

# Corpus orales de aprendientes de ELE para el estudio de la conversación en la L2 disponibles en Internet

Carlos GARCÍA RUIZ-CASTILLO

Instituto para la Investigación y la Enseñanza de Lenguas Extranjeras

Universidad de Hiroshima

En este trabajo pretendemos caracterizar aquellos corpus orales de aprendientes (CA) de español como lengua extranjera (ELE) que sean adecuados para el estudio de la interacción oral y de la conversación y que estén disponibles y accesibles en Internet actualmente. Para ello, en primer lugar, seleccionamos en indizadores y repertorios de CA de ELE aquellos corpus orales discursivos que incluyan muestras de conversación o interacción oral espontánea o semiespontánea y que ofrezcan acceso completo al texto mediante transcripciones y materiales audiovisuales. De los corpus seleccionados indicaremos el tipo de interacción que recogen, el objetivo seguido en su formación, las características de los participantes, ciertas particularidades en su diseño, la información ofrecida en las transcripciones y, en su caso, en sus anotaciones y etiquetado, y la disponibilidad de los materiales. Adicionalmente, basándonos en los rasgos con los que Albelda Marco (2022, p. 232) trata la rentabilidad de los corpus orales para la investigación de la pragmática, ofreceremos una visión de conjunto de los corpus seleccionados sobre su potencial uso para el estudio de la interacción y la conversación en aprendientes de ELE. En los CA orales de ELE seleccionados observamos cierta escasez de corpus formados por conversaciones espontáneas, ya que la mayoría se han obtenido mediante entrevistas semiestructuradas, además de ciertas limitaciones ya señaladas por otros autores (Rojo y Palacios, 2022, p. 85), tales como su limitada representatividad y la falta de homogeneidad en las pruebas para la obtención de los datos y en el tratamiento de los materiales.

## LA CONVERSACIÓN EN UNA LENGUA EXTRANJERA Y LOS CORPUS DE APRENDIENTES

En consonancia con el creciente número de CA de ELE desarrollados y publicados, actualmente contamos con algunos recursos que los recopilan y ofrecen una visión de conjunto. Entre estos recursos destacan ciertas publicaciones, como Rojo y Palacios (2022) y Rojo, Palacios, Sampedro Mella *et al.* (2022), así como los indizadores *Indexador de Corpus de aprendices de español*, de la Universidad Complutense de Madrid ([http://repositorios.fdi.ucm.es/corpus\\_aprendices\\_español/view/paginas](http://repositorios.fdi.ucm.es/corpus_aprendices_español/view/paginas)) y el *Learner corpora around the world*, de la Universidad Católica de Lovaina (<https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html>).

Por otro lado, se han producido avances en la investigación de la interacción oral inmediata y la conversación en aprendientes de ELE. En estos estudios encontramos objetivos y perspectivas teóricas variadas. A modo de ejemplo podemos citar los estudios sobre la gestión de la interacción y la estructura conversacional de aprendientes de ELE con diferentes L1 desarrollados principalmente bajo la dirección de Ana María Cestero Mancera y reunidos en un monográfico sobre la competencia conversacional de la revista *Lingüística en la Red* (volumen XIV, 2016), la perspectiva pragmática en el estudio de los actos de habla y de la atenuación adoptada en el *Corpus Español Multimodal de Actos de Habla* y en el *Corpus Oral de*

*Español como Lengua Extranjera de la Universidad de Alicante* (vid. *infra*) o el análisis de errores en la interlengua de aprendientes de ELE con diferentes L1 del *Corpus oral de Español como Lengua Extranjera* (vid. *infra*).

El ámbito de interés que guía este trabajo es la interacción oral inmediata y, en especial, la conversación de aprendientes de ELE. Justificaremos este interés y definiremos brevemente qué entendemos por conversación. De manera general, se puede considerar que la conversación es el modo de usar el lenguaje básico, primario, universal, independiente de la tecnología y que no requiere de habilidades especiales (Clark, 1996, pp. 8-9). La conversación cotidiana o coloquial se caracteriza por ciertos rasgos que la diferencian de otros tipos de interacciones orales: los participantes en la conversación tienen la misma categoría funcional o real, su organización general no está planificada ni convencionalizada, la alternancia de turnos se organiza en cada momento entre todos los participantes (sin estar predeterminada ni sometida al control o dirección de ningún participante en concreto) y se realiza con fines de comunicación interpersonal (Cestero Mancera, 2005, pp. 19-20). Algunos de estos rasgos y otros, como el tono informal, la proximidad entre los participantes y la cotidianidad del contexto de la interacción y de los temas tratados favorecen el grado de coloquialidad de la conversación (Briz Gómez y García-Ramón, 2021, p. 261). En cuanto a la interacción oral y la conversación, su importancia se reconoce, por ejemplo, en el *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación. Volumen complementario* del Consejo de Europa (2021), donde se afirma que “la interacción interpersonal es el origen del lenguaje” (p. 83) y se señala que las estrategias de interacción de la toma de turnos de palabra, de cooperación y de petición de aclaraciones, entre otras, son “tan importantes en el aprendizaje colaborativo como lo son en la comunicación del mundo real” (p. 83).

El estudio de la interacción oral y de la conversación en aprendientes de una lengua extranjera requiere de muestras reales recogidas en CA específicos. En este trabajo nos preguntamos cuáles de los CA orales de ELE que están actualmente disponibles en Internet son adecuados para dicho estudio.

