "02:33"/>), ya que sus transcripciones se transcriben directamente sobre un editor de textos y necesitan mantener una conexión entre el texto y el audio. También se pueden ver ejemplos de

etiquetado relativos a características de la oralidad como las palabras cortadas (<palabra_cortada/>), los alargamientos (<alargamiento/>) o las vacilaciones (<vacilación/>), entre otros fenómenos.

(2) **ESLORA - SCOM_H13_013**
hab2: <ininteligible/>
hab1: yo qué sé cómo están <ruido tipo="chasquido boca"/> <pausa_larga/>
hab1: están ampliando las aceras <pausa/> y están hacien y están pintando muchas fachadas de muchos edificios <pausa/> pues le están dando un toque más <pausa/> unificado a la zona nueva <pausa/> o sea algo más <pausa/> la están haciendo un poco más bonita <pausa/> de lo f <pausa/> o sea partiendo de lo fea que es <risa/> <pausa/>
hab2: intentan arreglándola un po
hab1: <ininteligible/> sí intentando arreglar un poco y <pausa_larga/>

El corpus ESLORA (Corpus para el estudio del español oral) también emplea el sistema de etiquetado XML. En (2), se observa el uso de etiquetas que marcan elementos de paralenguaje, como el chasquido de la lengua (<ruido tipo="chasquido boca"/>). Además, se indican dos tipos de pausas y se utilizan etiquetas específicas, como (<ininteligible/>), para señalar aquellas partes del audio que no pueden ser transcritas.

Al señalar fenómenos conversacionales, el etiquetado permite que las transcripciones no se limiten a registrar únicamente las palabras de los hablantes, sino que incluyan también la representación de diversos fenómenos lingüísticos y paralingüísticos, lo que facilita el procesamiento de la información, la búsqueda automatizada y la extracción de datos, de manera que los corpus permitan consultas complejas que van más allá del contenido literal del habla. Además, este sistema posibilita el recuento automático de los fenómenos identificados mediante las etiquetas, lo que simplifica la realización de estudios estadísticos.

2.2. De la transcripción al análisis de los datos orales

Un paso más en el proceso de elaboración de corpus consiste en añadir a la transcripción un análisis de los datos. Al incluir este objetivo, se supera la mera representación del material transcrito para adentrarse en el campo de la interpretación[3].

[3] En realidad, toda transcripción es, en cierta medida, una interpretación. Pero si en una transcripción la interpretación busca una representación más o menos fidedigna de la realidad, un análisis busca una explicación del material transcrito. Así, aunque la subjetividad permee ambas

Una de las razones para añadir esta capa extra de información es abordar el problema de la *sintaxis de lo hablado*. La obra de Antonio Narbona (en especial Narbona 1989a, 1989b, 1990) demuestra la inadecuación de la sintaxis oracional para dar cuenta de la organización de la materia hablada en los discursos orales coloquiales, prototípicamente representados por las conversaciones espontáneas. Se inicia así en la lingüística española la pregunta de cómo explicar la estructura de lo hablado desde una base no sintáctica. Los modelos de segmentación discursiva, desarrollados en su mayoría en las lenguas románicas (Pons Bordería 2014.), ofrecen respuestas a esta cuestión desde varios criterios, a menudo superpuestos: prosódico, semántico, sintáctico o pragmático.

El grupo Val.Es.Co ha desarrollado un modelo de este tipo (Briz *et al*. 2003, Briz y Grupo Val.Es.Co. 2014, Pons Bordería 2022) que divide la conversación coloquial en unidades y subunidades sin residuo de un modo similar a como se procede en un análisis sintáctico. Como resultado de este proceso, quedan situados en «ámbitos de estudio diferentes los fenómenos lingüísticos discursivos y, en concreto, del español hablado […] [Además,] se evita así la casuística y la descripción aislada» (Val.Es.Co. 2014: 12). Para comprobar la adecuación teórica del modelo al objeto de estudio, se ha analizado una parte del corpus 3.0 (más de treinta y cinco mil palabras) con dicho modelo y se ha incorporado al corpus en la web. El objetivo de este intento es mostrar la capacidad del modelo para responder a la pregunta de Narbona mediante una contrastación empírica amplia.

La incorporación de este modelo al corpus ha supuesto un reto teórico, pero también aplicado: ha sido necesario incluir en el diseño y elaboración del corpus una serie de procedimientos complejos para incluir este análisis interpretativo al puro y simple proceso de transcripción. La sección § 3 explica cómo se ha llevado a cabo este desarrollo desde un punto de vista técnico.

3. La elaboración del corpus Val.Es.Co. 3.0: decisiones técnicas

Aunque el corpus Val.Es.Co. es una obra colectiva que manifiesta su continuidad desde 1995 (Briz *et al*. 1995), su versión 3.0 ha introducido cambios importantes en la forma en la que se tratan las conversaciones del corpus, hasta el punto de que se podría hablar de una refundación del corpus mismo. Asimismo, la introducción del análisis ha obligado a modificaciones de calado en el diseño del corpus. Todo esto ha obligado a abordar las cuestiones que se detallan en esta sección, en la

operaciones, lo hace orientada a dos objetivos completamente diversos.

que se explicará el funcionamiento del *backend* del corpus Val.Es.Co. 3.0., que es la base de su visualización en la versión web. Para ello, se tratarán cuestiones relativas al sistema de transcripción y etiquetado (§ 3.1), a la configuración y organización de las líneas de análisis del programa de transcripción ELAN (§ 3.2.) y a la metodología empleada para segmentar y analizar una conversación (§ 3.3).

3.1. Cambios en la transcripción

El carácter oral, conversacional y coloquial del corpus Val.Es.Co hace necesario reflejar información que va más allá de lo meramente ortográfico:

1) El contenido de las intervenciones.

2) Información prosódica.

3) Información interactiva, que incluye fenómenos dialógicos, como los solapamientos o las interrupciones.

4) Los sucesos extralingüísticos que afectan directamente a la conversación.

A estas cuatro metas se ha añadido, en la versión 3.0, un quinto objetivo:

5) El análisis del contenido de los enunciados (segmentación de unidades de la conversación).

En 1995, el grupo Val.Es.Co. adaptó al español el sistema de transcripción jeffersoniano, que se especializa en los fenómenos conversacionales más relevantes en una conversación: solapamientos, toma de turno, pausas, silencios, tonemas o realizaciones paralingüísticas, entre otros. Este sistema aprovecha los símbolos que ofrece un teclado ASCII para asociar fenómenos conversacionales a signos del teclado, como se muestra en la Tabla 1:

Fenómeno	Sistema de transcripción Val.Es.Co.
Mantenimiento del turno de un participante en un solapamiento	=
Lugar donde se inicia un solapamiento o superposición	[
Final del habla simultánea	]
Entonación suspendida	→
Entonación ascendente	↑
Entonación descendente	↓
Entonación circunfleja	^
Fragmento ininteligible	(())
Pausa de menos de medio segundo	/

Tabla 1. Signos conversacionales del sistema de transcripción Val.Es.Co. (versión 1995).

(3) **Conversación 2011.PT.S2**
50A21: [es la-] es la entidad de la Comunidad Valenciana↑ ¡no! de Europa↑/ [que mejor] paga↓ tía↑ s[í]
51C27: [de la Unión EURO]PE que mejor [paga→]/ [a los monitores] [¡hombre!]=
52B26: [((¡qué barbaridad!))] [(( )) qué- ¡qué suerte tenéis!]
53C27: =[mil nove]cientos euros al mes↑ pues de lunes a viernes por hacer el ((tonto)) [por la mañana]

(4) **Conversación 2011.PT.S5**
24B13: ¿el qué? ¿el qué?
25A12: que a Vero le he cogido más aprecio↑/ desde que va con vosotras porque me la he encontrao más y aunque ahora vaya con los jarcoretas sigue siendo igual→

El sistema inicial de transcripción aprovechaba los símbolos del teclado para indicar, por ejemplo, los solapamientos con corchetes ([ ]) o el mantenimiento de un turno con el signo igual (=) , pero no disponía de marca asociada para otros elementos, como los fenómenos de fonética sintáctica (destacado en negrita en 4). Este sistema jeffersoniano es un instrumento muy adecuado para leer las conversaciones, algo necesario para los investigadores en pragmática o en análisis conversacional, ya que el comportamiento interactivo de los participantes se desarrolla durante todo el acontecimiento comunicativo y es preciso tener una visión detallada de dicha relación junto a la forma en la que se se está creando material lingüístico. Sin embargo, no es un método adecuado para su procesamiento informático, ya que no cumple las exigencias básicas de toda búsqueda automatizada. Una de las más evidentes es que no todos los símbolos se empleen de forma unívoca.

Un ejemplo claro de esto son los paréntesis, que, en Briz *et al*. (1995), se empleaban para la transcripción dudosa «((palabra dudosa))», la ininteligible «(( ))», los susurros «º(texto susurrado)º» o la delimitación de fenómenos paralingüísticos «(RISAS)». Otro problema consiste en que los símbolos pueden coincidir con la transcripción, como la duplicación de vocales o consonantes, que puede formar parte tanto del texto transcrito (en palabras como *creen* o *innegable*) como del sistema de transcripción (el fenómeno del alargamiento se representaba, precisamente, mediante dos vocales o consonantes idénticas (*Que yo no lo he heecho*).

A partir de 2019[4], y siguiendo los estándares en el campo, el sistema inicial se actualizó para convertirse en un sistema de base XML, en el que las convenciones de transcripción se asociaron a etiquetas. De esta forma, cada suceso podía ser reconocido de manera unívoca por un buscador.

