Círculo de Lingüística Aplicada a la Comunicación
ISSN: 1576-4737

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/clac.71174

Diseño e implementación del Corpus de Aprendientes de Español como Lengua Extranjera (CAELE)

Anita Alejandra Ferreira Cabrera[1], Jessica Elejalde Gómez[2] y Lorena Paulina Blanco San Martín[3]

Recibido: 22 de agosto de 2020/Aceptado: 10 de noviembre de 2021

Resumen. El desarrollo de los corpus de aprendientes ha permitido un avance significativo en los estudios de Adquisición de Segundas Lenguas (ASL) (Granger, 2012, 2015, 2017). El Corpus especializado de Aprendientes de Español como Lengua Extranjera (CAELE) es un inventario abierto que cuenta con 1217 textos producidos por 201 aprendientes de diferentes lenguas maternas y niveles de competencia en ELE A2 y B1. Los estudiantes provienen de universidades extranjeras en el contexto de programas de intercambio de nivel de pregrado y posgrado. La recolección de los textos se ha llevado a cabo a través de tareas de escritura bajo el enfoque metodológico basado en tareas. En este artículo se presenta el diseño e implementación del corpus acorde con los principios de Sinclair (2005). Los resultados evidencian el logro de un corpus representativo y homogéneo para realizar estudios tanto de análisis de errores como contrastivos de interlengua.
Palabras clave: corpus de aprendientes de lengua; CAELE; interlengua; taxonomía.

[en] Design and implementation of the corpus of Spanish learners as a foreign language (CAELE)

Abstract. The development of learner corpora has made a significant contribution to Second Language Acquisition (ASL) studies (Granger, 2012, 2015, 2017). The specialized Corpus of Learners of Spanish as a Foreign Language (CAELE) is an open inventory that has 1217 texts produced by 201 learners of different mother tongues and proficiency levels in ELE A2 and B1. Students come from foreign universities in the context of undergraduate and graduate level exchange programs. The collection of texts has been carried out through writing tasks under the task-based methodological approach. This article presents the design, and implementation of the corpus in accordance with the principles of Sinclair (2005). The results show the achievement of a representative and homogeneous corpus to carry out both error analysis and interlanguage contrast studies.
Keywords: corpus of language learners; CAELE; interlanguage; taxonomy.

Cómo citar: Ferreira Cabrera, A., Elejalde Gómez, J.; Blanco San Martín, L. (2022). Diseño e implementación del Corpus de Aprendientes de Español como Lengua Extranjera (CAELE). *Círculo de Lingüística Aplicada a la Comunicación* 90, 137-154.

1. Introducción

El interés por los corpus de aprendientes de lenguas se desarrolló en los años noventa y primera década del siglo XXI, inspirado en los trabajos de Granger (1998a, 1998b) en el *Centre for English Corpus Linguistics,* dando origen al primer corpus de aprendientes (se usa el término aprendiente definido por el Instituto Cervantes como la persona que se encuentra en proceso de aprendizaje de una lengua extranjera, al margen de otras consideraciones, como la edad o el contexto en que aprende (Instituto de Cervantes, 2020)) reconocido internacionalmente, el ICLE (*International Corpus of Learner English)*. ICLE es el corpus de aprendientes más extenso de todos los conocidos hasta el momento, incluye dos millones de palabras procedentes de trabajos escritos de alumnos de nivel avanzado con distintas lenguas maternas. El desarrollo de los corpus de aprendientes ha permitido un avance significativo en los estudios de Adquisición de Segundas Lenguas (ASL) (Granger, 2012, 2015, 2017). Mientras que los trabajos anteriores –basados en procedimientos tradicionales de obtención de datos– eran limitados en cuanto al número de sujetos estudiados para poder controlar las variables que afectan a la producción del aprendiente, los nuevos corpus computarizados permiten trabajar con más y mejor calidad de datos de lengua natural. Así pues, el poder

[1] Universidad de Concepción (Chile) Correo electrónico: aferreir@udec.cl ORCID https://orcid.org/0000-0001-7979-6467
[2] Universidad Católica de Temuco (Chile) Correo electrónico: jelejalde@uct.cl ORCID https://orcid.org/0000-0002-6024-2381
[3] Universidad de Concepción (Chile) Correo electrónico: lblanco@udec.cl ORCID https://orcid.org/0000-0002-9584-899X

analizar amplios corpus de aprendientes, que estén bien diseñados de acuerdo con los criterios establecidos por Sinclair (2005), proporciona una base empírica sólida para la descripción de la interlengua (IL), es decir, del sistema lingüístico cognitivo, específico, de un aprendiente de L2. Por ende, el principal objetivo es analizar qué sucede en la mente del aprendiente, cómo enjuicia esos datos y cómo los pone en práctica (Baralo, 2004).

Los corpus de aprendientes son colecciones de textos orales o escritos digitales de muestras auténticas de interlengua de aprendientes de lenguas extranjeras o segundas, los cuales han sido recolectados acorde con criterios y principios específicos para garantizar de cierta forma la calidad metodológica en su creación (Granger, 2017; Sinclair, 2005). Los corpus de aprendientes suelen ser utilizados en estudios sobre adquisición, de tal manera que se pueda conocer con exactitud cuáles son los procesos que sigue el aprendiente hasta alcanzar un determinado dominio de la lengua meta. Aportan con información relevante sobre las fases por las que pasa el aprendiente hasta la completa adquisición y las características de su interlengua en función del nivel de dominio en el que se encuentre. A través de los estudios, se puede conocer qué aspectos lingüísticos se aprenden antes que otros, cuáles son más difíciles de aprender y las diferencias con los hablantes nativos (HN), por medios de análisis contrastivos (Granger, 2015; 2017; Campillos, 2014; Lozano,2015).

Los corpus de aprendientes son especialmente útiles a la hora de investigar el proceso de aprendizaje y constituyen una fuente empírica de datos para líneas de investigación como el CEA (del inglés *Computer aided error analysis)* y los estudios en análisis de interlengua (Cruz Piñol, 2012). Al mismo tiempo, estas descripciones mejoradas de la lengua del aprendiente pueden ser utilizadas en la enseñanza de las L2 (Granger, 2012; Pastor Cesteros, 2004). Una de las problemáticas que articula estos enfoques interdisciplinares corresponde al estudio de los errores causados por transferencia lingüística en ELE. El proceso de transferencia lingüística es una de las temáticas más relevantes en el estudio de los procesos psicolingüísticos subyacentes al comportamiento interlingüístico (Selinker, 1972). El fenómeno de la transferencia suscribe todos aquellos vocablos, reglas y subsistemas que tienen lugar en la interlengua como resultado de la influencia de la lengua materna (Santos Gargallo, 1993; Pastor Cesteros, 2001; 2004).

En nuestro programa de investigación, Adquisición y Enseñanza del Español como Lengua Extranjera, L1 y L2, ADELE, se han adoptado diferentes enfoques como el análisis de errores asistido por el computador, el análisis contrastivo de la interlengua (Granger 2015; Gilquin 2008). Granger (1996) propone el Análisis Contrastivo de Interlengua (ACI) para analizar los corpus de aprendientes. La utilidad del ACI ha sido demostrada en una gran cantidad de estudios, como se refleja en Granger et al. (2002). El análisis comparativo ha permitido delimitar características distintivas del aprendiente de segunda lengua y sugerir tendencias de uso según niveles de competencia (Granger, 2015). Ante las críticas realizadas (Cook,1999; Hunston, 2002) a este tipo de contrastes, Granger (2015) sostiene que este procedimiento se sustenta en la validez psicológica de los procesos comparativos entre interlengua y L2 que realizan constantemente los aprendientes durante el proceso de adquisición. Granger (2009) sostiene que el ACI es una técnica heurística poderosa que permite descubrir rasgos de la lengua de los aprendientes y luego analizarlos desde una perspectiva de la L2. En este último tiempo, el modelo del ACI se ha convertido en el procedimiento estándar para llevar a cabo comparaciones entre la interlengua y la lengua meta (Ishikawa, 2013). En definitiva, los planteamientos más actuales pretenden dar cuenta de la competencia lingüística más allá del error, señalando los aciertos, frecuencia de uso, usos infrautilizados, y la sobreutilización de formas de la L2. En este sentido, los estudios de interlengua basados en corpus requieren también un corpus nativo comparable, esto es, diseñado para permitir comparaciones (Sánchez Rufat, 2015a). Este corpus de control de hablantes nativos (HN) sirve para medir las desviaciones de la lengua de los hablantes no nativos (HNN) con respecto a la lengua nativa (Granger, 2002).

El enfoque de investigación en torno al CAELE es "multidisciplinario". En el programa de investigación ADELE se busca a través del corpus de aprendientes CAELE indagar sobre los procesos y fenómenos de la interlengua, la adquisición y el aprendizaje del Español como L2 o LE. La necesidad de observar cómo los aprendientes usan de manera real la lengua meta, puede arrojar resultados que evidencien qué áreas presentan mayor dificultad, cuáles son persistentes o cuáles responden a diferentes fenómenos dentro del desarrollo de la interlengua. Por esta razón, la investigación en corpus de aprendientes contribuye a que la información obtenida de muestras a gran escala refleje el estado real y las tendencias de uso de la lengua objeto de estudio.

En este artículo presentaremos el diseño e implementación de dicho corpus y los avances en la investigación. CAELE es un corpus escrito de ELE de corte longitudinal recolectado diacrónicamente en ambientes de clases en contextos de tareas comunicativas (enfoque por tareas). Esto marca una diferencia con los corpus disponibles en el ámbito del ELE, por cuanto la mayoría de ellos son de corte transversal, sincrónico. Una de las críticas más frecuentes que se hace a dichos corpus es la ausencia del contexto. Puesto que, en general, no se tiene detalles del de dónde y cómo se realiza la recogida de datos, ni a veces tampoco de los aprendientes que producen dichos textos, todo lo cual supone un problema para los teóricos e investigadores en ASL. Este aspecto es muy relevante para realizar análisis consistentes de la interlengua y de los errores de los aprendientes. El corpus CAELE que aquí se describe no guarda relación con el corpus CÁELE/2 de Hincapié (2018). Existe similitud en el acrónimo, pero son investigaciones y corpus diferentes; el nuestro generado en Chile y el segundo generado en Colombia.

En este artículo el objetivo general es describir y explicar la construcción del corpus de aprendientes de Español como lengua Extranjera, CAELE. Para ello se han desarrollado dos objetivos específicos: (1) describir el diseño y

recolección de los corpus aprendientes de ELE, CAELE y (2) presentar los resultados y principales logros a la luz de los principios de Sinclair (2005). El artículo se organiza en las siguientes secciones: En la sección 2, nos referimos a los fundamentos teóricos del corpus. En la sección 3, abordamos la metodología del corpus CAELE. En la sección 4, presentamos el diseño y estadísticas del corpus CAELE. En la sección 5, nos referimos a algunas investigaciones realizadas sobre la base del corpus.

