http://dx.doi.org/10.1590/1984-6398201710927

Análisis de errores recurrentes en el Corpus de Aprendices de Español como Lengua Extranjera, CAELE[1]

Análise de erros recorrentes em um corpus de Aprendizes de Espanhol como Língua Estrangeira (Corpus CAELE)

Recurrent error analysis in a learner corpus of Spanish as a foreign language (Corpus CAELE)

Anita Ferreira Cabrera*
Universidad de Concepción
Concepción / Chile

Jéssica Elejalde Gómez**
Universidad de Concepción
Concepción / Chile

RESUMEN: El CAELE es una colección de 418 textos escritos producidos por 62 estudiantes de Español como Lengua Extranjera (ELE), recogidos entre los años 2014 y 2015 y guardados y procesados en formato digital. El objetivo principal de este estudio es determinar los errores más frecuentes y recurrentes de ese corpus, con el objetivo de apoyar a la selección de errores adecuados para los procesos de corrección gramatical mediante las estrategias de Feedback Correctivo Escrito (FCE). Los resultados indican que los errores más frecuentes y recurrentes en el CAELE son los de falsa selección de género gramatical y la omisión de la tilde ortográfica. En cuanto a los niveles de competencia A2 y B1, el error más recurrente en ambos niveles corresponde a la omisión de la tilde ortográfica en las palabras esdrújulas. Aunque hay algunas diferencias:

[1] El estudio de recurrencia de errores en ELE que se presenta en este artículo se desarrolló en el contexto del proyecto de investigación Fondecyt N.° 1140651 «El *feedback correctivo escrito* directo e indirecto en la adquisición y aprendizaje del Español como Lengua Extranjera». (Investigadora responsable: Dra. Anita Ferreira Cabrera)

* aferreir@udec.cl

** jelejalde@udec.cl

omisión de tilde ortográfica en las palabras llanas y en hiatos en el nivel A2, y falsa selección de género gramatical y omisión de acento en diacríticos en el nivel B1. Estos resultados sugieren una tendencia importante en el orden en que se deben tratar los errores según el nivel de competencia en el proceso de corrección por escrito.
PALABRAS CLAVE: corpus de aprendices de ELE; análisis de errores en ELE; errores recurrentes en ELE; feedback correctivo escrito.

RESUMO: O CAELE é uma coleção de 418 textos escritos por 62 aprendizes de Espanhol como Língua Estrangeira (ELE). Esses textos foram coletados entre os anos 2014 e 2015 e estão armazenados e processados em um formato digital. O objetivo principal deste estudo é determinar os erros linguísticos mais frequentes e recorrentes neste corpus, com a finalidade de apoiar a seleção de erros adequados para os processos de correção gramatical através de estratégias de *Feedback* Corretivo Escrito (FCE). Os resultados evidenciam que os tipos de erros mais frequentes e recorrentes no CAELE correspondem à falsa seleção do gênero gramatical e à omissão de acento nas palavras. Quanto aos níveis de proficiência A2 e B1, o erro mais frequente em ambos corresponde à omissão de acento nas palavras proparoxítonas. Porém, há algumas diferenças no nível A2, como a omissão de acento nas oxítonas (ou agudas) e nos hiatos e, no nível B1, a falsa seleção do gênero gramatical e a omissão de acentos nos diacríticos. Esses resultados sugerem uma tendência importante para a ordem que devem ser tratados os erros segundo o nível de proficiência num processo de correção escrita.
PALAVRAS-CHAVE: Corpus de aprendizes de ELE; análise de erros em ELE; erros recorrentes em ELE; *Feedback* Corretivo Escrito.

ABSTRACT: The CAELE is a collection of 418 written texts produced by 62 learners of Spanish as a foreign language (SFL). These texts were collected between years 2014 and 2015 and are saved and processed in digital format. The main objective of this study is to determine the most frequent and recurrent errors from that corpus with the aim of supporting the selection of adequate errors for the processes of grammatical correction with Written Corrective Feedback (WCF) strategies. Results indicate that the most frequent and recurrent errors in the CAELE are those of false selection of grammatical genre and the omission of orthographic word stress. Regarding proficiency levels A2 and B1, the most recurrent error in both levels corresponds to the omission of orthographic stress in last-syllable-stressed words. Although there are some differences: omission of orthographic stress in last-syllable-stressed words and hiatus in level A2, and false selection of grammatical genre and stress omission in diacritics at level B1. These results suggest an important tendency in the order in which errors should be treated according to proficiency level in the process of written correction.
KEYWORDS: SFL learner corpus; error analysis in SFL; recurrent errors in SFL; Written Corrective Feedback.

