UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **accepted version** of the book part:

Llisterri, Joaquim. «Corpus para investigar sobre el componente fónico en español como LE/L2». A: e-Research y español LE/L2: investigar en la era digital. 2021, p. 164-196. Abingdon: Routledge.

This version is available at https://ddd.uab.cat/record/305328

under the terms of the BY-NC-ND license

8

Corpus para investigar sobre el componente fónico en español como LE/L2

Joaquim Llisterri

Resumen

Este capítulo se centra en la relevancia de los corpus orales para el estudio del componente fónico en español como LE/L2. En primer lugar, se considera la especificidad de los corpus orales y se introducen los conceptos básicos relacionados con este campo, para explorar, a continuación, algunos de los ámbitos en los que los corpus orales, tanto de hablantes nativos como de hablantes no nativos, resultan de utilidad en la enseñanza y en la investigación. En la segunda parte del capítulo se discuten las posibilidades y las limitaciones de la consulta de corpus orales en línea, y se presentan sucintamente algunas investigaciones basadas en corpus llevadas a cabo sobre aspectos fonéticos de la producción y la percepción del habla en español como LE/L2. La tercera parte del capítulo resume las características esenciales, especialmente en lo que se refiere al acceso a los datos, de algunos corpus orales en español como L1 y como LE/L2 en los que el investigador o el profesor pueden contar con las grabaciones; finalmente, se mencionan las herramientas empleadas más habitualmente para el análisis y la gestión de un corpus oral.

8.1. Necesidades

El estudio del componente fónico en español como LE/L2 todavía no ha alcanzado su pleno desarrollo si se compara con las investigaciones llevadas a cabo sobre otros niveles del análisis lingüístico. Aunque existen varias razones que explican tal situación, la dificultad de contar con datos orales es, sin duda, una de ellas, pues, si la creación de recursos lingüísticos adecuados para caracterizar fonéticamente una primera lengua ya exige un esfuerzo considerable, como señala Granger (2008, 262-263), “The difficulty of collecting and transcribing speech is multiplied by a factor of 10 in the case of learner data”.

Pese a que recopilar textos escritos no es, como se ha visto en los capítulos 6 y 7 del presente volumen, una tarea trivial, obtener grabaciones que combinen la naturalidad y la espontaneidad de los participantes con un nivel de calidad que permita llevar a cabo un análisis fonético requiere una metodología, un entorno y unas herramientas específicas que aumentan la complejidad del trabajo del investigador; por otra parte, la transcripción ortográfica y, especialmente, el etiquetado fonético de un corpus oral, aunque puedan automatizarse hasta cierto punto, exigen una formación especializada y una notable inversión tanto en tiempo como en recursos humanos y económicos (Myles 2005; Delais-Roussarie y Yoo 2011; Caines, McCarthy y O’Keeffe 2016). Por todo ello, el conjunto de corpus escritos a disposición de la comunidad científica es notablemente mayor que el de corpus que contienen material sonoro susceptible de un estudio fonético, lo que, a su vez, repercute en la menor atención que se ha prestado al nivel fónico en comparación con el gramatical, el léxico o el discursivo.

8.1.1. Distinguir entre corpus orales y corpus de lengua oral

En la tipología de corpus presentada en el capítulo 6 se ha hecho referencia al criterio de la modalidad, que permite dividir los recursos lingüísticos en escritos y hablados, así como a la dicotomía entre escritos y orales en lo que respecta a la especificidad de los textos. Para los fines de este capítulo, conviene, sin embargo, distinguir entre los “corpus de lengua oral” que, en inglés, suelen denominarse *spoken language corpora*, y los “corpus orales”, conocidos en inglés como *speech corpora* (Llisterri 1996; Whichmann 2008; McCarthy y O’Keeffe 2013; Caines, McCarthy y O’Keeffe 2016). Los corpus de lengua oral ofrecen como principal

material de trabajo una transcripción ortográfica de las grabaciones originales —enriquecida para representar algunos fenómenos propios de la oralidad, como se explica en el apartado 8.1.2—, mientras que los corpus orales incluyen una transcripción fonética de los materiales sincronizada con la grabación y acompañada, en ocasiones, de algún tipo de anotación. La posibilidad de contar con la señal sonora condiciona, como es natural, el tipo de trabajo que puede llevarse a cabo. Por ello, en los corpus de lengua oral el análisis se aborda con los mismos procedimientos y herramientas que se emplean en un corpus de lengua escrita —de ahí que en ocasiones se utilice el término "textos orales" (*spoken texts*) para describir este tipo de materiales— ya que, sin acceso a la señal sonora, un corpus de lengua oral puede considerarse, a efectos prácticos, un conjunto de textos y tratarse tal como se ha explicado en los capítulos 6 y 7. En los corpus orales, en cambio, el interés del investigador se centra en el plano fónico, por lo que debe recurrirse a herramientas como las que se describen en el apartado 8.3.3.

En lo que se refiere específicamente a los corpus con materiales procedentes de hablantes no nativos que contienen algún componente relacionado con la oralidad, Ballier y Martin (2015) proponen una clasificación en la que se distinguen tres tipos de materiales: corpus que consisten únicamente en transcripciones ortográficas (*mute spoken corpora*), corpus con los documentos sonoros sincronizados con la transcripción ortográfica (*speaking corpora*) y corpus con la señal sonora sincronizada con un etiquetado en el nivel fonético segmental o suprasegmental (*phonetic corpora*). Gut (2014), por su parte, plantea una división entre bases de datos de habla en una L2 que contienen grabaciones de participantes que realizan tareas muy específicas y controladas (*databases of L2 speech*), corpus de lengua oral de hablantes no nativos, en los que se cuenta, básicamente, con una transcripción ortográfica enriquecida de la grabación, pero que, en general, no contienen la señal sonora (*corpora of spoken learner language*) y, finalmente, corpus fonológicos de hablantes no nativos (*phonological learner corpora*), consistentes en la señal sonora sincronizada con una transcripción fonética o fonológica.

Puesto que el presente capítulo se centra en el componente fónico, se consideran únicamente los recursos en los que se puede trabajar con la señal sonora, es decir, los corpus orales, excluyendo aquellos corpus de lengua oral que, en términos de Ballier y Martin (2015), se describirían como "mudos", puesto que el investigador no tiene acceso directo a las grabaciones. Cabe precisar también que se hace únicamente referencia a recursos que no son comerciales y, por ello, se pueden utilizar gratuitamente, aunque en algunos casos sea necesario crear una cuenta de usuario. El lector interesado en conocer otros materiales puede recurrir a los catálogos en línea de CLARIN (CLARIN 2018b), ELRA (ELRA 2018) o del LDC (LDC 2019); existen también catálogos especializados en corpus de hablantes no nativos como los de la Universidad Católica de Lovaina (Centre for English Corpus Linguistics 2019), del proyecto TalkBank (McWhinney, sin fecha) y de CLARIN (CLARIN 2018a), además del desarrollado específicamente para el español como LE/L2 por Díaz Sánchez (sin fecha); asimismo, los trabajos de Briz y Albelda (2009), Campillos (2012), Caballero (2015) o de Solís (2018) ofrecen información relevante sobre diversos corpus que contienen materiales orales en español como L1 y como LE/L2.

8.1.2. Reconocer la especificidad de los corpus orales

En la caracterización de los corpus orales se utilizan algunos términos cuyo significado resulta útil precisar antes de abordar los siguientes apartados. En primer lugar, suele emplearse el concepto de "transcripción ortográfica enriquecida" para referirse a una representación ortográfica en la que se señalan algunos elementos relacionados con la oralidad. En (1) se muestra un ejemplo de este tipo de transcripción, extraído de PRESEEA, en el que se marcan las pausas (mediante una barra inclinada), las risas, los ruidos producidos por el entrevistado,

los alargamientos, las dificultades encontradas en la transcripción, el discurso indirecto y las hesitaciones que se reflejan en la elisión de parte de una palabra.

(1) no / es que me molesta que <entre_risas> me traten de tú </entre_risas> por ejemplo / vas al banco / <ruido=respiración audible/> / y<alargamiento/> y l<alargamiento/>a empleada me trata de tú / y <transcripción_dudosa> yo diciendo </transcripcion_dudosa><cita> <entre_risas> por favor soy igual de cliente que mi padre </entre_risas></cita> / y<alargamiento/> <risas=I> mmm y no sé me mo <palabra_cortada/> me molesta /

Las convenciones utilizadas se encuentran documentadas en cada corpus y difieren en función de los objetivos de los investigadores pero, por lo general, se adaptan las definidas por la *Text Encoding Initiative* (Romary y Witt 2014; TEI Consortium 2018). Para facilitar el tratamiento informático de los datos, los fenómenos que se transcriben deben estar, además, codificados, y para tal fin se emplean marcas situadas entre ángulos, en las que < indica el principio y /> señala el final del fenómeno, propias del lenguaje de codificación denominado XML (*Extensible Markup Language*).

En un corpus oral orientado a los estudios sobre el plano fónico se realiza también una segmentación de la señal sonora, operación mediante la que se delimita el principio y el final de cada unidad de análisis en el nivel segmental (fonemas o alófonos), en el suprasegmental (por ejemplo, grupos fónicos o grupos entonativos) o en el de la representación ortográfica.

Una vez establecidas las fronteras, cada unidad se etiqueta para definir su contenido; pueden emplearse los símbolos del Alfabeto Fonético Internacional (IPA, sin fecha) o los de su adaptación para la transcripción de recursos electrónicos conocida como SAMPA (*Speech Assessment Methods Phonetic Alphabet*) (Wells 1999-2015), y también se utilizan etiquetas específicas para los fenómenos prosódicos, como las que se proponen, por ejemplo, en INTSINT (*International Transcription System for Intonation*) (Baqué y Estruch 2003) o en ToBI (*Tones and Break Indices*) (Hualde 2003); habitualmente se incluye también la representación ortográfica. Los términos "etiquetado" (*labelling*) y "anotación" (*annotation*) pueden considerarse, en la práctica, como sinónimos, si bien el primero se emplea a menudo en los estudios de tipo fonético y el segundo en los que tratan otros niveles del análisis lingüístico. Los diversos procedimientos para etiquetar o anotar corpus orales se discuten con detalle en Delais-Roussarie y Post (2014) y, específicamente en el caso de los corpus de hablantes no nativos, en Ballier y Martin (2015) y en Carranza (2016).

Finalmente, el concepto de "alineación" (*alignment*) remite a la sincronización temporal entre el etiquetado y la señal sonora, tal como se muestra en un ejemplo extraído del corpus CIEMPIESS, recogido en la Figura 1, en el que la señal se ha segmentado en palabras y la representación ortográfica —en la que el acento léxico se marca mediante una mayúscula y las pausas se indican con la etiqueta "++dis++"— se ha alineado con la grabación.

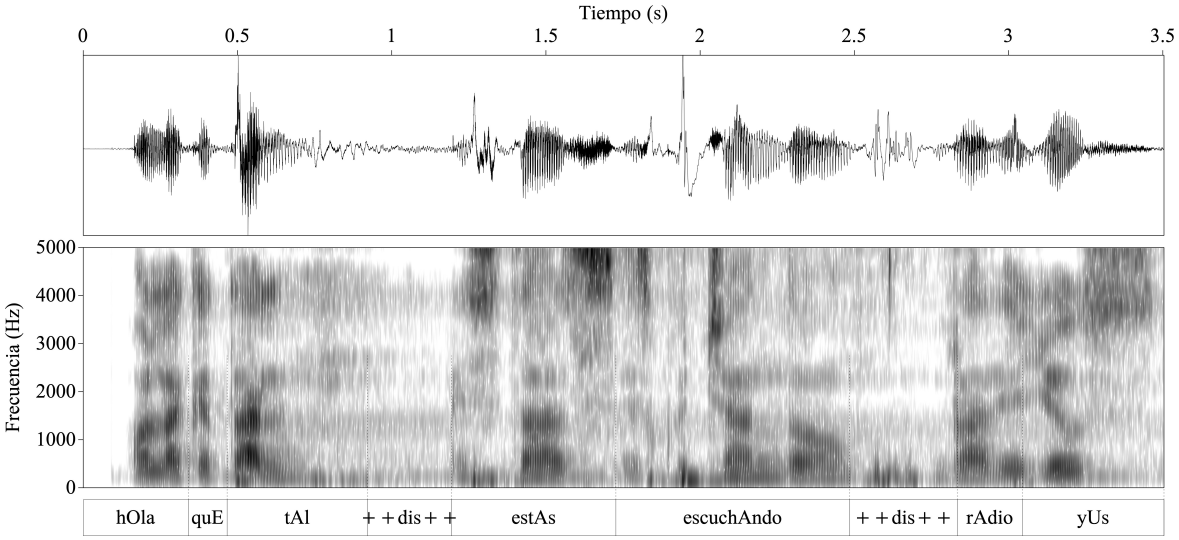

Figura 1. Transcripción ortográfica alineada con la señal sonora (oscilograma y espectrograma) mediante el programa Praat en el enunciado “Hola, qué tal. Estás escuchando radio Yus” extraído del corpus CIEMPIESS.

Cuando el etiquetado corresponde a distintos niveles de análisis, los datos del corpus quedan representados de modo semejante a una partitura de orquesta, tal como puede apreciarse en la Figura 2; en ella se reproduce el mismo enunciado de la Figura 1, al que se han añadido tres niveles de etiquetado que originalmente no estaban presentes en el corpus CIEMPIESS: el de los grupos fónicos, representados ortográficamente, el de las sílabas y el de los segmentos, ambos transcritos mediante el Alfabeto Fonético Internacional.