Por último, presentamos algunas conclusiones sobre los avances y limitaciones en esta investigación.

2. Fundamentos teóricos del corpus

La Lingüística de Corpus es una línea de investigación de la lingüística funcionalista actual y la lingüística textual, en la cual se distingue claramente su procedimiento y metodología. El carácter empírico que posee hace de este tipo de investigación, un procesamiento riguroso donde se analiza recolecciones extensas de textos naturales, ya sean orales o escritos, cuya denominación se orienta hacia *córporas de lenguas*. Estas muestras de textos, independientemente del canal de producción, son observadas y analizadas a través de programas computacionales, donde se destaca el empleo de las tecnologías recientes de la informática y la comunicación. Los estudios de corpus de aprendientes (Granger, 2015; 2017; Lozano, 2015; 2021) ponen de relieve una serie de aspectos metodológicos que deberían ser considerados en investigaciones basadas en corpus, entre ellas, se recomienda que los estudios deberían (1) estar teóricamente fundamentados y ser también explicativos (no sólo descriptivos), (2) usar sistemas de anotación más sofisticados (en vez de sistemas genéricos) y (3) construir corpus de aprendientes que estén bien diseñados y que controlen las variables del aprendiente. El cambio en el procedimiento es importante desde el punto de vista cuantitativo y cualitativo, pues los análisis deben apoyarse en volúmenes de datos más elevados y con una mayor rigurosidad en la homogeneización y categorización de los errores. Además de estas características de los corpus de aprendientes, existen otras referidas al 'tamaño', 'la representatividad' y 'la extensión del corpus'. Estos tres aspectos dependerán del propósito de la investigación y de la disponibilidad de la muestra (Granger, 2003, 2009; Parodi, 2008, 2010; Reppen, 2010). Granger (2003) y Reppen (2010) proponen que la representatividad debe adecuarse según la línea de investigación y el objeto de estudio.

2.1. Estudios sobre corpus de aprendientes

En el ámbito de los corpus de aprendientes de ELE, existen corpus orales y escritos en formato digital. La mayoría de estos corpus consideran variables como el nivel de competencia, la lengua materna y años de estudio de los aprendientes con el objeto de identificar aspectos diferenciadores del proceso de adquisición. En lo relativo a corpus orales de aprendientes, un corpus interesante es el *The Spanish Learner Language Oral Corpora* (SPLLOC) de la Universidad de Southampton (Mitchell et al., 2008). El corpus es de corte transversal y está conformado por descripciones orales de aprendientes de ELE (de niveles A1, A2, B1, B2, C1, C2), con L1 inglés. Los datos han sido recogidos de estudiantes de ELE en un entorno de instrucción formal. Se constituye en uno de los corpus más amplios disponibles de datos orales.

En cuanto a corpus escritos, uno de los más representativos por su rigor metodológico y tamaño es el Corpus CEDEL2 (Lozano, 2009a; Lozano y Mendikoetxea, 2013). Este surge en el seno del proyecto Word Order in Second Language Acquisition Corpora (WOSLAC), dirigido por la Dra. Amaya Mendikoetxea. El objetivo del programa de investigación viene marcado por interesantes hallazgos en L2 para determinar el papel que tienen las interfaces en el desarrollo de la interlengua (Lozano y Mendikoetxea, 2007). CEDEL2 es un corpus escrito de aprendientes anglófonos de español de niveles de competencia inicial, intermedio y avanzado, según los resultados obtenidos en una prueba de diagnóstico estandarizado. El diseño del corpus es de corte transversal estándar. Este corpus ha sido recolectado con fines contrastivos entre el español y el inglés por lo que considera un corpus de hablantes nativos (Lozano y Mendikoetxea, 2013). CEDEL2 está conformado por alrededor de 7 500 000 palabras en formato computacional y ha sido recolectado en línea de forma sistemática a través de estudiantes y profesores voluntarios. De acuerdo con Lozano (2009a, 2009b), el creciente interés por el estudio de la adquisición del español como L2 no ha venido acompañado por la creación de grandes corpus de español como L2 por lo que se requiere de mayores esfuerzos en esta línea de investigación. En su versión más reciente (Lozano, 2021) CEDEL2 (versión 2) presenta varios idiomas (español, inglés, alemán, holandés, portugués, italiano, francés, griego, ruso, japonés, chino y árabe) de estudiantes de español L2 en todos los niveles de competencia. Además, contiene varias subcorporas de control de hablante nativos (inglés, portugués, griego, japonés y árabe). En esta última versión contiene material de alrededor de 4.400 hablantes, lo que equivale a más de 1 100 000 palabras (Lozano, 2021).

Otro corpus escrito de aprendientes de ELE es el Corpus CORANE (Corpus para el Análisis de Errores de Aprendices de E/LE). Este es uno de los pocos corpus de corte longitudinal en ELE. Fue creado en la Universidad de Alcalá por el equipo integrado por Cestero Mancera y Penádez Martínez (2001). El corpus está integrado por 1091 composiciones escritas por 321 aprendientes de diversas lenguas (entre ellas: inglés, alemán, francés, japonés, pakistaní, húngaro) y de niveles de competencia elemental, intermedio, avanzado y superior. La recolección del corpus se llevó a cabo en el ambiente académico universitario.

En otro contexto, el corpus ESPALEX (Bustos, 2012) ha sido implementado sobre la base de los exámenes para la obtención de los Diplomas de Español como Lengua Extranjera (DELE). El propósito de la generación de este corpus es apoyar los estudios sobre la adquisición del español como lengua extranjera. El corpus está integrado por 12 576 textos de nivel de competencia B2 que versan sobre 50 temáticas diferentes agrupadas en diferentes tipos de textos: carta formal, carta informal, texto narrativo, texto descriptivo y texto expositivo. La longitud promedio de los textos es de 175 palabras con un total de 2 200 000 palabras. En Suecia, con el propósito principal de describir la interlengua del aprendiente sueco de ELE acorde con los lineamientos del MCERL (2002) se creó el corpus escrito SAELE (Corpus de Aprendientes Suecos de ELE) (Pino Rodríguez, 2009; 2013). El corpus fue recolectado en dos universidades suecas, entre los años 2008 y 2009. El corpus está constituido por una colección digital de textos argumentativos escritos. Los niveles de competencia de los aprendientes son A2 y B1. Los aspectos metodológicos más destacables en SAELE son: el contenido, la representatividad, el contraste, la estructura, la muestra, la documentación, el equilibrio, el tema y el etiquetado. El procesamiento de este corpus ha permitido obtener las frecuencias de usos en estructuras gramaticales y de coherencia textual, presentes en la interlengua de aprendientes suecos.

En el Instituto de Cervantes, el Corpus de Aprendices de Español como Lengua Extranjera (CAES) (Rojo y Palacios Martínez, 2016) corresponde a un conjunto de textos escritos producidos por estudiantes de ELE de todos los niveles de competencia. Este corpus ha sido elaborado en colaboración por investigadores y docentes de la Universidad de Santiago de Compostela y del Instituto de Cervantes, en España. Los textos han sido generados por aprendientes de diversas lenguas maternas: árabe, chino, mandarín, francés, inglés, portugués y ruso. El total de textos recolectados es de 3878 producidos por 1423 estudiantes que escribieron dos o tres textos cada uno. El objetivo del proyecto es permitir el uso de los datos para la investigación en la adquisición de ELE y está disponible para la consulta en línea.

Finalmente, mencionaremos el corpus longitudinal LANGSNAP (Tracy-Ventura, Mitchell y McManus, 2016), creado como parte del Proyecto de Idiomas y Redes Sociales en el Extranjero (LANGSNAP) que delimitó el aprendizaje tanto del español como del francés durante la pasantía de un año académico en el extranjero por parte de estudiantes universitarios angloparlantes británicos con especialización en idiomas. Se recopiló dos corpus de aprendientes longitudinales por un total de más de 300 000 palabras en cada lengua. Los participantes eran estudiantes de español y francés, que pasaban su tercer año (de una carrera de cuatro años) en el extranjero en España, México y Francia. En total, hubo un total de 57 participantes (provenientes de estudiantes de intercambio ERASMUS, asistentes de idiomas extranjeros y pasantes en el lugar de trabajo). También se reclutaron diez hablantes nativos de cada idioma.

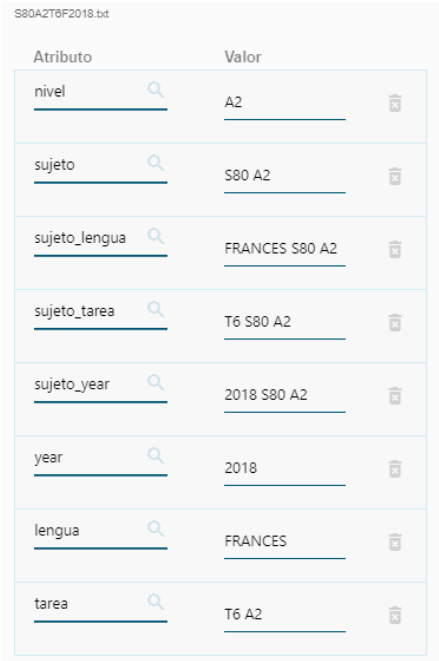
3. Metodología del corpus

En el modelo metodológico para el diseño de corpus, se consideraron los criterios establecidos por la lingüística de corpus (Sinclair, 2005) y la metodología de recolección de corpus para el estudio de la producción de aprendientes de lenguas (Granger, 2003, 2008). Estas dos líneas permitieron una construcción adecuada y equilibrada del corpus CAELE. Para ello, se consideró los siguientes procedimientos en la recolección del CAELE:

- ***Registro:*** Los aprendientes participantes de la investigación se registran a través de un formulario de antecedentes online, donde se recoge información sobre su nivel de estudios, lengua maternas y otras lenguas aprendidas, ámbitos de especialidad en sus carreras o posgrado, nivel de español, certificación de español, entre otros (Ferreira y Elejalde, 2019).
- ***Aplicación de la prueba multinivel de ELE*** (Ferreira, 2017): Se aplicó al inicio del proceso para determinar o corroborar el nivel de competencia del aprendiente de ELE
- ***Diseño de tareas comunicativas de destreza escrita:*** Terminada la etapa de inscripción y resultados del nivel de competencia, se diseñó un programa de tareas de escritura basado en el enfoque por tareas para elicitar los textos. De esta manera, el programa consideró una tarea de entrada para identificar el estado de la escritura, 8 tareas de desarrollo y dos tareas de salida. La finalidad de este programa es contar con una producción mínima de 2 textos por aprendiente con un máximo de 10 textos por sujeto para poder llevar a cabo estudios con enfoques longitudinales y poder analizar la evolución de la interlengua.
- ***Recolección del corpus de aprendientes de ELE***: Durante la producción textual para la recolección de las tareas se procedió a establecer algunos parámetros. El primero en relación con la escritura en el medio digital, esto es a través del programa bloc de notas con formato *txt* para cautelar la autenticidad del texto. Es decir, evitar la auto corrección del texto en otros programas como Word. El segundo parámetro corresponde al almacenamiento de los textos, dada la importancia del etiquetado y asignación de atributos para el posterior análisis. Para ello, el almacenamiento se realizó con codificación UTF-8 (codificación para el español, el cual permite el etiquetado gramatical y sintáctico automático en diferentes programas de etiquetado) y se alojó en la plataforma para corpus en línea *Sketch Engine.* Esta plataforma permite anotar el corpus de forma automática bajo los criterios del sistema lingüístico, estos son, la gramática (categorías gramaticales) y la sintaxis. Por otra parte, la asignación de atributos respondió a la organización de las variables que serían analizadas posteriormente para los estudios de interlengua contrastivos.

