doi: 10.15446/fyf.v31n2.74659

CORPUS DE APRENDIENTES DE ESPAÑOL COMO LENGUA EXTRANJERA Y SEGUNDA LENGUA (CAELE/2)*: EL COMPONENTE ESCRITO**

*Diana Hincapié****
Instituto Caro y Cuervo, Colombia

Resumen

Este artículo expone el proceso de diseño y construcción del Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CAELE/2) en su componente escrito. Este es el primer corpus de aprendientes de español como lengua extranjera y segunda lengua en Colombia. Tiene como objetivo ofrecer a docentes e investigadores muestras reales de lengua de aprendientes extranjeros y sordos, con el fin de llevar a cabo análisis y propuestas didácticas que beneficien a estas poblaciones. El CAELE/2 es un corpus abierto con 166 muestras sistematizadas, tomadas de 83 participantes quienes respondieron a las pruebas diseñadas para el proyecto.

Palabras clave: *corpus de aprendientes; español; aprendientes extranjeros; aprendientes sordos; lingüística de corpus.*

Cómo citar este artículo:
Hincapié, D. (2018). Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CAELE/2): el componente escrito. *Forma y Función*, *31*(2), 129-143.

Artículo de investigación. Recibido: 06-02-2018, aceptado: 19-06-2018

* El CAELE/2 no guarda ninguna relación con el proyecto denominado *Corpus de aprendices de español como lengua extranjera* (CAELE) de Ferreira y Elejalde, 2014, aunque son similares en su denominación.

** Artículo producto del proyecto de investigación CAELE/2 (Corpus de aprendientes de español como lengua extranjera/segunda lengua), Fase 1, avalado por el Instituto Caro y Cuervo, 2017.

*** Máster en Enseñanza de español como lengua extranjera de la Universidad Pontificia de Salamanca y FIDESCU, y Máster en Neuropsicología y educación de la Universidad Internacional de la Rioja. Docente-investigadora en la maestría en Enseñanza del Español como Lengua Extranjera y Segunda Lengua del Instituto Caro y Cuervo, y coordinadora del diplomado en Pedagogía y Didáctica para la Enseñanza del Español como Lengua Extranjera, modalidad virtual. diana.hincapie@caroycuervo.gov.co

LEARNER CORPUS OF SPANISH AS A FOREIGN LANGUAGE AND SECOND LANGUAGE (CAELE/2): THE WRITTEN COMPONENT

Abstract

The article explains the process of design and construction of the written component of the Learner Corpus of Spanish as a Foreign Language and Second Language (CAELE/2), which is the first of its kind in Colombia. Its objective is to provide teachers and researchers with actual language samples collected among foreign and deaf learners, so as to foster analyses and proposals that benefit these populations. CAELE/2 is an open corpus that includes 166 systematized samples provided by 83 participants who took the tests designed for the project.

Keywords: *learner corpus; Spanish; foreign learners; deaf learners; corpus linguistics.*

CORPUS DE APRENDIZES DE ESPANHOL COMO LÍNGUA ESTRANGEIRA E SEGUNDA LÍNGUA (CAELE/2): O COMPONENTE ESCRITO

Resumo

Este artigo expõe o processo de projeto e construção do Corpus de Aprendizes de Espanhol como Língua Estrangeira e Segunda Língua (CAELE/2) em seu componente escrito. Este é o primeiro corpus de aprendizes de espanhol como língua estrangeira e segunda língua na Colômbia. Tem como objetivo oferecer a docentes e pesquisadores mostras reais de língua de aprendizes estrangeiros e surdos, com a finalidade de executar análise e propostas didáticas que beneficiem essas populações. O CAELE/2 é um corpus aberto com 166 mostras sistematizadas, tomadas de 83 participantes que responderam às provas desenhadas para o projeto.

Palavras-chave: *aprendizes estrangeiros; aprendizes surdos; corpus de aprendizes; espanhol; linguística de corpus.*

Universidad Nacional de Colombia, Facultad de Ciencias Humanas, Departamento de Lingüística

Introducción

La academia y los docentes de español como lengua extranjera y segunda lengua han fijado su atención en la creación y explotación de corpus de aprendientes en los últimos años. Esto ha sido motivado por los avances investigativos en la lingüística de corpus y el desarrollo que ha tenido esta área en relación con la enseñanza de lenguas como el inglés.