## OBJETIVO Y CRITERIOS DE SELECCIÓN

El objetivo de este trabajo es caracterizar los CA orales de aprendientes de ELE en función de nuestro ámbito principal de interés: el estudio de la interacción oral y, en especial, la conversación. Consideraremos los corpus recogidos en varios recursos, parcialmente coincidentes. Consultamos dos publicaciones y repertorios de reciente elaboración (Rojo y Palacios, 2022; Rojo *et al.*, 2022, Rojo, Palacios, Sampedro Mella *et al.* (2022), los dos indizadores señalados en el apartado anterior (el *Indexador de Corpus de aprendices de español* y el *Learner corpora around the world*) y los corpus de aprendientes de ELE alojados en la sección SLABank del proyecto TalkBank (<https://slabank.talkbank.org/access/>).

Seleccionamos los corpus recogidos en los recursos señalados mediante los siguientes criterios:

1. Aspectos generales de los corpus: participantes y medio.
2. La forma de obtención de los datos y el tipo de interacción oral recogida en los corpus.
3. El tratamiento y acceso al corpus.
4. La disponibilidad del corpus.

En cuanto a los aspectos generales de los corpus, nos interesan aquellos formados con muestras de la producción de aprendientes de ELE, en oposición a corpus que recogen datos sobre la adquisición del español como lengua materna (Rojo *et al.*, 2022, p. 176). En cuanto al medio, han de ser corpus orales o multimodales.

Por lo que respecta al procedimiento de obtención de datos, este debe asegurar que la producción oral recogida corresponda realmente al tipo textual que se desea estudiar. Por ello, consideramos que un CA oral para el estudio de la conversación en ELE idealmente deberían estar formados por interacciones espontáneas entre participantes con la misma categoría funcional, sin planificación previa y, además, obtenidas mediante grabaciones secretas que aseguren la naturalidad y la ausencia de autorregulación en la producción (Pons Bordería, 2022, pp. 36-38). Es claro que, en el caso de los CA orales, este procedimiento es muy dificultoso.

Existen otros procedimientos que permiten obtener grabaciones de interacciones espontáneas o semiespontáneas de los participantes. En primer lugar, encontramos las conversaciones entre participantes con la misma categoría funcional (generalmente aprendientes de la L2) obtenidas de manera explícita por el investigador mediante una grabación no secreta, con cierta planificación y en contextos controlados. Este tipo de prueba preserva hasta cierto punto que sean los propios participantes los que gestionen la interacción y controlen la estructura conversacional, por lo que son apropiadas para el estudio de la conversación. En segundo lugar, podemos señalar las entrevistas semiestructuradas entre el investigador y el aprendiente en las que se permite cierta espontaneidad en las respuestas y en el desarrollo de la interacción. En este tipo de prueba, sin embargo, existe una asimetría en los roles de los participantes, ya que es el entrevistador quien controla la distribución de los turnos y la agenda temática. Como señala Solís García (2018, p. 118), “aunque las entrevistas llegan a ser espontáneas, una de las dos personas dirige el diálogo y el informante habla para que su testimonio lingüístico sea almacenado. Por este motivo no se dan las mismas dinámicas y estrategias comunicativas que en las conversaciones de tipo espontáneo en las que hay un reparto libre de papeles”. En tercer lugar, encontramos las simulaciones o juegos de roles, especialmente empleados en estudios pragmáticos sobre actos de habla. En los juegos de roles las interacciones suelen tener un fin pragmático determinado, por lo que tienden a finalizar una vez los participantes realizan el objetivo, al contrario que en las interacciones conversacionales, cuya principal finalidad es la comunicación interpersonal y tienden a prolongarse. Finalmente, consideramos que otras pruebas, entre ellas las narraciones a partir de elementos audiovisuales o las entrevistas estructuradas (tales como la compleción de enunciados cuya primera parte ofrece el entrevistador o la simple repetición de frases), son útiles para obtener datos acerca de la producción lingüística oral de los aprendientes, pero son escasamente adecuadas para el estudio de aspectos de la interacción.

Como tercer criterio de selección, seguimos a Albelda Marco (2022, pp. 230, 234) y Briz Gómez y Albelda Marco (2009, pp. 166-167), para distinguir entre “corpus discursivos o de acceso completo al texto” y los “corpus de acceso restringido a concordancias”. Los primeros ofrecen acceso a la grabación y a la transcripción completas, no solo a fragmentos, además de información contextual relativa a la situación de la interacción, rasgos sociolingüísticos de los participantes y desarrollo temático de la interacción. En contraste, los segundos no ofrecen acceso a la interacción completa, sino que se trata de bases de datos con tratamiento informático para el acceso a la información, como etiquetado y motores de búsqueda. Nos interesan, pues, los corpus discursivos o de acceso completo al texto.

Finalmente, en este trabajo solo tendremos en cuenta los corpus que, además de cumplir los otros criterios anteriormente indicados, estén disponibles actualmente de manera completa, abierta y gratuita en Internet.

## **CARACTERIZACIÓN DE LOS CADISPONIBLES PARA EL ESTUDIO DE LA CONVERSACIÓN EN ELE**

En los siguientes apartados caracterizaremos los corpus que hemos seleccionado. Según la forma de obtención de los datos y, por tanto, el tipo de interacción, los dividiremos entre corpus discursivos de interacciones diádicas semiespontáneas entre participantes con categoría funcional equivalente y corpus discursivos de interacciones obtenidas mediante entrevistas semiestructuradas. Con el objetivo de ofrecer una caracterización complementaria, y quizás más informativa y específica en función de nuestro ámbito de interés que la sintetizada en los recursos señalados en el apartado anterior, para cada corpus indicaremos con cierto grado de detalle el tipo de interacción, el objetivo seguido en la formación del corpus, características de los participantes, particularidades en su diseño, información ofrecida en las transcripciones y, en su caso, en las anotaciones y etiquetado, y disponibilidad de los materiales. La información que indicamos se ha obtenido principalmente de las páginas web de los corpus.