Introducción

La metodología basada en Lingüística de Corpus (LC) se ha ido implementando con mayor frecuencia en la investigación de Adquisición de Segundas Lenguas (ASL). Esto debido al reciente interés por el uso de la lengua en contextos reales, en que el investigador puede acceder a colecciones de producciones orales o escritas provenientes de diferentes situaciones comunicativas reales en formato digital. El corpus de aprendices constituye un valioso recurso para la investigación en el ámbito de la ASL y en la enseñanza de L2. Un corpus electrónico de aprendices de lengua (del inglés Computer Learner Corpora, CLC) consiste en la colección de datos lingüísticos auténticos (textos orales o escritos) en que se constata el uso de la lengua objeto de estudio (segunda lengua, L2, o lengua extranjera, LE) (GRANGER, 2003; 2004). Estas colecciones de textos pueden procesarse posteriormente por medio de herramientas de software especializado para la recuperación de información acorde con determinados criterios de búsqueda. En este contexto, si bien que uno de los objetivos principales en la ASL es construir modelos de las representaciones subyacentes al proceso de aprendizaje de un estadio en particular durante la adquisición de una L2, los métodos de recolección se han realizado a través de estudios experimentales, observacionales o métodos introspectivos. Así, como lo señala Granger (2004), la investigación en ASL ha favorecido los métodos con variables controladas y experimentales, relegando los datos lingüísticos del uso natural de la lengua.

La principal motivación para el uso de este tipo de procedimientos se relaciona con el hecho de que muchas veces no es posible elicitar las estructuras investigadas de forma natural. Por esta razón, la producción debe prepararse acorde con el interés del estudio y en muchos casos forzarse para conseguir los datos necesarios. También existen dificultades para controlar las variables que afectan la producción oral o escrita en un contexto no experimental (MACKEY; GASS, 2015). No obstante, en el ámbito de la enseñanza de la lengua, la necesidad de observar cómo usan de manera real la lengua meta los sujetos, por ejemplo, en una situación comunicativa en que utilicen lo aprendido en clase, puede arrojar resultados que indiquen qué áreas presentan mayor dificultad, cuáles son persistentes o cuáles responden a diferentes fenómenos dentro del desarrollo de la interlengua. Por esta razón, el surgimiento de la investigación basada en corpus de aprendientes ha permitido un acercamiento al uso de la L2 con facilidades de acceso a una mayor cantidad de muestras de lengua en formato digital. Por consiguiente,

CLC se centra en la descripción de la interlengua de los aprendices según el enfoque e interés del estudio en particular. Así esta línea de investigación 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.

El análisis de errores en general se ha centrado en la identificación de frecuencia, sistematicidad y gravedad de los errores. Sin embargo, los resultados de frecuencia de errores muestran solo una parte del análisis, dado que corresponden a la sumatoria total y repetida de cada error en un mismo texto. De esta forma, la frecuencia relega la importancia de identificar cuáles de estos errores son recurrentes a través del tiempo en los textos escritos por un mismo sujeto. En este artículo se presenta un estudio de Análisis de Errores Asistido por Computador (del inglés, *Computer Aided Error,* CEA) basado en los procedimientos de la lingüística de corpus, específicamente de la línea de corpus de aprendices. El estudio se fundamenta en investigaciones previas relacionadas con errores de interlengua y estrategias de *feedback correctivo* en Español como Lengua Extranjera (ELE) y en el proyecto de investigación por Ferreira Cabrera (2014) Fondecyt N.° 1140651: «*El feedback correctivo escrito directo e indirecto en la adquisición y aprendizaje del español como lengua extranjera*». El objetivo principal es determinar los errores lingüísticos más frecuentes y recurrentes en el CAELE y de esta forma apoyar la selección de errores adecuados para los procesos de corrección a través de estrategias de Feedback Correctivo Escrito (FCE).

El artículo se organiza en las siguientes secciones: En la sección 1, nos referimos a los principales fundamentos teóricos en el análisis de errores y corpus de aprendices. En la sección 2, abordamos el estudio sobre el análisis de errores recurrentes. En la sección 3, presentamos los resultados de las frecuencias y recurrencias de errores de todo el corpus CAELE. Finalmente, en la sección 4 presentamos algunos comentarios finales, proyecciones sobre los avances y logros obtenidos en esta investigación.