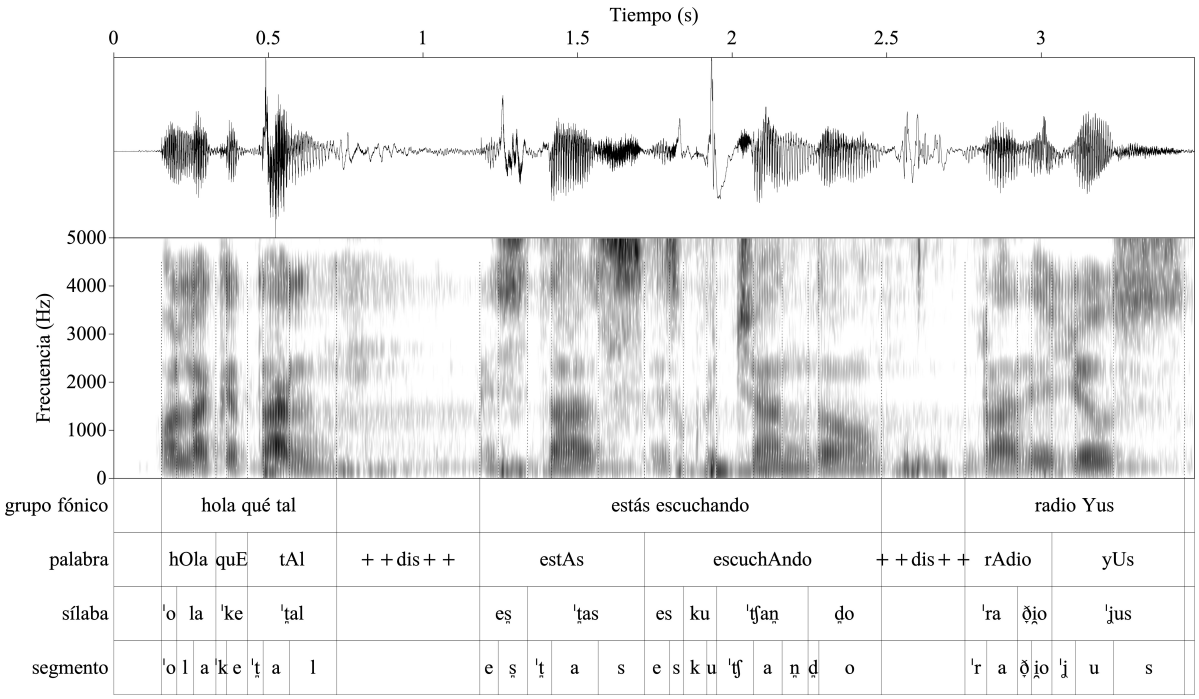

Figura 2. Etiquetado en cuatro niveles alineado con la señal sonora (oscilograma y espectrograma) realizado mediante el programa Praat en el enunciado “Hola, qué tal. Estás escuchando radio Yus” extraído del corpus CIEMPIESS.

En el proceso de creación de un corpus oral se deben tomar decisiones y definir criterios para cada uno de los aspectos que se acaban de mencionar —transcripción ortográfica, segmentación y etiquetado o anotación—, así como en lo que se refiere a las características de los participantes y del material lingüístico recogido y a los procedimientos para obtener los datos. El lector interesado puede encontrar indicaciones útiles sobre todas estas cuestiones en Durand, Gut y Kristoffersen (2014), Niebuhr y Michaud (2015) y Polo (2018) y, en el caso de los corpus de hablantes no nativos, en trabajos como los de Bonaventura, Howarth y Menzel (2000), Racine *et al*. (2011), Gilquin (2015) y de Pustka *et al*. (2018), así como en el capítulo 7 del presente volumen.

8.1.3. Reconocer la utilidad de los corpus orales

Desde un punto de vista conceptual, las aplicaciones de los corpus orales a la enseñanza del español como LE/L2 o al estudio de su adquisición no difieren sustancialmente de las ya descritas en los capítulos 6 y 7 aunque varían, lógicamente, tanto el nivel de análisis lingüístico en el que se centran el profesor o el investigador como las herramientas utilizadas. Se considerará, en primer lugar, la utilidad de los corpus orales de hablantes nativos (8.1.3.1) y, en segundo, la de los que recogen datos procedentes de hablantes que aprenden el español como LE/L2 (8.1.3.2).

8.1.3.1. Utilidad de los corpus orales de hablantes nativos

En los trabajos dedicados al papel de los corpus orales de hablantes nativos en la enseñanza de una LE/L2 se suelen mencionar principalmente dos grandes ámbitos: la descripción de la lengua en su uso real y la obtención de muestras de habla seleccionadas en función de criterios específicos (Albelda 2011; Bailini 2014; Secchi 2014; Caines, McCarthy y O'Keeffe 2016; Dong 2018). El primero de ellos tiene implicaciones en el diseño curricular y en el contenido de los materiales didácticos, mientras que el segundo se relaciona directamente con la creación de materiales, tanto para la enseñanza como para la evaluación, y con el empleo de corpus en el aula por parte de los estudiantes.

En lo que se refiere al primer ámbito, disponer de corpus orales como los que se presentan en el apartado 8.3.1 permite una investigación sistemática de los rasgos segmentales y suprasegmentales de la lengua basada en muestras representativas que reflejan una variada gama de situaciones comunicativas en las que interviene un número significativo de hablantes. Conviene recordar que el manual que todavía constituye una referencia ineludible en lo que respecta a la descripción fonética del español se centra en la pronunciación "castellana sin vulgarismos y culta sin afectación, estudiada especialmente en el ambiente universitario madrileño" (Navarro Tomás 2004, § 4); más recientemente, para obtener parte de la información contenida en el volumen dedicado a la fonética y a la fonología de la *Nueva gramática de la lengua española*, "se ha procedido a la selección rigurosa de informantes, titulados universitarios procedentes de las capitales de todos los países del mundo hispánico" (Real Academia Española y Asociación de Academias de la Lengua Española 2011, xx). Por otra parte, en los estudios experimentales más tradicionales sobre la descripción fonética del español se ha tendido a analizar el estilo leído de un número reducido de hablantes. Aunque en los últimos años se ha prestado más atención al habla espontánea, al tiempo que las tecnologías han hecho posible automatizar el análisis de una mayor cantidad de datos, todavía queda un largo camino por recorrer en lo que se refiere a la caracterización de determinados estilos de habla, a los estudios sociofonéticos o a la exploración de la interfaz entre el nivel fonético y otros niveles de análisis, por mencionar algunos ejemplos.

En cuanto al segundo ámbito al que se ha aludido al principio de este apartado, una de las necesidades de los creadores de materiales para la docencia o para la evaluación, así como también de los profesores que se proponen emplear metodologías basadas en corpus para el

trabajo en el aula, es obtener muestras reales del uso oral de la lengua. Puede argumentarse que, actualmente, Internet facilita el acceso a todo tipo de recursos hablados, pero el problema reside en que estos materiales no constituyen lo que propiamente se definiría como un corpus, es decir, no están, en la mayoría de los casos, transcritos, etiquetados y acompañados de una interfaz que facilite la consulta para encontrar exactamente lo que se necesita. El uso de corpus orales en línea, como se explica en el apartado 8.2.1, puede resultar de ayuda en este contexto si los materiales del corpus se han preparado de un modo adecuado.

Finalmente, contar con corpus orales diseñados, recogidos y anotados con criterios comparables permite realizar estudios de fonética contrastiva, tanto en el plano segmental como en el suprasegmental, superando así las limitaciones de los análisis basados en datos procedentes de manuales de fonética y fonología descriptivas de cada una de las lenguas comparadas.

8.1.3.2. Utilidad de los corpus orales de hablantes no nativos

Los corpus orales en los que se recogen datos procedentes de hablantes no nativos encuentran su principal aplicación en el estudio de los rasgos fónicos de la interlengua. La investigación basada en corpus permite obtener inventarios de errores que, al contrario de lo que sucede con las predicciones teóricas del análisis contrastivo clásico, proceden de datos reales, como se muestra, por ejemplo, en el trabajo de Carranza, Cucchiarini, Llisterri *et al*. (2014) presentado en el apartado 8.2.2. Un corpus amplio de habla no nativa también facilita detectar tanto regularidades como fenómenos de variación en la interlengua que pueden pasar desapercibidos en muestras de la producción en una L2/LE más reducidas, como las que se emplean habitualmente en los estudios fonéticos experimentales. Por otro lado, si en el corpus se incluyen hablantes de distintas L1, pueden llegar a determinarse los errores fónicos más frecuentes en una L2/LE con independencia de la primera lengua como se hace patente, por ejemplo, en los estudios perceptivos de Blanco (2016, 2017) resumidos en el apartado 8.2.4; si, además, los corpus contienen suficiente información sobre los hablantes, como la que se muestra en la Tabla 2 del capítulo 7, resulta posible correlacionar la frecuencia de aparición de determinados tipos de error con variables como el nivel de conocimiento de la lengua, el tiempo de exposición a la L2/LE o la edad en la que empezó su adquisición (Neri, Cucchiarini y Strik 2006; Cucchiarini *et al*. 2011; Gut 2014). Así mismo, mediante un corpus en el que se han recogido diversos estilos de habla puede determinarse si existen errores que se producen con mayor frecuencia en uno de ellos, contrastando, por ejemplo, la lectura con el habla espontánea. Finalmente, disponer de un corpus de hablantes nativos con un contenido y un procedimiento de anotación equivalentes al del corpus de hablantes no nativos permite realizar una comparación fonética o fonológica entre las realizaciones de ambos tipos de locutor.

Los aspectos anteriormente mencionados se estudian en el nivel segmental, considerando los errores que afectan a las vocales, las consonantes y a los grupos vocálicos y consonánticos, o en el suprasegmental, abordando los problemas relacionados con el acento, la melodía, las pausas, la velocidad de habla, el ritmo y la cualidad de voz, sea desde una perspectiva fonética —especialmente, la de la fonética acústica—, sea desde un punto de vista fonológico, analizando, por ejemplo, la estructura de la sílaba o los patrones entonativos. Cabe también, como apunta Gut (2014), investigar sobre las variables temporales que contribuyen a la fluidez en la LE/L2, relacionadas, en general, con las pausas y con la velocidad de habla.

El estudio de la adquisición del componente fónico de una LE/L2 puede llevarse a cabo desde un punto de vista longitudinal, lo que requiere contar con un corpus adecuadamente planificado, especialmente en lo que se refiere a la posibilidad de recoger datos de los mismos informantes durante un determinado período de tiempo.

Algunos de los resultados de las investigaciones basadas en corpus orales procedentes de hablantes de una LE/L2 tienen su aplicación en el diseño curricular y en la creación de materiales, así como, en niveles avanzados, en el trabajo que se realiza en el aula (Gut 2014). En este sentido, disponer de información sobre los errores fónicos más frecuentes —con independencia de la L1 del estudiante o en función de esta— facilita seleccionar los que necesariamente deberían tenerse en cuenta en la programación docente y establecer un orden de prioridades en la enseñanza de la pronunciación; los datos de naturaleza perceptiva como los que se presentan en los apartados 8.2.3 y 8.2.4 complementan los estudios basados en la producción.

Finalmente, los corpus orales de hablantes no nativos resultan imprescindibles para desarrollar las herramientas informáticas empleadas en el aprendizaje de lenguas extranjeras que proporcionan una evaluación automática de la pronunciación. Mediante un corpus en el que previamente se han etiquetado y clasificado los errores en el plano fónico se determinan los errores de pronunciación propios de cada grupo de estudiantes en función de su L1, se crean los modelos acústicos específicos para hablantes no nativos, se entrenan los sistemas de reconocimiento automático del habla con estos modelos para que puedan detectarse los errores y se evalúa el funcionamiento de las herramientas diseñadas (Neri, Cucchiarini y Strik 2003; Carranza, Cucchiarini, Burgos *et al*. 2014; O'Brien *et al*. 2018).

8.2. Cómo ayudan las tecnologías

En este apartado se ofrecen dos perspectivas sobre el modo en el que pueden llevarse a cabo investigaciones basadas en corpus orales. La primera de ellas se centra en la consulta de recursos en línea (8.2.1) para obtener muestras de determinados fenómenos, en tanto que la segunda, expuesta en los apartados 8.2.2, 8.2.3 y 8.2.4, implica el uso de corpus completos, generalmente almacenados localmente. Conviene aclarar que en los trabajos experimentales dedicados a la producción o la percepción del español como LE/L2 se cuenta siempre con un corpus de datos, aunque su diseño se ciñe al fenómeno estudiado —se trata de lo que Gut (2014) describe como *databases of L2 speech*—, en contraste con los recursos de naturaleza más amplia que se tratan a continuación. En todos los casos, sin embargo, el componente fónico se aborda desde una perspectiva basada en los datos (conocida, en inglés, como *data-driven research*), tal como se expone en los capítulos 1 y 6 del presente volumen.

8.2.1. La consulta de corpus orales en línea

Como se ha explicado en el apartado anterior, en muchas ocasiones el profesor o el investigador necesitan acceder a datos orales; la opción más frecuente es recurrir a los corpus que pueden consultarse en línea, como los que se recogen en la Tabla 4 y en la Tabla 7. A continuación, se muestran algunas posibilidades basadas en ejemplos procedentes de corpus de hablantes nativos de español.

La diferencia más importante con respecto a las búsquedas de materiales que puedan realizarse en la red (por ejemplo, en plataformas como YouTube o en los archivos de emisoras de radio y televisión) reside en el hecho de que un corpus se organiza a partir de un conjunto de variables relacionadas con las características de los hablantes, con los procedimientos que se utilizan para obtener las producciones orales y con el estilo de habla que de ellos resulta. Esta información se incorpora a cada uno de los documentos del corpus y constituye lo que se denominan "metadatos" (Broeder y Uytvanck 2014). Así, como se aprecia en la Figura 3, en el corpus PRESEEA los metadatos asociados a cada documento permiten realizar búsquedas en función de la localidad, el sexo, el grupo de edad y el nivel de estudios del informante.