En un apartado final, caracterizaremos estos corpus mediante los rasgos relativos a la información contextual, atributos técnicos y acceso y disponibilidad que emplea Albelda Marco (2022, p. 232) para tratar la potencialidad de estos corpus en relación con los estudios pragmáticos.

En anexo ofrecemos un listado de los corpus analizados pero no incluidos en este trabajo y especificamos la razón de su exclusión.

### **CA de ELE discursivos formados por interacciones semiespontáneas entre participantes con categoría funcional equivalente y de acceso completo al texto y a la transcripción de la interacción**

Solo hemos encontrado tres corpus de aprendientes de ELE disponibles que recojan interacciones diádicas espontáneas o semiespontáneas y que ofrezcan acceso completo al texto y a la transcripción de la interacción.

#### *Corpus de conversaciones en italiano (CIELE)*

El CIELE (Pascual Escagedo, 2014; [https://linred.web.uah.es/numero12\\_corpus-1.html](https://linred.web.uah.es/numero12_corpus-1.html)) recoge conversaciones en ELE de aprendientes italianos y conversaciones en la lengua materna (L1) de los participantes. En total, el corpus está formado por 45 interacciones diádicas de 60 participantes de dos universidades situadas en la región de Campania que cuentan con facultades de Lenguas.

El corpus se organiza en tres grupos o subcorpus: a) un grupo de 15 conversaciones diádicas en ELE de estudiantes de segundo curso de universidad, con un nivel B1 en la L2, b) un grupo de 15 conversaciones diádicas en ELE de estudiantes de cuarto curso de universidad, con un nivel C1 y c) un grupo de 15 conversaciones en italiano y diádicas de los mismos participantes que los del grupo anterior. Este diseño permite realizar estudios sobre fenómenos de la conversación en dos grupos de diferente nivel de dominio de la lengua objeto y entre la lengua objeto y la lengua materna de los participantes. No se especifica si el nivel de dominio de la L2 está controlado formalmente.

Las conversaciones grabadas son espontáneas, sin organización temática o directrices establecidas previamente por el investigador, ausente durante la grabación. Sin embargo, no son secretas. Se pide a los participantes que interactúen en un espacio controlado, si bien familiar para ellos, con el dispositivo de grabación a la vista. Cada interacción dura 10 minutos aproximadamente. Las grabaciones se realizaron en 2010 y 2011.

Las transcripciones, realizadas con las convenciones del grupo Val.Es.Co. y del proyecto PRESEEA, ofrecen indicaciones de fenómenos de la oralidad y de la interacción, tales como los turnos de palabra, reinicios, pausas y su duración aproximada (en intervalos de medición de 0,5 segundos), superposiciones de habla, turnos de apoyo, alargamientos de sonido, etc. Se ofrece la transcripción de 5 minutos de cada interacción. Cada transcripción se acompaña de una ficha con los rasgos sociolingüísticos de los participantes y situacionales; los primeros son especialmente completos.

Las transcripciones están disponibles para descarga en formato PDF y los archivos de audio de las grabaciones en formato MP4 son reproducibles desde el navegador y descargables. Las transcripciones y archivos de audio no están alineados.

#### *Spanish Learner Language Oral Corpora (SPLLOC)*

El SPLLOC (Mitchell, Domínguez, Arche *et al.*, 2008; <http://www.splloc.soton.ac.uk/>) está formado por dos subcorpus, el SPLLOC1 y el SPLLOC2. Si bien ambos se recogieron mediante entrevistas semiestructuradas, solo el primero incluye también muestras de interacción semiespontánea entre pares. Nos centraremos en este trabajo en el SPLLOC1 y en la tarea de interacción semiespontánea, dado su mayor potencialidad para el estudio de la conversación.

El SPLLOC1 recoge textos orales de aprendientes de ELE cuya L1 es el inglés, organizados en tres grupos de 20 participantes cada uno según el nivel de dominio: a) inicial (de 13-14 años de edad y nivel aproximado A2), b) intermedio (de 17-18 años y nivel B1-B2) y c) avanzado (de 21-22 años y nivel C1-C2). El corpus permite la comparación con el español como L1, ya que incluye muestras de 10 hablantes nativos. El nivel de dominio no está controlado formalmente. Las grabaciones se realizaron entre 2006 y 2007 en instituciones educativas del Reino Unido.

La obtención de datos se realiza mediante tareas de producción semiespontánea y tareas estructuradas. Para cada tarea se ofrecen de manera independiente las transcripciones y los archivos de audio. Una de las tareas de producción semiespontánea es una conversación en parejas, entre dos aprendientes, en la que eligen uno o más temas de discusión (por ejemplo, el calentamiento global) y deben ponerse de acuerdo para ordenar una serie de afirmaciones o argumentos relativos a cada tema. Tanto los temas como las afirmaciones o argumentos son ofrecidos por el investigador. Esta tarea está diseñada, según los investigadores responsables del corpus, para obtener datos no solo sobre los recursos lingüísticos que los aprendientes emplean para expresar opinión y preferencia, sino también sobre su competencia en aspectos conversacionales tales como la toma de turnos y la reparación.

El corpus incluye 10 grabaciones de discusiones diádicas entre participantes pertenecientes al grupo intermedio y 11 grabaciones del grupo avanzado, además de 5 grabaciones del grupo de hablantes nativos. La duración de las discusiones es variable; en el caso del grupo avanzado, la grabación más breve y la más larga duran, respectivamente, 7:16 y 23:15 minutos. Las grabaciones no son secretas y el investigador

interviene en determinadas ocasiones: para ayudar en búsquedas de palabras, para confirmar si el interlocutor está de acuerdo, para organizar la discusión en función del orden asignado por cada participante a las ideas o argumentos de cada tema, etc.