Asimismo, se aprovechó este cambio de sistema para incluir en la transcripción algunos fenómenos que no contaban con ningún tipo de marca explícita previa, como la fonética sintáctica, o que presentaban ambigüedades de lectura, como los alargamientos:

Fenómeno	**Etiquetas TEI**	**Transcripción tradicional**
Estilo directo	<cita>palabras</cita>	*palabra*
Alargamientos	palabra<al/>	palabraa
Fragmentos ininteligibles	<in/>	(( ))
Fonética sintáctica	<fsr t=”l’almohada”>la almohada</fsr>	l’almohada
Tonema ascendente	<ta/>	↑
Fragmentos entre risas	<e_risas>palabras</e_risas>	Palabras(ENTRE RISAS)

Tabla 2. Ejemplo del sistema de transcripción con etiquetas y su traducción al modelo tradicional.

Esta ventaja, sin embargo, se enfrenta a un inconveniente: la transcripción se sobrecarga con etiquetas que dificultan la lectura frente al sistema tradicional, tal y como ilustran los ejemplos (5a) y (5b):

(5) **Conversación 1994.PT.3**

(5a) **187A88**: [sí<al/>] <e_risas><in/> un pueblerino</e_risas> // [<e_risas>y para decir masovero</e_risas>] no podía decir ni campo ni nada y decía <cita>los que están más <fsr t=”p’allá”>para allá</fsr><ta/></cita> y todo el mundo

[4] Como resultado del proyecto de investigación UDEMADIS (FFI-2016-77841-P).

<cita>ovni estraterrestre</cita> <risas/> <cita>¡no no!</cita> <cita>más <fsr t="p'allá">para allá</cita> pero un poco más <fsr t="p'acá">para acá</fsr> [y mira]

(5b) **187A88**: [síí] (( )) un pueblerino(ENTRE RISAS)// [y para decir masovero(ENTRE RISAS)] no podía decir ni campo ni nada y decía *los que están más p'allá*↑ y todo el mundo *ovni estraterrestre* (RISAS) *¡no no! más p'allá [y mira]*

La versión etiquetada, más completa desde el punto de vista de la transcripción, plantea, sin embargo, un dilema: una transcripción como la jeffersoniana, perfectamente adaptada a su objeto de estudio, no permite un tratamiento informatizado; por su parte, una transcripción de fenómenos conversacionales con el etiquetado XML convierte el texto resultante en algo difícil para la lectura y el análisis del investigador, y todo esto aunque se empleen etiquetas comprensibles como «<ininteligible/>» para fragmentos de audio no comprensibles.

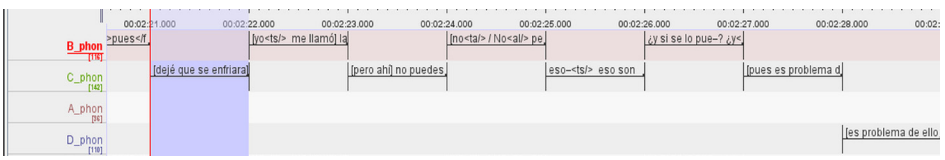
La forma de resolver dicho dilema en el corpus 3.0 ha sido la de mantener ambos tipos de transcripciones, situando cada una en dos partes distintas del corpus: un *backend*, en el que las conversaciones están etiquetadas y sus elementos constitutivos pueden ser recuperados mediante un buscador, y un *frontend* que devuelve al usuario final la conversación transcrita mediante el sistema jeffersoniano original, como se verá en § 4.1. Para lograrlo, a través de la aplicación web del corpus se ha desarrollado un sistema de traducción que vincula cada etiqueta con el símbolo o formato correspondiente en la transcripción tradicional, de manera que en la web del corpus las conversaciones tengan un formato legible, pero sin perder las ventajas informáticas del etiquetado.

No obstante, las novedades de la versión 3.0 no se limitan solo a la combinación de la transcripción tradicional con el sistema TEI/XML; la transcripción ha incorporado herramientas como el programa ELAN, que permite una organización más detallada y flexible de los datos. A continuación, se describe cómo se ha estructurado la hoja de trabajo de ELAN en el corpus 3.0, que, como se verá, presenta una complejidad considerable.

3.2. Una hoja de ELAN compleja

El programa ELAN es, hoy en día, uno de los estándares más utilizados para la transcripción de audio y video. Esta herramienta permite integrar información de diferentes formatos (audio, video y texto) y separar los datos de la transcripción por niveles, lo que permite exportar la información para su cuantificación y posterior procesamiento. Para

ello, este *software* permite la creación de líneas (*tiers*) en las que pueden clasificarse los fenómenos que se quiere estudiar. Al estar asociadas dichas líneas a una marca temporal, se hace posible asociar fenómenos lingüísticos al lugar de la transcripción en que aparezcan y al hablante



que los produce:
Imagen 1. Interfaz de ELAN con anotaciones.

Una transcripción en ELAN utiliza un sistema *en pentagrama* (Vázquez *et al*. 2021), que tiene la forma de un papiro (Pons Bordería 2022) que se despliega de forma horizontal sobre el eje temporal (en la parte superior) y que consta de una línea por cada hablante. Dentro de cada una de las líneas se crean cajas alineadas temporalmente con el audio para cada fragmento de transcripción.

Estas cajas pueden diseñarse según diferentes criterios, pero el estándar comúnmente empleado, tanto en corpus alineados como en tradicionales, ha sido identificar la intervención o turno como la unidad base para la transcripción. Esta elección se fundamenta en la necesidad de segmentar el discurso en unidades que permitan analizar tanto las dinámicas de toma de turno como la estructura interna de la interacción (Briz 2000; Pons Bordería 2022). Además, la intervención permite delimitar cada una de las contribuciones del hablante, aspecto esencial para comprender fenómenos como la organización de los turnos de habla.

Sin embargo, aunque las transcripciones realizadas por el grupo Val.Es.Co. se presentan tradicionalmente a través de la unidad intervención, siempre se ha tenido en cuenta el grupo entonativo, dado su interés en la segmentación del discurso (Pons Bordería 2016). Este constituye la unidad física reconocible, definida por una pausa o un tonema marcado (Cabedo 2009 y 2011) que, a su vez, conforma las intervenciones. En la versión 3.0, gracias al uso del programa ELAN y al sistema de análisis descrito en § 3.3., ha sido posible crear *cajas*[5] que separen los grupos entonativos, manteniendo al mismo tiempo su identificación dentro de las intervenciones que los contienen. En el ejemplo siguiente se muestra la línea de transcripción de C (C_PHON) con cuatro grupos entonativos que están recogidos dentro de una única intervención (ver § 3.3.2.).

[5] En el programa ELAN, una caja es un espacio de trabajo digital vinculado a un fragmento temporal de un audio o video. En ellas, no solo es posible escribir la transcripción de lo que se dice en la grabación, sino también analizar la información en diferentes niveles, como se verá en § 3.3.

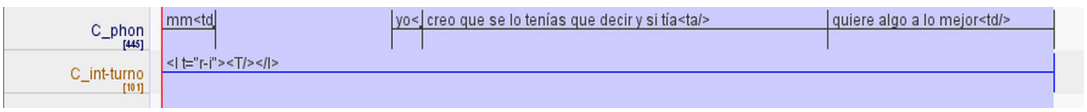


Imagen 2. Grupos entonativos recogidos dentro de una intervención.

La línea de transcripción con cajas que separan cada grupo entonativo se tokeniza en una segunda línea (HABLANTE_palabras). Este proceso utiliza el espacio gráfico como frontera y genera un espacio para cada una de las palabras que conforman un grupo entonativo (imagen 3).

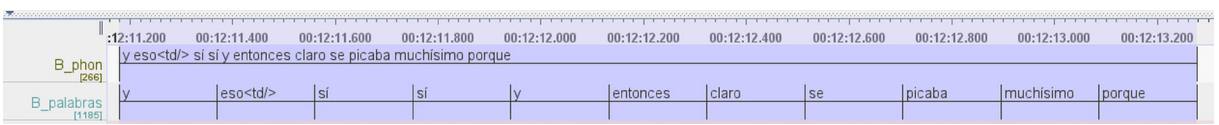


Imagen 3. Fragmento de transcripción en ELAN con tokenización de palarbas - Conversación 011.PT.S4.

De este modo, se crea una caja para cada una de las palabras que componen el grupo entonativo, con sus correspondientes identificadores temporales de inicio y fin[6]. Estos identificadores son básicos, puesto que son los que se emplean para asociar las cajas de las distintas líneas de análisis (§ 3.3.2).

La utilidad de este proceso se aprecia en la imagen 4, en la que los límites temporales del subacto sustantivo directivo (SSD) no se establecen desde la línea _PHON, que incluye todo el grupo entonativo, ya que el subacto solo afecta a una parte de este. Sin embargo, los límites del subacto coinciden exactamente con el inicio de la caja de la primera palabra que compone el subacto (*aquí*) y con el fin de la que lo termina («((aquí))»). Por esta razón, la tokenización, al dividir por palabras, permite una división exacta por unidades discursivas.