- ***Anotación del corpus de aprendientes de ELE:*** Los textos están anotados automáticamente por el software *Sketch Engine* para el análisis lingüístico acorde con los objetivos de investigación. El programa *Sketch Engine* ofrece la posibilidad de realizar diferentes tipos de análisis con los textos anotados. En la siguiente figura 1 se delimita la asignación de atributos para realizar los análisis en función de los estudios de interlengua. Los atributos corresponden a: 1) nivel, para identificar el nivel de competencia del aprendiente, 2) sujeto, de acuerdo con la nomenclatura para identificar el sujeto dentro del corpus, este se utiliza con una S mayúscula de sujeto y el número de asignación, 3) sujeto-lengua, este atributo permite analizar la variable sujeto por lengua materna para identificar la producción de textos por cada sujeto según su lengua materna, 4) sujeto-tarea, permite identificar el número de tareas que realiza un sujeto, 5) sujeto-*year*, referido a la variable de cuántos sujetos produjeron textos por año, 6) año, atributo que permite identificar la producción textual por año, 7) lengua, permite identificar el número de lenguas en el corpus y el total de sujetos por lengua, y 8) tarea, atributo que permite identificar los tipos de tarea y el número de producción por cada una.

Figura 1. Asignación de atributos en el etiquetado del CAELE.
Elaboración propia.



4. Diseño y estadísticas del corpus CAELE

4.1 Aplicación de los criterios de diseño de corpus

Como ya se mencionó en la sección metodológica, el diseño e implementación del corpus se basa en los diez criterios y principios propuestos por Sinclair (2005) que garantizan de cierta forma la calidad metodológica en la construcción de un corpus de aprendientes en formato computacional: (1) el contenido y (2) los temas del corpus se enfocarán en la función comunicativa de los textos del corpus. (3) El corpus debe ser lo más representativo posible; contener datos suficientes para los objetivos, considerando muestras de interlengua en diferentes estadios de desarrollo. (4) Debe considerar también un corpus de hablantes nativos de español para ser contrastado. (5) Debe tener una estructura que considere la división del corpus en un subcorpus de aprendices (dividido en niveles de competencia) y un subcorpus de nativos (dividido en niveles de escolaridad). (6) El corpus debe ser etiquetado e implementado computacionalmente. (7) Debe estar constituido por textos completos. (8) El diseño y la composición deben estar documentados en cada una de sus fases de desarrollo. (9) Se debe tener presente no sólo la representatividad del corpus sino también el equilibrio como guía en el diseño de los distintos componentes del corpus. (10) Se debe resguardar en los textos la homogeneidad y estandarización de sus componentes y, al mismo tiempo, mantener una cobertura adecuada y evitar textos atípicos (Sinclair, 2005). Todo ello con el objeto de que pueda ser reutilizado para otros fines científicos o pedagógicos.

A continuación, se describe cómo se aplicaron dichos principios en el diseño y recolección del corpus CAELE.

- ***Principio sobre la selección del contenido***

De acuerdo con Sinclair (2005) el contenido de un corpus debe seleccionarse acorde con las funciones comunicativas de un texto, y no centrado en las estructuras lingüísticas de la lengua. En el caso del CAELE se ha definido como eje central del contenido la recolección de textos de tipo escrito con variadas temáticas y secuencias textuales

correspondientes al ámbito académico universitario. En este sentido, el contenido busca elicitar estructuras de la interlengua que sean tendencias de uso, al igual que realizar análisis de los errores provenientes tanto de la lengua materna como la conformación de la interlengua. Las secuencias textuales recolectadas en los textos responden a la estructura narrativa, descriptiva y argumentativa (Ferreira et al., 2014).

El CAELE es un corpus especializado en el ámbito académico cubriendo en mayor grado el ámbito científico y humanístico, dado que los estudiantes que generaron los textos pertenecen a distintas carreras y durante su estadía de intercambio universitario necesitaban mejorar su fluidez oral y escrita para atender a las clases en español (Ferreira y Elejalde, 2019). El contenido del CAELE a su vez, sigue los principios de diseño y los criterios estructurales de corpus de aprendientes, donde se considera además del nivel de competencia, la lengua materna del sujeto, el año de generación de los textos, el número de textos producidos por cada sujeto, entre otros. De igual forma, los sujetos que participaron en esta recolección debieron contestar un cuestionario simple para recabar la información respecto de su procedencia y atributos lingüísticos pertinentes al estudio (Ferreira y Elejalde, 2019).

- ***Principio sobre el tema***

Considerando que la construcción de este corpus responde a la recolección de producciones textuales realizadas en aula y bajo la instrucción de un profesor de ELE, se procedió en primera instancia a identificar las áreas de interés de los aprendientes, de acuerdo con los principios metodológicos del enfoque en tareas (Estaire, 2009). Para ello, se solicitó a los sujetos a partir de un cuestionario simple que escribieran sus temas de interés para poder abordarlos durante las tareas de escritura. Para ello, se les planteó la siguiente pregunta: ¿En relación con tu carrera y práctica del español, qué temas te gustaría trabajar en este curso?. En este contexto, se pudo determinar varias áreas temáticas en el género discursivo académico 1) humanístico, 2) científico 3) cultural, 4) anécdotas o historias personales y 5) contextos sociopolíticos de cada país (Ferreira y Elejalde, 2019). En la Tabla No.1 se ilustra con ejemplos los diferentes tipos de temas cubiertos en el CAELE y se describe brevemente cada tipo. La variedad de temas del CAELE se relaciona tanto con el principio del enfoque basado en tareas de respetar el interés y necesidades de los aprendientes que realizan el intercambio universitario como con la función comunicativa de los textos. De esta manera, es posible realizar tendencias de uso de acuerdo con las temáticas tratadas, análisis de errores según las estructuras textuales, así como también análisis de la interlengua.

Tabla 1. Temática de los textos del CAELE. Elaboración propia

Tema	Descripción	Ejemplo
Académico-humanístico	Referidos a diversos temas de tipo humanístico.	Historia geopolítica, aspectos de comunicación, relaciones entre diferentes países.
Académico-científico	Referidos a los aspectos en materia de ciencias, ingeniería, entre otros.	El cambio climático, medicina, ingeniería de materiales.
Culturales	Tratan sobre las costumbres y aspectos culturales de Latinoamérica, Chile y de sus países de origen.	Fiestas, arte, música, costumbres.
Anécdotas o historias personales	Textos para afianzar estructuras lingüísticas, por ejemplo, tiempos del pretérito, subjuntivo, entre otros.	Historias personales durante la pasantía en Chile, historias de la universidad.
Contexto socio-políticos de cada país	Referidos a lo sociopolítico, geopolítico o económico de cada país en contraste con Latinoamérica y Chile.	Jubilaciones, educación, salud pública, sistemas políticos y sociales de cada país.

- ***Principio sobre la representatividad***

El contenido de los textos que integran el CAELE debería evidenciar la interlengua de los aprendientes de español como lengua extranjera que realizan un intercambio semestral universitario en un país de habla latinoamericana. En este contexto, en el CAELE hay una mayor representatividad de las lenguas maternas de los estudiantes que realizan más sistemáticamente intercambios académicos en la universidad, como es el caso de la lenguas maternas inglesa, francesa y alemana. Por otra parte, los sujetos que decidieron participar de este intercambio y tomar los cursos de español, aceptaron voluntariamente la recolección de tareas de escritura en beneficio de identificar sus errores o

producciones complejas para mejorar su fluidez comunicativa en la destreza escrita. Los estudiantes firmaron un consentimiento informado para poder revisar y analizar sus textos.

En cuanto a la representatividad de la composición escrita, los criterios del CAELE consideraron tres aspectos relevantes que permitirían la observación, análisis e interpretación tanto de errores como de tendencias: 1) el número de textos producido por sujeto, ya que permitiría identificar si la tendencia o error corresponde a un lapsus, una falta o un error sistemático en el tiempo (Ferreira y Elejalde, 2017), 2) el tipo de temática para elicitar los textos y estructuras gramaticales, para lo cual los temas culturales o históricos permiten indagar sobre tendencias narrativas o descriptivas, por el contrario, temas geopolíticos o científicos tienden a elicitar estructuras argumentativas, de opinión o que generan opiniones diversas y 3) clasificación por nivel de competencia, dado que permiten determinar con mayor precisión la transición o estadio de adquisición de una lengua extranjera y la conformación de la interlengua. La determinación de estos criterios permitió una mayor incorporación de estructuras posibles durante la descripción de la interlengua en ELE.

- ***Principio sobre el contraste***

De acuerdo con este principio, en nuestro programa de investigación se ha considerado también la recolección de un corpus de hablantes nativos de español para ser contrastado a nivel de las problemáticas, fenómenos o estructuras en estudio. La recolección de dicho corpus ha seguido la misma metodología. Se trata de un corpus longitudinal que considera un promedio de 7 tareas de escritura (1 de pre-test, 4 de escritura de proceso y una de posttest). se ha delimitado un mínimo de 2 tareas por aprendiente. El corpus de español como lengua materna está conformado por 1219 textos escritos, los cuales han sido producidos por 154 estudiantes chilenos. 613 textos (50.3 %) fueron escritos por 99 estudiantes de enseñanza media y equivalen al nivel de competencia B1 de ELE, mientras que 606 textos (4) %) fueron producidos por 55 estudiantes de enseñanza básica y equivalen al nivel de competencia A2. El número de tokens asciende a 277 320 caracteres y el número de palabras a 256 028. En cuanto a la sintaxis, este corpus contiene 14 902 oraciones.