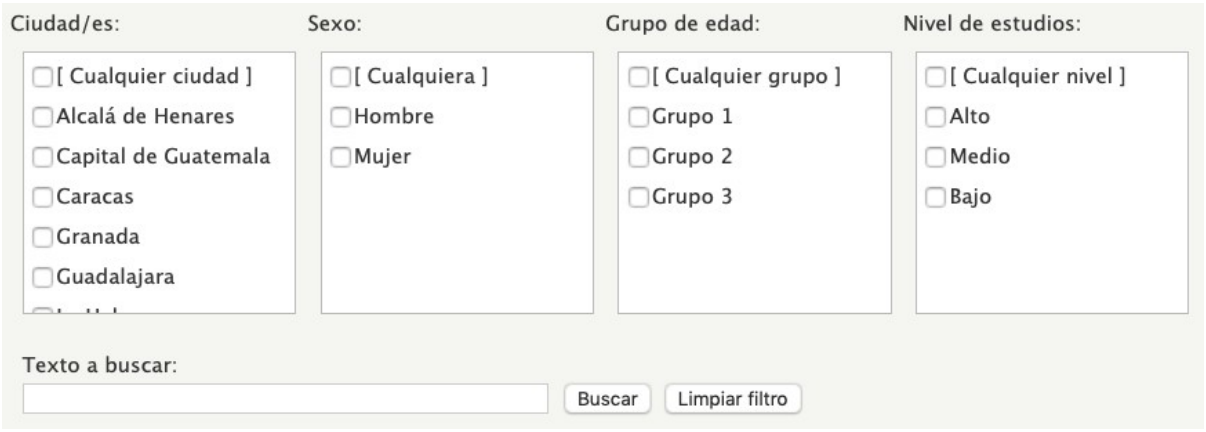

Figura 3. Criterios empleados en la consulta del corpus PRESEEA.

En recursos especializados como el *Atlas interactivo de la entonación del español* los metadatos incluyen características prosódicas de los enunciados (la modalidad oracional, el carácter neutro o marcado del enunciado, el tipo de foco o el valor pragmático), de modo que una búsqueda como la que se muestra en la Figura 4 permite encontrar ejemplos de frases interrogativas absolutas de carácter neutro con un valor pragmático de confirmación producidos por hablantes del dialecto andino.

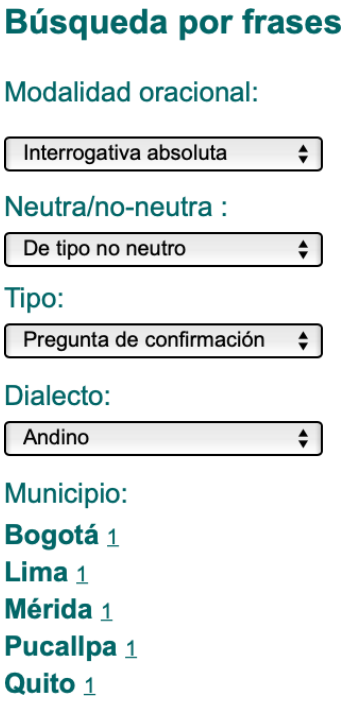

Figura 4. Búsqueda de frases interrogativas absolutas de tipo neutro con un valor pragmático de confirmación en el dialecto andino en el *Atlas interactivo de la entonación del español*.

Un segundo elemento esencial viene dado por la posibilidad de buscar en el corpus partiendo de los elementos que en él se han etiquetado o anotado. Una transcripción ortográfica enriquecida con la anotación de aspectos propios de la lengua oral como la que se ha mostrado en (1) o la que se encuentra, por ejemplo, en ESLORA, facilita localizar fenómenos como los alargamientos segmentales o el énfasis en determinados elementos del discurso. El resultado de una búsqueda de la palabra ‘bueno’ seleccionando los casos en los que se

produce un alargamiento (en la opción ‘Buscar en’), permite obtener resultados como los que se recogen en la Figura 5.

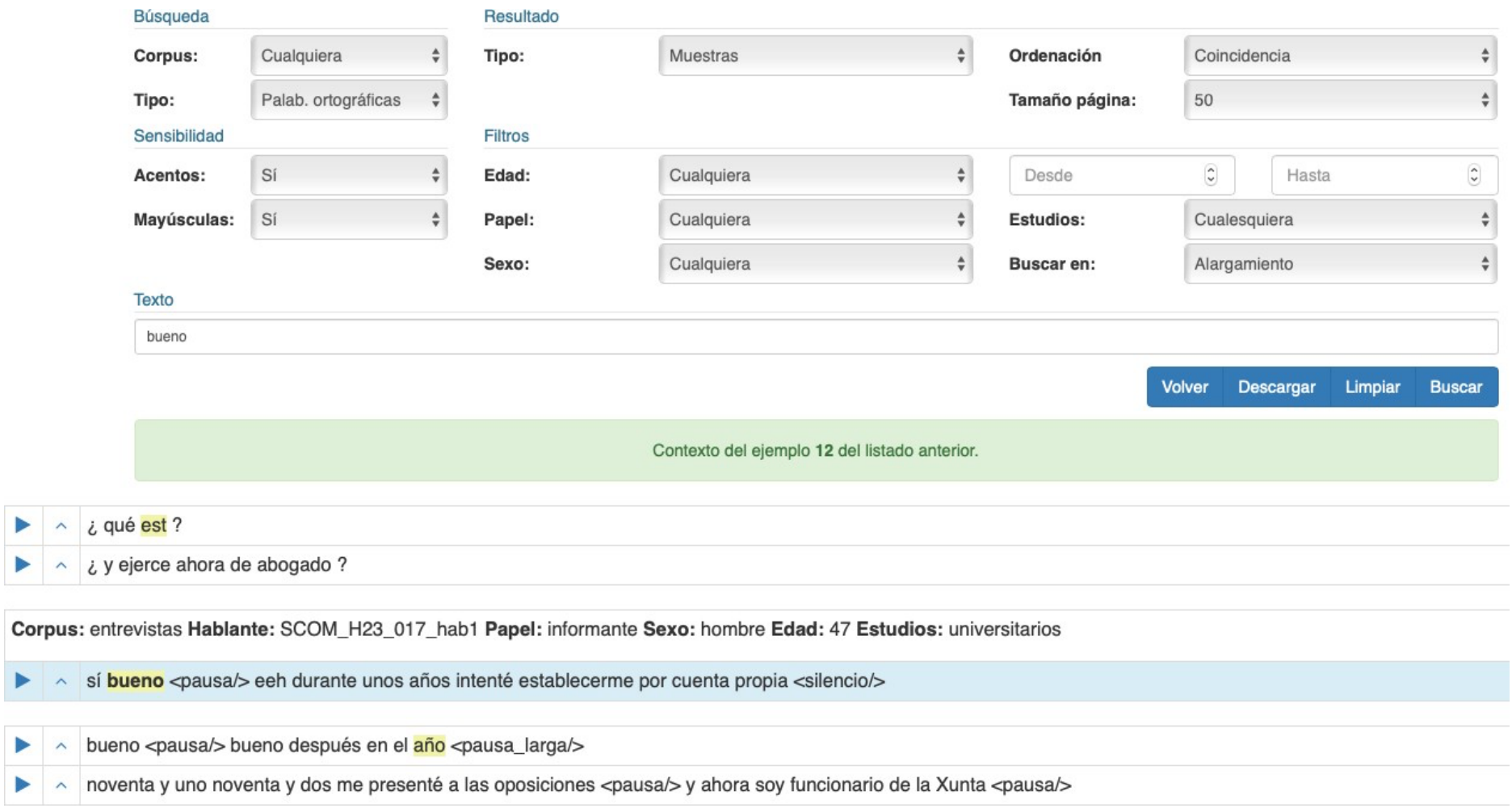

Figura 5. Muestra de los resultados de la búsqueda de la palabra ‘bueno’ con alargamiento en el corpus ESLORA.

Puesto que en ESLORA es posible descargar la grabación de cada fragmento (véase la Tabla 3), el enunciado puede visualizarse y analizarse mediante programas de análisis acústico, como se muestra en la Figura 6.

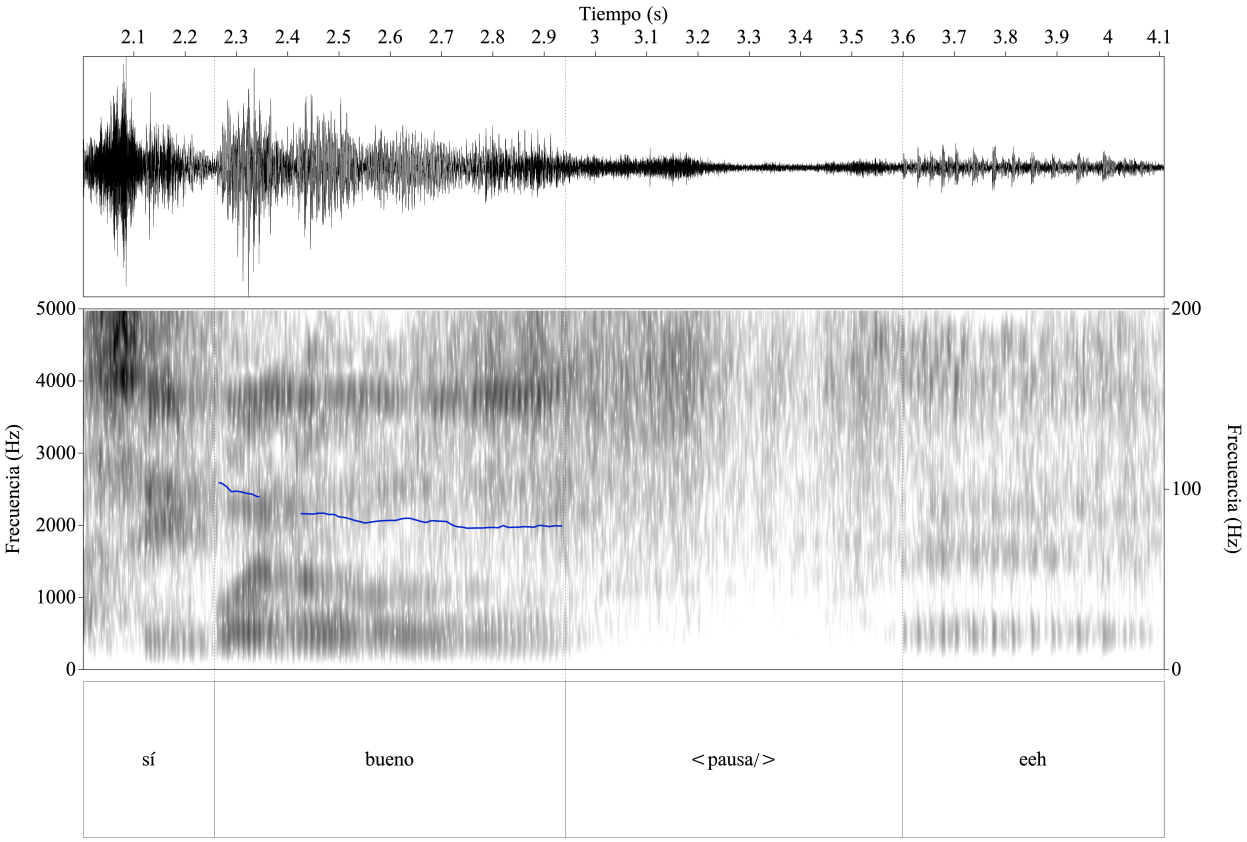

Figura 6. Oscilograma, espectrograma y curva melódica obtenidos mediante el programa Praat del enunciado 'sí bueno eeh' extraído del corpus ESLORA.

En la Figura 6 se ejemplifica también una de las principales dificultades del análisis fonético de muestras de corpus orales que se obtienen en entornos naturales para facilitar que el hablante se exprese con espontaneidad: si se compara con la Figura 7, procedente del corpus DiEspa, grabado en una sala acústicamente acondicionada como las que se encuentran en un laboratorio de fonética, se aprecian las diferencias en la señal en los fragmentos silenciosos y en el conjunto de la grabación.