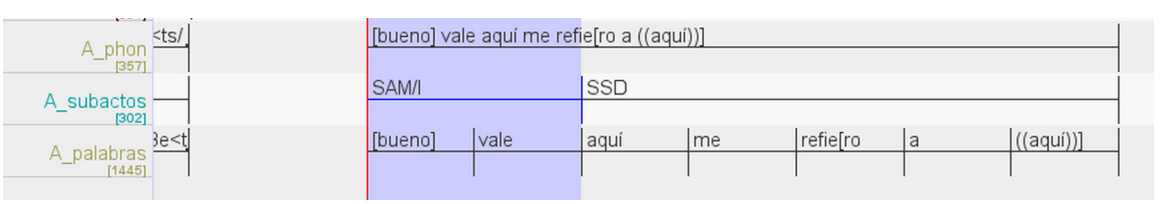


Imagen 4. Ejemplo de segmentación de subactos a través de ELAN – Conversación 2018. PT.S11

[6] Cabe reseñar que las cajas creadas automáticamente para las palabras no se corresponden con el fragmento exacto en el que esta ha sido pronunciada, sino que sus marcas temporales son virtuales dentro del grupo entonativo en el que se han pronunciado. En otras palabras, aunque los identificadores temporales de los tokens creados no sean reales, sí que se mantienen dentro del tiempo del grupo entonativo en el que fueron pronunciadas.

Con estos elementos ya se cuenta con las herramientas necesarias para crear un corpus que no solo permita la búsqueda de concordancias o de determinados fenómenos, sino que también pueda mostrar los fragmentos exactos de audio que se corresponden con la transcripción, tal y como implementan los grupos ESLORA o AMERESCO. El corpus Val.Es.Co. 3.0 introduce además la novedad de consultar en línea todas las conversaciones (§ 4.2.) y, algunas de ellas, completamente segmentadas. En § 3.3 se detallará cómo se han adaptado las funcionalidades del programa para alcanzar este objetivo.

3.3. Introducción del análisis jerárquico de unidades discursivas

La segmentación del discurso propuesta por el grupo Val.Es.Co. se basa en un modelo jerárquico y recursivo (§ 2). *Jerárquico* implica que las unidades menores están incluidas dentro de las superiores. Al mismo tiempo, el análisis es recursivo, lo que significa que un nivel puede estar compuesto por unidades del mismo nivel o de nivel superior. Por ejemplo, los subactos directores, que suelen ser el núcleo de los actos, pueden contener en su interior subactos adyacentes, como se observa en (6), donde un subacto adyacente modal integra un subacto sustantivo director :

(6) **(2011.PT.S2)**
126B76: #{$_{SSD}$ qué bueno {$_{SAM}$ ¿eh? $_{SAM}$} el español coloquial(ENTRE RISAS)$_{SSD}$}# #(RISAS)#

Este sistema teórico se tiene que adaptar al programa ELAN y esto supone adoptar una serie de decisiones que, vistas en conjunto, han supuesto un desarrollo técnico considerable. En ELAN cada línea está asociada a un hablante, lo que crea un esquema del tipo {A, B, C…}, donde {A, B, C…} son los hablantes. En la versión 3.0 se han asociado los distintos niveles de análisis (subacto, acto, intervención, turno, diálogo y discurso) a cada uno de los hablantes, lo que produce un esquema del tipo {A(l_1, l_2, l_3, …l_n), B(l_1, l_2, l_3, …l_n)…}, donde {A, B, C…} son los hablantes y {(l_1, l_2, l_3, …l_n} son las diferentes líneas que se asocian a este, lo que implica, tanto una línea por nivel de análisis, como las líneas PHON_ y PALABRAS_, descritas en (§ 3.2).

Además de por el fenómeno que contienen, las líneas de la transcripción se pueden dividir en función del tipo de contenido que reproduzcan; así, se distinguen líneas de contenido libre y líneas de contenido cerrado. Las primeras se asocian directamente con la transcripción y son las que aparecen en primer lugar en la hoja de ELAN: X_ PHON, X_ PALABRAS y _OBSERVACIONES. Las segundas no permiten la escritura manual, sino la selección del vocabulario programado

dentro de ellas, ya que ELAN permite crear vocabularios controlados que optimizan el proceso de etiquetado.

Con estos vocabularios se reduce en buena medida el riesgo de errores humanos debidos a la escritura incorrecta de las etiquetas. Las líneas de contenido cerrado se han utilizado para el análisis en unidades discursivas: HABLANTE_subactos, HABLANTE_actos, HABLANTE_ intervención-turno, HABLANTE_diálogos y HABLANTE_discurso.

En resumen, el formato de la hoja de ELAN desarrollada para la transcripción de las conversaciones del corpus 3.0 está formado por los siguientes elementos (Pons Bordería 2022: 138-140 [adaptación]):

Tipo de contenido	**Contenido**	**Línea**	**Tipo**	**Vocabulario**
Libre	Transcripción por grupos entonativos del hablante	A/ B/ C... _phon	phon	-
	Palabras del hablante	A/ B/ C... palabras	word-tokenización	-
	Acciones, gestos o información que influencia o transcurre durante la conversación	Observaciones	obs	-
Cerrado	Subacto del hablante	A/ B/ C…_subactos/ subactos_II	subactos	SSD SSS SAM SAI SAT SS/SA
	Actos del hablante	A/ B/ C…_actos / actos_II	actos	Acto
	Intervenciones del hablante	A/ B/ C…_int-turno	Int_turno	<I t="i"><T/></I> <I t="i-r"><T/></I> <I t="r"><T/></I> <I t="r"> <I t="ind">
	Diálogos de la conversación	Diálogos_I Y _II	dialogo	<Di t="n"/>
	Discursos de la conversación	Discurso	discurso	<DSC/>

Tabla 3. Formato de las conversaciones en ELAN para el corpus Val.Es.Co.

Así, cada línea posee unos rasgos distintivos que permiten su posterior procesamiento informático. A continuación, se describirá el funcionamiento de estos dos tipos de líneas.

3.3.1. Líneas de contenido abierto

Las líneas de contenido libre son tres: HABLANTE_phon, HABLANTE_palabras y Observaciones. Se caracterizan por rellenarse de forma libre.

La línea HABLANTE_phon incluye el texto de la transcripción, como se muestra en la imagen 5. Se trata de la más importante del sistema de transcripción y análisis, dado que es la primera en alinearse con el audio y sirve como base para las demás, al contener los referentes temporales reales asociados a cada fragmento de habla. El resto de las líneas, en cambio, se alinean sobre las cajas de las palabras que corresponden a cada grupo entonativo (§ 3.2.), de manera que cada elemento de las líneas de análisis se vincula al contenido de la transcripción y no al audio de la conversación.

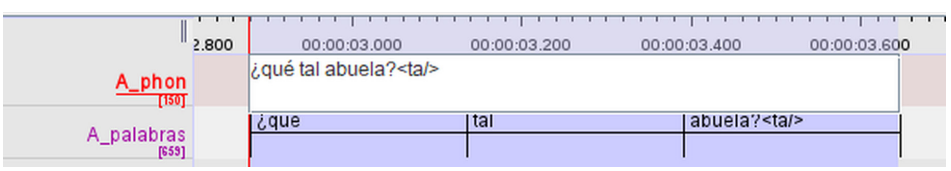


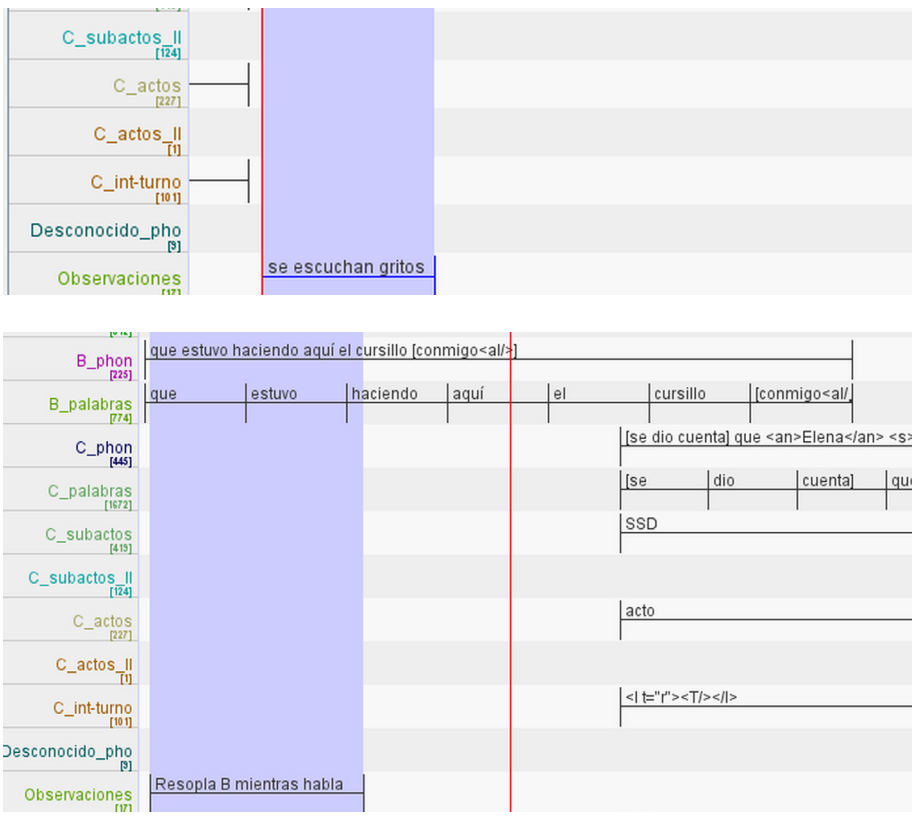
Imagen 5. Forma de transcripción del audio a través de ELAN

A pesar de que los fenómenos prosódicos o discursivos recurrentes en la conversación cuentan con un sistema de etiquetas estable, estas se introducen manualmente en este campo, ya que se trata de fenómenos que afectan directamente al contenido de la grabación y, por lo tanto, se codifican en el texto. En otras palabras, no se ha definido un vocabulario para estas etiquetas en ELAN porque se emplean dentro del contenido de la caja de transcripción asociadas a las palabras que afectan.