1 Fundamentos teóricos

El Análisis de Errores Asistido por Computador (del inglés Computer-aided Error Analysis, CEA) es un enfoque de investigación basado en corpus electrónico de aprendices (del inglés Computer Learner Corpora, CLC) y en LC para la identificación, clasificación y explicación de los errores de forma digital (DAGNEAUX; DENNESS; GRANGER, 1998). Este tipo de estudio permite procesar los errores de lengua y su frecuencia en

el contexto de un texto (oral o escrito) a través de herramientas de análisis computacional (GRANGER, 2004). La metodología de análisis de errores asistido por computador se basa en los procedimientos establecidos por Corder (1975): la identificación, la descripción, la clasificación, la explicación y la evaluación de la gravedad del error (CORDER, 1975; ELLIS, 1997; FERREIRA CABRERA; ELEJALDE GÓMEZ; VINE JARA, 2014; TORIJANO, 2006). Junto con dichos procedimientos, se realiza un etiquetado semiautomático con software especializado para etiquetado de corpus, de acuerdo con un sistema de anotación previo. Finalizado el etiquetado, se realiza el procesamiento de los datos para explorar el corpus y obtener los resultados según las variables del estudio.

En español como lengua extranjera, han sido pocos los estudios llevados a cabo con la metodología asistida por computador. Algunos de los estudios destacados son el de Campillos Llanos (2014b) relacionado con la observación del uso de las preposiciones en el habla no nativa de nivel intermedio de diferentes lenguas maternas contrastadas con el español. En otro estudio de Campillos Llanos (2014a) se aborda el análisis de errores léxicos en el español oral como LE. Los resultados muestran que los errores formales (ortografía y morfología) son más frecuentes en el nivel A2[2] y que persisten en el nivel B1[3] con menos frecuencia. Ambos estudios fueron procesados y analizados por medio del software GRAMPAL y SketchEngine, en que la obtención de listas de frecuencias y concordancias de forma automática permitió evidenciar los fenómenos investigados.

Por otra parte, Ferreira Cabrera, Elejalde Goméz y Vine Jara (2014), presentan un estudio sobre análisis de errores asistido por computador basado en un corpus de aprendientes de ELE. El corpus está compuesto por 84 resúmenes escritos por 22 estudiantes del nivel B1 en modalidad expositiva, narrativa y argumentativa. Se examinan los errores de escritura cometidos por estudiantes de ELE al desarrollar una tarea de producción escrita de un resumen por medio del computador. El objetivo es determinar los tipos de errores más frecuentes de aprendientes de ELE de nivel B1.

[2] En el nivel A2 el alumno puede comunicarse en actividades simples y cotidianas que requieren de intercambios sencillos y directos de información sobre cuestiones que le son conocidas o habituales.

[3] En el nivel B1 el aprendiz es capaz de producir textos sobre experiencias, acontecimientos, deseos y aspiraciones, así como justificar brevemente sus opiniones o explicar sus planes.

Los resultados del etiquetado y procesamiento del corpus sugieren una mayor tendencia en el uso incorrecto de las categorías gramaticales, la falsa selección de género y número en la concordancia gramatical y la omisión de tildes en la ortografía acentual.

Los estudios aquí expuestos corroboran la utilidad que ofrece un análisis computarizado por permitir una exploración rápida y precisa con datos a gran escala almacenados electrónicamente. La recuperación del etiquetado puede configurarse a partir de los criterios (lingüístico frente a lengua materna o nivel de competencia) según las variables y objetivos del estudio.

1.1 Corpus de Aprendices de ELE

En el ámbito de la lingüística aplicada a la enseñanza y aprendizaje del español como LE o L2, existen corpus orales y escritos recolectados en formato electrónico. 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. Algunos de estos corpus son, por ejemplo, el Corpus Escrito del Español como L2 (CEDEL2) de aprendices de inglés como L1. Este corpus considera producciones textuales en todos los niveles de competencia y ha sido recolectado con fines contrastivos entre el español y el inglés (LOZANO; MENDIKOETXEA, 2013). CEDEL2 está conformado por alrededor de 750 000 palabras en formato electrónico y ha sido recolectado en línea de forma sistemática mediante estudiantes y profesores voluntarios. El corpus está disponible como fuente de datos para los investigadores de ELE y con abundancia de ejemplos para la enseñanza de ELE.