La preocupación actual en este campo recae en la falta de investigación descriptiva, teórica y aplicada, razón por la cual se hace cada vez más necesaria e inaplazable la construcción de corpus de aprendientes. Marta Baralo describe esta necesidad de manera detallada, por medio de preguntas que surgen en el campo de la enseñanza de lenguas:

> A medida que se amplía y se intensifica la investigación en todos los aspectos de la enseñanza y el aprendizaje de ELE, surgen preguntas de investigación que requieren descripción y análisis a partir de datos fiables: ¿Qué tipos de errores son más frecuentes? ¿En qué nivel de dominio de la lengua meta se encuentran? ¿Son errores de desarrollo? ¿Por qué existen ciertas estructuras fonológicas, morfológicas, sintácticas, léxicas, pragmáticas que se fosilizan? Estos procesos de fosilización, ¿son iguales en todos los que aprenden español o son diferentes según la lengua materna del que aprende? ¿En qué medida influye y de qué modo la lengua materna?, ¿en los errores?, ¿en el tiempo necesario para alcanzar determinado nivel de dominio en ELE?, ¿en el tipo de estructuras que deben trabajarse durante más tiempo?, ¿en la formación de nuevos hábitos lingüísticos? ¿Son iguales de válidos todos los métodos de enseñanza de ELE para todos los aprendientes, o varía su eficacia según otras variables relacionadas con las estrategias de aprendizaje y de comunicación, con las tradiciones culturales de cada comunidad idiomática? ¿Es universalmente válida la corrección de errores? ¿Influyen otros factores en la eficacia de diferentes formas de reparar el error? Si existen esos otros factores, ¿cuáles son los más eficaces para cada tipo de alumno o para cada contexto de aprendizaje de ELE? (Baralo, 2010, p. 2)

Estas y otras preguntas surgen en la práctica e investigación docente, y las herramientas para resolverlas son limitadas, por lo que se solucionan desde la subjetividad y la experiencia. Uno de los caminos para enfrentar esta dificultad es la construcción de corpus de aprendientes de español que, al ofrecer gran cantidad de textos orales y escritos producidos en contextos reales, permiten el análisis de los datos y la reflexión sobre ellos. Con su uso, se llega a respuestas y soluciones a muchas de las preguntas y dificultades que surgen acerca de los procesos de enseñanza-aprendizaje de la lengua.

¿Qué sucede con los corpus de aprendientes de español?

La lingüística de corpus es un conjunto de principios metodológicos apoyados en técnicas estadísticas y computacionales que se emplean para estudiar datos reales de la lengua (Parodi, 2010). Esta metodología es la encargada de diseñar y construir corpus. Por tal motivo, no se puede hablar de un corpus de aprendientes sin antes tener claro qué es un corpus.

Las definiciones que aún son vigentes aparecieron con el auge de los computadores en las décadas de 1970 y 1980, ya que el desarrollo tecnológico permitió almacenar grandes cantidades de datos. Según Sinclair (1991), un corpus corresponde a una colección de textos o muestras de lengua natural que se escogen para caracterizar una variedad específica de una lengua o el estado de la misma; por supuesto, de manera computacional.

Por tanto, un corpus de aprendientes es aquel que reúne muestras de lengua producidas por personas que se encuentran en el proceso de adquisición/aprendizaje de una lengua distinta a la materna. Estas muestras pueden ser de naturaleza oral o escrita y se encuentran sistematizadas, de manera tal que sus metadatos o información adicional permitan la realización de búsquedas y el cruce de información sobre el hablante y la producción lingüística.

Aunque el uso de corpus lingüísticos para la investigación y la enseñanza de idiomas es cada vez más frecuente, en español no existe un número vasto de corpus que soporten las necesidades de la lengua y de la comunidad hispanohablante. Los únicos corpus escritos de aprendientes de español registrados hasta la fecha se pueden visualizar en la Tabla 1.

Seguramente, existen corpus privados en español que no se encuentran en línea. También puede haber docentes e instituciones que han compilado un gran número de producciones de sus estudiantes que hasta el momento no han sido sistematizadas, pero tienen gran valor investigativo y pueden llegar a convertirse (con el trabajo y el interés necesarios) en corpus de aprendientes. Por esta razón, el Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CAELE/2) es una propuesta emergente que quiere proveer a docentes e investigadores de material lingüístico que facilite su labor.