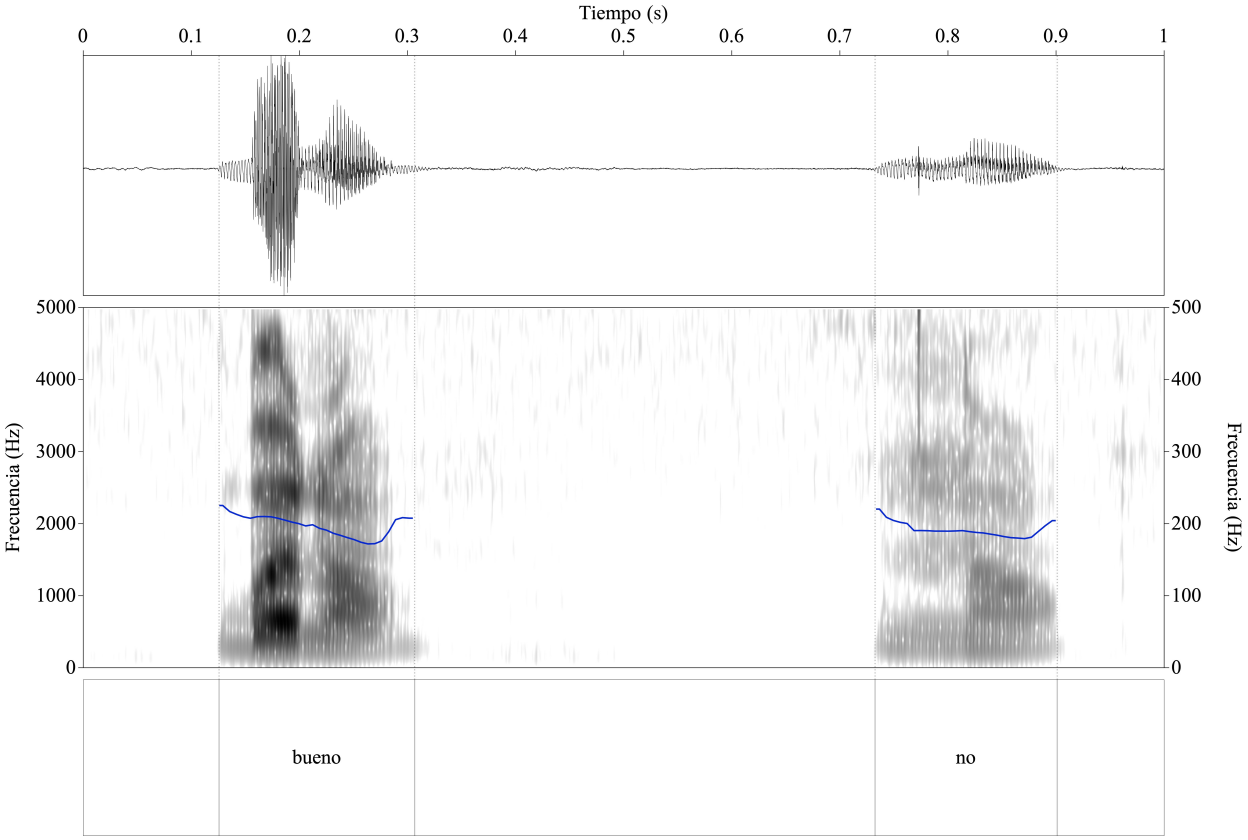

Figura 7. Oscilograma, espectrograma y curva melódica obtenidos mediante el programa Praat del enunciado 'bueno, no' extraído del corpus DiEspa.

Para combinar el grado de espontaneidad propio de una conversación o de una entrevista con la calidad de la grabación necesaria para un análisis fonético acústico se recurre en ocasiones a tareas que requieran la cooperación de dos participantes para alcanzar un objetivo y que puedan llevarse a cabo en un entorno acústicamente controlado. Entre las más conocidas se encuentran la tarea del mapa (*Map Task*), usada en el *Atlas interactivo de la entonación del español* y el juego de las diferencias, empleado en DiEspa y en DiapixFL; el objetivo es conseguir que los hablantes se concentren en la realización de la tarea —trazar un camino en un mapa siguiendo las indicaciones del interlocutor o descubrir las diferencias entre dos imágenes similares— y actúen de la forma más espontánea posible. Una discusión detallada sobre las distintas tareas empleadas en la constitución de un corpus oral se encuentra en Birch (2014); desde un punto de vista más práctico, en IRIS y, específicamente en el caso del español como LE/L2, en Fono.data, se puede acceder a diversos instrumentos útiles para la recogida de datos.

Finalmente, si el corpus se ha etiquetado fonética o fonológicamente es posible realizar búsquedas específicas para localizar enunciados en los que aparezca un determinado segmento. Aunque esta no es la situación más frecuente, debido a los recursos humanos y económicos que implica una tarea de este tipo, el proyecto *Romance Phonetics Database* permite ejemplificar las posibilidades de búsqueda en un corpus etiquetado en el nivel fonético, como se muestra en la Figura 8.

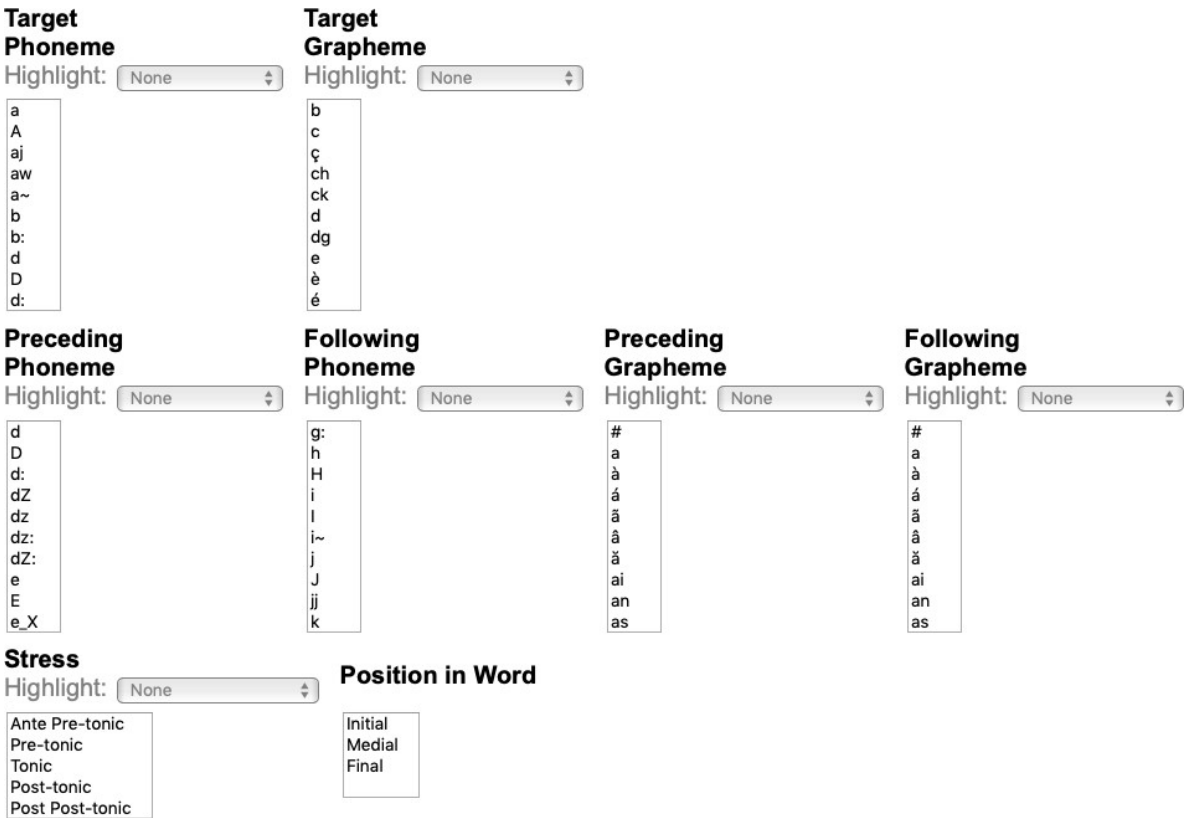

Figura 8. Variables consideradas en la búsqueda de enunciados que contengan un determinado segmento en función del segmento precedente, del siguiente, del acento y de la posición en la palabra en la *Romance Phonetics Database*.

En resumen, el potencial de un corpus en línea para la investigación o para la obtención de materiales que puedan usarse en el aula depende, al menos, de tres factores: el grado de detalle de los metadatos, la posibilidad de realizar búsquedas en el etiquetado o anotación del corpus y la calidad de las grabaciones. En lo que se refiere a las aplicaciones didácticas, los trabajos de Drange (2008), Bailini (2014), Nicolás (2014) y de Martín Sánchez, Pascual y Paz (2017) ofrecen muestras de actividades orientadas al nivel fónico y basadas en el uso de corpus de hablantes nativos que puede llevarse a cabo con estudiantes de español como LE/L2.

8.2.2. El estudio de la producción basado en corpus orales en español como LE/L2

Una de las primeras tareas que se plantean al estudiar los rasgos fonéticos y fonológicos de la producción de hablantes no nativos es el establecimiento de una tipología de errores que permita, una vez etiquetado el corpus, obtener datos lingüísticamente pertinentes. En los trabajos de Blanco y Nogueroles (2013, 2014), basados en el corpus Fono.ele, se propone una clasificación de errores fónicos en español como LE/L2 realizada a partir de cuatro criterios: la naturaleza lingüística del error (fonológico, fonético o fónico), el tipo de elemento afectado (segmental o prosódico), los procesos que subyacen al error (inserción, elisión, sustitución, modificación y desplazamiento) y el efecto comunicativo del error (impide, dificulta o no dificulta la comunicación). En cambio, en los proyectos más orientados a la creación de corpus que permitan entrenar sistemas de reconocimiento automático del habla para la enseñanza de la pronunciación suelen considerarse únicamente tres tipos de errores segmentales (inserción, elisión y sustitución), así como la naturaleza de los segmentos que siguen y preceden al error (Carranza, Cucchiarini, Llisterri *et al*. 2014).

Cuando se dispone de un corpus en el que se han etiquetado los errores en el plano fónico es posible realizar, como se ha explicado en el apartado 8.1.3.2, análisis estadísticos relativos a la frecuencia de aparición de cada tipo de error teniendo en cuenta factores como la L1 del estudiante (Blanco y Nogueroles 2013, 2014), su nivel de conocimiento de la L2, el estilo de habla o el contexto fonético en el que se produce el error (Carranza, Cucchiarini, Llisterri *et al*. 2014). Una muestra de este tipo de trabajo puede encontrarse en el estudio que Carranza, Cucchiarini, Llisterri *et al*. (2014) llevaron a cabo partiendo de las grabaciones de los exámenes orales realizados cada seis meses durante los dos primeros cursos académicos por 20 estudiantes universitarios japoneses de español; el corpus contiene diferentes estilos de habla correspondientes a las distintas tareas llevadas a cabo durante las pruebas y los errores se anotaron mediante un procedimiento que permite conocer el segmento correcto, el incorrecto, el tipo de error y el contexto de aparición. Así, por ejemplo, en la Figura 9 puede apreciarse un error de sustitución (codificado mediante la letra 'a') de [e] (codificado como '02') por [i] que se da entre [β̞] (codificado como '26') y [s] (codificado como '23') en la segunda sílaba de la palabra *vives*.

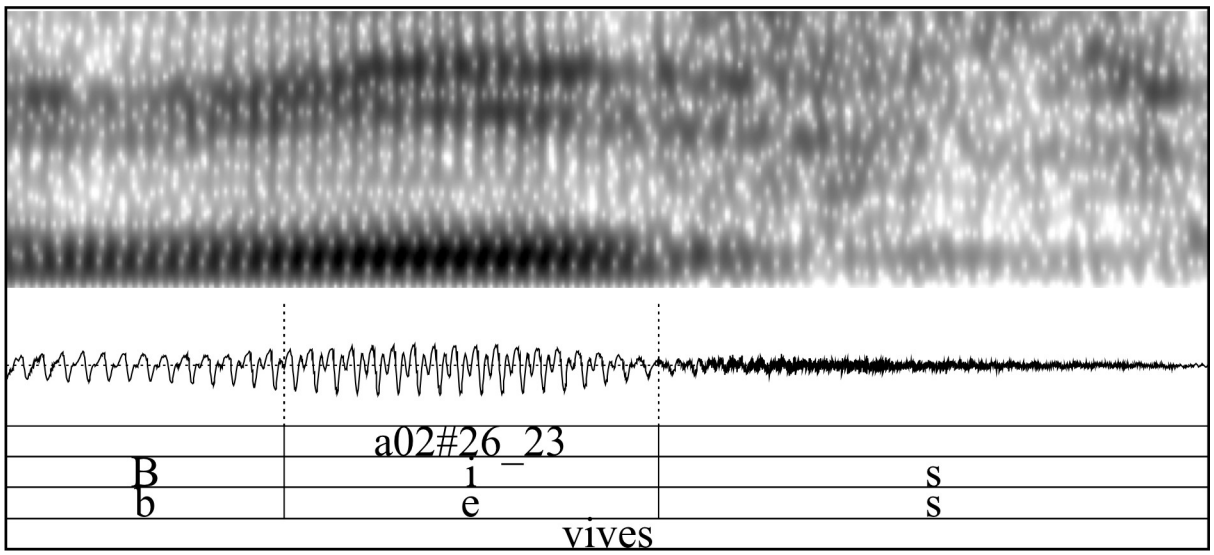

Figura 9. Etiquetado de un error de sustitución de [e] por [i] en el contexto [β̞_s] (transcrito como [B_s] en SAMPA) realizado mediante el programa Praat. En la primera fila aparece la codificación del error, en la segunda, la transcripción de la realización canónica [i] y, en la tercera, la transcripción de la realización errónea [e] (Carranza 2013).

En el etiquetado fonético de los errores surge, en ocasiones, el problema de lo que puede describirse como "realizaciones intermedias". Uno de los ejemplos que se discuten en Carranza (2013) en el caso de estudiantes japoneses de español son las realizaciones de /ɾ/ que presentan la estructura formántica típica de una consonante lateral y la oclusión propia de una rótica simple. Cuando se dan estas situaciones, el investigador debe tomar decisiones apoyadas en la estructura acústica de la señal o bien recurrir a una validación externa de la transcripción realizada por hablantes nativos con conocimientos de fonética (Carranza, Cucchiarini, Llisterri *et al*. 2014).

El tipo de análisis descrito permite, tomando como referencia un corpus de 8,9 horas de duración que contiene 13 410 errores producidos por estudiantes japoneses de español, determinar, por ejemplo, que los errores de sustitución aparecen con mayor frecuencia que los de inserción, mientras que los de elisión son los menos habituales y que, entre las sustituciones, los tres errores más frecuentes afectan a las consonantes [ð̞], [l] y [β̞] , como se

muestra en la Tabla 1, adaptada de Carranza, Cucchiarini, Llisterri *et al*. (2014).