A partir de HABLANTE_phon se lleva a cabo el proceso de tokenización en la línea HABLANTE_palabras. Gracias a este proceso, el análisis realizado en las líneas de contenido cerrado puede hacerse palabra por palabra (§ 3.3.2). Por ejemplo, para vincular la información de la línea «subactos» al segmento correspondiente, se asocia el punto de inicio de la primera palabra con el punto final, como se ha visto en la imagen 4 del apartado anterior.

Por último, dentro de las líneas de contenido abierto, se encuentra (_Observaciones), destinada a incluir comentarios sobre el contexto o

las acciones de los participantes[7]. Esta línea puede alinearse de forma libre o en relación con _phon, ya que su contenido puede estar relacionado o no con la intervención de un hablante (imágenes 6 y 7, respectivamente). Cabe señalar que, aunque dicha información se utiliza para facilitar una mejor comprensión de la situación, no se emplea directamente en el análisis.



Imágenes 6 y 7. Transcripción de observaciones a través de ELAN

3.3.2. Líneas de contenido cerrado

Las líneas de contenido cerrado están vinculadas a las unidades de segmentación definidas por el modelo Val.Es.Co. (2014, Pons Bordería 2022). Cada línea cuenta con un vocabulario predefinido (ver tabla 3), que permite seleccionar directamente la unidad correspondiente sin necesidad de escribirla manualmente. La jerarquía estricta del sistema de unidades se reproduce en la página de ELAN, donde cada línea representa un nivel de información que abarca, desde la unidad estructural mínima (los subactos) hasta la unidad dialógica máxima (los discursos). Este diseño asegura la coherencia del modelo, de modo que

[7] Es la traducción, en la versión 3.0, de las notas a pie de página de las versiones en papel del corpus.

un subacto no puede abarcar dos actos, ni un acto puede distribuirse entre dos intervenciones, como se muestra en la imagen 8.

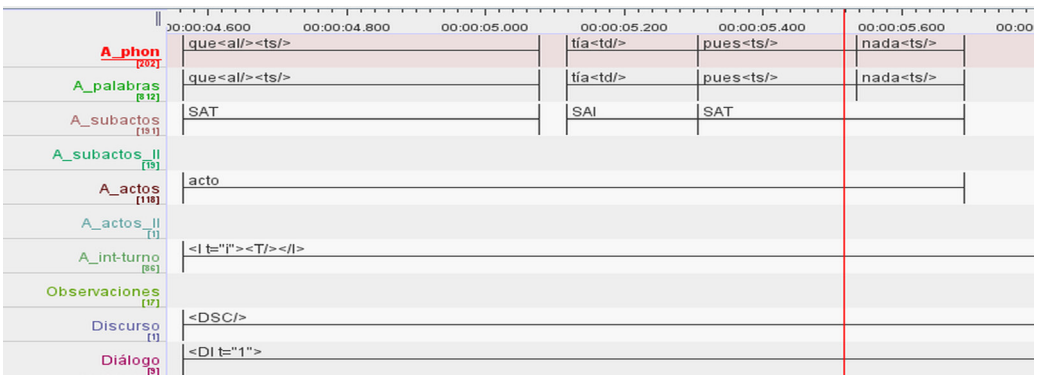


Imagen 8. Muestra de segmentación de unidades – Conversación 1994.PT.S1.

Esta estructura, por así decirlo, vertical de las líneas de análisis, se completa con una doble distinción basada en sus características: según el número de participantes a los que afectan las líneas de análisis y según el tipo de vocabulario implementado en ellas.

La primera clasificación se basa en la distinción entre unidades monologales y dialogales del modelo Val.Es.Co. Las unidades monologales (subactos, actos e intervenciones) están asociadas a un único participante, ya que se derivan exclusivamente de las emisiones de un solo hablante. Por lo tanto, cada una de estas unidades se repetirá tantas veces como participantes haya en la conversación, al igual que ocurre con las líneas de transcripción y tokenización, tal como se ilustra en la imagen 9 con los hablantes A y B:

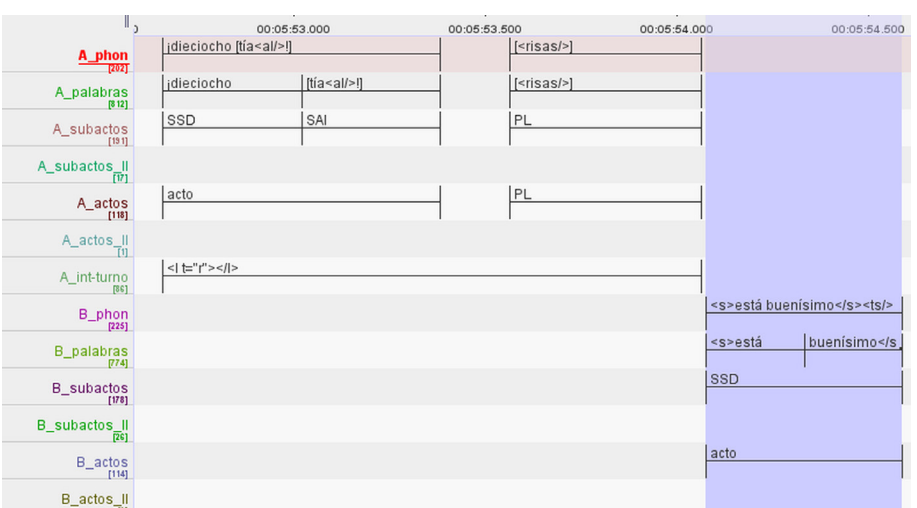


Imagen 9. Interfaz de ELAN con líneas monologales para A y B. Extraído de 1992.PT.S1.

Asimismo, los subactos y los actos presentan recursividad, de modo que un acto puede estar dentro de otro acto, o un subacto dentro de otro subacto. Así, estas líneas se pueden duplicar añadiendo el sufijo _II, lo que permite incorporar una línea más de análisis dentro de la categorización general de una estructura, como se observa en la imagen 10:

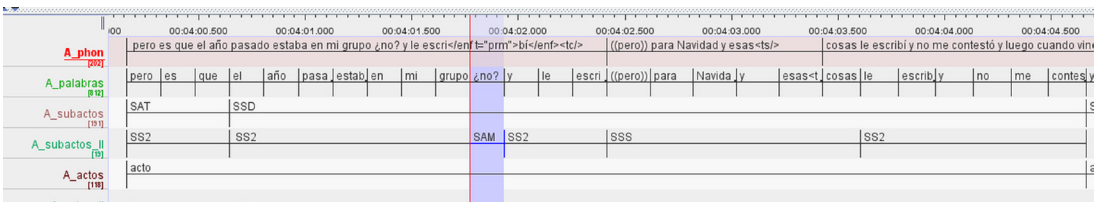
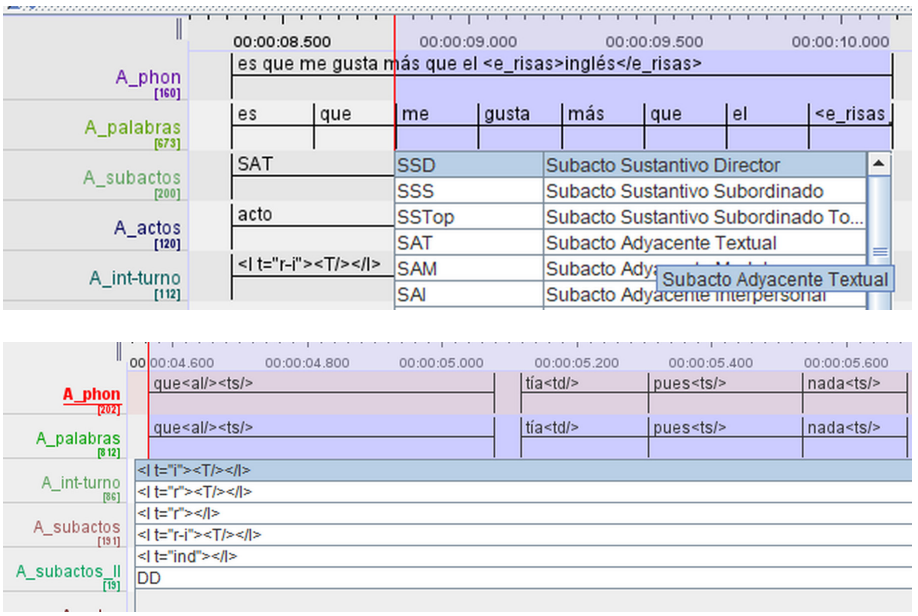


Imagen 10. Interfaz de ELAN con análisis de segundo nivel para los subactos de A.

En la línea HABLANTE_SUBACTOS_II de la Figura 10 se puede observar una etiqueta denominada SS2. Esta etiqueta se emplea para satisfacer una exigencia inherente a este sistema de análisis: garantizar que todas las líneas de análisis monologales estén completamente rellenas. En otras palabras, son etiquetas de carácter exclusivamente técnico, sin relación directa con el análisis lingüístico.

La segunda clasificación necesaria para describir las líneas se basa en el tipo de vocabulario que conforman las líneas de contenido cerrado. Según este criterio, se pueden distinguir tres tipos de líneas. En primer lugar, aquellas que incluyen una tipología, como sucede con los subactos y las intervenciones, ya que estas unidades se subdividen en varios tipos. A cada una de ellas le corresponde una etiqueta que permite su identificación en ELAN y se muestra en un desplegable (imágenes 11, 12 y 13, siguiente página).