Uno de los proyectos en español con gran relevancia es el Corpus de Aprendices de Español como Lengua Extranjera (CAELE) elaborado por Ferreira y Elejalde (2014, 2017). Este recopila un corpus de 84 resúmenes escritos, realizados por 22 estudiantes extranjeros en el nivel B1.

Tabla 1. Corpus de aprendientes de español–modalidad escrita

Corpus	L1	Medición de muestras	Metadatos
Aprescrilov	Neerlandés, francés	1 000 000 palabras	Nacionalidad, año académico, contacto con el español, nivel de lenguas extranjeras.
CAELE (Ferreira & Elejalde, 2017)	Alemán, francés, inglés, portugués, sueco, checo, italiano, ruso	418 textos	Nacionalidad, nivel de español, temática.
CAES	Árabe, portugués, inglés, francés, mandarín, chino, ruso	3878 textos	Sexo, edad, país, nivel educativo, número de años estudiando español, países hispanos que ha visitado, lenguas extranjeras.
CEDEL2	Inglés	750 000 palabras	Sexo, edad, número de años estudiando español, lenguas habladas en casa, países hispanos que ha visitado, lengua de los padres.
LANGSNAP	Inglés	300 000 palabras, entre muestras orales y escritas.	Sexo, edad, otras lenguas estudiadas, número de años estudiando español.
SAELE	Sueco	135 textos	Nivel de español, edad, sexo, conocimiento de otras lenguas, tiempo estudiando español.

El CAELE/2 no guarda ninguna relación con el anterior proyecto. Aunque su denominación es similar, su gran diferencia radica en la inclusión de muestras de español como segunda lengua, y no únicamente como lengua extranjera. A continuación, se presenta la descripción del CAELE/2, sus objetivos, la metodología de elaboración y los resultados obtenidos.

¿Qué es el CAELE/2–componente escrito?

El CAELE/2 es un proyecto del Instituto Caro y Cuervo que busca diseñar, recopilar, sistematizar y construir un corpus escrito y oral de muestras de aprendientes de español, extranjeros, sordos e indígenas en Colombia. Tiene el fin de ofrecer a la comunidad académica y educativa (docentes, investigadores, aprendientes, instituciones y comunidades de habla) un material que permita detectar, analizar y desarrollar propuestas pedagógicas y didácticas para la enseñanza del español[1].

1 La población seleccionada para la recolección de muestras del corpus se relaciona directamente con las necesidades investigativas de la Maestría en Enseñanza de Español como Lengua Extranjera y

A diferencia de otros corpus de aprendientes extranjeros, este proyecto recopila muestras de segunda lengua respondiendo a la diversidad lingüística del país. En este sentido, se tiene presente que las necesidades educativas no son exclusivas de estudiantes extranjeros, sino de colombianos indígenas y sordos que no tienen como L1 el español.

El CAELE/2 tiene 4 grandes fases: componente escrito para informantes extranjeros y sordos, componente escrito para informantes indígenas, componente oral para informantes extranjeros e indígenas, y componente de interacción para aprendientes sordos.

Para llegar a la construcción del componente escrito, ha sido necesario pasar por distintos procesos, siguiendo parte de la metodología de Torruela y Llisterri (1999):

- Revisión teórica, práctica y metodológica para el diseño del corpus.
- Propuesta de metadatos necesarios para el componente escrito, teniendo en cuenta las características de los informantes, los niveles de lengua y los distintos escenarios de aprendizaje.
- Diseño, implementación y verificación de los instrumentos para la recolección de muestras del componente escrito, según las características de los aprendientes, las producciones y el contexto.
- Recolección y sistematización de las muestras escritas (lo que da como resultado un corpus digital).

En términos generales, es un componente formado por muestras escritas producidas por aprendientes de ELE y EL2 en Colombia, abierto y sin un número definido de informantes o palabras. Los aprendientes tienen distintas lenguas maternas y diversos niveles de español. Además, las muestras son recolectadas mediante pruebas específicas diseñadas para este proyecto[2].