Para cada discusión se ofrece una transcripción ortográfica, realizada con una versión adaptada del sistema CHAT del proyecto CHILDES. Las transcripciones recogen aspectos de la interacción estructurales, semánticos y de la oralidad, tales como el hablante que produce cada turno, habla solapada, pausas internas (sin indicación de su duración), final de turno interrumpido o suspendido, uso de estilo indirecto, corrección de errores en la producción, etc. La transcripción no está alineada con el audio ni ofrece marcado de tiempo. Además de la transcripción ortográfica en el sistema CHAT, en el corpus se ofrece una transcripción con etiquetado morfológico realizada de manera automática con el sistema MOR de CHILDES, pero solo para las conversaciones de nivel intermedio. Finalmente, los autores también ofrecen una transcripción en formato XML. La información contextual es muy limitada. En cuanto a los rasgos sociolingüísticos, apenas se nos ofrece el sexo de cada participante y el rango de edad de los grupos a los que pertenecen.

Las transcripciones están disponibles para su descarga en archivos de formato de texto (en el sistema CHAT y en XML). Los audios, en formato MP3, se pueden descargar.

#### *Corpus Español Multimodal de Actos de Habla (COR.E.M.A.H.)*

El COR.E.M.A.H. (<https://coremah.com/>), realizado por Marta Vacas Matos, es un corpus multimodal, con grabaciones audiovisuales, creado con la finalidad de estudiar los actos de habla y la cortesía en la producción oral de aprendientes de ELE cuya L1 es el inglés. Está organizado en dos grupos, uno de nivel intermedio (B1), formado por estudiantes universitarios que han curso al menos tres semestres de español, y uno de nivel avanzado (C1) con estudiantes de posgrado. El corpus incluye un tercer grupo como grupo de control, formado por hablantes nativos, por lo que permite estudios comparativos entre el español como L2 y como L1.

Los datos se obtienen mediante juegos de roles basados en tres actos de habla: rechazo a recibir ayuda (en relación con una tarea académica), cumplido (en concreto, en cuanto al dominio de la L2) y petición de disculpas (por no haber acudido a una cita). En cada grupo del corpus participaron 24 sujetos, de manera que el corpus recoge 12 interacciones por grupo, 36 por acto de habla y 108 en total.

Las instrucciones ofrecidas a los participantes para desarrollar las conversaciones basadas en estos juegos de roles son muy escuetas. Las interacciones, por ello, tienen un alto grado de libertad. Sin embargo, dado su fin pragmático, son de cierta brevedad. La duración media de las interacciones es de 177 segundos en el grupo de nivel intermedio, 135 en el grupo avanzado y 228,33 en el grupo de hablantes nativos.

En las transcripciones se han empleado las convenciones del grupo Val.Es.Co, con ciertas simplificaciones y modificaciones. Incluyen aspectos de la oralidad y las interacciones tales como alargamiento de sonidos, falsos inicios y palabras truncadas, solapamientos, pausas y silencios, pronunciación marcada, pronunciación silabeada, etc. Además, en cada acto de habla se han etiquetado las estrategias y los mitigadores e intensificadores empleados por los hablantes.

El corpus ofrece acceso completo a las grabaciones audiovisuales y a las transcripciones. Es posible realizar búsquedas por palabras y etiquetas en todas las transcripciones. Dentro de cada transcripción se pueden visualizar los fragmentos etiquetados mediante un código visual de colores.

Los rasgos sociolingüísticos recogidos para cada participante, especialmente completos, son: edad, sexo, datos sobre los estudios de L2 en el extranjero, nivel de ELE, tiempo de estudio de ELE, otras lenguas habladas, contacto con otras culturas y procedencia de sus profesores de ELE.

### **Corpus orales discursivos de aprendientes de ELE obtenidos mediante entrevistas semiestructuradas**

En comparación con los escasos corpus orales de interacción espontánea o semiespontánea disponibles, hemos encontrado un número mayor de corpus orales cuyos datos han sido obtenidos mediante una entrevista semiestructurada. En ellos, los procedimientos de obtención varían en cuanto al grado de libertad y espontaneidad que permiten en las respuestas. Como indicábamos, las entrevistas son un tipo de interacción en el que la toma de turnos está predeterminada y existe asimetría funcional entre los participantes, por lo que su utilidad para el estudio de la gestión de la interacción y de la estructura de la conversación se reduce. Sin embargo, sí permiten estudiar algunos aspectos relevantes de la interacción, como los mecanismos de reparación, el uso pragmático de los silencios, determinadas vocalizaciones como pausas oralizadas, entre otros. Las entrevistas se realizan en entornos controlados y no son secretas.

#### *Corpus Díaz Rodríguez*

El *Corpus Díaz Rodríguez* (<https://slabank.talkbank.org/access/Spanish/DiazRodriguez.html>) está formado por entrevistas semiestructuradas entre el investigador y ocho aprendientes con cinco L1 diferentes: islandés, chino, sueco, dos participantes de alemán y tres de coreano. Las grabaciones se realizaron en 1996 en un contexto de inmersión en Barcelona. Cada participante realizó la entrevista en varias ocasiones (entre dos y tres) y con un intervalo aproximado entre cada entrevista de uno o dos meses. Este diseño longitudinal permitiría observar el desarrollo de determinados fenómenos de la oralidad en la interlengua de los aprendientes.

Las entrevistas constan de tareas estructuradas, consistentes en descripción de dibujos y fotografías, repetición de frases, producción de frases interrogativas y relativas, entre otras pruebas, y una conversación semiespontánea con el entrevistador en el que este realiza determinadas preguntas sobre la profesión presente o futura del aprendiente y su familia. Esta prueba semiespontánea, sin embargo, parece haberse realizado solamente en una de las entrevistas en cada aprendiente, por lo que no permite analizar el desarrollo en la interlengua de aspectos relacionados con la interacción. Es, además, de duración relativamente breve.