El segundo grupo lo conforman las unidades sin tipología, como el acto y el discurso. Estas cuentan solo con un tipo de etiqueta que se corresponde con la misma unidad, como se ve en la imagen 13 (siguiente página):



Imágenes 11 y 12. Interfaz de ELAN con el desplegable del vocabulario de la línea de subactos y de intervenciones.

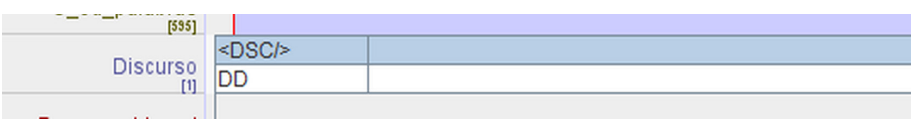


Imagen 13. Interfaz de ELAN con el desplegable del vocabulario de la línea de discurso

Por último, aunque las líneas de diálogos no tienen una tipología específica, se numeran desde el análisis debido a un fenómeno que sucede en conversaciones con más de dos participantes: es posible que un diálogo no haya finalizado y que, antes de su conclusión o de manera simultánea, comience otro. Para dar cuenta de esta situación, se introduce una segunda línea de análisis para los diálogos (_Diálogo II). De este modo, puede indicarse el inicio de un nuevo diálogo antes de que haya finalizado el anterior, como se observa en la imagen 14:

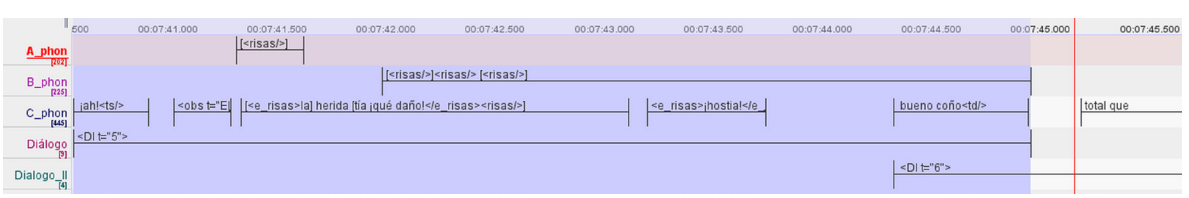


Imagen 14. Interfaz de ELAN con dos diálogos solapados.

Esta configuración en cajas y en líneas de vocabulario cerrado resulta muy útil para el investigador, ya que solo debe preocuparse por

crear y ajustar las cajas alineadas de acuerdo con las palabras a las que afecte cada unidad. Una vez creada la caja, el programa no permite la escritura directa como en las líneas de contenido libre, sino que ofrece un desplegable con las opciones disponibles, de entre las que el segmentador solo debe elegir la etiqueta correcta. Esta cuestión es de vital importancia para la construcción del corpus: un error de escritura puede provocar que la base de datos del corpus genere un fallo de lectura, lo que derivaría en un procesamiento de la información incorrecto y, consecuentemente, en una visualización errónea en el corpus.

Por último, toda la estructura de unidades descrita hasta el momento se replica para el estilo directo[8], que se analiza como un nuevo discurso. Para ello, se traslada el contenido de los fragmentos de estilo directo a una nueva línea llamada «HABLANTE_ed» y se duplica el resto de líneas con la marca «ed»:

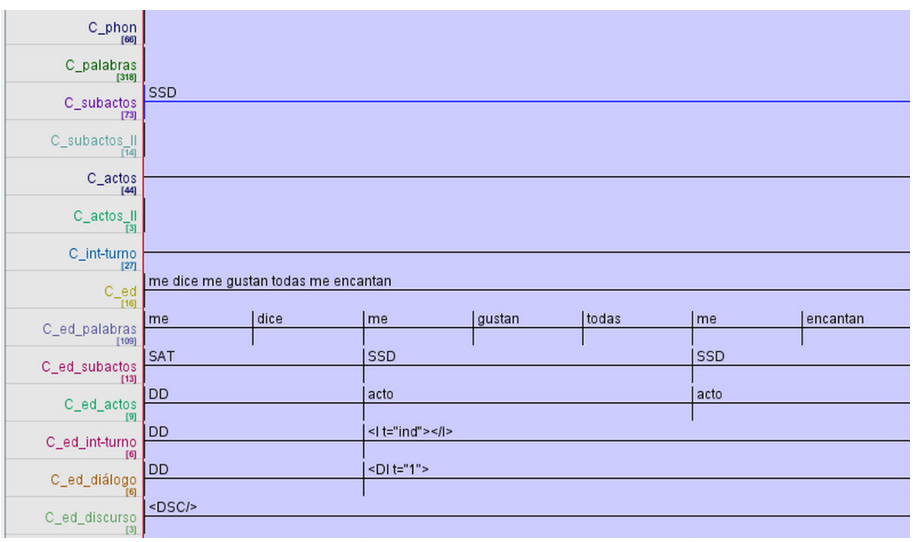


Imagen 15. Fragmento de transcripción de estilo directo – conversación 2016.PT.(20).S6.

Como resultado, el análisis del contenido de la transcripción en fragmentos de estilo directo se aborda en un doble nivel. Con respecto al marco que los contiene en el estilo principal, estos fragmentos suelen categorizarse como un SSD, ya que narran algo dicho por otro hablante en un momento previo, es decir, se trata de contenido puramente conceptual. Por otro, debido a que implican un desplazamiento de los ejes deícticos de la conversación (*yo*, *aquí*, *ahora*), se analizan de manera específica como un nuevo discurso, compuesto por diálogos, intervenciones, actos y subactos. Este enfoque permite reflejar la dualidad y complejidad inherente al discurso reportado.

[8] El análisis del estilo directo siguiendo a Benavent (2024) es más complejo y desborda los límites del presente trabajo. Al igual que con el estilo principal, solo se ha explicado su parte técnica.

En conclusión, si se consideran todas las líneas que se han presentado en este apartado, una conversación completamente segmentada incluye 7 líneas por hablante para el estilo principal (2 de contenido libre y 5 de contenido cerrado), 7 para el estilo directo y 4 líneas grupales. Esto da como resultado que la plantilla de cada grabación cuente con un mínimo de 32 líneas, asumiendo que cualquier conversación involucra, al menos, dos hablantes[9].

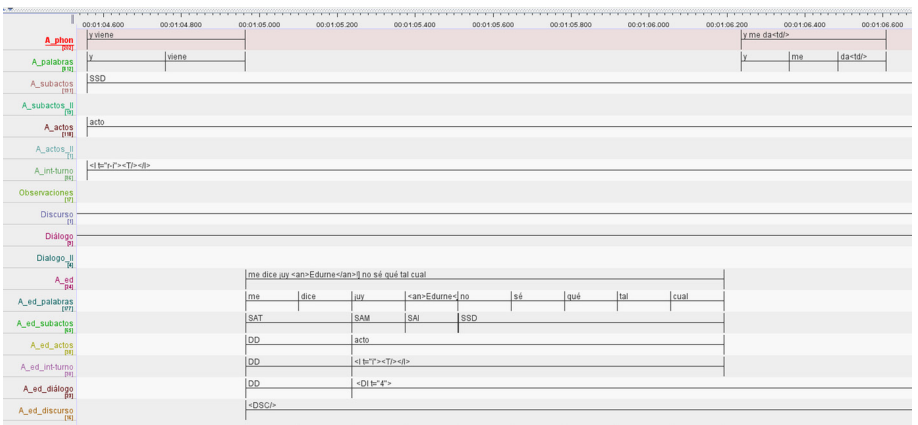


Imagen 16. Fragmento de transcripción con todas las líneas para un hablante – conversación 1994.PT.S1.

Esta sección ha explicado los aspectos fundamentales desarrollados en el corpus Val.Es.Co. 3.0 para crear una transcripción segmentada desde el punto de vista técnico. Estos aspectos incluyen, desde la estructura básica del sistema de transcripción, hasta la organización inicial en ELAN. Se trata de una metodología ciertamente laboriosa que ha permitido reflejar en el corpus un análisis tan complejo como la segmentación del discurso oral. En el siguiente apartado, se expondrá cómo todo este trabajo de *backend* se muestra a los usuarios a través de la web del corpus.

4. El corpus Val.Es.Co. 3.0 en la web: el *frontend*

La sección 3 ha detallado cómo el *backend* del corpus Val.Es.Co. 3.0 se ha adaptado al objeto de estudio, lo que ha implicado importantes adaptaciones en todos los aspectos de la transcripción. Sin embargo, las etiquetas y marcas de este sistema no pueden mostrarse de manera literal a los usuarios, ya que harían imposible el estudio de las transcripciones y exigirían un conocimiento técnico específico del formato

[9] De acuerdo con el carácter flexible de toda conversación, se podría ampliar el número de líneas si se quisiera incluir, por ejemplo, los gestos o el paralenguaje. El análisis de los gestos en las conversaciones coloquiales (Cabanes 2023) añade siete líneas adicionales:

de análisis en ELAN, lo que reduciría enormemente su utilidad. En este apartado, se explicará cómo este trabajo técnico se ha adaptado para ofrecer una visualización funcional y accesible en la web que permita a los usuarios consultar las transcripciones y explorar los datos segmentados sin enfrentarse a la complejidad del *backend*, optimizando así la experiencia de consulta y análisis del corpus Val.Es.Co. 3.0.