Metadatos del componente escrito: informantes extranjeros y sordos

Gabrielatos (2005) asegura que la investigación lingüística basada en corpus ha proporcionado descripciones de la L1 y la lengua meta mucho más claras y apropiadas

Segunda Lengua del Instituto Caro y Cuervo. Uno de los objetivos es ofrecer a los estudiantes de dicho programa muestras reales de lengua provenientes de las poblaciones con las cuales trabajan o planean trabajar.

2 El de Hincapié y Rubio (2017) es el primer texto que habla sobre el actual corpus. En él, se puede encontrar la descripción del primer estadio del corpus, el cual se puede comparar con los resultados del actual artículo. Además, es un documento que permite la ampliación del tema en cuanto a la planificación curricular, tema que no se trata en el actual artículo.

que aquellas basadas en las intuiciones. Uno de los factores que permite lograr tal grado de descripción y especificidad es la elección de los metadatos adecuados. Un metadato corresponde a información estructurada que describe el contenido y las características de los datos, los textos y los corpus, y que, a su vez, permite realizar búsquedas dentro de la colección (Hincapié & Bernal, en prensa).

Para la construcción del CAELE/2, en su componente escrito, los metadatos corresponden a la muestra (Tabla 2) y al informante (Tabla 3). En algunos casos, un informante proporcionó distintas muestras, por lo cual el metadato denominado *IdAprendiente* va conectado con el *IdMuestra*. Aunque hay unas etiquetas comunes a todos los subcorpus (extranjeros y sordos), existen también metadatos específicos para cada subcorpus. Las Tablas 2, 3 y 4 presentan los metadatos y su respectiva explicación cuando es necesario.

Tabla 2. Metadatos para cada muestra escrita

Metadato	Descripción
IdMuestra	Nombre del subcorpus (ELE/L2S) + me (muestra escrita) + Número de la muestra (0001)
Ubicación Id .txt/.docx/.pdf	Ruta de acceso de la muestra almacenada
Subcorpus	ELE (español como lengua extranjera) o L2S (segunda lengua para aprendientes sordos)
Tipo de texto	Descriptivo/Argumentativo/Narrativo
Temas	Palabras clave extraídas del contenido de las muestras
Tipo de tarea	Descripción o instrucción de la tarea
Ejercicio	Número del ejercicio
Número de palabras	
Número de párrafos	
Ayuda para el desarrollo del ejercicio	Sí/No
¿Qué tipo de ayuda?	Traductor, diccionario, otra persona, etc.
Encuestador/Trascriptor	Persona encargada de aplicar y sistematizar la prueba
Revisor	Persona encargada de revisar los datos una vez sistematizados

La mayoría de estos metadatos permiten realizar búsquedas desde una plataforma web, aunque no todos son visibles para los usuarios, ya que algunos son utilizados para controlar la sistematización y organización del corpus. En cuanto a los informantes, muchos metadatos se comparten entre aprendientes extranjeros y aprendientes sordos, pero se han adicionado algunos, dependiendo de la población, con el fin de brindar mayor información extralingüística. Los metadatos de informantes se pueden observar en las Tablas 3 y 4.

Tabla 3. Metadatos de informantes extranjeros

Metadato	Descripción
IdAprendiente	IdMuestra + Sexo (F/M)+ Nivel de español (A1/A2/B1/B2/C1) + IdNacionalidad
Nombre	Esta información no es necesaria
Edad	
Sexo	F (femenino) / M (masculino)
Nacionalidad	
IdNacionalidad	Tres primeras letras del país de nacimiento
Nivel educativo	Primaria/Secundaria/Pregrado/Posgrado/Doctorado
Profesión	
Lengua materna	
Segunda lengua	
Lenguas extranjeras	Lenguas extranjeras que domina en algún nivel
Nivel de español	Según MCER (A1/A2/B1/B2/C1)
Tiempo aprendiendo español	
Países hispanohablantes que ha visitado	
Ciudad colombiana en la que permaneció más tiempo	
Tiempo de estancia en Colombia	
Contacto con el español	Experiencias de aprendizaje/Amigos/Medios de comunicación/Trabajo/Etc.