- ***Principio sobre los criterios estructurales***

Con respecto a la estructura del corpus CAELE se determinó la construcción de subcorpus para responder a los objetivos de análisis contrastivos de interlengua (análisis de errores y tendencias de uso). Para ello se consideró el nivel de competencia, la lengua materna y el número de textos realizados por los sujetos. En relación con el nivel de competencia, la recolección se ha centrado mayormente en aprendientes de nivel A2 y B1, dado que la gran mayoría de aprendientes de intercambio universitario chileno presentan dichos niveles de lengua. Además, estos niveles de competencia evidencian que los estudiantes tienen una base de español para poder interactuar en las clases durante su periodo de intercambio en la universidad. Para ratificar los niveles de competencia que los aprendientes declaran en la encuesta de registro (Ferreira y Elejalde, 2019), previamente se aplica la prueba de multinivel de competencia (Ferreira, 2017). Por otra parte, como ya se señaló, la lengua materna de los estudiantes corresponde mayoritariamente al alemán, francés e inglés. Esto se condice con los convenios actuales de la universidad con universidades extranjeras provenientes de Europa y América del Norte (ver Tabla 2). Estas tres lenguas se han convertido en las centrales dada la cantidad de textos producidos. No obstante, en el futuro se proyecta incrementar el número de textos de otras lenguas como el portugués y el chino mandarín, entre otras. En cuanto al número de textos producidos por aprendiente se ha delimitado un número mínimo de 2 textos con la finalidad de poder observar y describir en un determinado lapso de tiempo cuál es el comportamiento y los procesos subyacentes a la conformación de la interlengua en ELE. Es decir, la escritura de más de un texto permite observar en el tiempo que estructuras, tendencias de uso y aparición de errores ocurren de manera sistemática. En este caso se trata de un análisis de recurrencia donde se puede detectar qué formas lingüísticas se mantienen con el tiempo, cuáles podrían fosilizarse y cuáles responden al estadío de evolución de la interlengua (para mayor información, Ferreira y Elejalde, 2017).

- ***Principio sobre la anotación***

El corpus CAELE ha sido procesado, etiquetado y anotado en dos etapas. La primera etapa se anotaron un total de 450 textos y se etiquetaron errores para realizar estudios de interlengua en el marco del análisis de errores. Para etiquetar y analizar los errores se procesó el corpus en el programa *UAM corpus Tool* (O´Donnell, 2009). Se identificaron errores frecuentes y recurrentes en relación con las variables de nivel de competencia y la lengua materna de los aprendientes (Ferreira y Elejalde, 2017). El propósito era identificar procesos subyacentes a la conformación de la interlengua y a la explicación del origen de los errores sistemáticos y su conexión con la influencia de la lengua materna. En la segunda etapa, se han implementado los textos en el software *Sketch Engine* (sin incluir la anotación de errores, correspondientes al trabajo en UAM corpus Tool) y se han procesado en línea para realizar análisis de interlengua contrastivos que involucran el contraste entre L2 vs. L2 y análisis de interlengua para la identificación de la influencia de la L1 sobre ELE (Granger, 2008).

Tabla 2. Taxonomía para el etiquetamiento de errores. Fuente: Ferreira y Elejalde (2020)

Clasific.	Tipo	Profundidad	Nivel	Descripción LO
Interlingüístico	Transferencia directa	Categorías gramaticales	Palabra	Omisión, adición, falsa selección y forma errónea
		Concordancia sintáctica	Oración	
		Estructura morfológica	Palabra	
		Léxico	Palabra	
		Coherencia textual	Párrafo/texto	
		Cambio de código	Palabra	
		Falsos cognados	Palabra	
		Traducción literal	Oración	
		Interferencia de otras lenguas aprendidas	Palabra	
Intralingüístico	Neutralización			
	Sobregeneralización			
	Hipercorrección			
	Simplificación	Aplicación incompleta de la regla		
		Aplicación incorrecta de la regla		
		Desconocimiento de la regla		
	Léxico creado por derivación			

Para el etiquetamiento de los errores se desarrolló una taxonomía basada en primera instancia en los principios y criterios de anotación y etiquetado propuestos por Granger (2004), Díaz-Negrillo y Domínguez (2006) y Alexopolou (2006) y Payrató (1995). Posteriormente, se produjo un sistema de etiquetamiento y anotado del corpus con una taxonomía actualizada con un criterio etiológico (ver Tabla No.2) para identificar errores en el sistema lingüístico de la lengua meta para su explicación y origen de estos (Ferreira y Elejalde, 2020).

Últimamente, el corpus ha sido procesado en *Sketch Engine* para identificar tendencias de uso correcto y realizar los estudios de análisis de interlengua contrastivos. Para ello, se ha procesado con las etiquetas gramaticales y de concordancia disponibles en *Sketch Engine* para gestionar el corpus (anotación automática). Asimismo, se ha generado un sistema de etiquetas para realizar los análisis según las variables del corpus, estas son sujeto, lengua, nivel, tarea, entre otros.

▪ *Principio sobre el tamaño del corpus*

El corpus CAELE está conformado por 1217 textos, un inventario creciente de textos, los cuales mayormente corresponden a aprendientes de niveles de competencia A2 y B1, como ya se dijo. La extensión de los textos en el nivel de competencia A2 presenta un promedio entre 150 y 200 palabras aproximadamente, mientras que en el B1 la extensión es de 200 a 300 palabras. De acuerdo con el principio de Sinclair (2005), es importante considerar que los textos deben ser completos independiente de la longitud o número de palabras. Es decir, el tamaño y la representatividad de un texto completo es más importante que la recolección de textos con variación de palabras o longitud de estos. En este contexto, en el corpus CAELE la recolección se orientó a la producción de textos completos y clasificados según su nivel de competencia.

▪ *Principio sobre la documentación*

En materia de corpus de aprendientes es relevante documentar todo el proceso de recolección de los textos, así como la procedencia de los aprendientes de la lengua. En el corpus CAELE se procedió a organizar la información recabada considerando un registro inicial, una encuesta a todos los participantes, y una rigurosa identificación y anotación de cada texto acorde con los datos más relevantes para el programa de investigación. En la encuesta de registro, elaborada el año 2014 según los datos más relevantes de los estudiantes extranjeros de movilidad estudiantil, se recogió la siguiente información: 1) nacionalidad de los estudiantes, 2) otras lenguas segundas o extranjeras aprendidas, 3) nivel de escolaridad (escolar, universitaria, posgrado), 4) años de estudio del español como lengua extranjera, 5) nivel de competencia declarado, 6) certificación del español como LE, 7) carreras de los sujetos (Ferreira y Elejalde, 2019). Los objetivos principales de recabar esta información son son: 1) describir un perfil lingüístico-comunicativo y 2) identificar las necesidades por cada destreza en ELE para fines académicos. Al procesar esta encuesta y registros realizados puede evidenciarse un perfil del estudiante caracterizado por un nivel intermedio entre los niveles A2 y B1 en ELE, con un dominio del inglés como lengua franca, con estudios presenciales de español como LE durante más de un año en una variante peninsular, sin certificación de nivel de competencia. Esta documentación permitió dar respuesta a la estructura del corpus para posteriormente realizar los análisis de interlengua contrastivos. (Ferreira y Elejalde, 2019).

- ***Principio sobre el balance***

Se debe tener presente no sólo la representatividad del corpus sino también el equilibrio que debe orientar el diseño de los distintos componentes del corpus. En este sentido, se cauteló que la recolección de los textos considerara dos ejes centrales, el nivel de competencia y la lengua materna. En este último, si bien se busca documentar y recolectar distintas lenguas maternas, solo tres han tenido mayor número de sujetos y de textos. Asimismo, estamos siempre procurando recolectar un mayor número de textos en las otras lenguas en la medida que se incrementa el número de estudiantes que llegan a la universidad por movilidad estudiantil. La Tabla No.3 muestra la distribución de la producción textual según las variables de lengua materna, nivel de proficiencia y sujetos participantes.

Tabla 3. Distribución del corpus de acuerdo con las variables de lengua, producción textual, nivel de proficiencia y número de sujetos participantes

L1	Nivel	No. Sujetos	No. Textos	Total	% L1	% Balance
Alemán	A2	27	177	312	57 %	26 %
	B1	27	135		43 %	
Inglés	A2	26	174	325	54 %	27 %
	B1	21	151		46 %	
Francés	A2	23	147	384	38 %	32 %
	B1	38	237		62 %	
Portugués	A2	10	22	48	46 %	4 %
	B1	4	26		54 %	
Filandés	A2	2	20	27	74 %	2 %
	B1	1	7		26 %	
Italiano	A2	1	5	22	23 %	2 %
	B1	2	17		77 %	
Holandés	A2	3	17	21	81 %	2 %
	B1	1	4		19 %	
Sueco	A2	1	7	14	50 %	1 %
	B1	1	7		50 %	
Checo	A2	1	6	13	46 %	1 %
	B1	1	7		54 %	
Chino	A2	1	4	12	33 %	1 %
	B1	2	8		67 %	
Yugoslavo	A2	1	5	10	50 %	1 %
	B1	1	5		50 %	
Persa	A2	1	9	9	100 %	1 %
	B1		0		0 %	
Danés	A2		0	9	0 %	1 %
	B1	2	9		100 %	
Ruso	A2		0	7	0 %	1 %
	B1	1	7		100 %	
Neerlandés	A2	1	4	4	100 %	0 %
	B1		0		0 %	
Totales		**201**			**1217**	**100 %**

Como se ilustra en la Tabla No.3, el alemán, inglés y francés muestran un equilibrio tanto en la producción de textos como en el nivel de proficiencia, lo que se muestra en un porcentaje distribuido entre el 26 % y el 32 % respecto del total de textos del corpus. En el alemán, se observa un 57 % de producción en el B1 respecto del A2 con un 43 %, lo que en términos de diferencia marca una ligera tendencia hacia un nivel intermedio del español. Por otro lado, el inglés muestra la misma tendencia marcando un 54 % en la producción del B1 y un 46 % del A2. Así, la tendencia refleja que la producción se mantiene en el nivel intermedio del español con aproximadamente 10 puntos de diferencia en ambas lenguas. Al contrario, se observa un contraste más marcado en el francés, donde la producción en el B1 es del 62 % a diferencia del A2 con un 38 %. Esta distancia en la producción, si bien no comprende un desbalance total, refleja una inclinación en la producción de nivel básico avanzado, lo que significa mayor número de sujetos de estos niveles interesados por aprender español. En el contexto del alemán e inglés, los participantes han tenido mayor participación en el nivel B1, sin embargo, se ha recibido una cantidad cercana al porcentaje de producción en el nivel A2.