Otro corpus interesante es el Corpus de Aprendices de Español como Lengua Extranjera (CAES) (ROJO; PALACIOS MARTÍNEZ, 2015). Corresponde a un conjunto de textos escritos producidos por estudiantes de español como LE 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 diferentes 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.

El Corpus de Interlengua Oral de italiano y español como LE (CORINÉI) comprende textos orales (grabación de las interacciones por Skype, transcripción y almacenamiento de las conversaciones) a partir del desarrollo del proyecto colaborativo *Teletándem* (GONZÁLEZ ROYO, 2012). Dicho proyecto está orientado a facilitar el aprendizaje colaborativo entre hablantes nativos y no nativos de español e italiano en la enseñanza aprendizaje de lenguas para la traducción. Los resultados han permitido investigar las estructuras oracionales de las conversaciones generadas en este tipo de ambiente de aprendizaje y, de este modo, contribuir al mejoramiento de las habilidades orales en ambas lenguas.

Con el propósito principal de describir la interlengua del aprendiz sueco acorde con los lineamientos del MCERL (2002), el Corpus de Suecos Aprendices de ELE SAELE (PINO RODRIGUEZ, 2009) está constituido por una colección digital de textos argumentativos escritos por aprendices de ELE provenientes de Suecia. Este corpus ha sido recopilado en dos universidades suecas, entre el 2008 y el 2009. Los niveles de competencia de los aprendices son A2 y B1. La exploración de este corpus ha permitido obtener las frecuencias de usos en estructuras gramaticales o de coherencia textual, presentes en la interlengua de aprendientes suecos. El estudio realizado por Pino Rodríguez (2012) se centró en el análisis de los conectores «porque, por eso y entonces» en la interacción con diferentes tipos de palabras.

El creciente interés por este tipo de investigaciones en corpus pone de manifiesto la necesidad de identificar las problemáticas o fenómenos que atañen al uso de la lengua meta en diversos contextos comunicativos. Los corpus de aprendices 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 y los estudios en análisis de interlengua (CRUZ PIÑOL, 2012).

1.2 El error: la frecuencia, la sistematicidad y la recurrencia

El error se define como la desviación sistemática de la norma de la lengua meta producida por el hecho de que un aprendiente todavía no ha internalizado la forma correcta o desconoce la regla (CORDER, 1975; ELLIS, 1997; LONG; LARSEN-FREEMAN, 1991). Los errores son parte natural del aprendizaje de la lengua meta y revelan patrones de desarrollo de la interlengua, donde se puede observar fenómenos relacionados con la sobregeneralización de reglas, transferencia de la lengua materna o evasión

de la norma (FERREIRA CABRERA; ELEJALDE GÓMEZ; VINE JARA, 2014). En ese sentido, los errores se identifican de acuerdo con el número de veces que aparecen, el contexto y la repetición de estos. Para ello, existen dos formas de precisar la identificación de los errores en valor numérico, las cuales describe Torijano (2006):

> **1. La frecuencia**: entendida como el número de veces que ocurre un error en cualquier muestra de textos orales o escritos. Esta frecuencia, en términos generales, describe la interlengua de la muestra que se observa y puede determinarse según dos modalidades, la absoluta y la relativa. La primera es el conteo por cada error y su aparición, a diferencia de la segunda que corresponde al número de ocurrencias que aparece un error en proporción con la extensión de un texto. El resultado será un criterio numérico útil para determinar la gravedad de los errores en relación con la producción escrita u oral.
> **2. La sistematicidad**: a la hora de valorar los errores, la sistematicidad está relacionada con qué tipo de errores son sistemáticos y repetitivos dentro de la misma producción textual de un sujeto. Este término permite evaluar la gravedad del error para determinar si son desviaciones sistemáticas, lapsus o equivocaciones.

En lingüística de corpus, la necesidad de identificar la recurrencia de errores está relacionada con dicha sistematicidad de errores que especifica Torijano (2006). Es decir, analizar los errores que se mantienen a lo largo del tiempo en un determinado proceso de aprendizaje. Este tipo de medición permite registrar los errores reiterativos en distintas producciones de un mismo sujeto. El análisis de recurrencia de los errores de un corpus contribuye a identificar qué errores aún persisten incluso durante o después de la instrucción formal en la clase de ELE. En ese sentido, el procedimiento metodológico de este tipo de análisis permite discernir si un estudiante o grupo domina o no una regla.

2 El estudio

El diseño de investigación de este estudio es descriptivo y longitudinal con un enfoque de análisis de datos cuantitativo. El objetivo general es determinar los errores más frecuentes y recurrentes en el CAELE. Para ello, se definen 3 objetivos específicos.