El corpus está alojado en la sección SLABANK del proyecto TALKBANK. Es posible acceder a la transcripción de todas las entrevistas a los ocho participantes, pero solo al audio de las correspondientes a cuatro participantes. Las transcripciones están realizadas según el sistema CHAT y solo las de un participante están completamente alineadas con los archivos de audio.

La información contextual de cada grabación se limita a la fecha de la grabación y a los siguientes datos sociolingüísticos: sexo, edad y lengua materna de los participantes. Todos los participantes son mujeres. No se ofrece información sobre el nivel de dominio.

#### *Corpus oral de Español como Lengua Extranjera*

El *Corpus oral de Español como Lengua Extranjera* ([http://cartago.llf.uam.es/corele/home\\_es.html](http://cartago.llf.uam.es/corele/home_es.html)) es un corpus creado con el fin de analizar los errores en la producción oral de aprendientes de ELE con

diferentes L1. En él participan 40 aprendientes de ELE de nueve lenguas diferentes y con niveles de dominio de la lengua meta entre A2 y B1.

Los datos se obtienen mediante una entrevista semiestructurada entre el investigador y cada aprendiente, organizada en cuatro partes o pruebas diferentes: 1) presentación del estudiante, en la que el entrevistador realiza algunas preguntas sobre temas personales, tales como las lenguas que habla el aprendiente, la razón de su interés por el español o sus preferencias respecto al ocio; b) una narración basada en viñetas, procedente del Diploma intermedio de Español como Lengua Extranjera, en la que se incluye una pregunta sobre funciones comunicativas; c) descripción de dos fotografías sobre alimentación, con el objetivo de obtener datos sobre el vocabulario de dicho campo, y d) conversación y discusión con el entrevistador sobre la alimentación, realizada con el fin de obtener una interacción espontánea.

Las entrevistas tienen una duración de entre 13 y 30 minutos, aproximadamente, de forma que el corpus recoge más de una hora de grabaciones para cada una de las lenguas y más de 13 horas respecto al corpus en total. El nivel de dominio de la L2 no parece estar controlado formalmente.

De cada entrevista se ofrece una transcripción adaptada tanto del sistema CHAT como del corpus SPLLOC. Por ello, en las transcripciones se anotan fenómenos propios de la interacción, como interrupciones por otro hablante, autointerrupciones, solapamientos, turnos de apoyo, apoyos vocálicos o elementos paralingüísticos. Además, dado el interés en la interlengua de este corpus, en las transcripciones se añaden marcas para palabras en otros idiomas, formas verbales mal conjugadas, creaciones léxicas o pronunciaciones erróneas o no estándares. Por otro lado, en el corpus se han etiquetado los errores de la interlengua. La información y criterios empleados incluyen la categoría de la unidad lingüística, el tipo de error y nivel lingüístico en el que se produce (pronunciación, léxico, gramática, y nivel pragmático-discursivo) y el mecanismo de cambio implicado en el error.

Además del acceso a las transcripciones completas, gracias a su etiquetado lingüístico, el corpus ofrece un motor de búsqueda de errores y de palabras. Las transcripciones están alienadas con el audio e incluyen marcas de tiempo. Sin embargo, dado que la página web del corpus no ha sido actualizada, el sistema de audio integrado no funciona en las versiones actuales de los navegadores.

En el corpus se incluye información sobre rasgos sociolingüísticos de los participantes: sexo, edad, lengua materna, lenguas que habla, tiempo de estudio del español y tiempo de estancia en países hispanohablantes.

#### *Corpus Oral de Español como Lengua Extranjera de la Universidad de Alicante (CORELE-UA)*

El objetivo del corpus CORELE-UA (Medina Soler, 2017; <https://slabank.talkbank.org/access/Spanish/Nebrija-CORELE-UA.html>), es el de proporcionar datos para el estudio de la atenuación en el discurso oral de aprendientes de ELE. Las entrevistas, con un alto grado de espontaneidad, tienen la forma de conversación dirigida por el entrevistador sobre los siguientes temas relacionados con aspectos sociales y culturales: los españoles, costumbres y vida en la ciudad, comida y gastronomía, los jóvenes y la universidad y la situación económica. No incluyen pruebas estructuradas.

Recoge la producción discursiva de 10 aprendientes procedentes de varios países y con diferentes L1: alemán, polaco, italiano y lituano. Las grabaciones se realizaron en 2012 en la Universidad de Alicante. Los aprendientes tienen un nivel de dominio B1.

La duración de las grabaciones varía entre 3:43 y los 12:20 minutos, con un tiempo total del corpus de 1 hora y 27 minutos. El corpus está alojado en la sección SLABANK del proyecto TALKBANK. Es posible acceder a la transcripción de la interacción, realizada con el sistema CHAT, y al audio completo y alineado.

El corpus ofrece la siguiente información contextual y sociolingüística de cada grabación y participante: fecha de la grabación, lugar de nacimiento, lengua materna y sexo (seis mujeres y cuatro hombres).

### Corpus LANGSNAP y LANGSNAP 3.0

El corpus LANGSNAP (Tracy-Ventura, Huensch y Mitchell, 2021; <https://web-archive.southampton.ac.uk/langsnap.soton.ac.uk/index.html>) forma parte de un proyecto cuyo objetivo es el estudio longitudinal del aprendizaje del español y del francés como L2, así como de la posible relación entre las relaciones sociales y las oportunidades de uso de la lengua objeto y su aprendizaje en contexto de inmersión, en aprendientes cuya L1 es el inglés. Los participantes son estudiantes universitarios de especialidad, con un nivel avanzado de la L2 (no se especifica de otra forma el nivel de dominio) que realizan en el tercer o cuarto curso un programa de estancia en el extranjero. Se trata de un corpus oral y escrito. Su diseño longitudinal incluye seis fases en la recogida de datos: una previa a la estancia, tres durante la estancia de nueve meses y dos tras la estancia. Las grabaciones se realizaron entre 2011 y 2013. Además, el corpus permite la comparación entre la L2 y la L1, ya que incluye la producción de hablantes nativos que realizaron las mismas pruebas que los aprendientes.