Tabla 1. Datos cuantitativos sobre los tres errores de sustitución más frecuentes en un corpus de estudiantes japoneses de español, adaptado de Carranza, Cucchiarini, Llisterri *et al*. (2014).

Número de casos	Realización correcta	Realización incorrecta	Frecuencia de aparición en %	Segmento precedente (frecuencia de aparición en %)	Segmento siguiente (frecuencia de aparición en %)
921	[ð̞]	[d]	82	vocal (79)	vocal (89)
		[t]	16	vocal (49) [s] (40)	vocal (94)
815	[l]	[ɾ]	66	vocal (68) silencio (8)	vocal (76) silencio (8)
806	[β̞]	[b]	89	vocal (84)	vocal (87)
		[p]	9	vocal (68) [s] (16)	vocal (88)

Como es lógico, los datos que puedan obtenerse dependerán de las decisiones que se hayan tomado en el momento de diseñar y de etiquetar el corpus. Las variables que intervienen en la creación del corpus formarán parte de los metadatos que permiten realizar búsquedas, como se ha expuesto en el apartado 8.2.1, mientras que los niveles de etiquetado (segmental, suprasegmental, fonético, fonológico, etc.), el detalle con el que se etiquetará cada nivel y el procedimiento mediante el cual se codifiquen los errores condicionarán la naturaleza de los resultados del análisis. Por ello, conviene insistir en que el diseño del corpus y la definición de los criterios de etiquetado constituyen una etapa de primordial importancia en cualquier investigación.

8.2.3. El estudio de la percepción basado en corpus orales en español como LE/L2

Conocer cómo los hablantes nativos valoran perceptivamente las realizaciones no nativas resulta de interés tanto desde un punto de vista teórico como en lo que respecta al establecimiento de prioridades en la enseñanza de la pronunciación. Las dimensiones que habitualmente se consideran son la inteligibilidad, la facilidad de comprensión (*comprehensibility*) y el grado de acento extranjero (*accentedness*) (Munro y Derwing 1995), que pueden analizarse en función de la tipología de errores fónicos empleada en el etiquetado del corpus.

Un estudio que ilustra la metodología empleada en este tipo de investigación es el que, basado en el corpus de estudiantes japoneses de español al que se ha hecho referencia en el apartado 8.2.2, se describe en Carranza (2015). En este experimento se presentan muestras de realizaciones no nativas, tanto erróneas como correctas, a un grupo de hispanohablantes para que transcriban ortográficamente los enunciados —lo que permite evaluar la inteligibilidad— y valoren en una escala numérica el grado de dificultad para entenderlos, así como el grado de acento extranjero percibido. Los resultados muestran, por ejemplo, que las valoraciones más negativas se encuentran en los errores consistentes en la epéntesis de una vocal, puesto que se altera la estructura silábica, seguidos por los errores fonológicos derivados de la confusión entre [l] y [ɾ]; en cambio, los errores de naturaleza fonética como las sustituciones de [x] o de [r] parecen incidir en menor medida en las tres dimensiones analizadas.

8.2.4. El estudio de la percepción por parte de hablantes no nativos de español

Así como es posible crear recursos para estudiar la producción de hablantes no nativos, resulta también factible recoger de forma sistemática datos que reflejen su percepción fónica. Una muestra del tipo de análisis que puede llevarse a cabo se encuentra en los trabajos de Blanco (2016, 2017), en los que se presenta la vertiente perceptiva del corpus Fono.ele. La primera etapa consiste en el diseño de una prueba de percepción en la que se incluyen un total de 10 ejercicios centrados en los siguientes aspectos: contrastes de sonidos en pares mínimos, identificación de sonidos en palabras, identificación de la posición del acento en pares mínimos acentuales, del número de sílabas, de la sílaba tónica y del patrón acentual de la palabra, identificación del patrón entonativo en pares y en listas de oraciones, identificación del tonema final como ascendente o descendente e identificación de la modalidad oracional como afirmativa, interrogativa o exclamativa. Esta prueba, con la que se obtienen 300 respuestas por participante, se aplicó a 204 estudiantes de español como LE/L2, hablantes de 10 lenguas maternas diferentes, con un nivel de conocimiento del español que se sitúa entre el A2 y el C1, con distintos grados de contacto con el español y con diferentes experiencias de aprendizaje en lo que se refiere al trabajo sobre los aspectos fónicos en el aula.

El corpus de respuestas (61 200 en total) permite realizar una amplia diversidad de estudios, tanto en lo que se refiere a los ítems de la prueba que plantean más dificultades como relacionando los resultados con las variables relativas a los hablantes. Así, por ejemplo, se constata que "a nuestros estudiantes de español, independientemente de su lengua materna, solo les plantea dificultad notoria la percepción de las oposiciones ɾ/r, ð/ d̪, ɲ/n + j/n y diptongo/vocal" (Blanco 2016, 10), elementos a los que cabe añadir "el cómputo de sílabas … diferenciar tipos de palabras según la posición de la sílaba tónica y, por último, identificar el esquema entonativo cuando entre las opciones aparece la suspensión tonal" (Blanco 2017, 80).

8.3. Casos concretos

Realizar un inventario exhaustivo de los corpus orales existentes en español como L1 y como L2/LE y exponer detalladamente sus características excede los límites del presente capítulo, por lo que a continuación se consideran únicamente las posibilidades de acceso a los datos que ofrecen algunos corpus orales, tanto de hablantes nativos de español (8.3.1) como de estudiantes de español como LE/L2 (8.3.2); con ello se pretende proporcionar una aproximación al tipo de materiales disponibles y, si se requiere más información sobre cada recurso, pueden consultarse los enlaces que se recogen en la Tabla 2 y en la Tabla 5. Se ofrecen, finalmente, unas breves indicaciones sobre herramientas de dominio público que permiten gestionar y analizar corpus orales (8.3.3).

8.3.1. Corpus orales de hablantes nativos de español

Como ya se ha señalado en el apartado 8.1.1, se hace únicamente referencia a los corpus orales en los que es posible disponer directamente de las grabaciones y que constituyen recursos de dominio público, aunque en algunos casos se requiera crear una cuenta de usuario. Entre los recursos que cumplen con estos dos requisitos y que contienen datos procedentes de hablantes nativos de español se cuentan los que se muestran en la Tabla 2.

Tabla 2. Corpus orales de hablantes nativos de español, de dominio público y con acceso a las grabaciones.

AIEE	Atlas interactivo de la entonación del español	http://prosodia.upf.edu/atlasentonacion/
Backbone	Pedagogic Corpora for Content and Language Integrated Learning	http://projects.ael.uni-tuebingen.de/backbone/moodle/
CallFriend	CallFriend Spanish corpus of telephone conversations	https://doi.org/10.21415/T5ZC76
CallHome	CallHome Spanish corpus of telephone conversations	https://doi.org/10.21415/T51K54
CCC	Columbia corpus de conversaciones para E/LE	https://edblogs.columbia.edu/corpusdeconversaciones/
CET	Corpus del español en Texas: Corpus del español en Texas, SpinTX video archive	http://corpus.spanishintexas.org
CHA	Corpus del habla en Almería	http://nevada.ual.es/otri/ilse/corpus.asp
CIEMPIESS	Corpus de investigación en español de México del Posgrado de Ingeniería Eléctrica y Servicio Social: CIEMPIESS, CIEMPIESS Light, CIEMPIESS Balance	http://www.ciempiess.org/downloads https://catalog.ldc.upenn.edu/LDC2017S23 https://catalog.ldc.upenn.edu/LDC2018S11
CLICC	Corpus lingüísticos del Instituto Caro y Cuervo: Corpus del español hablado en Bogotá, Corpus del habla culta de Bogotá	http://clicc.caroycuervo.gov.co/corpus/EHB http://clicc.caroycuervo.gov.co/corpus/HCB
COLA	Corpus oral de lenguaje adolescente	http://www.colam.org
COrDiAL	Corpus oral didáctico anotado lingüísticamente	http://lablita.it/app/cordial/index.php

COREC	Corpus oral de referencia de español en contacto	http://espanolcontacto.fe.uam.es/wordpress/corpus-oral-de-referencia/
COREMAH	Corpus español multimodal de actos de habla	https://coremah-1a37f.firebaseapp.com
COSER	Corpus oral y sonoro del español rural	http://www.corpusrural.es
DdE	Dialectoteca del español	http://dialects.its.uiowa.edu
DiapixFL	A bi-directional task-based corpus of learners' conversational speech	http://doi.org/10.7488/ds/139
DiEspa	Corpus de diálogos en español	http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/792-corpus-diespa-dialogos-en-espanol
ESLORA	Corpus para el estudio del español oral	http://galvan.usc.es/eslora/
Fonocortesía	Mecanismos fónicos para la expresión de cortesía y descortesía verbales en español coloquial	http://fonocortesia.es
HaCASpa	Hamburg Corpus of Argentinean Spanish	http://hdl.handle.net/11022/0000-0000-5F0B-B
MAVIR	Mejorando el acceso y visibilidad de la información multilingüe en red	http://www.lllf.uam.es/ESP/CorpusMavir.html
OpenProDat	Open Source Speech Database	https://hdl.handle.net/11403/openprodat/v2
PRESEEA	Proyecto para el estudio sociolingüístico del español de España y de América	http://preseea.linguas.net
RPD	The University of Toronto Romance Phonetics Database:	http://rpd.chass.utoronto.ca

	Romance Language Survey, Dialect Atlas of Argentina, Experimental Phonology	http://rpd.chass.utoronto.ca/docs/corpora_1.html http://rpd.chass.utoronto.ca/docs/corpora_2.html http://rpd.chass.utoronto.ca/docs/corpora_3.html
SPLLOC	Spanish Learner Language Oral Corpora: SPLLOC, SPLLOC1, SPLLOC2	http://www.splloc.soton.ac.uk/ http://doi.org/10.21415/T5TW27 http://doi.org/10.21415/T5RG7C
VHW	Voices of the Hispanic World	http://dialectos.osu.edu

Los recursos recogidos en la Tabla 2 ofrecen una amplia gama de contenidos, que responde a los diferentes objetivos de cada corpus. Algunos de ellos se orientan al análisis del discurso oral, otros se centran en aspectos sociolingüísticos, mientras que en algunos casos el corpus se ha creado para desarrollar tecnologías del habla; también se cuenta con corpus concebidos para la enseñanza del español como LE/L2, con recursos multilingües en los que el español es una de las lenguas que se han tomado en consideración y con materiales que contienen muestras de hablantes nativos de español, a modo de grupo de control, junto con datos procedentes de hablantes no nativos. Cabe destacar, en relación con el tema de este capítulo, los corpus más específicamente adaptados a los estudios relacionados con el plano fónico, sea en el nivel segmental o en el suprasegmental: el *Atlas interactivo de la entonación del español*, DiapixFL (García Lecumberri, Cooke y Wester 2017), DiEspa (Alfano *et al*. 2018), Fonocortesía (Hidalgo 2013), HaCASpa (Gabriel 2012), OpenProDat y la *Romance Phonetics Database*.

Como se puede observar en la Tabla 3, en algunos recursos es posible contar con el corpus completo para su uso local, mientras que en otros se accede a los datos mediante la consulta en la web ("en pantalla"), aunque existe la posibilidad de descargar documentos individuales en los que se recoge la transcripción ortográfica, la anotación del corpus o las grabaciones, sean en audio o en vídeo; en este último caso, para realizar un análisis fonético con herramientas como las descritas más adelante en el apartado 8.3.3 es necesario utilizar programas que extraen la señal sonora de un archivo de vídeo. El método de descarga de los datos puede variar en función del recurso, pues si algunos corpus ofrecen procedimientos de exportación automática, en otros es preciso recurrir a las opciones de descarga de documentos propias del sistema operativo que se esté empleando.

Tabla 3. Posibilidades de acceso a los datos en corpus orales de hablantes nativos de español, de dominio público y con acceso a las grabaciones.

Corpus	Transcripción ortográfica			Anotación			Grabaciones		
	En pantalla	Descarga		En pantalla	Descarga		En pantalla	Descarga	
		Documentos individuales	Corpus completo		Documentos individuales	Corpus completo		Documentos individuales	Corpus completo
AIEE	✓						✓	✓	
Backbone	✓	✓		✓			✓	✓	
CallFriend	✓		✓	✓		✓	✓	✓	
CallHome	✓		✓	✓		✓	✓	✓	
CCC	✓						✓	✓	
CET	✓	✓	✓		✓	✓	✓	✓	
CHA							✓	✓	
CIEMPIESS			✓			✓			✓
CLICC	✓						✓	✓	
COLA	✓			✓				✓	
COrDiAL		✓			✓			✓	
COREC	✓						✓		
COREMAH	✓	✓		✓	✓		✓	✓	
COSER	✓	✓		✓	✓		✓	✓	
Dde	✓						✓		

DiapixFL			✓			✓				✓
DiEspa			✓			✓				✓
ESLORA	✓		✓	✓		✓	✓	✓	✓	
Fonocortesía	✓	✓		✓	✓		✓	✓		
HaCASpa			✓			✓				✓
MAVIR	✓		✓	✓		✓	✓	✓	✓	
OpenProDat			✓							✓
PRESEEA	✓	✓		✓	✓				✓	
RPD									✓	
SPLLOC	✓	✓			✓				✓	
VHW	✓						✓		✓	

En los corpus que se encuentran en línea y disponen de una interfaz para su consulta, existen, como se muestra en la Tabla 4, tres posibilidades básicas de encontrar muestras de habla: a partir de los metadatos del corpus, ya ejemplificada en el apartado 8.2.1, localizando fenómenos específicos anotados en el corpus —por ejemplo, alargamientos, modalidad oracional, patrón melódico, segmentos que siguen y preceden al buscado o tipo de acto de habla— y a partir de la transcripción ortográfica; en este último caso, en algunos recursos se muestran los resultados en forma de concordancias, un procedimiento de presentación y análisis de los datos que se explica en los capítulos 6 y 7 del presente volumen.