Tabla 4. Metadatos de informantes sordos

Metadato	Descripción
IdAprendiente	IdMuestra + Sexo (F/M)
Nombre	Esta información no es necesaria
Edad	
Sexo	F (Femenino)/M (Masculino)
Ciudad de nacimiento	
Ciudad de residencia	
Nivel educativo	Primaria/Secundaria/Pregrado/Posgrado/Doctorado
Profesión	
Lengua materna	
Segunda lengua	
Lenguas extranjeras	Lenguas extranjeras que domina en algún nivel

Edad de adquisición de la lengua de señas colombiana (LSC)	
Tipo de deficiencia auditiva	Congénita/Aparición tardía
Oralizado	Sí/No
Padre/Madre	Sordo(a)/Oyente
Tiempo de instrucción en español	
Nivel de español	Básico/Intermedio/Avanzado
Contacto con el español	Experiencias de aprendizaje/Amigos/Medios de comunicación/Trabajo/Etc.

Diseño de los instrumentos para la recolección de muestras escritas de informantes extranjeros y sordos

Para la recolección de los metadatos de los aprendientes, se diseñó, en ambos casos, un formulario con preguntas sobre datos personales que respondían a los metadatos estipulados en el corpus. Estos formularios se redactaron de tal manera que las preguntas fueran lo suficientemente transparentes, ya que existían informantes con un nivel de español básico. Los formatos podían diligenciarse en papel o en línea, mediante los formularios de Google Drive, que no tienen autocorrector.

El segundo instrumento corresponde a la prueba lingüística que da como resultado la muestra de lengua. Para el caso de las muestras de español para extranjeros, se creó un formulario con siete ejercicios distintos. Estas actividades se plantearon tras revisar los criterios de evaluación por nivel, según el Marco Común Europeo de Referencia (MCER, 2002), y con base en el análisis de criterios que realizan Gozalo y Martín (2009).

Los ejercicios corresponden a las tareas de describir, narrar o argumentar, y se dividen en los niveles A1, A2, B1, B2 y C. Los informantes son libres de escoger el ejercicio o los ejercicios que a su criterio se relacionan con el nivel de lengua que poseen, por tanto, las pruebas no especifican el nivel al que corresponden ni el tipo de tarea a la que pertenecen. Esta información es conocida únicamente por los investigadores. Algunas de las actividades cuentan con un *input* visual, y los temas son variados (desde redes sociales hasta política), con el fin de motivar al estudiante a desarrollar la tarea.

Cabe aclarar que la finalidad del corpus no es el análisis del discurso, sino de las construcciones lingüísticas que los aprendientes realizan en la lengua meta. Por tanto, no se realizan juicios de valoración sobre el contenido de las muestras.

En cuanto a las pruebas diseñadas para aprendientes sordos, el planteamiento se hizo desde cuatro actividades. Como no existe un sistema de evaluación para conocer el nivel en español escrito, los ejercicios se plantearon desde las mismas tareas expuestas para los aprendientes extranjeros: describir, narrar y argumentar.

Cada tarea tiene un grado mayor de complejidad que la anterior: partiendo de las necesidades de esta población, el ejercicio número uno corresponde a la creación de un perfil personal en una red social; los dos ejercicios siguientes consisten en narrar una historia a partir de una secuencia de imágenes o de un video; por último, el ejercicio argumentativo parte de un *input* visual, en donde se presenta a una persona sorda en la portada de una revista de gran circulación en Colombia.

El último instrumento está enfocado a los procedimientos éticos de la investigación y corresponde a un consentimiento informado, en el cual se explica el proyecto y sus objetivos, y se pide autorización para el manejo de los datos personales y las muestras proporcionadas para uso científico y académico, sin fines lucrativos.