En 2016, tres años después de las grabaciones que conforman el corpus LANGSNAP, se recogió de nuevo la producción oral y escrita de parte de los mismos aprendientes y se formó el corpus LANGSNAP 3.0 con el objetivo de estudiar el mantenimiento y desarrollo de la L2 tras la estancia de inmersión.

En el caso del español como L2, en el corpus LANGSNAP participan 27 aprendientes de la L2 y 10 hablantes nativos (8 de España y 2 de México), mientras que en el corpus LANGSNAP 3.0 participan 15 de los aprendientes de LANGSNAP.

La producción discursiva oral de los aprendientes se obtiene en ambos corpus mediante las siguientes tareas: una entrevista semiestructurada en la L2 centrada en las opiniones y experiencias sobre la estancia en el país de la lengua objeto, con preguntas adaptadas a cada ocasión, y narraciones basadas en ilustraciones. Las entrevistas tienen una duración de unos 20 minutos en el caso de LANGSNAP y de unos diez minutos en el corpus LANGSNAP 3.0.

Ambos corpus están alojados en la sección SLABANK del proyecto TALKBANK. Es posible acceder a la transcripción de la interacción, realizada con el sistema CHAT, y al audio completo y alineado. También es posible descargar los audios. La transcripción ofrece anotaciones sobre fenómenos interaccionales y de la oralidad, tales como pausas (no medidas), solapamientos y vocalizaciones. Además, las transcripciones están anotadas lingüísticamente de manera automática.

### *Spanish Corpus Proficiency Level Training (SPT)*

El objetivo del SPT (<https://www.laits.utexas.edu/spt/training/>) es ayudar a los usuarios del corpus a familiarizarse con la evaluación del nivel de dominio de ELE. Para alcanzar este objetivo, la página web del corpus propone una rúbrica y unos vídeos con los que emplearla. Dicha rúbrica se basa, de manera general, en las guías elaboradas por el American Council on the Teaching of Foreign Languages. Está formada por

cinco niveles y los correspondientes criterios relativos a la exactitud gramatical y léxica, fluidez y habilidades relacionadas con la conversación de los aprendientes. Las grabaciones audiovisuales con la producción de los estudiantes y su transcripción forman el corpus propiamente dicho.

El corpus recoge entrevistas semiestructuradas a 38 participantes con diferente nivel de dominio de ELE. Si bien no se ofrece esta información de manera explícita, se puede suponer que la L1 de los participantes es el inglés norteamericano y que estos tenían alguna relación con la universidad en la que se llevó a cabo la investigación. Sí se especifica que, de estos 38 participantes, 16 son hablantes de español como lengua de herencia. Las grabaciones se realizaron posiblemente entre 2006 y 2011.

Los datos se obtienen mediante una entrevista semiestructurada sobre los siguientes temas: familia, ciudad natal, comparación de la ciudad natal con la ciudad de la universidad, estudios en la universidad, otros países, deportes, vivienda, política y acontecimientos en el pasado. Por cada participante y tema se ofrece la grabación audiovisual (disponible para descarga) y la transcripción en formato texto. La duración de las grabaciones es variable, aproximadamente entre uno y dos minutos en cada tema. En las transcripciones se anotan el interlocutor que produce el turno y ciertos aspectos de la oralidad, como pausas (sin medición de su duración), vocalizaciones, autointerrupciones y solapamientos de habla. No hemos encontrado fichas técnicas de las grabaciones ni datos sociolingüísticos de los participantes.

### **Rentabilidad de los CA disponibles para el estudio de la conversación en ELE**

En la tabla 1 sintetizamos algunos rasgos de los CA orales de ELE detallados en los apartados anteriores. Los rasgos sobre la caracterización contextual y técnica de los CA están basados en los que emplea Albelda Marco (2022, p. 232) para los corpus discursivos del español hablado. Esperamos que esta tabla sirva como aproximación a la rentabilidad o potencial uso de los CA orales para el estudio de la conversación e interacción oral en aprendientes de ELE.

**TABLA 1. Rentabilidad de los CA orales de ELE para el estudio de la interacción oral y la conversación**

CIELE	SPLLOC1	CORE.M.A.H.	C. Díaz Rodríguez	C. Oral ELE	CORELE-UA	LANGSNAP	SPT
<i>Caracterización contextual</i>							
Tipo de grabación	No secreta	No secreta	No secreta	No secreta	No secreta	No secreta	No secreta
Tipo de interacción	Conversación libre con alto grado de espontaneidad	Juegos de roles	Entrevista semi-estructurada	Entrevista semi-estructurada	Entrevista semi-estructurada	Entrevista semi-estructurada	Entrevista semi-estructurada
Transcripción con signos de la oralidad de la interacción completa	✓ Convenciones Val.Es.Co. y PRESEEA adaptadas	✓ Convenciones CHAT	✓ Convenciones CHAT	✓ Convenciones CHAT adaptadas	✓ Convenciones CHAT	✓ Convenciones CHAT	✓
Acceso a la grabación oral de la interacción completa	✓	✓	Solo de algunas grabaciones	No (desactualización de la página web)	✓	✓	✓
Fichas técnicas o metadatos con información contextual (rasgos sociolingüísticos y situacionales)	✓ Escasos rasgos socio-lingüísticos	✓	✓ Escasos rasgos socio-lingüísticos	✓	✓ Escasos rasgos socio-lingüísticos	✓ Escasos rasgos socio-lingüísticos	✓
<i>Caracterización técnica</i>							
Incorporación de motor de búsqueda	✓	✓	✓ TalkBank	✓	✓ TalkBank	✓ TalkBank	✓
Alineación audio-transcripción			✓ Solo algunas grabaciones	No (desactualización de la página web)	✓	✓	
Incorporación de vídeo		✓					✓