Tabla 4. Posibilidades de búsqueda en corpus orales de hablantes nativos de español, de dominio público, con interfaz de consulta y con acceso a las grabaciones.

Corpus	Consulta a partir de los metadatos del corpus	Búsqueda de fenómenos anotados en el corpus	Búsqueda en la transcripción ortográfica	Concordancias
AIEE	✓			
Backbone	✓	✓	✓	✓
CCC	✓		✓	
CET	✓	✓	✓	
CHA	✓			
CLICC	✓		✓	✓
COLA	✓		✓	✓
COrDiAL	✓	✓		
COREMAH		✓	✓	
COSER	✓	✓	✓	
DdE	✓	✓		
ESLORA	✓	✓	✓	✓
Fonocortesía	✓	✓		
MAVIR			✓	
PRESEEA	✓		✓	
RPD	✓	✓	✓	
SPLLOC	✓		✓	
VHW	✓	✓		

8.3.2. Corpus orales de hablantes no nativos de español

Al igual que en el apartado anterior, los recursos mencionados a continuación son de dominio público —aunque en algún caso sea necesario crear una cuenta de usuario— y permiten la consulta de las grabaciones. En la Tabla 5 se recopilan algunos corpus que cumplen ambos requisitos y contienen datos procedentes de estudiantes de español como LE/L2.

Tabla 5. Corpus orales de estudiantes de español como LE/L2, de dominio público y con acceso a las grabaciones.

CIELE	Corpus de conversaciones en italiano y en español LE	http://www.linred.es/numero12_corpus-1.html
CORELE	Corpus oral de español como lengua extranjera (ELE)	http://cartago.lllf.uam.es/corele/index.html
COREMAH	Corpus español multimodal de actos de habla	https://coremah-1a37f.firebaseapp.com
DiapixFL	A bi-directional task-based corpus of learners' conversational speech	http://doi.org/10.7488/ds/139
DRC	Díaz Rodríguez Corpus	http://doi.org/10.21415/T5MW2C
Fono.ele	Adquisición y aprendizaje del componente fónico del español como lengua extranjera/segunda lengua	http://www3.uah.es/fonoele/corpus.php
INMIGRA	Nebrija-INMIGRA Corpus	http://doi.org/10.21415/T5GM44
LANGSNAP	Languages and Social Networks Abroad Project: LANGSNAP, LANGSNAP 3.0	http://langsnap.soton.ac.uk https://scholarcommons.usf.edu/langsnap/
Nicolás	Nicolás Corpus	http://doi.org/10.21415/T5J31F
OAP	Nebrija-OAP-English (Oral Academic Presentations)	http://doi.org/10.21415/T5GQ45
OCAE	Nebrija-OCAE (Oral Chinese Academic Emails)	http://doi.org/10.21415/T5BT4Z
RPD	The University of Toronto Romance Phonetics Database: Romance Language Survey, L2A of French and Spanish obstruent-liquid clusters	http://rpd.chass.utoronto.ca http://rpd.chass.utoronto.ca/docs/corpora_1.html http://rpd.chass.utoronto.ca/docs/corpora_a3.html
SPLLOC	Spanish Learner Language Oral Corpora: SPLLOC, SPLLOC1, SPLLOC2	http://www.splloc.soton.ac.uk/ http://doi.org/10.21415/T5TW27 http://doi.org/10.21415/T5RG7C
SPT	Spanish Corpus & Proficiency Level Training	http://www.laits.utexas.edu/spt/

Los corpus recogidos en la Tabla 5 ofrecen una amplia diversidad, tanto en lo que se refiere a las primeras lenguas de los estudiantes como a los estilos de habla recogidos. Algunos de los recursos se centran en una única L1 (en general, el inglés, aunque en CIELE es el italiano, en Nicolás, el árabe y en OCAE, el chino), mientras que en otros (CORELE, DRC, Fono.ele e INMIGRA) se encuentra representado el español de estudiantes con distintas lenguas maternas. Los estilos de habla que pueden encontrarse incluyen, principalmente, entrevistas semidirigidas y conversaciones semiespontáneas, aunque en algunos corpus también se recoge el estilo propio de la lectura. En ciertos casos se han utilizado tareas específicas como el juego de las diferencias, ya mencionado en el apartado 8.2.1, empleado en DiapixFL, la narración de una historia a partir de imágenes (en CORELE, LANGSNAP y SPLLOC, por ejemplo) o juegos de rol (COREMAH).

Al igual que sucede con los corpus de hablantes nativos, en los corpus de estudiantes de español como LE/L2 existen diversas posibilidades de acceso a los datos, tal como se recoge en la Tabla 6.

Tabla 6. Posibilidades de acceso a los datos en corpus orales de estudiantes de español como LE/L2, de dominio público y con acceso a las grabaciones.

Corpus	Transcripción ortográfica			Anotación			Grabaciones		
	En pantalla	Descarga		En pantalla	Descarga		En pantalla	Descarga	
		Documentos individuales	Corpus completo		Documentos individuales	Corpus completo		Documentos individuales	Corpus completo
CIELE			✓					✓	✓
CORELE	✓			✓			✓		
COREMAH	✓	✓		✓	✓		✓	✓	
DiapixFL			✓			✓			✓
DRC	✓		✓	✓		✓	✓	✓	
Fono.ele							✓	✓	
INMIGRA	✓		✓	✓		✓	✓	✓	
LANGSNAP		✓		✓				✓	
Nicolás	✓		✓	✓		✓	✓	✓	
OAP	✓		✓	✓		✓	✓	✓	
OCAE	✓		✓	✓		✓	✓	✓	
RPD								✓	
SPLLOC	✓	✓		✓				✓	
SPT	✓						✓		

Algunos recursos cuentan con una interfaz de consulta que permite realizar búsquedas de diversa naturaleza, como se muestra en la Tabla 7. En la mayoría de los casos se puede partir de los metadatos del corpus o de la representación ortográfica, mientras que la posibilidad de localizar los fenómenos anotados en el corpus resulta menos frecuente.

Tabla 7. Posibilidades de búsqueda en corpus orales de estudiantes de español como LE/L2, de dominio público, con interfaz de consulta y con acceso a las grabaciones.

Corpus	Consulta a partir de los metadatos del corpus	Búsqueda de fenómenos anotados en el corpus	Búsqueda en la transcripción ortográfica
CORELE	✓	✓	✓
COREMAH		✓	✓
Fono.ele	✓		✓
RPD	✓	✓	✓
SPLLOC	✓		✓
SPT	✓		

En lo que se refiere a los corpus específicamente orientados al estudio del nivel fónico, la *Romance Phonetics Database* ofrece un sistema de consulta muy detallado, tanto en lo que se refiere a las características de los hablantes como a las variables fonéticas (mostradas en la Figura 8), para acceder a dos corpus que contienen producciones en español como LE/L2: el *Romance Language Survey* y el *L2 acquisition of French and Spanish obstruent-liquid clusters*. Por su parte, en la versión pública de Fono.ele (Adquisición y aprendizaje del componente fónico del español como lengua extranjera/segunda lengua) (Blanco 2012) es posible seleccionar grabaciones —que pueden escucharse y descargarse usando el procedimiento habitual en el sistema operativo que se emplee— combinando las variables consideradas en el diseño del corpus (es decir, los metadatos) con la búsqueda en la representación ortográfica, como se muestra en la Figura 10.

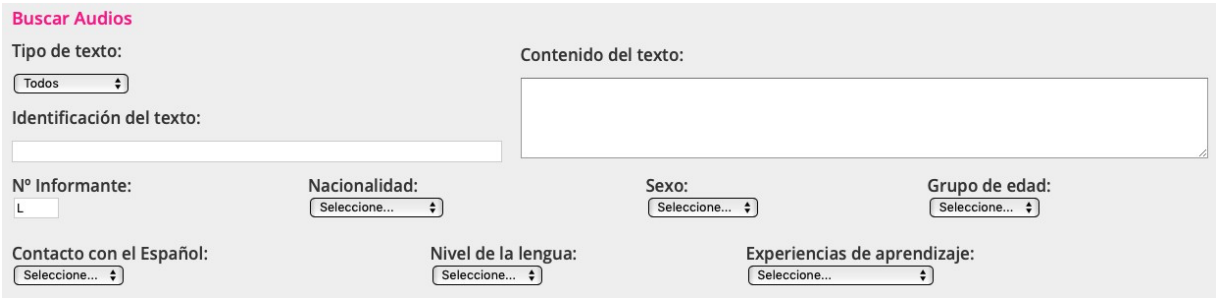

Figura 10. Variables consideradas en la búsqueda de grabaciones en el corpus Fono.ele.

En CORELE (Corpus oral de español como lengua extranjera) (Campillos 2012, 2013a), en cambio, pueden encontrarse muestras de habla partiendo del tipo de error. En el corpus se transcribieron y se etiquetaron los errores de pronunciación, distinguiendo entre los segmentales y los suprasegmentales y, entre estos últimos, los relacionados con el acento y los que se producen en la entonación; también se etiquetó el mecanismo que causa el error

(elisión, epéntesis y sustitución, entre otros), por lo que es posible obtener resultados como los que se reproducen en la Figura 11, en la que se muestra el primer resultado de una búsqueda de errores de pronunciación debidos a elisiones.

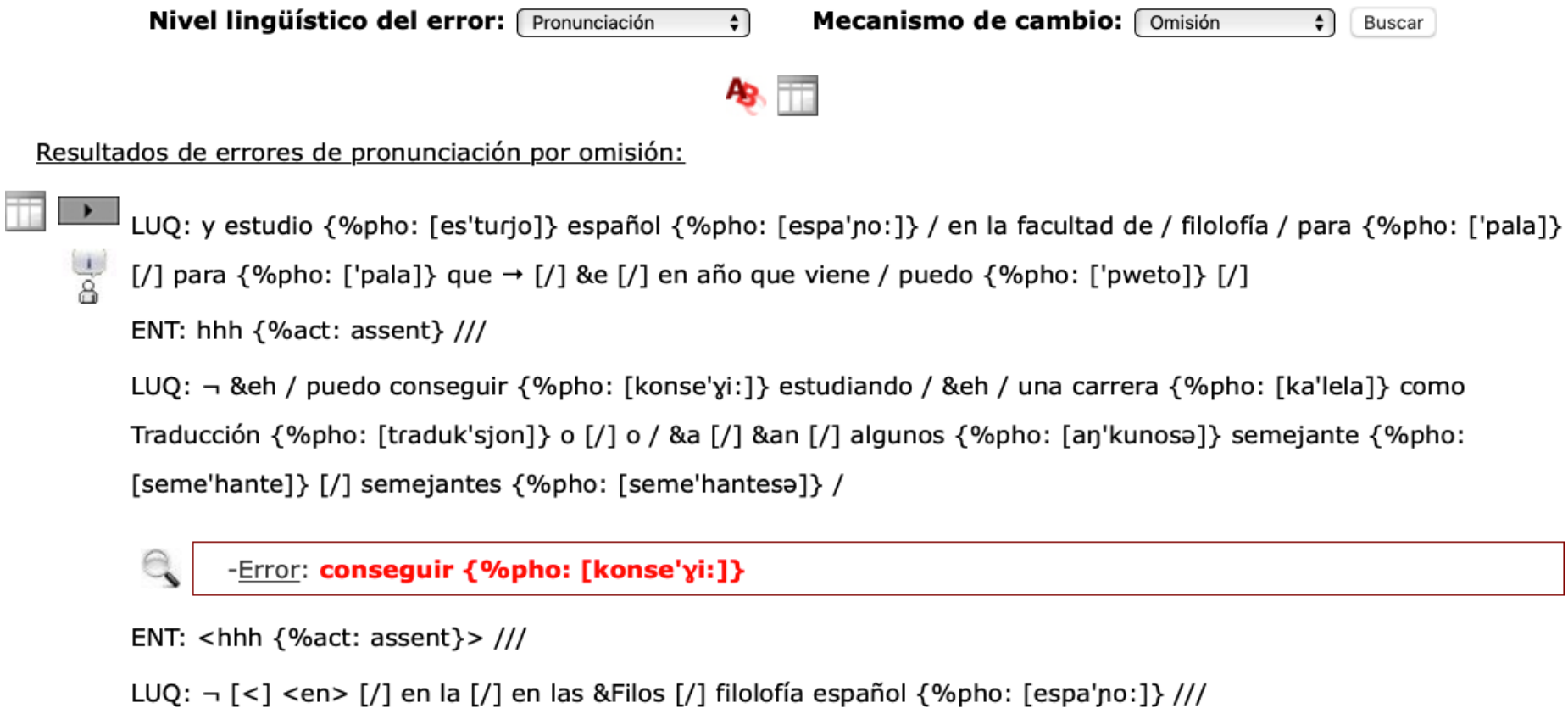

Figura 11. Primer resultado de la búsqueda de errores de pronunciación debidos a elisiones en el corpus CORELE. Haciendo clic en los iconos se puede escuchar el fragmento y obtener información sobre el estudiante y sobre el error, así como la frecuencia de aparición del tipo de error en el documento.

8.3.3. Herramientas para el tratamiento de corpus orales

Muy probablemente, la herramienta más utilizada en la actualidad para el análisis fonético de corpus orales sea Praat (Boersma 2014; Correa 2014), por su carácter abierto, multiplataforma y gratuito y por la posibilidad de automatizar tareas e incorporar diversos tipos de análisis con la ayuda de unos programas complementarios, también de dominio público, que se conocen como *scripts*.