Respecto de las otras lenguas, la producción dista en proporción dadas las mismas condiciones expresadas anteriormente, cuya participación se concentra más en estudiantes de las lenguas descritas. Según los datos del CAELE, una lengua que ha ido acrecentando el número de interesados corresponde al portugués con un total de 14 sujetos participantes, no obstante, en las demás lenguas, la participación y producción no supera el 4 % de toda la muestra total (1217 textos).

▪ ***Principio sobre la homogeneidad***

Al realizar el análisis del corpus, se pudo identificar que no existían textos atípicos o que no correspondiera a la cantidad mínima o máxima según el nivel de competencia. Por otra parte, en cuanto a la homogeneidad acorde con el número de sujetos por nivel, se ha buscado mantener un adecuado número de textos por cada nivel de competencia. Es decir, la escritura de textos por cada sujeto en número fue similar, lo cual dio como resultado un equilibrio entre la producción del nivel A2 y B1.

4.2. Dimensiones de los datos del CAELE

El conjunto de textos recolectados, anotados y procesados a través del software Sketch Engine arrojó como principales resultados un Corpus de Aprendientes de Español como Lengua Extranjera constituido por una colección de 1217 textos producidos en un ambiente presencial en un aula virtual. Los aprendientes que escribieron dichos textos provienen de diferentes países, tales como, Francia, Bélgica, Alemania, Sudáfrica, Estados Unidos, Inglaterra, Brasil, Portugal, entre otros. Los estudiantes pertenecen a diferentes áreas de estudio: ingeniería, matemáticas e industria, administración y astrofísica, áreas más relacionadas con las ciencias exactas y el uso del lenguaje matemático. También provienen de las áreas de las humanidades y ciencias sociales.

A continuación, se presentan los resultados logrados a partir de la implementación en el software Sketch Engine y el procesamiento de los textos acorde con las variables de la investigación.

▪ ***Temáticas de los textos***

Las temáticas a las que se refirieron los estudiantes son variadas. Los temas se clasifican en cuatro ámbitos como se observa en la Tabla No.4: 1) académico-humanístico, 2) académico-científico 3) culturales, 4) anécdotas o historias personales y 5) contexto socio-políticos de cada país.

Tabla 4. Ejemplos de los textos recolectados. Elaboración propia.

Temática	**Ejemplo**
Académico-humanístico	*"...Los jóvenes tienen que manifestar sobre una tema como equidad de género, porque es una tema muy importante en todo el mundo..."* **Sujeto A2 alemán.**
Académico-científico	*"...Se puede evitar diabetes fácilmente con un simple cambio de dieta que apoya la medicina natural..."* **Sujeto A2 inglés.**
Cultural	*"...En mi ciudad hay una otra especialidad antigua, el teatro de marionetas de madera i tejido que se llaman "les guignoles"..."* **Sujeto B1 francés.**
Anécdotas o historias personales	*"...Hasta ahora no he visitado muchos lugares pero la ciudad me gusta mas es Santiago. Santiago es la capital de Chile y la ciudad mas grande..."* **Sujeto A2 alemán.**
Contexto socio-político de cada país	*"...Hay muchas problemas sociales que afectan ciudadanos en mi país, y estas problemas parecen similares a muchas problemas de lo chilenos, aunque hay algunas problemas únicas de los Estados Unidos..."* **Sujeto B1 inglés.**

▪ ***Número de textos que integran el corpus***

Como resultado de la anotación y etiquetamiento de las variables se ha podido delimitar un total de 1217 textos producidos por aprendientes de ELE, recolectados durante los periodos de clases entre los años 2014 y 2019. A partir del procesamiento de los textos vía Sketch Engine se pudo determinar, además el número total de palabras del corpus correspondiente a 265 362, el número de *tokens (*palabras) *compuesto* por 296 136, el número de oraciones 14 902. y finalmente la extensión promedio de los textos correspondiendo a un total de 218 palabras por texto aproximadamente.

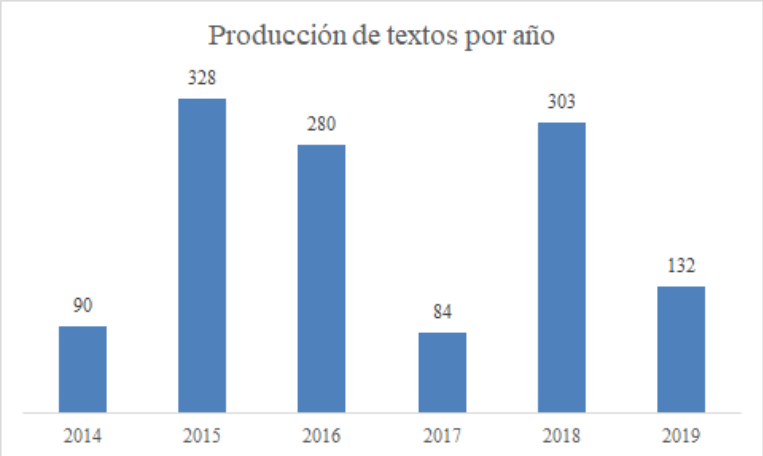
▪ ***Número de textos producidos por año***

La recolección de los textos del corpus CAELE se ha estado realizando a lo largo de los últimos 6 años en forma longitudinal y sistemática. Es decir, durante cada año se realizaron las tareas de escritura en el contexto de la asignatura de enseñanza del español como lengua extranjera con fines académicos y de inserción académica. Los estudiantes debían realizar un mínimo de 2 textos, correspondientes al pre-test y una tarea de proceso del curso hasta un total de 12 textos en un período de 3 meses, correspondientes a un pretest, 8 tareas de escritura de proceso del curso y a 3 evaluaciones de finalización del curso (ver Gráfico 7). La recolección longitudinal realizada en diferentes momentos del proceso de enseñanza aprendizaje de un aprendiente puede evidenciar los fenómenos de interlengua que se presentan durante el proceso de aprendizaje de la lengua a través del tiempo esta carcaterística del Corpus CAELE es una ventaja en relación con otros corpus de ELE de corte transversal. Con respecto al aspecto diacrónico permite que se pueda recolectar en una misma población, en este caso estudiantes extranjeros de intercambio universitario.

Actualmente, la mayoría de los corpus de aprendientes disponibles son transversales, es decir, comprenden datos recopilados generalmente de un gran número de informantes en un solo punto en el tiempo. Los corpus de aprendientes también podrían potencialmente permitir desarrollar investigaciones sobre los procesos de desarrollo que subyacen al aprendizaje una L2 (Callies, 2015). Sin embargo, el número de corpus longitudinales es escaso y requiere de esfuerzos por parte de los investigadores para incluir datos recopilados de aprendientes en intervalos de tiempo durante un período prolongado de recolección de datos.

La recolección de los datos se realizó durante cada semestre del año, dado que los estudiantes podían realizar intercambio por un semestre o por un año completo. El promedio de textos por año asciende a 204 textos, lo que implica que se obtuvo por cada año un número de textos no menor de 204 textos entre el año 2015, el 2016 y 303 el año 2018. En los años que hubo menor recolección de textos (2014 y 2017) se debió a situaciones y dificultades tanto a nivel país como universitario, dándose la imposibilidad de recoger o terminar los procesos de las clases en general. También es importante señalar, que algunos años durante los semestres de otoño (inicio de año académico en Chile durante el mes de marzo), había un menor número de estudiantes que realizaban el intercambio, pero en el segundo semestre por corresponder a la primavera había más posibilidades de recepción de estudiantes. El año más productivo corresponde al año 2015, seguido del año 2018 y 2016, tal como se muestra en el Gráfico 1.

Gráfico 1. Distribución de textos por año. Elaboración propia.

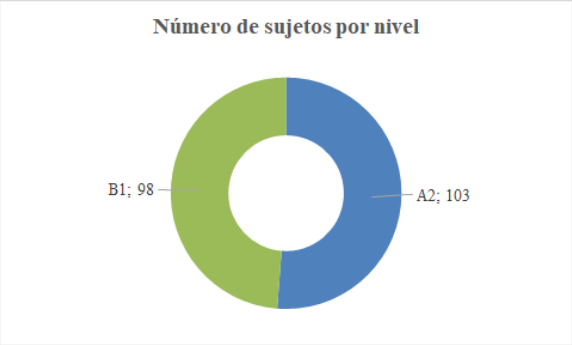


▪ ***Documentación de los aprendientes y su perfil lingüístico***

En total los sujetos que realizaron las tareas de escritura y generaron los textos del corpus CAELE son 201 aprendientes de ELE. Los niveles de competencia que presentaron como resultado de la Prueba de Multinivel de

Competencia (Ferreira, 2017) corresponden a nivel A2 y nivel B1. La distribución de los aprendientes corresponde a 103 pertenecientes al nivel A2 y, 98 al nivel B1 (Ver gráfico 2).

Gráfico 2. Distribución de los sujetos por nivel de competencia. Elaboración propia.



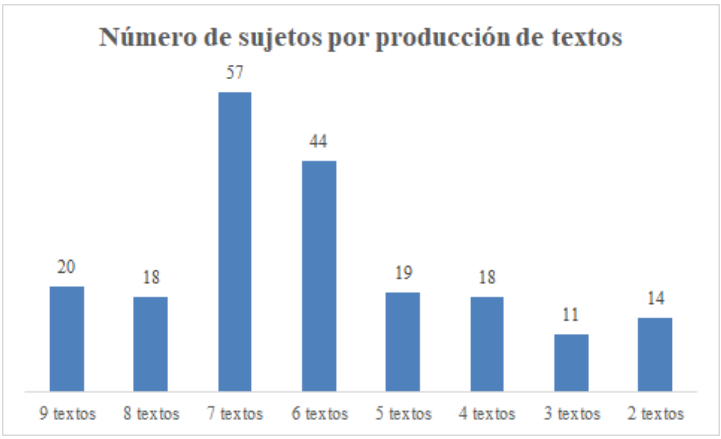
Los sujetos tienen una edad que oscila entre los 18 a los 35 años aproximadamente, asimismo, son estudiantes en diferentes niveles de la educación superior, aunque mayoritariamente de nivel de pregrado y en menor número de posgrado (magíster, doctorado y posdoctorado). La mayoría realizó intercambio por los diferentes convenios que tiene la universidad con otros países. Los países de procedencia en su mayoría corresponden a Estados Unidos, Reino Unido, Francia, Alemania y Canadá con respecto a los estudiantes de no hablantes de español. En menor cantidad proceden de países como India, Portugal, Brasil, China, Rusia, Irán, entre otros.

En síntesis, atendiendo a los datos expuestos, se ha podido determinar la homogeneidad del corpus CAELE. Por otro lado, en cuanto a la representatividad, se puede decir que las muestras textuales del corpus CAELE corresponden a la población de estudiantes extranjeros que realizan intercambios en las universidades chilenas.