Recolección y sistematización de las muestras escritas

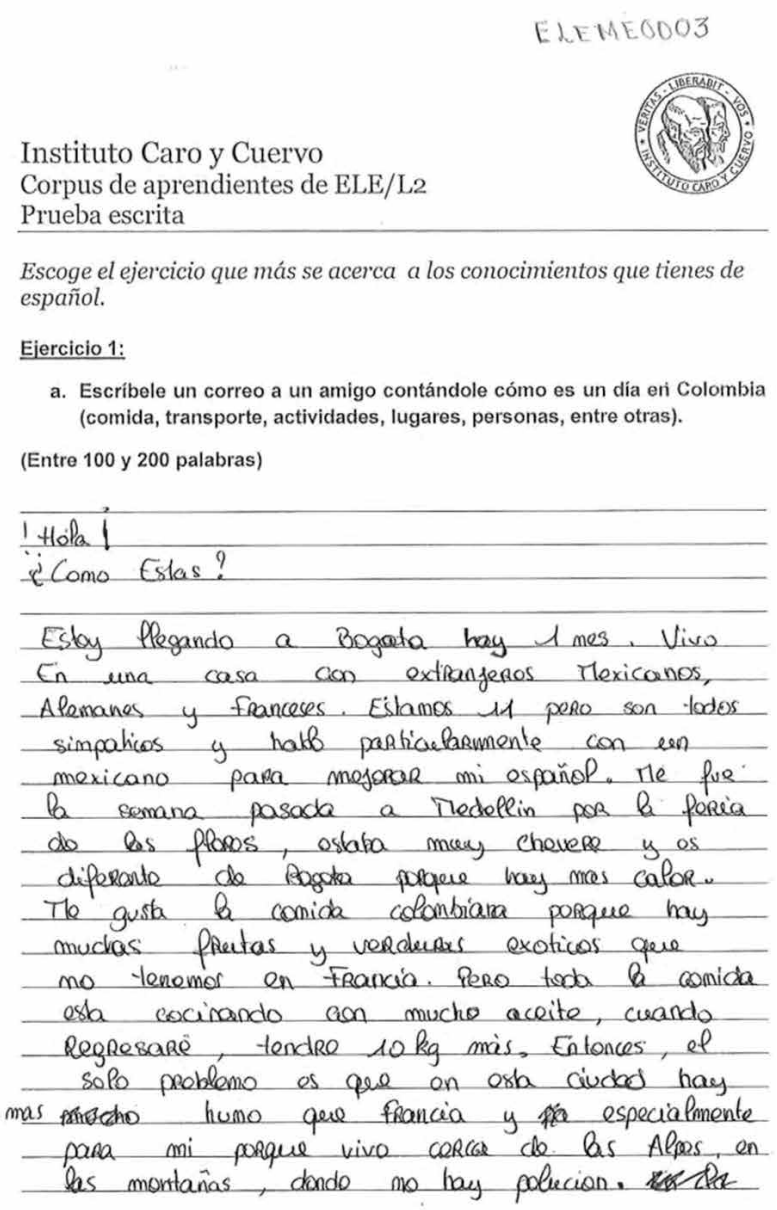
La recolección se realizó de manera presencial y virtual, según las facilidades de los participantes. Cada participante debía responder a tres formularios: consentimiento informado, formulario de datos personales y prueba escrita. Aquellas muestras que no contaban con toda la información fueron descartadas. Ya que el corpus es de carácter abierto, no se limitó la recolección a un número específico de muestras, pero sí se pilotearon las pruebas recolectando 25 muestras para cada subcomponente: aprendientes extranjeros y aprendientes sordos.

A medida que se hacía la recolección, se iban sistematizando las muestras en una hoja de cálculo en línea que permitía seguir el consecutivo del *IdMuestra*. Para facilitar los procesos de explotación en un futuro, cada una de las muestras debe ser digitalizada y almacenada en distintos formatos: *.pdf*, *.docx* y *.txt*. Estos formatos permiten hacer estudios de distinta índole, ya que el *.pdf* muestra la escritura real de los participantes, el documento *.docx* permite hacer análisis manual de los datos, y el archivo *.txt* es el que se procesa computacionalmente para permitir análisis y búsquedas automáticas.

La sistematización de la información contenida en el formulario de datos personales es la que permite la búsqueda cruzada dentro del corpus (por ejemplo, filtrar todas las muestras de nivel básico de español de participantes con padres sordos, o las muestras de nivel de español A2 con portugués como lengua materna). Por otro lado, los metadatos de la muestra (como número de palabras, temas, etc.) se extraen en el momento en que los investigadores sistematizan las pruebas.

Ya que el proceso de sistematización se realiza manualmente por medio de la ayuda de investigadores, hay un paso adicional: el de revisión. Tras almacenar y organizar las pruebas en una computadora, disco o en la web, los datos son revisados por un segundo investigador, quien corrobora la veracidad de los metadatos, la asociación

correcta entre *IdMuestra* y el *IdAprendiente*, la ruta de almacenamiento, y revisa que cada muestra cuente con los archivos correspondientes, el consentimiento informado y el formulario de datos personales. Los ejemplos de las muestras tomadas se presentan en las Figuras 1, 2 y 3.