## CONSIDERACIONES FINALES

En este trabajo hemos caracterizado de manera individualizada y en su conjunto los CA orales de ELE disponibles y adecuados para el estudio de la interacción oral y de la conversación. Uno de los primeros aspectos de los corpus que deseamos destacar se relaciona con la distinción que establece Pons Bordería (2022, pp. 15-16) entre *corpus con anzuelo* y *corpus de arrastre*. Según este autor, el desarrollo de los primeros responde a la tarea particular de un investigador o grupo de investigadores que buscan alcanzar un objetivo de investigación determinado, por lo que suelen ser de pequeño tamaño; los segundos, por el contrario, son proyectos de mayor envergadura que involucran a numerosos equipos y que buscan ofrecer datos y materiales a los investigadores para que realicen diferentes estudios. En el caso de los CA de ELE, si bien existen corpus escritos que corresponden a la segunda categoría, como el *Corpus de aprendices de español* (CAES) o el *Corpus Escrito del Español L2* (CEDEL2), todos los orales pertenecen a la primera.

El carácter de *corpus con anzuelo* resulta, sin duda, en ciertas limitaciones de los CA orales de ELE que han señalado otros autores. Así, Rojo y Palacios (2022, p. 85) consideran que, frente a CA realizados en otras L2, los CA orales de ELE presentan escasa variedad de L1 y de familias lingüísticas, limitada representatividad por el reducido número de participantes, ausencia de anotación y lematización y falta de homogeneidad en cuanto a la forma de obtención de los materiales y su transcripción.

Además de estas carencias, la principal que nosotros observamos es la de muestras obtenidas mediante grabaciones secretas o, en su defecto, de corpus en los que se recoja la conversación espontánea o semiespontánea en ELE entre participantes con igual categoría funcional. En el primer apartado de este trabajo señalábamos la importancia que tiene la conversación tanto en el aprendizaje como en el uso de la L2. Es necesario contar con más CA orales de conversaciones espontáneas de aprendientes de ELE en los que estén representadas más L1, de más familias lingüísticas y con participantes con diferentes niveles de dominio. Si disponemos de esas herramientas, podremos profundizar nuestros conocimientos sobre ciertos aspectos fundamentales de la interlengua y de la competencia comunicativa de aprendientes de ELE, tales como el desarrollo de la capacidad de interactuar oralmente en los diferentes niveles de dominio, o la posible influencia de las características lingüísticas de la L1 y de las prácticas sociales de la cultura de origen en la forma de conversar de estos aprendientes.

## AGRADECIMIENTOS

Esta investigación ha sido financiada por *kakenhi* (23K00698) *Grant-in-Aid for Scientific Research* de la Japan Society for the Promotion of Science.

## REFERENCIAS

- Albelda Marco, M. (2022). Los corpus del español hablado y los estudios pragmáticos. En G. Parodi, P. Cantos-Gómez, y C. Howe (Eds.), *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics* (pp. 223–238). Routledge.
- Briz Gómez, A., y Albelda Marco, M. (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D. En *Anuario del Instituto Cervantes 2009* (pp. 165–226). Instituto Cervantes.
- Briz Gómez, A., y García-Ramón, A. (2021). La conversación coloquial como prototipo de lo dialogal. En Ó. Loureda y A. Schrott (Eds.), *Manual de lingüística del hablar* (pp. 261–286). De Gruyter. <https://doi.org/10.1515/9783110611111-014>

org/doi:10.1515/9783110335224-014

- Cestero Mancera, A. M. (2005). *Conversación y enseñanza de lenguas extranjeras*. Arco Libros.
- Clark, H. H. (1996). *Using language*. Cambridge University Press.
- Consejo de Europa. (2021). *Marco común europeo de referencia para las lenguas: Aprendizaje, enseñanza, evaluación. Volumen complementario*. Ministerio de Educación y Formación profesional e Instituto Cervantes.
- Medina Soler, I. (2017). *La atenuación en el discurso oral de estudiantes de E/LE universitarios con nivel B1 en contexto de inmersión para los actos de habla disintivos* [Tesis doctoral]. Universidad Antonio de Nebrija.
- Mitchell, R., Domínguez, L., Arche, M. J., Myles, F., y Marsden, E. (2008). SPLLOC: A new database for Spanish second language acquisition research. *EUROSLA Yearbook*, 8, 287–304. <https://doi.org/10.1075/eurosla.8.15smit>
- Pascual Escagedo, C. (2014). Corpus de conversaciones en italiano y en español LE (CIELE). *Linred: Lingüística en la Red*, 12. [https://linred.web.uah.es/numero12\\_corpus-1.html](https://linred.web.uah.es/numero12_corpus-1.html)
- Pons Bordería, S. (2022). *Creación y análisis de corpus orales: Saberes prácticos y reflexiones teóricas*. Peter Lang.
- Rojo, G., y Palacios, I. (2022). Los corpus de aprendientes de español como L2. En G. Parodi, P. Cantos-Gómez, y C. Howe (Eds.), *Lingüística de corpus en español / The Routledge Handbook of Spanish Corpus Linguistics* (pp. 74–88). Routledge.
- Rojo, G., Palacios, I., Sampedro Mella, M., y Marsily, A. (2022). Los corpus de aprendices de español LE/L2: Panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching*, 9(2), 174–189. <https://doi.org/10.1080/23247797.2022.2157085>
- Solís García, I. (2018). Corpus españoles dialógicos para el análisis de la conversación. *CHIMERA: Revista de Corpus de Lenguas Romanes y Estudios Lingüísticos*, 5(1). <https://doi.org/10.15366/chimera2018.5.1.010>
- Tracy-Ventura, N., Huensch, A., y Mitchell, R. (2021). Understanding the Long-Term Evolution of L2 Lexical Diversity: The Contribution of a Longitudinal Learner Corpus. En B. Le Bruyn y M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 148–171). Cambridge University Press. <https://doi.org/10.1017/9781108674577.008>