4.1. La visualización de la transcripción

La transcripción constituye el punto de partida del corpus Val.Es.Co. 3.0, ya que sobre ella se construye el resto de las líneas de análisis y segmentación detalladas en el apartado anterior. Sin embargo, el formato técnico empleado para la transcripción no resulta adecuado para su consulta directa: por un lado, la visión horizontal de la conversación no resulta amigable para la lectura y, por otro, el etiquetado de fenómeno vuelve la lectura de las transcripciones casi incomprensible.

Por esta razón, cuando una conversación se sube a la aplicación web, esta sufre tres transformaciones: primero, toda la transcripción, agrupada en grupos entonativos, se reestructura en sus correspondientes intervenciones, gracias a su análisis en el *backend*[10]. Luego, estas intervenciones se ordenan verticalmente de acuerdo con sus índices temporales, un formato mucho más familiar para la lectura. Por último, la aplicación traduce las etiquetas XML al sistema de transcripción jeffersoniano que, como se ha mencionado anteriormente, resulta más legible. Como consecuencia de estas transformaciones, el usuario accede a la conversación de la imagen 17 en el formato de la imagen 18 (siguiente página).

La visualización de la transcripción en el corpus Val.Es.Co. 3.0 responde a la necesidad de adaptar el formato técnico original a un entorno más accesible y familiar para los usuarios. Sin embargo, este corpus, además de permitir el acceso a las conversaciones mediante la búsqueda de concordancias, también hace posible la consulta en línea de las transcripciones completas.

[10] Esta característica obliga a que, como mínimo, todas las conversaciones del corpus estén segmentadas en intervenciones.

Imagen 17. Interfaz de ELAN para la conversación 1989.PT.56(1).

213C87	1 - Ir	al principio (RISAS)
214A103	2 - Ii	en fin mañana a las ocho te levantarás tu madre ↑ te va a levantar a las ocho- a las siete
215C88	3 - Iri	mi madre se aburre y se levanta a las siete [(RISAS)] todo el mundo lo sabe(ENTRE RISAS)
B58	4 - Ir	[(RISAS)]

Imagen 18. Muestra de la conversación 1994.PT.3 mediante la visualización de la búsqueda.

4.2. Los modos del corpus: de la búsqueda en conversaciones a su consulta

En la actualidad, la mayoría de los corpus lingüísticos orales incorporan un buscador que permite localizar palabras o fenómenos en los documentos que contienen, ya que el usuario consulta las transcripciones para encontrar elementos concretos, ya sean palabras, estructuras lingüísticas o fenómenos discursivos (Rojo 2024).

Al igual que otros corpus, Val.Es.Co. 3.0 cuenta con un sistema de búsqueda por concordancias que incorpora cuatro tipos de filtros, organizados según distintos criterios: la consideración lingüística del término de búsqueda, los metadatos de las conversaciones, las características prosódicas y las unidades discursivas. El primer tipo de filtro permite definir cómo debe interpretarse el término introducido, especificando aspectos como forma o lema, la distinción entre mayúsculas y minúsculas, la inclusión o no de acentos, y si el fragmento está precedido o seguido de signos como interrogaciones, exclamaciones o pausas. Además, permite delimitar la búsqueda por la categoría gramatical a la que pertenece el término[11] (imagen 19).

[11] Esta clasificación es posible gracias a la implementación del etiquetador morfológico XIADA (Centro Ramón Piñeiro para la investigación en humanidades), desarrollado por Mario Barcala y el corpus ESLORA.

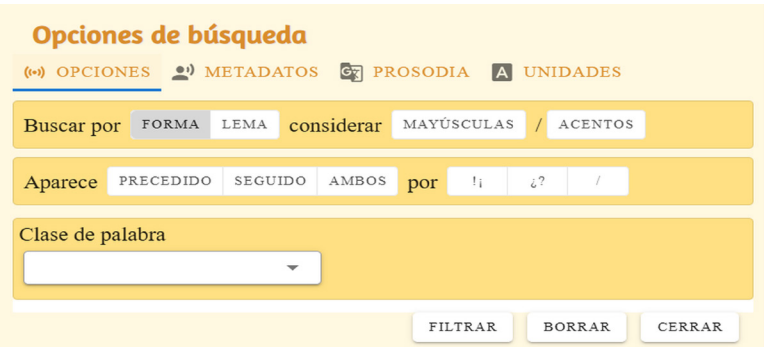
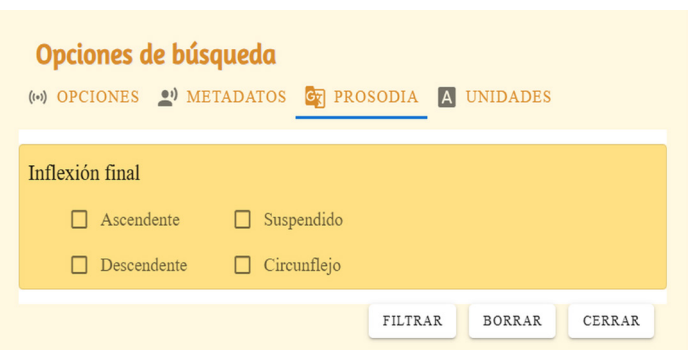


Imagen 19. Interfaz de filtros de búsqueda - opciones generales.

El segundo tipo de filtro se refiere a los datos asociados a las conversaciones y permite seleccionar dos categorías de información: los aspectos relacionados con la identificación general de la conversación, como el año de registro o si esta es considerada prototípica, y los datos sociolingüísticos de los hablantes, como su género, nivel educativo o grupo etario (imagen 20)[12]. Por su parte, el filtro Prosodia permite buscar fragmentos de transcripción que finalicen con un determinado tonema (imagen 21).



Imágenes 20 y 21. Interfaz de búsqueda – filtro de metadatos y de prosodia.

El último filtro está relacionado con la segmentación de unidades. Gracias al análisis del *backend*, realizado durante el proceso de segmentación en ELAN (§ 3.3), el sistema de búsqueda permite al usuario seleccionar y filtrar por el tipo específico de subacto o intervención que desee consultar. Por ejemplo, es posible buscar únicamente subactos sustantivos directores (SSD) o intervenciones de tipo independiente, como se muestra en la imagen 22 (siguiente página):

[12] Esta es la traducción de la tradicional «ficha técnica», tal y como se presentaba en las versiones previas en papel.

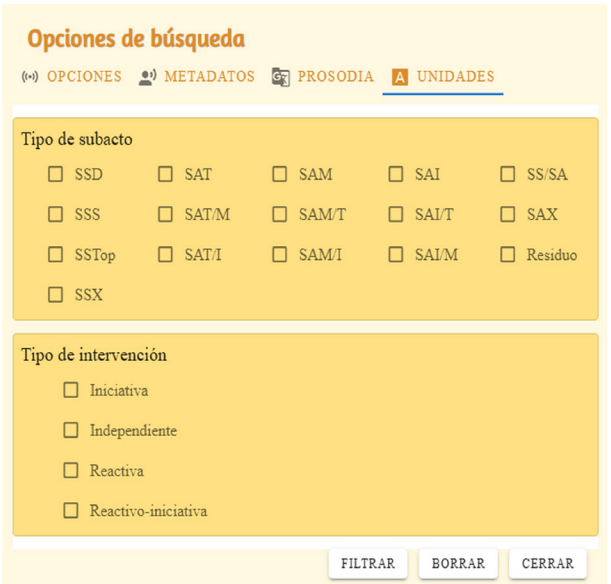


Imagen 22. Interfaz de búsqueda - filtro de unidades.

Los resultados obtenidos en el buscador se presentan en una tabla de concordancias que muestra la intervención completa a la que pertenece cada resultado. Cada ejemplo incluye la opción de reproducir el fragmento de audio correspondiente y de ampliar el contexto hasta en cinco intervenciones, previas o posteriores. Esta información es exportable en formato .docx o .xml para facilitar su análisis o uso externo. Además, las conversaciones segmentadas, identificadas con el marcador «S», permiten la exportación adicional con el nivel de análisis que el usuario seleccione.

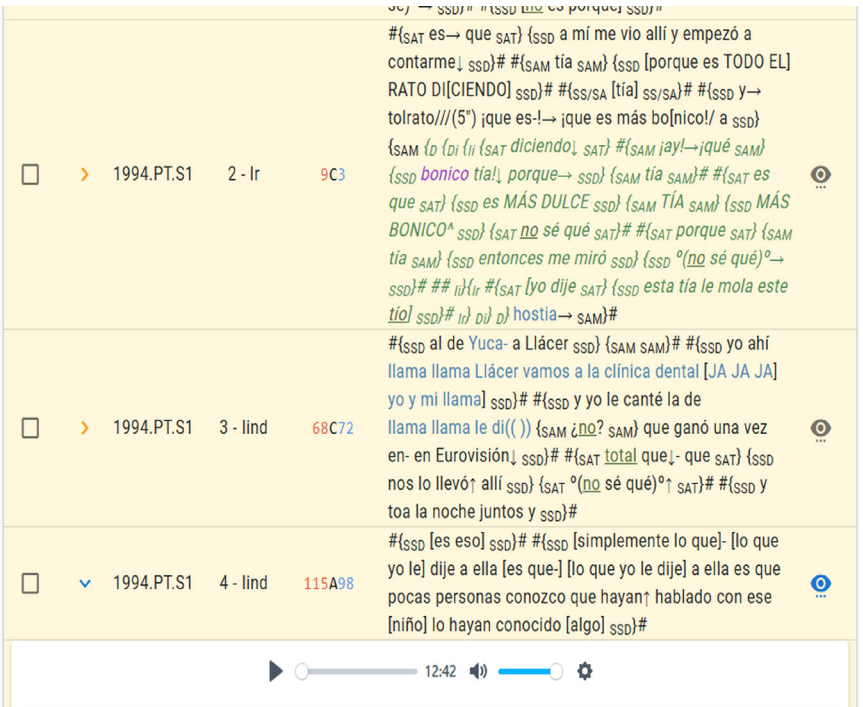


Imagen 23. Interfaz de búsqueda – resultado de concordancias y ampliación de contexto con análisis de subactos.