ELEME0003

Instituto Caro y Cuervo
Corpus de aprendientes de ELE/L2
Prueba escrita

Escoge el ejercicio que más se acerca a los conocimientos que tienes de español.

Ejercicio 1:

a. Escríbele un correo a un amigo contándole cómo es un día en Colombia (comida, transporte, actividades, lugares, personas, entre otras).

(Entre 100 y 200 palabras)

¡Hola!
¿Como Estas?

Estoy llegando a Bogota hay 1 mes. Vivo En una casa con extranjeros Mexicanos, Alemanes y Franceses. Estamos 11 pero son todos simpaticos y hablo particularmente con un mexicano para mejorar mi español. Me fue la semana pasada a Medellin por la feria de las flores, estaba muy chevere y es diferente de Bogota porque hay mas calor. Me gusta la comida colombiana porque hay muchas frutas y verduras exoticos que no tenemos en Francia. Pero toda la comida esta cocinando con mucho aceite, cuando regresare, tendre 10 kg más. Entonces, el solo problemo es que en esta ciudad hay mas mucho humo que Francia y especialmente para mi porque vivo cerca de las Alpes, en las montañas, donde no hay polucion.

Figura 1. Muestra de aprendiente extranjero

Ejercicio 2: Esta es una portada de la revista SOHO que se publicó hace algún tiempo. ¿Qué piensas de la portada? ¿Estás de acuerdo con el uso de la palabra "sordomuda"? ¿Por qué?
La palabra en la época tradicionalmente continuo mismo palabra entonces se transforma palabra alta o tecnica varios palabra adecuados: no oyente, sordo, limite de auditivo, no oir… Todavia cuantas ciudadanos creencia o usa palabra es así…

Figura 2. Muestra 1 de aprendiente sordo

Ejercicio 4: Mira el video y escribe una historia sobre lo que pasó.
Piramide
Un hombre saco una piedra azul miro con jeroglífica y camello va aparecio una cara de Egipto y el salio de hueco metio la puerta de l boca Egipto se voltio la cabeza, camello hocico coge piedra azul se mueva mientras cabeza de Egipto movimiento y se encuentra la misma forma de camello luego cambia y se metio al fondo transforma piramide. Un rato piramide salio hombre caída al frente con camello.

Figura 3. Muestra 2 de aprendiente sordo

Resultados y conclusiones

Finalizado el proceso de recolección y sistematización, el CAELE/2 cuenta con un total de 166 muestras de interlengua, 78 muestras provenientes de aprendientes extranjeros y 88 muestras escritas por aprendientes sordos, y un total de 83 participantes, 56 estudiantes extranjeros y 27 pertenecientes a la comunidad sorda colombiana. Los resultados de las muestras de aprendientes extranjeros se presentan en la Tabla 5.

Tabla 5. CAELE/2 componente escrito: muestras de aprendientes extranjeros

Datos	Resultado
Informantes	Hombres: 30
	Mujeres: 26
	Total: 56
	Edad: 18 a los 60 años
Información lingüística	Nivel de español C: 20 Nivel de español B2: 10 Nivel de español B1: 12 Nivel de español A2: 8 Nivel de español A1: 6
	Lenguas maternas: 10 (alemán, chino, filipino, francés, hindi, inglés, italiano, lituano, portugués y vietnamita)
Muestras	Número de muestras: 78
	Número de palabras: 19 500 palabras aproximadamente

Aunque el número de palabras del CAELE/2 aún no es comparable con corpus como el CAES o el CEDEL2, expuestos anteriormente (ya que superan las 500 000 palabras), es el primer corpus de aprendientes de la variedad colombiana de español que se crea en Colombia. Además, tiene la ventaja de que es un corpus abierto, sin un número fijo de palabras y está diseñado bajo un estricto procedimiento de sistematización, lo que señala que puede ser alimentado cada vez que se quiera.

La información de las muestras escritas por aprendientes sordos se presenta en la Tabla 6.