1. Delimitar la frecuencia total de errores en el CAELE.
2. Determinar la recurrencia de errores en textos del CAELE, producidos por los mismos sujetos.
3. Precisar la recurrencia de errores en ELE acorde con los niveles de competencia A2 y B1.

2.1 Muestra de aprendientes de ELE

La muestra se constituyó por un total de 62 estudiantes universitarios extranjeros de una universidad chilena, correspondientes a la cohorte 2014 (15 sujetos= 24 %) y 2015 (47 sujetos= 76 %) de cursos de español general como lengua extranjera.

Los sujetos fueron evaluados mediante el examen de Certificación del Español como Lengua Extranjera (CELE) (FERREIRA CABRERA; VINE JARA; ELEJALDE GÓMEZ, 2013) para determinar el nivel de competencia y se distribuyeron en dos grupos según el nivel de competencia A2+ con 26 (42 %) sujetos y B1 con 36 (58 %) sujetos.

Las lenguas maternas de los sujetos corresponde al alemán con 20 sujetos (32 %), francés 17 (27 %), inglés 17 (27 %), portugués 2 (3 %), sueco 2 (3 %), checo 2 (3 %), italiano 1 (2 %) y ruso 1 (2 %).

TABLA 1: Muestra del CAELE

Sujetos por nivel		No. Sujetos	Total sujetos	No. Textos	Total textos	%=418
A2+	2014	5	26	29	176	42%
	2015	21		147		
B1	2014	10	36	60	242	58%
	2015	26		182		
Totales		**62**	**62**	**418**	**418**	**100%**

Como se observa en la Tabla 1, cada sujeto escribió un total de 6 textos en la cohorte 2014 y 7 textos en la cohorte 2015. El número de textos producidos en el nivel A2 corresponde a un total de 176 textos (42 %) y a 242 textos (58 %) en el nivel B1. Cada texto se generó en el marco del desarrollo normal de las clases de ELE.

2.2 El corpus de aprendices de español como lengua extranjera, CAELE

El CAELE es una colección de 408 textos digitales de aprendices de español como lengua extranjera generados en el contexto de los proyectos de investigación Fondecyt N.° 1110812 y N.° 1140651 (FERREIRA CABRERA, 2011; 2014). Estos textos se recolectaron durante tres periodos de clases de ELE entre los años 2014 y 2015. Los sujetos de la cohorte 2014 escribieron 6 textos cada uno con un total de 89 textos. La cohorte 2015 escribió 7 textos por cada sujeto con un total de 329 textos. De los 418 textos, los aprendientes del nivel A2+ produjeron un total de 176 (42 %) textos y los del nivel B1 242 (58 %). La modalidad textual predominante es la narrativa.

2.3 Metodología de la investigación

La metodología de este estudio se sustenta en el enfoque de investigación del Análisis de Errores asistido por Computador en lo que corresponde a la recopilación del corpus y al procesamiento de los datos a través del software *Uam Corpus Tool* versión 3.2[4]. Para la identificación, clasificación y anotación de los errores, se procedió con los criterios y taxonomía de errores del modelo metodológico desarrollado por Ferreira Cabrera (2011; 2014).

2.3.1 Recopilación del corpus

Para la recolección del CAELE se consideró una serie de pasos preliminares con el objeto de asegurar que los textos fueran auténticos y pertinentes a la investigación (FERREIRA CABRERA; ELEJALDE GÓMEZ; VINE JARA, 2014). Estos textos escritos se produjeron en un período de tres meses con un total de 12 semanas. La escritura de los textos estuvo circunscrita en el desarrollo normal de las clases presenciales de los cursos de ELE general en los niveles A2+ y B1, los aprendices escribieron en promedio un texto por semana. La extensión de los textos fue de 150 a 200 palabras en el nivel A2 y de 250 a 300 en el B1.

[4] Este Software fue desarrollado en el contexto del Ministerio de Educación y Ciencia de España bajo el número de licencia HUM2005-01728/FILO (The WOSLAC Project). Para mayor información consultar www.wagsoft.com/Corput/Tool

La elección de los temas para la escritura se basó en el análisis de necesidades de los sujetos realizado en el estudio Ferreira Cabrera, Elejalde Gómez, Vine Jara (2014). A partir de dicho análisis, se eligieron los siguientes temas: (1) música folclórica del país de origen; (2) música latinoamericana; (3) artesanía chilena; (4) pueblos originarios de Latinoamérica y Chile; (5) literatura latinoamericana; (6) música folclórica latinoamericana, y (7) música juvenil chilena (para la cohorte 2015).