Desde mediados de 2024, se ha incluido en el corpus una modalidad de búsqueda cronológica que, utilizando las funciones del buscador previamente descrito, posibilita realizar búsquedas dobles dentro de un rango específico de años. Esta búsqueda responde al carácter microdiacrónico del corpus Val.Es.Co., que, con casi treinta años de existencia, contiene datos correspondientes a dos generaciones de hablantes. Siguiendo con los objetivos de los proyectos de investigación DIAXX y DIA20, citados al inicio de este artículo, el corpus Val.Es.Co. 3.0 ha transformado su estructura para convertirse en el primer corpus diacrónico oral del dominio hispánico.

Gracias a esta doble búsqueda, los resultados se pueden presentar en paralelo, facilitando de este modo su comparación, tal y como se muestra en la imagen 24:

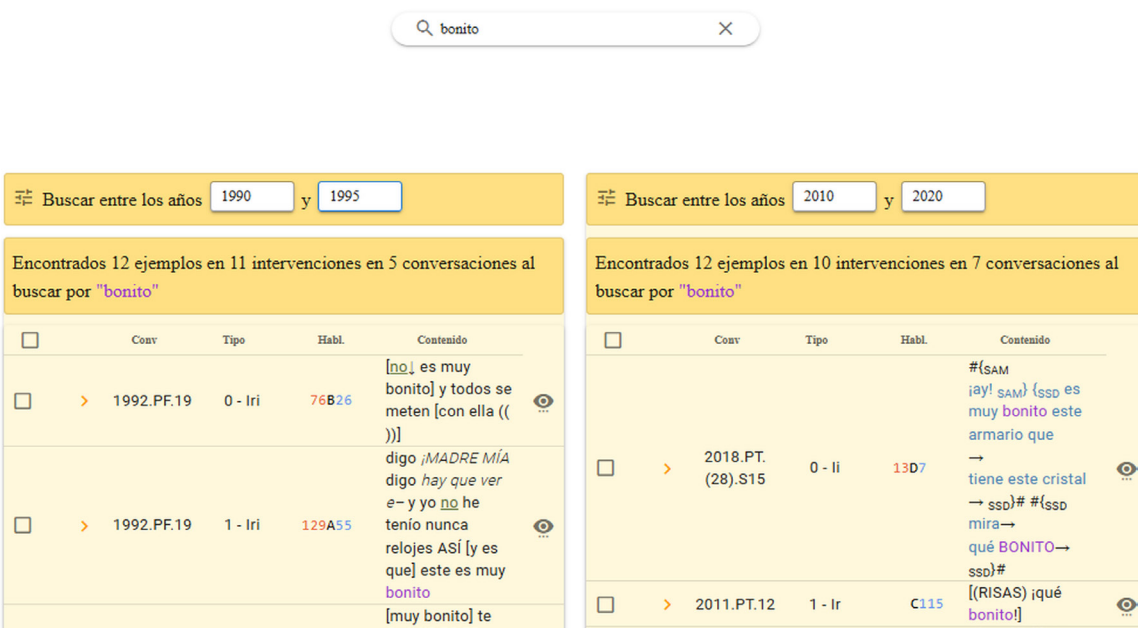


Imagen 24. Resultados de la búsqueda cronológica

Además de la búsqueda, el corpus también ofrece la posibilidad de consultar las conversaciones completas en línea mediante el modo REPRODUCCIÓN AVANZADA. Este sistema no solo presenta el contenido de cada transcripción en formato vertical (§ 4.1), sino que permite seguir el desarrollo de la grabación en tiempo real en modo *karaoke*: a medida que el audio avanza, la intervención correspondiente se resalta, como se muestra en la imagen 25 con la intervención 5A3. Asimismo, al igual que en el modo búsqueda, la reproducción avanzada permite visualizar distintos niveles de análisis de segmentación, siempre que dicha información esté disponible en la conversación, como se ilustra en 26.

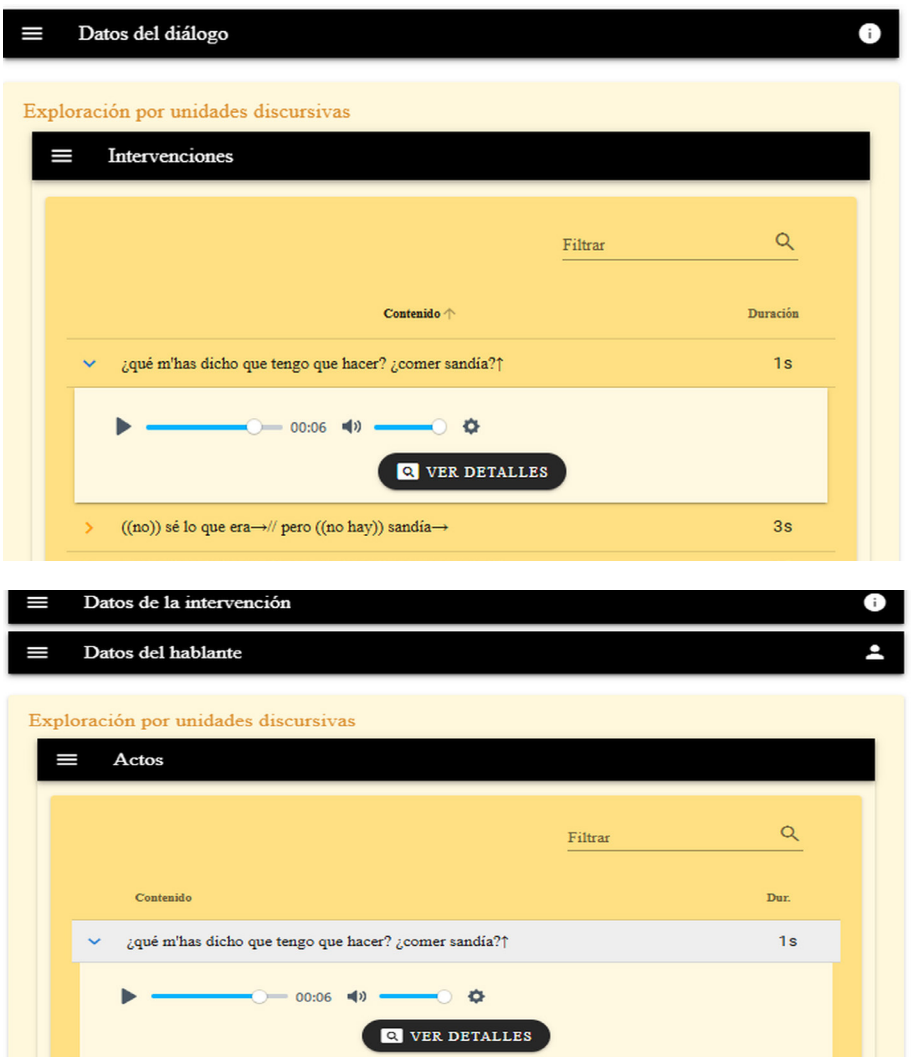
Imágenes 25 y 26. Interfaz de reproducción avanzada – modo karaoke y transcripción segmentación de subactos

Esta funcionalidad, junto con la posibilidad de visualizar distintos niveles de análisis, facilita una exploración detallada y dinámica del corpus, de modo que el usuario pueda acceder casi de manera completa a la totalidad del trabajo del grupo de investigación Val.Es.Co.

Por último, cada conversación ofrece una opción de VER DETALLES que no solo proporciona los datos identificativos de la conversación y de los hablantes, sino que también permite acceder a la transcripción a través de las distintas unidades discursivas que se han segmentado (ver § 3.3.2.). Gracias a este trabajo y al proceso de lectura que hace la aplicación web, es posible consultarla desde los discursos, diálogos, intervenciones, actos y subactos que la conforman, manteniendo su correspondencia con el fragmento de audio asociado. Además, cada unidad incluye información sobre su duración, el tiempo de inicio y fin

dentro del audio, y, en el caso de las unidades monologales, los datos del hablante que las ha pronunciado.

Cabe señalar que esta consulta respeta la estructura jerárquica de las unidades descrita en el apartado §3.3.2. Así, al explorar un diálogo, solo es posible acceder a sus intervenciones; al consultar una intervención, únicamente se pueden visualizar sus actos, y así sucesivamente, tal como se ilustra en las imágenes 27 y 28.



Imágenes 27 y 28. Consulta de la conversación a partir de las unidades diálogos e intervención.

5. Conclusión

En conclusión, el propósito de este artículo ha sido mostrar los objetivos del corpus Val.Es.Co. 3.0, las decisiones técnicas necesarias para su creación y el resultado final, con el fin de ofrecer una guía para el desarrollo de nuevos corpus en línea que quieran ir más allá de la transcripción. Con ello, se ha tratado de destacar la importancia de integrar técnicas de transcripción con análisis cualitativos, como la segmentación del discurso. El resumen del trabajo expuesto en el artículo puede sintetizarse en la siguiente tabla:

<table>
<tr><th>Fenómeno</th><th>Tratamiento de datos</th><th>Codificación informática</th><th>Visualización web del corpus</th></tr>
<tr><td>Contenido</td><td rowspan="4">Transcripción (líneas de contenido libre en ELAN)</td><td colspan="2" rowspan="2">Ortotipográfica</td></tr>
<tr><td>Paralenguaje y sucesos</td></tr>
<tr><td>Prosodia</td><td rowspan="3">Etiquetas XML</td><td rowspan="2">Símbolos (sistema de base jeffersoniana)</td></tr>
<tr><td>Fenómenos discursivos</td></tr>
<tr><td>Análisis de contenido</td><td rowspan="2">Segmentación (líneas con vocabulario cerrado en ELAN)</td><td rowspan="2">Símbolos ortotipográficos</td></tr>
<tr><td>Análisis de interacción</td></tr>
</table>

Tabla 4. Resumen metodología de trabajo y visualización en la web.