- ***Promedio de textos producidos por los sujetos***

El Gráfico 3 indica la cantidad de textos escritos por los 201 sujetos. Se observa que los aprendientes produjeron en promedio entre 6 y 7 textos, 57 sujetos produjeron 7 textos y 44 que escribieron 6 textos. Los estudiantes que produjeron un menor número de textos, 2 y tres textos, correspondían a aprendientes quienes por circunstancias académicas (clases, horario o pruebas en las asignaturas de su carrera) no podían asistir de manera sistemática al curso de ELE. No obstante, pese a dichas situaciones, lo relevante es que ha sido posible lograr una recolección mayor a partir de 4 textos por aprendiente. Esto permite realizar estudios de interlengua que den cuenta sistematicidad y longitudinalidad de los fenómenos o problemáticas en estudio. En el gráfico 3 puede apreciarse la distribución de los sujetos por el número total de textos producidos.

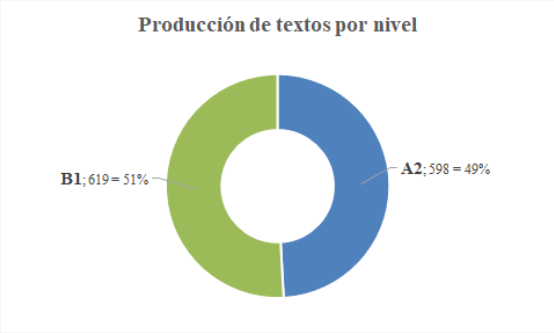
Gráfico 3. Promedio de textos. Elaboración propia.



- ***Número de textos producidos por nivel de competencia***

De los 1217 textos recolectados, en lo que respecta a la distribución de textos por nivel de competencia, en el Gráfico 4 se puede observar que 598 textos (49 %) corresponden a la producción de los aprendientes del nivel de competencia A2 y 619 textos (51 %) corresponden a los sujetos del nivel B1.

Gráfico 4. Distribución de textos por nivel de competencia. Elaboración propia.



Considerando los principios de la homogeneidad y la representatividad, se puede destacar que en ambos niveles hay un equilibrio entre la producción textual, en la cual puede observarse una diferencia de 21 textos entre los dos niveles. Además, es posible también delimitar en los análisis que se realicen por cada nivel, distintas generalizaciones en los resultados acordes con la procedencia de los aprendientes y su nivel de estudios declarados en el perfil lingüístico de cada sujeto del corpus CAELE.

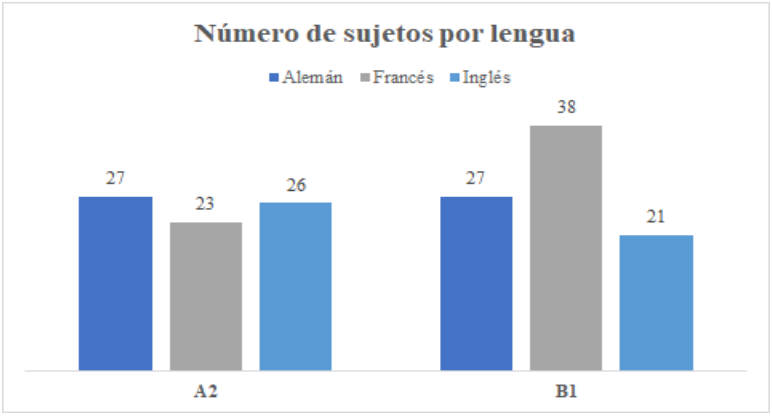
- ***Lengua materna de los sujetos***

En el Gráfico 5, se observa las lenguas maternas de los aprendientes. Las más representativas corresponden al francés (384 textos), seguido del inglés (325 textos) y del alemán (312 textos). También se incluye un grupo de lenguas con un menor número de textos (196 textos), liderado por el portugués (48 textos) y seguido por el finlandés (27 textos), italiano (22 textos), holandés (21 textos), sueco (14 textos), checo (13 textos), chino (12 textos), yugoslavo (10 textos), persa (9 textos), danés (9 textos), ruso (7 textos) y neerlandés (4 textos).

Gráfico 5. Distribución de textos por lengua materna. Elaboración propia.

Con respecto al número de sujetos y su lengua materna, puede evidenciarse en el Gráfico 6 que en el nivel A2 hay un total de 27 sujetos con L1 alemán, 23 con L1 francés y 26 con L1 inglés. En el nivel de competencia B1 38 aprendientes presentan como lengua materna el francés, seguido por 27 aprendientes de lengua alemana y finalmente 21 con la lengua inglesa.

Gráfico 6. Número de sujetos por lenguas con mayor producción. Elaboración propia.

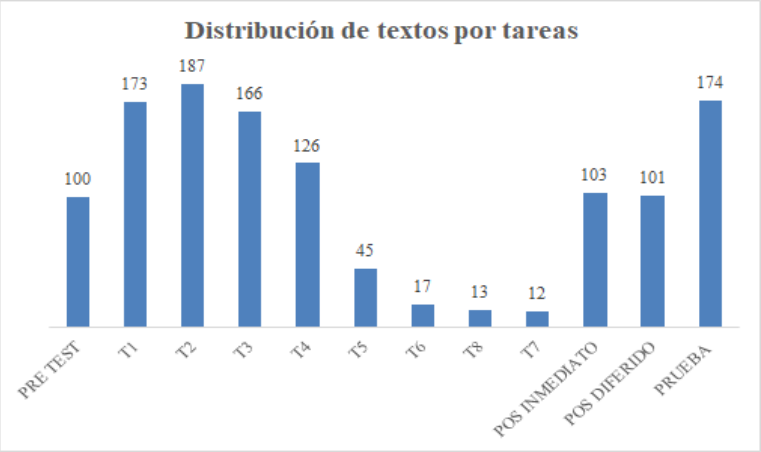


Como se aprecia en el Gráfico 6, la distribución de las lenguas por nivel está sujeta al número de sujetos por procedencia, inclinándose hacia el alemán en el nivel A2 y hacia el francés en el nivel B1. Con respecto a las otras lenguas, el número de sujetos por cada una de estas corresponde a: 1) portugués con 10 sujetos en el A2 y 4 en el B1, 2) finlandés con 2 sujetos en el A2 y 1 en el B1, 3) italiano con 1 en A2 y 2 en B1, 4) holandés con 3 sujetos en el A2 y 1 en B1, 5) sueco con un sujeto en cada nivel, 6) checo con 1 sujeto en A2 y 3 en B1, 7) chino con 1 sujeto en A2 y 2 en B1, 8) yugoslavo con un sujeto en cada nivel, 9) persa con un sujeto en el A2, 10) danés con 2 sujetos en el B1, 11) ruso con 1 sujeto en el B1 y neerlandés con 1 sujeto en el A2.

- ***Tipos de tareas de escritura***

En relación con los tipos de tareas, la producción textual abarcó un número de 12 tareas distintas, las cuales incluyen tareas en orden correlativo desde la tarea 1 hasta la tarea 8, tareas de escritura incluidas en instrumentos como pretest, postest inmediato y postest diferido y pruebas de evaluación. En el Gráfico 4, se observa que la tarea 2 (T2) presenta un mayor número de textos, 187 textos, seguida de los textos producidos en la tarea de escritura de las pruebas de evaluación con 174 y la tarea 1 (T1) con 173 textos.

Gráfico 7. Distribución de tareas según tipo. Elaboración propia.



Las tareas del número 1 a la 8 corresponden a la producción de textos generados en las distintas instancias de contenido y temas cubiertos en el curso. Las temáticas fueron organizadas de tal manera que se escribiera de manera equilibrada, es decir, se distribuyeron de forma que no todas quedaran recargadas en lo humanístico o lo científico, sino alternando las temáticas para elicitar diferentes formas lingüísticas y cautelando que los sujetos no se agobiaran con una misma estructura de texto. Asimismo, las tareas se realizaron de acuerdo con el proceso formativo de la lengua, en este caso el español, que se enmarcan durante el proceso de aprendizaje (T1 a la T8) y el proceso evaluativo (pruebas, pretest, postest) que consistía también en tareas de escritura.

5. Investigaciones a partir del CAELE

La recolección del corpus CAELE ha permitido generar una serie de investigaciones en el ámbito de proyectos con financiamiento externo (de la Agencia Nacional de Investigación y Desarrollo, ANID- CHILE), así como tesis de magíster en lingüística aplicada y doctorado em lingüística.

En el contexto del Programa de Investigación de Adquisición y Enseñanza del Español como LE, L2 y L1 (ADELE), se han llevado a cabo investigaciones basado en CAELE sobre análisis de errores asistido por el computador y feedback correctivo escrito (Ferreira, Elejalde et al. 2014). El objetivo en dicho estudio fue delimitar los errores gramaticales más frecuentes que cometen los estudiantes de ELE de nivel B1. El corpus se compuso por 84 resúmenes escritos por 22 estudiantes del nivel B1 en modalidad expositiva, narrativa y argumentativa. A partir del análisis de errores se examinaron los errores de escritura cometidos por estudiantes de ELE. Los procedimientos para la identificación, clasificación y procesamiento de los datos se realizaron a través del software NVIVO. Los resultados evidenciaron que los errores de mayor frecuencia correpondían a los de ortografía acentual, seguido por las preposiciones, la concordancia gramatical, el verbo y los artículos.

En un estudio posterior, Ferreira y Elejalde (2017) determinaron los errores lingüísticos más sistemáticos en un subcorpus del corpus CAELE compuesto por 408 textos escritos, producidos por 62 estudiantes de ELE de los niveles A2 y B1. Los errores fueron etiquetados y procesados con el software UAM *Corpus Tool*. Los resultados arrojaron que los errores más sistemáticos en el corpus fueron los de falsa selección de género gramatical y la omisión de la tilde ortográfica esdrújula. En cuanto a los niveles de competencia A2 y B1, se observaron algunas diferencias a nivel

de la omisión de tilde ortográfica en las palabras llanas y en hiatos en el nivel A2, y en la falsa selección de género gramatical y omisión de acento en diacríticos en el nivel B1.

Con un enfoque descriptivo en el estudio de Ferreira y Elejalde (2019) se definió el perfil lingüístico-comunicativo de los aprendientes que han producido los textos del Corpus CAELE, se identificó las principales necesidades por cada destreza en ELE para fines académicos. Los resultados evidenciaron un perfil del estudiante caracterizado por un nivel intermedio en ELE, con un dominio del inglés como lengua franca, (la L2 para comunicación formal ante la ausencia del español es el inglés) con estudios presenciales durante más de un año en una variante peninsular, sin certificación. En cuanto a las necesidades lingüísticas en ELE con fines académicos, los discursos académicos con mayor dificultad son las clases expositivas, las entrevistas y las presentaciones orales, la comprensión lectora de textos de la especialidad y la escritura de artículos científicos e informes de investigación.