Mediante Praat se lleva a cabo, en primer lugar, la segmentación de la señal sonora y, a continuación, se realiza el etiquetado, tal como se ha explicado en el apartado 8.1.2, de modo que se obtiene un documento, denominado *TextGrid*, que contiene tanto la anotación del corpus como la información necesaria para alinear la señal con las etiquetas. En este sentido, en recursos como CIEMPIESS (Hernández Mena y Herrera 2014) o DiapixFL (García Lecumberri, Cooke y Wester 2017) o HaCASpa (Gabriel 2012) el investigador puede descargar las grabaciones y los *TextGrid* asociados para disponer, así, de un corpus ya segmentado y etiquetado.

En el análisis de corpus multimodales es de uso muy habitual el programa ELAN (Sloetjes 2014), una herramienta también de dominio público y multiplataforma, especialmente concebida para anotar grabaciones en vídeo y que es compatible con Praat.

Por último, cabe considerar el empleo de herramientas que, además de segmentar y etiquetar un corpus, permiten organizar los archivos, realizar búsquedas o tratamientos estadísticos y obtener informaciones de diverso tipo. Entre las de dominio público y multiplataforma cabe mencionar Phon vinculada a las bases de datos de TalkBank (Rose y MacWhinney 2014), Dolmen, desarrollada en el marco del proyecto IPFC (*Interphonologie du Français Contemporain*) (Eychenne y Paternostro 2016), SPPAS (Bigi 2015) y las herramientas de EXMARaLDA (Schmidt y Wörner 2014).

El grado de complejidad y las prestaciones de cada programa difieren en ciertos aspectos, por lo que corresponde al investigador elegir el que mejor se adapte a sus propósitos. Esta elección requiere, como es lógico, contar con los conocimientos y el tiempo necesarios para explorar las posibilidades que ofrecen las tecnologías.

8.4. Conclusión

El uso de corpus orales abre un amplio abanico de posibilidades en la investigación del componente fónico en español como LE/L2, al tiempo que ofrece al profesor un conjunto de recursos que resultan de utilidad para la clase.

Sin embargo, conviene no olvidar que, frente a las innegables ventajas que presentan la investigación y la docencia basadas en corpus, existen una serie de limitaciones, entre las que destaca el número todavía insuficiente de recursos orientados al estudio del plano fónico. En este sentido, en un corpus adaptado a las necesidades del campo que nos ocupa se debería, idealmente, poner a disposición del usuario las grabaciones sincronizadas con las anotaciones, en un formato compatible con las herramientas y los estándares de mayor difusión, todo ello acompañado de una documentación exhaustiva y, si se trata de un recurso en línea, de una interfaz de consulta que facilite la búsqueda tanto en los metadatos como en las anotaciones.

Lograr este objetivo no es tarea fácil, pero contar con corpus orales adecuadamente diseñados, anotados, documentados y distribuidos, así como con profesionales específicamente formados para utilizarlos, constituye una necesidad ineludible para que la investigación sobre el nivel fónico en español como LE/L2 alcance su pleno desarrollo.

8.5. Bibliografía

Albelda, M. 2011. “Rentabilidad de los corpus discursivos en la didáctica de lenguas extranjeras”. En *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE. Actas del XXI Congreso Internacional de ASELE,* eds.

J. de Santiago, H. Bongaerts, J. J. Sánchez Iglesias y M. Seseña, 1:83-96. Salamanca: Asociación para la Enseñanza del Español como Lengua Extranjera. https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/21/21_0083.pdf

Alfano, I., R. Savy, S. Sbranna y L. Schettino. 2018. "Strategie discorsive in spagnolo L1 ed L2 a confronto: un'indagine su corpora dialogici". *CHIMERA: Romance Corpora and Linguistic Studies* 5 (1): 25-57. https://revistas.uam.es/index.php/chimera/article/view/9731

Bailini, S. 2014. "Los corpus como recursos didácticos para la enseñanza de las variedades diatópicas del español". En *¿Qué español enseñar y cómo? Variedades del español y su enseñanza. V Congreso Internacional de la Federación Internacional de Profesores de Español,* 1-17. Cuenca, España. https://www.educacionyfp.gob.es/dam/jcr:9879b92d-9541-4b3f-b2bd-67a58a86f26e/3--los-corpus-como-recursos-didacticos-para-la-ensenanza-de-las-variedades-diatopicas-del-espanol--bailinisonia-pdf.pdf

Ballier, N. y P. Martin. 2015. "Speech annotation of learner corpora". En *The Cambridge handbook of learner corpus research,* eds. S. Granger, G. Gilquin y F. Meunier, 107-134. Cambridge: Cambridge University Press.

Baqué, L. y M. Estruch. 2003. "Modelo de Aix-en-Provence". En *Teorías de la entonación,* ed. P. Prieto, 123-153. Barcelona: Ariel. https://6898baab-a-62cb3a1a-s-sites.googlegroups.com/site/lorrainebaqueuab/publis/ModeloAix-en-ProvenceV3.pdf

Bigi, Brigitte. 2015. "SPPAS – Multi-lingual approaches to the automatic annotation of speech". *The Phonetician* 111-112:54-69. http://www.isphs.org/Phonetician/Phonetician_111-112.pdf#page=54

Birch, B. 2014. "Data collection". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 27-45. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.004

Blanco, A. 2012. "Corpus oral para el estudio de la adquisición y aprendizaje del componente fónico del español como lengua extranjera". *Revista de Lingüística Teórica y Aplicada* 50 (2): 13-37. http://dx.doi.org/10.4067/S0718-48832012000200002

.2016. "La influencia de la lengua materna en la percepción fónica del español/L2". *Loquens* 3 (1): 1-11. https://doi.org/10.3989/loquens.2016.028

. 2017. "Habilidades de percepción fónica en español / L2 y su relación con el nivel de dominio lingüístico". *Estudios de Lingüística Aplicada* 65:59-82. https://doi.org/10.22201/enallt.01852647p.2017.65.727

Blanco, A. y M. Nogueroles. 2013. "Descripción y categorización de errores fónicos en estudiantes de español/L2. Validación de la taxonomía de errores AACFELE". *Logos. Revista de lingüística, filosofía y literatura* 23 (2): 196-225. https://revistas.userena.cl/index.php/logos/article/view/365

Blanco, A. y M. Nogueroles. 2014. "Errores fónicos de producción en español/L2: una propuesta de categorización". *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 27 (2): 255-274. https://doi.org/10.1075/resla.27.2.01can

Boersma, P. 2014. "The use of Praat in corpus research". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 342-360. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.016

Bonaventura, P., P. Howarth y W. Menzel. 2000. "Phonetic annotation of a non-native speech corpus". En *InSTIL 2000. Proceedings of the Workshop on Integrating Speech Technology in (Language) Learning,* 10-17. Dundee, Scotland, UK. https://pdfs.semanticscholar.org/129a/d09f0300fb6a43e62767fc2516ad0e94d271.pdf

Briz, A. y M. Albelda. 2009. "Estado actual de los corpus de lengua española hablada y escrita; I+D". En *El español en el mundo. Anuario del Instituto Cervantes 2009.*

Madrid: Instituto Cervantes. https://cvc.cervantes.es/lengua/anuario/anuario_09/default.htm

Broeder, D. y D. van Uytvanck. 2014. "Metadata formats". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 150-165. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.008

Caballero, M. 2015. "La variabilidad lingüística nativa y no nativa en escenarios comunicativos. La función de transacción en una situación cotidiana: corpus y descripción para el español (L1/LE)". Tesis doctoral, Universidad de Barcelona. http://hdl.handle.net/2445/101702

Caines, A., M. McCarthy y A. O'Keeffe. 2016. "Spoken language corpora and pedagogical applications". En *The Routledge handbook of language learning and technology,* eds. F. Farr y L. Murray, 348-361. Londres: Routledge.

Campillos, L. 2012. "La expresión oral en español lengua extranjera: interlengua y análisis de errores basado en corpus". Tesis doctoral, Universidad Autónoma de Madrid. http://hdl.handle.net/10486/660336

_____. 2013a. "Análisis de la producción y de errores en un corpus oral de español como lengua extranjera". *Revista Iberoamericana de Lingüística* 8:5-43. http://www.lllf.uam.es/ING/pdf/Campillos_RIL8_2013.pdf

Carranza, M. 2013. "Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus". En *Proceedings of SLaTE 2013. Interspeech 2013 Satellite workshop on Speech and Language Technology in Education,* eds. P. Badin, T. Hueber, G. Bailly, D. Demolin y F. Raby, 168-171. Grenoble, France. https://www.isca-speech.org/archive/slate_2013/sl13_168.html

Carranza, M. 2015. "The influence of intelligibility, comprehensibility and degree of foreign accent in evaluating and categorizing non-native pronunciation errors". En *Book of extended abstracts. Workshop on Phonetic Learner Corpora. Satellite Workshop of the 18th International Congress of Phonetic Sciences,* eds. J Trouvain, F. Zimmerer, M. Gósy y A. Bonneau, 17-19. Glasgow, Scotland. http://www.ifcasl.org/docs/Carranza_final.pdf

_____. 2016. "Transcription and annotation of spontaneous non-native spoken corpora". En *Technology-enhanced language learning for specialized linguistic domains: Practical applications and mobility,* eds. E. Martín Monje, I. Elorza y B. García Riaza, 216-227. Abingdon: Routledge.

Carranza, M., C. Cucchiarini, P. Burgos y H. Strik. 2014. "Non-native speech corpora for the development of computer-assisted pronunciation training systems". En *Edulearn14 Proceedings. 6th International Conference on Education and New Learning Technologies. July 7th-9th, 2014, Barcelona, Spain,* 3624-3633. València: IATED Academy.

Carranza, M., C. Cucchiarini, J. Llisterri, M. Jesús Machuca y A. Ríos. 2014. "A corpus-based study of Spanish L2 mispronunciations by Japanese speakers". En *Edulearn14 Proceedings. 6th International Conference on Education and New Learning Technologies. July 7th-9th, 2014, Barcelona, Spain,* 3696-3705. València: IATED. http://liceu.uab.cat/~joaquim/publicacions/Carranza_et_al_14_Corpus_Spanish_L2.pdf

Centre for English Corpus Linguistics. 2019. "Learner corpora around the world". Université Catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

CLARIN. 2018a. "L2 corpora". European Research Infrastructure for Language Resources and Technology. https://www.clarin.eu/resource-families/L2-corpora

_____. 2018b. "Language resource inventory". European Research Infrastructure for Language Resources and Technology. https://www.clarin.eu/content/language-resource-inventory

Correa, J. A. 2014. *Manual de análisis acústico del habla con Praat.* Bogotá: Instituto Caro y Cuervo. http://bibliotecadigital.caroycuervo.gov.co/id/eprint/998

Cucchiarini, C., H. van den Heuvel, E. Sanders y H. Strik. 2011. "Error selection for ASR-based English pronunciation training in 'My Pronunciation Coach'". En *Interspeech 2011. Proceedings of the 12th Annual Conference of the International Speech Communication Association,* eds. P. Cosi, R. de Mori, G. di Fabbrizio y R. Pieraccini, 1165-1168. Florence, Italy. https://www.isca-speech.org/archive/interspeech_2011/i11_1165.html

Delais-Roussarie, E. y B. Post. 2014. "Corpus annotation: Methodology and transcription systems". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 46-88. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.002

Delais-Roussarie, E. y Hi-Yon Yoo. 2011. "Learner corpora and prosody: From the COREIL corpus to principles on data collection and corpus design". *Poznań Studies in Contemporary Linguistics* 47 (1): 26-39. https://doi.org/10.2478/psicl-2011-0004

Díaz Sánchez, A. Sin fecha. "Indexador de corpus de aprendices de español". Universidad Complutense de Madrid. http://repositorios.fdi.ucm.es/corpus_aprendices_espa%c3%b1ol/view/paginas/view_paginas.php?id=1

Dong, Y. 2018. "El uso de corpus orales de español espontáneo para la enseñanza de dicha lengua". En *Monográficos SinoELE. Núm. 17. Actas del IX Congreso de la Asociación Asiática de Hispanistas (Bangkok, 2016),* 226-237. http://www.sinoele.org/images/Revista/17/monograficos/AAH_2016/AAH_2016_yang_dong.pdf

Drange, E.M. 2008. "Un corpus oral en línea como recurso didáctico". En *Multiculturalidad y norma policéntrica: aplicaciones en el aula de ELE. II Congreso Nacional de la Asociación Noruega de Profesores de Español,* 1-11. Bergen, Noruega. http://www.culturaydeporte.gob.es/dam/jcr:c1a7c40d-54bd-414d-a254-7d99fefa230b/2009-esp-09-03drange-pdf.pdf

Durand, J., U. Gut y G. Kristoffersen, eds. 2014. *The Oxford handbook of corpus phonology.* Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.001.0001

ELRA. 2018. "ELRA Catalogue". European Language Resources Association. http://catalog.elra.info/en-us/

Eychenne, J. y R. Paternostro. 2016. "Analyzing transcribed speech with Dolmen". En *Varieties of spoken French,* eds. S. Detey, J. Durand, B. Lacks y C. Lyche, D35-D52. Oxford: Oxford University Press.

Gabriel, C. 2012. "The Hamburg Corpus of Argentinean Spanish (HaCASpa)". En *Multilingual corpora and multilingual corpus analysis,* eds. T. Schmidt y K. Wörner, 183-198. Amsterdam: John Benjamins.