Creemos que estas decisiones abordan (y, en algunos casos, resuelven) problemas a los que se pueden enfrentar quienes deseen adentrarse en el complejo, duro y laborioso mundo de la creación de corpus, que es un campo de creación colectiva en el que la lingüística española brilla por méritos propios y que la sitúa muy por delante de cualquier otra lengua románica. Este artículo se propone como una invitación para utilizar estos hallazgos en futuras investigaciones que sigan desarrollando nuevos métodos de transcripción, análisis y trabajo para capturar y, en última instancia, comprender ese inasible apoyado en el tiempo que es el lenguaje oral coloquial.

Bibliografía

Albelda, Marta, y Maria Estellés (dirs.), *Corpus Ameresco*. Disponible en: https://corpusameresco.com. [Fecha de consulta: 8 de septiembre de 2024].

Alcaraz Martínez, Rubén, y Elisabet Vázquez Puig (2016), «TEI: un estándar para codificar textos en el ámbito de las humanidades digitales», *BiD: Textos Universitaris de Biblioteconomia i Documentació*, 37: s.p. DOI: 10.1344/BiD2016.37.24.

Bolaños Cuéllar, Sergio (2015), «La lingüística de corpus: perspectivas para la investigación lingüística contemporánea», *Forma y Función*, 28 (1): 31-54. DOI: 10.15446/fyf.v28n1.51970.

Briz, Antonio (1996), *El español coloquial: situación y uso*, Barcelona, Ariel.

Briz Antonio. (2010), «Lo coloquial y lo formal, el eje de la variedad lingüística», en Castañer, R. M. y Lagüéns, V. (eds.): *«De moneda nunca usada»: Estudios dedicados a José Ma Enguita Utrilla*, Zaragoza, Instituto Fernando El Católico: 125-133.

Briz, Antonio *et al*. (1995), *La conversación coloquial: materiales para su estudio*, Valencia, Universitat de València.

Briz, Antonio y Carcelén, A. (2019): «El futuro iberoamericano del español: la investigación del español oral y en español», en Richard Bueno Hudson (dir.), El español en el mundo: anuario del Instituto Cervantes 2019, Madrid, Bala Perdida/Instituto Cervantes: 189-217.

Brun, Rircardo Eíto (2005). «XML y la gestión de contenidos», *Hipertext. net: Revista Académica sobre Documentación Digital y Comunicación Interactiva*, 3: s.p.

Cabedo Nebot, Adrián (2011). «El reajuste tonal en la delimitación de grupos entonativos», en Antonio Hidalgo Navarro, Yolanda Congosto Martín y Mercedes Quilis Merín (eds.), *El estudio de la prosodia en España en el siglo XXI: Perspectivas y ámbitos*, Valencia, Universitat de València, 209-222.

Cabanes Pérez, Sandra (2023), *Análisis multimodal en la distinción entre intervención y turno: efectos en la segmentación de la conversación desde el modelo Val.Es.Co.*, tesis doctoral, Universitat de València.

Cestero Mancera, Ana M.ª (2014), «Comunicación no verbal y comunicación eficaz», *ELUA*, 28: 125-150.

CORPES = Real Academia Española, *Corpus del Español del Siglo XXI*. Disponible en: https://www.rae.es/corpes. [Fecha de consulta: 8 de septiembre de 2024].

Criado de Val, Manuel (1964), *Fisonomía del español y de las lenguas modernas*, Madrid, Aguilar.

Del Rio Riande, Gimena, y Susanna Allés-Torrent (2023). «Treinta años de TEI en español: usos y comunidad». *Journal of the Text Encoding Initiative*, 16: 1-8.

ESLORA = *Corpus para el estudio del español oral*, versión 2.3. Disponible en: <http://eslora.usc.es>. [Fecha de consulta: octubre de 2024].

García-Miguel, José M. (2022), «Lingüística de corpus», *Estudios de Lingüística del Español*, 45: 11-42.

Garfinkel, Harold (1967), *Studies in ethnomethodology*, Englewood Cliffs, Prentice-Hall.

Jefferson, Gail (2004), «Glossary of transcript symbols with an introduction», en Gene Lerner (ed.), *Conversation analysis: studies from the first generation,* Amsterdam (Phil.), John Benjamin: 13-31. DOI: 10.1075/pbns.125.02jef.

Llamazares, Milka Villayandre (2008), «Lingüística con corpus (I)», *Estudios Humanísticos. Filología*, 30: 329-349. DOI: 10.18002/ehf.v0i30.2847.

Lope Blanch., Juan M. (1971), «El léxico de la zona maya en el marco de la dialectología mexicana», *Nueva Revista de Filología Hispánica*, 20 (1): 1-63. DOI: 10.24201/nrfh.v20i1.1557.

Lope Blanch, Juan M. (1976), «Algunos casos de polimorfismo fonético en México», *Revista de Dialectología y Tradiciones Populares*, 32 (1): 247-262.

Lope Blanch, Juan M. (1986), E*l estudio del español hablado culto: historia de un proyecto*, Ciudad de México, Universidad Nacional Autónoma de México.

Marcos Marín, Francisco (dir.), *Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)*. Disponible en: https://cvc.cervantes.es/lengua/corlec.htm. [Fecha de consulta: 8 de septiembre de 2024].

Narbona, Antonio (1989), *Sintaxis española: nuevos y viejos enfoques*, Barcelona, Ariel.

O'Keefe, Daniel J. (1979), «Ethnomethodology», *Journal for the Theory of Social Behaviour*, 9 (2): 187-219.

Pons Bordería, Salvador (dir.), *Corpus Val.Es.Co*. Disponible en: http://www.valesco.es. [Fecha de consulta: 8 de septiembre de 2024].

Pons Bordería, Salvador (ed.) (2014): *Discourse segmentation in Romance languages*. Amsterdam (Phil.), John Benjamins.

Pons Bordería, Salvador (2016). «Cómo dividir una conversación en actos y subactos», en Antonio Miguel Bañón *et al.* (eds.), *Oralidad*

y análisis del discurso: homenaje a Luis Cortés Rodríguez, Almería, Universidad de Almería, 545-566.

Pons Bordería, Salvador (2022), *Creación y análisis de corpus orales: saberes prácticos y reflexiones teóricas*, Berna, Peter Lang.

Poyatos, Fernando (1994), *La comunicación no verbal*, Madrid, Istmo.

Poyatos, Fernando (2018), *Advances in non-berbal communication*, Amsterdam (Phil.), John Benjamins.

PRESEEA = *Proyecto para el estudio sociolingüístico del español de España y América*. Disponible en: https://preseea.linguas.net. [Fecha de consulta: 8 de septiembre de 2024].

Rojo, Guillermo (2016), «Los corpus textuales del español», *Enciclopedia lingüística hispánica*, 2: 285-296. DOI: 10.4324/9781315792942.

Rojo, Guillermo (2024). «El futuro de los corpus de referencia», *Studia Linguistica Romanica*, 12: 18-33.

Roulet, Eddy, Laurent Fillietaz, y Anne Grobet (2002), «Un modèle et un instrument d'analyse de l'organisation du discours», en Patrick Charaudeau y Dominique Maingueneau (eds.), D*ictionnaire d'analyse du discours*, París, Seuil.

Roulet, Eddy, *et al*. (1981), *L'articulation du discours en français contemporain*, Berna, Peter Lang.

Sacks, Harvey, Emanuel A. Schegloff, y Gail Jefferson (1974), «A simplest systematics for the organization of turn-taking for conversation», *Language*, 50 (4): 696-735.

Sacks, Harvey, y Gail Jefferson (2000), «Convenciones de transcripción», en Teun A. Van Dijk (comp.), *El discurso como estructura y proceso. Estudios del discurso: introducción multidisciplinaria*, Barcelona, Gedisa: 442-444.

Torruella, Joan, y Joaquim Llisterri (1999), «Diseño de corpus textuales y orales», en José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella (eds.): *Filología e informática: nuevas tecnologías en los estudios filológicos*, Barcelona, Milenio/Universidad Autónoma de Barcelona: 45-77.

Val.Es.Co. (2014), «Las unidades del discurso oral: la propuesta Val. Es.Co. de segmentación de la conversación (coloquial)», *Estudios de Lingüística del Español*, 35: 11-71.

Vázquez Rozas, Victoria, *et al*. (2020), «Codificación y anotación del habla en un contexto bilingüe: el corpus ESLORA de español de

Galicia» en Ángel Gallego y Francesc Roca (eds.), *Dialectología Digital del Español*, Santiago de Compostela, Universidade de Santiago de Compostela, 189-224.

Venegas, Rene, Iris Viviana Bosio, y Constanza Ceda-Canales (2022), «Los corpus sincrónicos del español: descripción y potencialidades para la investigación teórica y aplicada de la lengua», *Revista de Lexicografía y Lingüística Aplicada*, 22 (3): 45-67.

Zimmerman, Don H. (1978), «Ethnomethodology», *The American Sociologist*, 13 (1), 6-15.