Desde una perspectiva contrastiva de la interlengua, Oportus y Ferreira (2019) se centran en las colocaciones gramaticales verbo + preposición en ELE a partir de un subcorpus del CAELE de aprendientes anglófonos de niveles A2 y B1. El propósito fue examinar la producción de colocaciones gramaticales verbo + preposición (ColGram v+p) en relación con la producción de un corpus comparable de hablantes nativos (HN) de español. Para ello, se identificaron las ocurrencias colocacionales, se determinó su corrección desde un punto de vista de la combinatoria colocacional, y se describieron los usos aceptables acorde el patrón sintáctico-semántico que las caracteriza. Basados en los datos de frecuencia, se pudo determinar que los aprendientes muestran rasgos de naturalidad restringida en el uso de las unidades estudiadas, en términos que el nivel A2 produce un menor número y menor variedad de unidades que los HN; mientras que el nivel B1, aunque alcanza una frecuencia similar, presenta también un índice de variedad inferior a los HN, tal como el nivel A2. Lo anterior se explicaría por el uso recurrente de algunas colocaciones en ambos grupos de aprendientes, las que contribuyen a engrosar la frecuencia; no así, la variedad. Este estudio aporta desde un punto fraseológico a la investigación de las combinaciones verbo + preposición de fijación intermedia de la interlengua colectiva de aprendientes anglófonos de ELE.

En Ferreira y Elejalde (2020) se propone una taxonomía del corpus CAELE que puede adaptarse a diferentes contextos de aprendizaje y análisis de la lengua en estudio, lo cual supone un avance relevante en materia de investigación a partir de clasificaciones más acertadas de la interlengua de un aprendiente. En el estudio se aborda la importancia de la problemática de categorización y etiquetado de errores y describe una propuesta hacia una taxonomía estandarizada para la clasificación de errores de interlengua con un criterio etiológico. El sistema de taxonomía y anotación propuesta responde a etiquetas: 1) *informativas*, considerando que las etiquetas tienen tres partes en la taxonomía propuesta, entregan la información necesaria para detectar rápidamente el nivel de profundidad, la localización del error y el criterio que se aplica, 2) *reutilizables*, en relación con la posibilidad de utilizar las etiquetas en otras lenguas, para este caso lenguas romances e indoeuropeas como la alemana, 3) *flexibles* para permitir la adición o eliminación de etiquetas considerando nuevas problemáticas identificadas, 4) *neutras*, evitando en este tipo la modalidad de revisión subjetiva, para así determinar con mayor precisión la problemática y 5) *universales,* que involucra la compatibilidad con otros sistemas de anotación como por ejemplo, el etiquetario del Corpus CAES (2018) propuesto por el proyecto del Instituto cervantes u otros etiquetarios que permitan dar cuenta de errores de interlengua.

Posteriormente, Carrillo y Ferreira (2020) en un estudio mixto (cuantitativo y cualitativo) analizaron los contextos de aparición de los errores más frecuentes por transferencia negativa de la L1 inglés en las preposiciones “a”, “de”, “en”, “para” y “por” en un subcorpus de del CAELE. Los resultados evidenciaron que los cuatro errores más frecuentes de este estudio son la falsa selección de la preposición “en” para expresar pertenencia y la falsa selección de “para” en infinitivo como complemento de sustantivo, la adición de “a” por falsa regencia preposicional y la omisión de “a” en pronombres de objeto directo con verbos del tipo gustar.

En otro estudio contrastivo Blanco y Ferreira (2020) sobre verbos de apoyo presentes en colocaciones léxicas verbo-nominales se contrastó las colocaciones léxicas verbo-nominales con verbos de apoyo de un sub-corpus de 236 tareas de nivel B1, perteneciente al corpus CAELE, con las combinaciones presentes en un sub-corpus de nativos de español. Se analizaron las 236 tareas de escritura compuestas por 70 mil palabras aproximadamente mediante UAM corpus. Los principales hallazgos indicaron que 1) los colocados más frecuentes en el primer sub-corpus fueron “tener” y “dar” y 2) en el segundo sub-corpus, “hacer”, “tomar” y “tener”. Si bien existió una coincidencia considerable en cuanto a la cantidad de colocaciones encontradas en ambos sub-corpus lingüísticos, esta medida no se vio reflejada en la selección léxica de los tipos de colocados y de bases.

6. Conclusiones: proyecciones y limitaciones

En este artículo, hemos presentado un estudio descriptivo con un enfoque de análisis de datos mixto con el propósito de describir el diseño e implementación del corpus de aprendientes de español como lengua extranjera, denominado CAELE. Con respecto a la metodología, hay ciertos aspectos que deben tenerse en cuenta al momento de recolectar, almacenar y procesar un corpus de lengua extranjera. Por ejemplo, el criterio para la recolección debe responder principalmente a la temática o contenido de este, considerando además el enfoque de recolección. Es decir, cuando se planifica la intervención lingüística, la recolección debe reflejar el procedimiento del enfoque, que en el caso del CAELE, respondió al enfoque por tareas. En este sentido, la producción textual se produjo en un contexto planteado

para la elaboración de tareas de forma sistemática, con temáticas basadas en el interés de los aprendientes y acorde con los criterios del nivel de competencia.

Por otra parte, los textos recolectados deben ser almacenados de tal manera que puedan procesarse en diferentes softwares y, además, que permita mantener y cautelar la autenticidad de los textos. Por esta razón, la opción más adecuada es el almacenamiento en el formato *txt*, en el cual puede evitarse el uso de correctores gramaticales u ortográficos y, por otro lado, permite utilizar una codificación apropiada para los caracteres del español, este es la codificación UTF-8.

En cuanto al procesamiento, es imprescindible contar con una taxonomía y sistema de notación definida inicialmente para organizar el análisis de acuerdo con las variables y atributos que se le asignen tanto a los sujetos como los textos recolectados. En este último aspecto, algunos programas no permiten el cruce de variables de forma independiente, sino que están supeditados a la organización y jerarquía de los atributos, así como ocurrió con el CAELE procesado en el Skecth Engine. Basándonos en los resultados del CAELE, 1) las variables distribución de los sujetos por nivel de competencia lingüística y distribución de textos por nivel de competencia lingüística, son consistentes con los principios de homogeneidad, balance y representatividad de corpus de aprendientes. En consonancia con lo anterior y en el ámbito de ELE, se pueden proyectar tendencias en los resultados y hallazgos estudios realizados sobre la base del CAELE. 2) En relación con la variable número de textos producidos por lengua materna, el CAELE da cuenta de la representatividad de las tareas de escritura en función de tres lenguas (francés, inglés y alemán). En este sentido, la investigación generada y que se proyecta seguir desarrollando contempla la realización de estudios contrastivos y de interlengua. 3) En cuanto a la variable número de sujeto por producción de textos, los datos observados muestran un grado de longitudinalidad en función de las distintas tareas de escritura; dado que la producción fluctúa entre 6 y 7 textos por aprendiente. En consecuencia, este resultado permite ilustrar y explicar la sistematicidad de un determinado fenómeno en estudio.

Dado que el campo de 'learner corpora' o corpus de aprendientes está aún en fase incipiente en el mundo hispanohablante, este corpus es una contribución al área y suma a los corpus ya existentes. Se destaca particularmente, sus características de ser un corpus longitudinal, que son complejos de lograr dado el tiempo y metodología involucrados en dicho tipo de recolección textual. Otra característica relevante lo constituye su configuración por un mínimo de dos textos y hasta 12 producidos por un mismo aprendiente, esto facilita estudiar la variación textual a nivel de género discursivo o tipo de tarea manteniendo como constante un mismo aprendiente. Además, cubre distintas áreas en el género discursivo académico lo que también permitirá realizar análisis con dicha variable en perspectiva. CAELE se constituye en un inventario abierto que año tras año considera la recolección y el procesamiento de nuevos textos.

No obstante, estos avances significativos a partir del CAELE que permiten contribuir en el ámbito de los errores, feedback correctivo, colocaciones, concordancias, tendencias de uso y aspectos cognitivos en el Español como lengua Extranjera, se debe señalar al mismo tiempo las limitaciones metodológicas cuando se trata de recolectar un corpus de corte longitudinal. Por un lado, el diseño de tareas de escritura acordes con el programa de la asignatura y una mínima intervención para no transgredir los principios metodológicos de la enseñanza de la lengua ha sido todo un desafío para el equipo de investigación. Por otro lado, como se trata además de estudiantes extranjeros mayoritariamente con L1 inglés, francés y alemán, se ha hecho arduo el poder incrementar los textos de aprendientes de otras lenguas para estudio como el portugués o el italiano.

Así también cabe mencionar como otra limitación del corpus el número de textos por nivel de competencia y por cada aprendiente, está dentro de las proyecciones el poder incorporar otros niveles como el B2 o C1, sin embargo, esto depende de la competencia que presentan los estudiantes que llegan a la universidad. Por esa razón, se está estudiando la posibilidad de establecer alianzas con un par de otras universidades chilenas con el objeto de incrementar el CAELE acorde con las variables L1, nivel de competencia y géneros discursivos disciplinares.

Agradecimientos

Este artículo se ha desarrollado en el contexto del Proyecto Fondecyt 1180974 "*Diseño e implementación de un corpus escrito de aprendientes de Español como Lengua Extranjera (ELE) para el análisis de la interlengua"*, 2018-2021. La investigadora responsable, Dra. Anita Ferreira agradece a ANID-FONDECYT por su patrocinio.

Referencias

Alexopoulou, Angélica. (2006). Los criterios descriptivo y etiológico en la clasificación de los errores del hablante no nativo: una nueva perspectiva. *Porta Linguarum. Revista Internacional de Didáctica de las Lenguas Extranjeras*, 5, 17-35. https://dialnet.unirioja.es/ejemplar/131295

Baralo, M. (2004). La interlengua del hablante no nativo. En J. Sánchez Lobato e I. Santos Gargallo (Eds.). *Vademécum para la formación de profesores: enseñar español como segunda lengua (L2)/ lengua extranjera (LE)* (pp. 369-387). SGEL.

Bustos, J. y Sánchez, J. (2012). Espalex: un corpus para el estudio de la adquisición del español como lengua extranjera. En C. Hernández, A. Carrasco y E. Álvarez (Eds.). *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera* (pp. 149 a 159). *ASELE.*

Blanco, L. Ferreira, A. y Blanco E. (2020). Colocaciones léxicas verbo-nominales en un corpus de aprendices de Español como Lengua Extranjera de nivel A2 y B1. *Letras de Hoje*, 54(3), 407-416. https://doi.org/10.15448/1984-7726.2019.3.33243

Callies, M. (2015). Learner Corpus Methodology. En S. Granger, G. Gilquin, & F. Meunier (Eds.). *The Cambridge Handbook of Learner Corpus Research* (pp. 35-55). Cambridge University Press.