El CAELE/2 es el primer corpus en español que incluye muestras de EL2 de aprendientes sordos. El número total de palabras aún es bajo, pero es posible alimentarlo constantemente.

Aunque los corpus son lingüísticos, los instrumentos y los metadatos arrojan información que no necesariamente tiene que ver con la lengua. El componente de aprendientes sordos contiene material interesante para investigaciones futuras de carácter interdisciplinar. Gracias a estos elementos, existe información que puede abrir temas en áreas como cultura y comunidad sorda, educación, y política lingüística, entre otros. La sistematización de las muestras, sin ser sometidas a un análisis profundo arroja datos como:

Tabla 6. CAELE/2 componente escrito: muestras aprendientes sordos

Datos	Resultado
Informantes	Hombres: 9
	Mujeres: 18
	Total: 27
	Edad: 21 a los 61 años
Información lingüística	Nivel de español básico: 7 Nivel de español intermedio: 13 Nivel de español avanzado: 7
	Lengua de señas colombiana como lengua materna: 18 aprendientes Español como lengua materna: 5 aprendientes No reconocen cuál es su lengua materna: 4 aprendientes
	Oralizados: 21 aprendientes
Deficiencia auditiva	Congénita: 15 aprendientes Aparición tardía: 12 aprendientes
Muestras	Número de muestras: 88
	Número de palabras: 4275

- La extensión de las muestras escritas por aprendientes sordos es mucho menor que la extensión de las muestras de aprendientes extranjeros. Esta situación puede deberse a factores contrastivos entre las lenguas maternas y el español, o a condiciones educativas.
- A la pregunta ¿cuál es su lengua materna? 4 informantes (14 % de la muestra) respondió que no sabría decir si es el español o la LSC. Aunque la muestra no es representativa, sí deja entrever una dificultad de identidad lingüística. Sería interesante profundizar más en este tema.
- En términos de adquisición-aprendizaje, la mayoría de aprendientes sordos adquieren su lengua materna (LSC) a una edad avanzada. Esto puede tener repercusiones en el aprendizaje de la L2 (en este caso del español), lo que lleva a pensar que se necesitan

políticas y estrategias lingüísticas, educativas y didácticas más eficaces y acordes a la realidad de esta población.

El CAELE/2 continúa en su proceso de construcción. Por tanto, pretende ser un material completo que reúna muestras escritas y orales del español como lengua extranjera y segunda lengua, con el fin de aportar a la investigación lingüística aplicada directamente en áreas como la educación. Sus muestras pueden servir para caracterizar la interlengua de los aprendientes, como material para el diseño de currículos, de material pedagógico y didáctico, como evidencia para la mejora de políticas lingüísticas y educativas, entre muchas otras posibilidades.

Referencias

Baralo, M. (2010). La investigación del español como lengua segunda y extranjera. En: *Actas Congreso internacional de la lengua española Valparaíso*. Consultado en: http://congresosdelalengua.es/valparaiso/ponencias/lengua_educacion/baralo_marta.htm

Consejo de Europa. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. (Instituto Cervantes, trad.). Estrasburgo: Consejo de Europa, Ministerio de Educación y Grupo Anaya.

Ferreira, A., & Elejalde, J. (2017). Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE. *Revista Brasileira de Linguística Aplicada*, *17*(3), 509-538.

Gabrielatos, C. (2005). Corpora and Language Teaching: Just a fling or wedding bells? *TESL-EJ, 8*(4), 1-37.

Gozalo, P., & Martín, M. (2009). *Pruebas de nivel. Modelos de examen de clasificación*. Madrid: SGEL.

Hincapié, D., & Rubio, R. (2017). Diseño y construcción del CAELE2: Base para una planificación curricular. *Hechos y Proyecciones del Lenguaje*. *23*(1), 42-52.

Hincapié, D., & Bernal, J. (en prensa). *Lingüística de corpus*. Bogotá: Instituto Caro y Cuervo.

Parodi, G. (2010). *Lingüística de corpus: de la teoría a la empiria*. Madrid/Frankfurt: Iberoamericana.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Torruela, J., & Llisterri, J. (1999). *Diseño de corpus textuales y orales*. En *Filología e informática: Nuevas tecnologías en los estudios filológicos* (pp. 45-77). Barcelona: Milenio.