García Lecumberri, M.L., M. Cooke y M. Wester. 2017. "A bi-directional task-based corpus of learners' conversational speech". *International Journal of Learner Corpus Research* 3 (2): 175-195. https://doi.org/10.1075/ijlcr.3.2.04gar

Gilquin, G. 2015. "From design to collection of learner corpora". En *The Cambridge handbook of learner corpus research,* eds. S. Granger, G. Gilquin y F. Meunier, 9-34. Cambridge: Cambridge University Press.

Granger, S. 2008. "Leaner corpora". En *Corpus linguistics. An international handbook,* eds. A. Lüdeling y M. Kytö, 1:259-275. Berlin: Walter de Gruyter.

Gut, U. 2014. “Corpus phonology and second language acquisition”. En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 286-301. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.027

Hernández Mena, C. D. y J. A. Herrera. 2014. “CIEMPIESS: A new open-sourced Mexican Spanish radio corpus”. En *LREC 2014. Proceedings of the 9th International Conference on Language Resources and Evaluation,* 371-375. Reykjavik, Iceland. http://www.lrec-conf.org/proceedings/lrec2014/pdf/182_Paper.pdf

Hidalgo, A. 2013. “La fono(des)cortesía. Marcas prosódicas (des)corteses en español hablado. Su estudio a través de corpus orales”. *Revista de Lingüística Teórica y Aplicada* 51 (2): 127-150. http://hdl.handle.net/10550/33931

Hualde, J. I. 2003. “El modelo métrico y autosegmental”. En *Teorías de la entonación,* ed. P. Prieto, 155-184. Barcelona: Ariel.

IPA. Sin fecha. “The International Phonetic Alphabet and the IPA Chart”. International Phonetic Association. https://www.internationalphoneticassociation.org/content/ipa-chart

LDC. 2019. “LDC Catalog”. Linguistic Data Consortium. https://catalog.ldc.upenn.edu/

Llisterri, J. 1996. *Preliminary recommendations on spoken texts.* EAGLES Document EAG-TCWG-CTYP/P. LRE-61100 EAGLES, Expert Advisory Group on Language Engineering Standards. http://liceu.uab.cat/~joaquim/publicacions/EAGLES_86_Preliminary_recommendations_spoken_texts.pdf

Martín Sánchez, M. T., C. Pascual y M. Paz. 2017. “Creación de material didáctico para nivel A2 de ELE, a partir de conversaciones procedentes del Corpus Corinéi (Corpus oral de interlengua español/italiano)”. En *Investigación en docencia universitaria. Diseñando el futuro a partir de la investigación educativa,* ed. R. Roig-Vila, 969-979. Barcelona: Octaedro. http://hdl.handle.net/10045/71081

McCarthy, M. y A. O’Keeffe. 2013. “Analyzing spoken corpora”. En *The encyclopedia of applied linguistics,* eds. C. A. Chapelle, 104-112. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781405198431.wbeal0028

McWhinney, B. Sin fecha. “Index to TalkBank SLABank data on second language acquisition (SLA)”. Carnegie Mellon University. https://slabank.talkbank.org/access/

Munro, M. J. y T. M Derwing. 1995. “Foreign accent, comprehensibility, and intelligibility in the speech of second language learners”. *Language Learning* 45 (1): 73-97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Myles, F. 2005. “Interlanguage corpora and second language acquisition research”. *Second Language Research* 21 (4): 373-391.

Navarro Tomás, T. 2004. *Manual de pronunciación española.* 28.ª edición. Madrid: Consejo Superior de Investigaciones Científicas.

Neri, A., C. Cucchiarini y H. Strik. 2003. “Automatic Speech Recognition for second language learning: How and why it actually works”. En *Proceedings of the 15th International Congress of Phonetic Sciences,* eds. M. J. Solé, D. Recasens y J. Romero, 1157-1160. Barcelona, Spain. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1157.html

. 2006. “Selecting segmental errors in non-native Dutch for optimal pronunciation training”. *IRAL. International Review of Applied Linguistics in Language Teaching* 44 (4): 357-404. https://doi.org/10.1515/IRAL.2006.016

Nicolás, C. 2014. “Propuestas prácticas para el uso en el aula de C-Or-DiAL (Corpus Oral Didáctico anotado lingüísticamente)”. En *La enseñanza del español como LE/L2 en el siglo XXI. XXIV Congreso Internacional de ASELE,* ed. N. M. Contreras, 903-912.

Málaga: Asociación para la Enseñanza del Español como Lengua Extranjera. https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/24/24_903.pdf
Niebuhr, O. y A. Michaud. 2015. "Speech data acquisition: The underestimated challenge". *KALIPHO – Kieler Arbeiten zur Linguistik und Phonetik, Special Issue: "Theoretical and empirical foundations of experimental phonetics"* 3: 1-42. http://www.isfas.uni-kiel.de/de/linguistik/forschung/arbeitsberichte/kalipho-2-3
O'Brien, M. G., T. M. Derwing, C. Cucchiarini, D. M. Hardison, H. Mixdorff, R. I. Thomson, H. Strik *et al*. 2018. "Directions for the future of technology in pronunciation research and teaching". *Journal of Second Language Pronunciation* 4 (2): 182-207. https://doi.org/10.1075/jslp.17001.obr
Polo, N. 2018. "Recomendaciones para la confección de un corpus oral válido para el análisis fonético". *e-Scripta Romanica* 5: 71-79. https://doi.org/10.18778/2392-0718.05.07
Pustka, E, C. Gabriel, T. Meisenburg, M. Burkard y K. Dziallas. 2018. "(Inter-)Fonología del Español Contemporáneo (I)FEC: metodología de un programa de investigación para la fonología de corpus". *Loquens* 5 (1): 1-16. https://doi.org/10.3989/loquens.2018.046
Racine, I, F. Zay, S. Detey y Y. Kawaguchi. 2011. "De la transcription de corpus à l'analyse interphonologique: enjeux méthodologiques en FLE". En *Transcrire, écrire, formaliser: Actes du 24ème colloque du CERLICO, Université de Tours, juin 2010,* eds. G. Col y S. N. Osu, 13-30. Rennes: Presses Universitaires de Rennes.
Real Academia Española y Asociación de Academias de la Lengua Española. 2011. *Nueva gramática de la lengua española. Fonética y fonología.* Madrid: Espasa.
Romary, L. y A. Witt. 2014. "Data formats for phonological corpora". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 166-190. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.005
Rose, Y. y B. MacWhinney. 2014. "The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 308-401. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.023
Schmidt, T. y K. Wörner. 2014. "EXMARaLDA". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 402-419. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.030
Secchi, D. 2014. "Los corpus discursivos orales como recurso didáctico: la oralidad a través del texto en contexto". *Foro de profesores de E/LE* 10:241-250. https://ojs.uv.es/index.php/foroele/article/view/6673
Sloetjes, H. 2014. "ELAN: Multimedia annotation application". En *The Oxford handbook of corpus phonology,* eds. J. Durand, U. Gut y G. Kristoffersen, 305-320. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199571932.013.019
Solís, I. 2018. "Corpus españoles dialógicos para el análisis de la conversación". *CHIMERA: Romance Corpora and Linguistic Studies* 5 (1): 117-129. http://dx.doi.org/10.15366/chimera2018.5.1.010
TEI Consortium. 2018. "8 Transcription of speech". En *TEI P5: Guidelines for Electronic Text Encoding and Interchange,* Version 3.4.0. https://tei-c.org/release/doc/tei-p5-doc/en/html/TS.html
Wells, J. C. 1999-2015. "SAMPA Computer Readable Phonetic Alphabet". University College London. https://www.phon.ucl.ac.uk/home/sampa/index.html

Whichmann, A. 2008. “Speech corpora and spoken corpora”. En *Corpus linguistics. An international handbook,* eds. A. Lüdeling y M. Kytö, 187-207. Berlín: Walter de Gruyter.

8.6. Recursos en línea

- *Atlas interactivo de la entonación del español:* http://prosodia.upf.edu/atlasentonacion/
- *Backbone. Pedagogic Corpora for Content and Language Integrated Learning*: http://projects.ael.uni-tuebingen.de/backbone/moodle/
- *CallFriend – Spanish Corpus.* Trascripciones y grabaciones: http://doi.org/10.21415/T5ZC76.
- *CallHome – Spanish Corpus.* Transcripciones y grabaciones*:* http://doi.org/10.21415/T51K54
- *CIEMPIESS Balance.* Transcripciones y grabaciones: https://catalog.ldc.upenn.edu/LDC2018S11
- *CIEMPIESS Corpus.* Transcripciones y grabaciones: http://www.ciempiess.org/downloads
- *CIEMPIESS Light.* Transcripciones y grabaciones: https://catalog.ldc.upenn.edu/LDC2017S23
- *Columbia Corpus de Conversaciones para E/LE:* https://edblogs.columbia.edu/corpusdeconversaciones/
- *C-Or-DiAL. Acceso a las sesiones:* http://lablita.it/app/cordial/index.php
- *Corpus de conversaciones en italiano y en español LE (CIELE)*. Transcripciones y grabaciones: http://www.linred.es/numero12_corpus-1.html
- *Corpus del español en Texas*: http://corpus.spanishintexas.org/es
- *Corpus del español hablado en Bogotá:* http://clicc.caroycuervo.gov.co/corpus/EHB
- *Corpus del habla culta de Bogotá:* http://clicc.caroycuervo.gov.co/corpus/HCB
- *Corpus del habla en Almería*: http://nevada.ual.es/otri/ilse/corpus.asp
- *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América*: http://preseea.linguas.net/
- *Corpus DiEspa (Diálogos en Español).* Transcripciones y grabaciones: http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/792-corpus-diespa-dialogos-en-espanol
- *Corpus Español Multimodal de Actos de Habla*: https://coremah-1a37f.firebaseapp.com/
- *Corpus MAVIR*: http://www.lllf.uam.es/ESP/CorpusMavir.html
- *Corpus oral de español como lengua extranjera (CORELE):* http://cartago.lllf.uam.es/corele/index.html.
- *Corpus Oral de Lenguaje Adolescente (COLA):* http://www.colam.org/
- *Corpus Oral de Referencia de Español en Contacto (COREC)*: http://espanolcontacto.fe.uam.es/wordpress/corpus-oral-de-referencia/
- *Corpus Oral y Sonoro del Español Rural (COSER)*: http://www.corpusrural.es/
- *Corpus para el estudio del español oral (ESLORA)*: http://eslora.usc.es/
- *Dialectoteca del español*: http://dialects.its.uiowa.edu/
- *DiapixFL.* Transcripciones y grabaciones: http://doi.org/10.7488/ds/139
- *Díaz Rodríguez Corpus.* Transcripciones y grabaciones*:* http://doi.org/10.21415/T5MW2C
- *Dolmen: A program for the analysis of speech corpora*: http://www.dolmen-ling.org/

- *ELAN:* https://tla.mpi.nl/tools/tla-tools/elan/
- *EXMARaLDA*: https://exmaralda.org/en/
- *Fono.data*: http://www3.uah.es/fonoele/fono-data.php
- *Fono.ele*: http://www3.uah.es/fonoele/corpus.php
- *Fonocortesía. Mecanismos fónicos para la expresión de cortesía y descortesía verbales en español coloquial:* http://fonocortesia.es/
- *Hamburg Corpus of Argentinean Spanish (HaCASpa).* Transcripciones y grabaciones: http://hdl.handle.net/11022/0000-0000-5F0B-B
- *IRIS. A digital repository of data collection instruments for research into second language learning and teaching*: http://www.iris-database.org
- *LANGSNAP 3.0.* Transcripciones y grabaciones: https://scholarcommons.usf.edu/langsnap/
- *LANGSNAP*: http://langsnap.soton.ac.uk/
- *Nebrija-INMIGRA Corpus*. Transcripciones y grabaciones: http://doi.org/10.21415/T5GM44
- *Nebrija-OAP-English.* Transcripciones y grabaciones: http://doi.org/10.21415/T5GQ45
- *Nebrija-OCAE.* Transcripciones y grabaciones: http://doi.org/10.21415/T5BT4Z
- *Nicolas Corpus.* Transcripciones y grabaciones: http://doi.org/10.21415/T5J31F
- *OpenProDat – Open Source Speech Database.* Transcripciones y grabaciones: https://hdl.handle.net/11403/openprodat/v2
- *Phon:* https://phon.ca
- *Praat: Doing phonetics by computer:* http://www.praat.org
- *Spanish Corpus & Proficiency Level Training:* http://www.laits.utexas.edu/spt/
- *SpinTX. Authentic Spanish videos for language learning:* http://www.coerll.utexas.edu/spintx/
- *SPLLOC, Spanish Learner Language Oral Corpora:* http://www.splloc.soton.ac.uk/
- *SPLLOC1 Corpus.* Transcripciones y grabaciones: http://doi.org/10.21415/T5TW27
- *SPLLOC2 Corpus.* Transcripciones y grabaciones: http://doi.org/10.21415/T5RG7C
- *SPPAS: The automatic annotation and analysis of speech:* http://www.sppas.org/
- *The University of Toronto Romance Phonetics Database:* http://rpd.chass.utoronto.ca/
- *The University of Toronto Romance Phonetics Database: Dialect Atlas of Argentina:* http://rpd.chass.utoronto.ca/docs/corpora_2.html
- *The University of Toronto Romance Phonetics Database:Experimental phonology:* http://rpd.chass.utoronto.ca/docs/corpora_3.html
- *The University of Toronto Romance Phonetics Database: L2A of French and Spanish obstruent-liquid clusters:* http://rpd.chass.utoronto.ca/docs/corpora_a3.html
- *The University of Toronto Romance Phonetics Database: Romance Language Survey:* http://rpd.chass.utoronto.ca/docs/corpora_1.html
- *Voices of the Hispanic World:* http://dialectos.osu.edu/