Campillos Llanos, L. (2014). Errores léxicos en el español oral no nativo: análisis de la interlengua basado en corpus. *Revista ELUA*, Alicante, 28, 85-124. https://doi.org/10.14198/ELUA2014.28.04

Carrillo, A. y Ferreira, A. (2020). Contexto de aparición de los errores más frecuentes por transferencia negativa del inglés como L1 en el uso de las preposiciones "a", "en", "de", "para" y "por" en ELE. *Revista FOLIOS,* 51, 151-16. https://doi.org/10.17227/folios.51-8612

Cestero, A. M., Penadés, I., Blanco, A., Camargo, L., y Granda, J.F.S. (2001). Corpus para el análisis de errores de aprendices de E/LE (CORANE). En. A. Gimeno (Ed.). *Actas del XII Congreso Internacional de ASELE: tecnologías de la información y de las comunicaciones en la enseñanza de la E/LE* (pp. 527-534). Centro Virtual Cervantes.

Cook, V. (1999). Going Beyond the Native Speaker in Language Teaching. *TESOL trimestral*, 2, 185-209.

Cruz Piñol, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Arco/Libros.

Díaz-Negrillo, A. y Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, 19, 83-102.

Estaire, S. (2009). *El aprendizaje de lenguas mediante tareas: de la programación al aula*. Edinumen.

Ferreira, A., y Elejalde, J. (2020). Propuesta de una taxonomía etiológica para etiquetar errores de interlengua en el contexto de un corpus escrito de aprendientes de ELE. *Forma y Función,* 33(1), 115-146. https://revistas.unal.edu.co/index.php/formayfuncion/article/view/84182

Ferreira, A., y Elejalde, J. (2019). Hacia un perfil lingüístico-comunicativo del estudiante de Español como Lengua Extranjera para fines Académicos. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 13(27), 145-165. https://doi.org/10.26378/rnlael1327326

Ferreira, A., y Elejalde, J. (2017). Análisis de errores recurrentes en un Corpus de aprendices de español como lengua extranjera (Corpus CAELE). *Revista Brasileira de Lingüística Aplicada*, 17(3), 509-537. https://doi.org/10.1590/1984-6398201710927

Ferreira, Anita. (2017). *Prueba de Multinivel con Fines Específicos Académicos*, Proyecto Fondecyt 1040500, Universidad de Concepción, Chile.

Ferreira, A. Elejalde, J, y Vine, A. (2014). Análisis de Errores Asistido por Computador basado en un Corpus de Aprendientes de Español como Lengua Extranjera. *Revista Signos*, 47(86), 385-411. http://dx.doi.org/10.4067/S0718-09342014000300003

Gilquin, G. (2008). Combining contrastive and interlanguage analysis to apprehend transfer: detection, explanation, evaluation. En Gilquin Gaëtanelle (Ed.). *Linking up Contrastive and Learner Corpus Research* (pp. 3-33). Rodopi.

Granger, S. (2017). Learner corpora in foreign language education. En S. Thorne and S. May (Eds.). *Language and Technology. Encyclopedia of Language and Education*. 3rd edition (pp. 427-440). Springer International Publishing.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, 1(1), 7-24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S. (2012). How to use foreign and second language learner corpora. En Mackey, Alison, y Gass, Susan M. (Eds.). *Research methods in Second Language Acquisition: A practical guide* (pp.7-29). Blackwell Publishing.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. En Granger, S., Gilquin. G. y Meunier, F. (Eds.). *The Cambridge Handbook of Learner Corpus Research* (pp.485-510). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.022

Granger S. (2008). Learner corpora. En Lüdeling, A. y Kytö, M. (Eds.). *Corpus Linguistics. An International Handbook.* Volume 1 (pp. 259-275). Walter de Gruyter. https://doi.org/10.1002/9781405198431.wbeal0669

Granger, S. (2003). The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3), 538-546. https://doi.org/10.1080/23247797.2015.1

Granger, S. (2002). A bird's-eye view of learner corpus research. En S. Granger, J. Hung y S. Petch-Tyson (Eds.). *Computer Corpora, Second Language Acqui- sition and Foreign Language Teaching* (pp.3-33). Benjamins. https://doi.org/10.1075/lllt.6.04gra

Granger, S., Hung, J. y Petch-Tyson, S. (2002*). Computer Corpora, Second Language Acquisition and Foreign Language Teaching*. Benjamins.

Granger, S. (1998a). *Learner English on Computer*. Longman.

Granger, S. (1998b). Prefabricated patterns in advanced EFL writing: Collocations and formulae. En A. P. Cowie (Ed.) *Phraseology: Theory, analysis and applications* (pp. 145-160). Clarendon Press.

Granger, S. (1996). Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. En K. Aijmer, B. Altenberg, & M. Johansson (Eds.). *Languages in contrast* (pp. 37-51). University Press

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Instituto Cervantes. (2020a). *Plan curricular del Instituto Cervantes. Niveles de Referencia para el español,* Madrid: Instituto Cervantes-Biblioteca Nueva.

Ishikawa, K. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. En S. Ishikawa (Ed.). *Learner corpus studies in Asia and the world* Kobe: Kobe University. 91 -118.

Lozano, C. (2021). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*. https://doi.org/10.1177/02676583211050522

Lozano, C. (2015). Leaner corpora as a research tool for the investigation of lexical competence in L2 Spanish. *Journal of Spanish Language Teaching*, 2(2): 180-193. DOI: https://10.1080/23247797.2015.1104035

Lozano,C**.** (2009a). "CEDEL2: Corpus Escrito del Español L2". En Bretones Callejas, C. et al. (eds.) *Applied Linguistics Now: Understanding Language and Mind*, Universidad de Almería editorial: Almería, pp. 80-93.

Lozano, C. (2009b). "Selective deficits at the syntax-discourse interface: Evidence from the CEDEL2 corpus", Leung Y. *et al*. (eds.): *Representational Deficits in Second Language Acquisition*, Amsterdam: Benjamins, pp. 127-166. DOI:10.1075/lald.47.09loz

Lozano, C., y Mendikoetxea, A. (2013). Learner corpora and SLA: the design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier & P. Thompson (Eds.), Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins, pp. 65-100.

Lozano, C. y Mendikoetxea, A. (2007). "Learner corpora and the acquisition of word order: A study of the production of Verb-Subject structures in L2 English". Eds. M. Davies, P. Rayson, S. Hunston, y P. Danielsson. *Proceedings of the Corpus Linguistics Conference*. Birmingham: University of Birmingham.

Ministerio de Educación, Cultura y Deporte. (2002) Marco Común Europeo de Referencia para las Lenguas: aprendizaje, enseñanza, evaluación. Madrid, Secretaría General Técnica del MECD y Grupo Anaya, trad. y adap. por el Instituto Cervantes.

Moreno, A. (2002): "La evolución de los corpus de habla espontánea: la experiencia del LLI-UAM". En: *Actas de las Segundas Jornadas de Tecnologías del Habla*, Granada: 16-18 de diciembre de 2003, Universidad de Granada.

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, (1), 287-304.

O'Donell, M. (2009): "The UAM Corpus Tool: software for corpus annotation and exploration", Bretones C. et al.: Applied Linguistics Now: Understanding Language and Mind/La Lingüística Aplicada actual: Comprendiendo el Lenguaje y la Mente, Almería: Universidad de Almería, pp.197–212

Oportus, R. y Ferreira, A. (2019). Colocaciones gramaticales verbo + preposición en ele: aspectos de naturalidad en aprendientes anglófonos de nivel a2 y b11. *Trabalhos em lingüística, (TLA)*. V. 58 n.2,826-858. https://doi.org/10.1590/01031813865395 7451621

Payrató, L. (1995). Transcripción del discurso coloquial. En L. Cortés Rodríguez (ed.), El español coloquial. Actas del I Simposio sobre Análisis del Discurso Oral. Almería: 23-25 de noviembre España.

Parodi, G. (2008). Lingüística de corpus: Una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada*, *46*(1), 93-119. http://dx.doi.org/10.4067/S0718-48832008000100006

Parodi, G. (2010). *Lingüística de Corpus: De la teoría a la empiria.* Frankfurt: Iberoamericana/Vervuert. DOI:10.31819/9783865278715

Pastor Cesteros, S. (2004). *Aprendizaje de segundas lenguas: Lingüística aplicada a la enseñanza de idiomas*. Alicante: Universidad de Alicante.

Pastor Cesteros, S. (2001). "La concordancia en la interlengua de los aprendices de español como lengua extranjera", Pastor Cesteros, S. y Salazar, V. (eds.): *Tendencias y Líneas de Investigación en adquisición de segundas lenguas. Anexo I*, Alicante: Universidad de Alicante, pp. 5-60.

Pino Rodríguez, A. (2009). Palabras en interacción: un corpus de aprendices suecos de E/LE. *A survey of corpus-based research*, p. 470-487.

Pino Rodríguez, A. (2013). El uso de combinaciones de palabras en un corpus de aprendices suecos de español como lengua extranjera. *Annuari de filología. Estudis de lingüística*, 2, 211-212.

Reppen, R. (2010). Building a corpus: What are the key considerations. En A. O´Keeffe y M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* Nueva York: Routledge, 31-38.

Rojo, G. y Palacios Martínez, I. (2016). Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project. En M. Alonso Ramos (Ed.), Spanish Learner Corpus Research: Current Trends and Future Perspectives (pp. 55-87). John Benjamins.

Sánchez Rufat, A. (2015). *El verbo dar en el español escrito de aprendientes de L1 inglés: estudio comparativo entre hablantes no nativos y hablantes nativos basado en corpus*. Tesis doctoral de la Universidad de Extremadura.

Santos Gargallo, I. (1993). *Análisis contrastivo, análisis de errores e interlengua en el marco de la Lingüística contrastiva*. Madrid: Síntesis.

Selinker, L. (1972). "Interlanguage". En: *International Review of Applied Linguistics in Language Teaching*, 10, 3, pp. 209-231.

Sinclair, J. M. (2005). "How to build a corpus", Wynne, M. (ed.): *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford: Oxbow Books, pp. 79-83.

Tracy-Ventura, N., Mitchell, R., y McManus, K. (2016). The LANGSNAP longitudinal learner corpus: Design and use. En M. Alonso-Ramos, *Spanish Learner Corpus Research: State of the Art and Perspectives*. Amsterdam: John Benjamins.