## ANEXO

Enumeramos a continuación los corpus que hemos examinado pero que no han sido incluidos en este trabajo. Indicamos el nombre del corpus, su abreviatura, la autoría y la institución, y especificamos la razón de la exclusión según sea una de las siguientes: a) corpus de acceso restringido o publicado solo parcialmente en estudios; b) corpus reseñado como de acceso abierto, pero no accesible por razones técnicas en el momento de la realización de este trabajo; c) corpus que incluye datos orales que, por su forma de obtención, no permiten estudiar la interacción espontánea o semiespontánea; d) corpus no discursivos o restringidos a bases de datos o concordancias.

- *Corpus de Conversaciones en Español como Lengua Extranjera*. M. García (Universitt Bayreuth). a)
- *The University of Toronto Romance Phonetics Database* (RPD). L. Colantoni y J. Steele (University

- of Toronto). a), d)
- *Corpus Oral de Peticiones en Interacciones Naturalizadas en Español* (COPINE) A. Marsily (Université catholique de Louvain). a)
  - *Corpus Oral del Español en Taiwán* (COET). M. Rubio Lastra (Tunghai University). a)
  - *Corpus de Aprendices Taiwaneses de Español* (CATE). H.C. Lu (Universidad de Cheng Kung). b)
  - *The New England Corpus of Heritage and Second Language Speakers*. L. Amaral y P. Gubitosi, (University of Massachusetts Amherst). b)
  - *Multilingual Corpus of Second Language Speech* (MUSSEL), University of Utah Second Language Teaching & Research Center. b)
  - *Corpus de español hablado y escrito por sinohablantes*. (CorSinoELE). Xingxing Yin, Rosa Ana Martín Vegas (Universidad de Salamanca). c)
  - *Corpus Escrito del Español L2* (CEDEL2). C. Lozano y Grupo Woslac (Universidad de Granada). c)
  - *BCN-L2 Spanish Corpus*. A. Bel Gaya (Universitat Pompeu Fabra). c)
  - *Leiden Learner Corpus*. E. Mauder, M.C. Parafita Couto y J. Caspers (Universiteit Leiden). c)
  - *Corpus Fono.ele* (Fono.ele). Equipo Fono.ele. d)
  - *Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua*. (CAELE/2). D. Hincapié (Instituto Caro y Cuervo). d)

## ABSTRACT

### **Characteristics of Publicly Available Learner Corpora for the Study of Oral Interaction and Conversation in Spanish as a Foreign Language**

Carlos GARCÍA RUIZ-CASTILLO

Institute for Foreign Language Research and Education

Hiroshima University

In this article, my aim is to categorize the oral corpora of Spanish as a Foreign Language (SFL) learners that are currently available and accessible on the Internet. To do this, I first select from learner corpora indexes and repertoires those discursive oral corpora that include samples of spontaneous or semi-spontaneous oral interaction and that offer complete access to the text through transcriptions and audiovisual materials. Of the selected corpora, I indicate the type of interaction they collect, the objective followed in their formation, the characteristics of the participants, certain particularities in their design, the information offered in the transcriptions, and the availability of the materials. Additionally, I give an overview of their potential use in studies related to interaction and conversation. I build on the features with which Albelda Marco (2022, p. 232) discusses the scope of oral corpora in relation to pragmatic research. In this study, aspects already noted by Rojo and Palacios (2022, p. 85) are observed concerning the limited representativeness of SFL oral learner corpora and their lack of homogeneity in the collection and treatment of materials. In addition to these aspects, I discuss the scarcity of corpora formed by spontaneous conversations of SFL learners, since the main form of elicitation in the analyzed oral corpora is the semi-structured interview.

## 要 旨

### 一般公開されているスペイン語学習者コーパスの特徴と オーラル・インタラクションおよび会話研究における有用性

ガルシア ルイス カステージョ・カルロス  
広島大学外国語教育研究センター

本論文では、現在インターネット上で一般に公開されているスペイン語学習者の会話コーパスを分類・整理する。そのためにまず、学習者コーパスの索引や標題の中から、自発的／半自発的な会話のサンプルを含み、文字起こしや視聴覚資料により完全なテキストにアクセス可能なものを抽出した。そして、それらのコーパスで収集された会話の種類、作成の目的、被験者の属性・適格条件、文字起こしや視聴覚資料により提供される情報、資料の可用性といった情報を示す。加えて、オーラル・インタラクションや会話研究における潜在的有用性について述べる。これらは Albelda Marco (2022, p. 232) が語用論的研究の観点から論じる特徴に基づくものである。本研究を通して、Rojo and Palacios (2022, p. 85) が既に指摘しているスペイン語学習者コーパスの限定的代表性および資料収集とその扱いにおける不均質性の問題に加えて、スペイン語学習者の自発的な会話から作成されたコーパスの不足が確認できた。これは分析したオーラルコーパスにおける誘発形式の大半が半構造化されたインタビューであることに起因すると考えられる。