Los textos se escribieron en el bloc de notas con dos propósitos: facilitar la digitalización del corpus en formato electrónico y cautelar el uso de correctores ortográficos que intervinieran en el uso real por parte del sujeto (por ejemplo, Word). También se limitó el uso de internet, de diccionario traductor o celular con el mismo fin de evitar ayudas externas que interfirieran en los datos reales de la escritura. Una vez finalizada la escritura, cada texto fue guardado en formato *.txt* con codificación UTF-8*,* ordenada y compilada para procesarlos en el software *Uam Corpus Tool* versión 3.2.

2.3.2 Identificación y clasificación de los errores

Según los procedimientos establecidos por Corder (1975), esto es, identificar, clasificar y explicar los errores encontrados, se procedió al análisis del corpus a partir de una taxonomía para la identificación y clasificación de errores de tipo lingüístico. La taxonomía se desarrolló durante la ejecución del proyecto Fondecyt N.° 1110812 (FERREIRA CABRERA, 2011) y se depuró en el proyecto N.° 1140651 (FERREIRA CABRERA, 2014).

Como se observa en la Tabla 2, los niveles de categorización en la taxonomía corresponden a: gramática (categorías gramaticales y coherencia textual), léxico (a nivel morfológico y de significado) y ortografía (acentual, literal, dierética y diacrítica). El criterio lingüístico se fundamenta en la visión del error como un proceso sistemático que todo aprendiente de lenguas experimenta durante su aprendizaje de L2 (ELLIS, 1997). Autores como Vázquez (2009), Quiñones (2009a, 2009b) han desarrollado taxonomías para ELE basadas en el sistema lingüístico del español, para dar cuenta de los errores por cada categoría y nivel de escritura (palabra u oración).

TABLA 2: Taxonomía para la identificación y clasificación de errores, proyecto Fondecyt N.° 1140651 (FERREIRA CABRERA, 2014)

Criterio	Nivel	Categoría	Subcategoría	Tipo		Descripción del error
Lingüístico	Palabra y oración	Gramática	Categorías gramaticales	Preposición		ADI (adición) FS (falsa selección) OMI (omisión) FE (forma errónea) EE (elección errónea)
				Artículo		
				Pronombre		
				Verbo		
				Adjetivo		
				Sustantivo		
				Adverbio		
			Coherencia textual	Concor. Sintáctica	Nominal	
					Verbal	
					Atributo	
				Puntuación		
				Conectores		
	Palabra	Léxico	Léxico creado por derivación			
			Léxico morfología	Verbal		
				Sustantivo-adjetivo		
				Derivación		
			Léxico significado	Forma errónea		
				Elección errónea significado		
				Léxico inncesario		
		Ortografía	Acentual			
			Literal			
			Dierética			
			Diacrítica			

La anotación de errores se basa en la identificación de las modificaciones o usos de la L2, las cuales clasifican el error en: (1) omisión de elementos oracionales o categorías gramaticales (OMI: "…_____ * *Gente siempre hablan rapido*…"); (2) falsa selección (FS: "…*voy* ***con** bus a Viña*…") de cualquier elemento dentro del sintagma nominal o verbal; (3) forma errónea (FE: "… *esto pasa* ***por qué** no conoce la biblioteca*…") en la elección de la forma acorde con el significado; (4) adición de elementos gramaticales (ADI: "… *he decidido* ***de** aprender*…"), y (5) elección errónea de la realización semántica de una palabra o la ortografía de una palabra en específico (EE: "…*los* ***musicos** chilenos son importante para la gente*…") (ALEXOPOULOU, 2006; QUIÑONES, 2009a).

El proceso de identificación se realizó en primera instancia de forma manual en formato digital en Word. Cada texto escrito en *txt*, fue copiado en un documento Word y se le asignó a cada error la etiqueta correspondiente según las categorías establecidas. De esta forma se aseguró un primer proceso de identificación para probar la taxonomía (FERREIRA CABRERA, 2014),

identificar nuevas tendencias o determinar nuevos fenómenos que fuera necesario incluir en el etiquetado.

FIGURA 1: Ejemplo para la identificación de errores

	Fecha realización: 30-09-2015
Sujeto: 33	Grupo: 1
Nivel: B1	Tarea: Tarea 1
L1: Francés	Tema: Artesanía chilena
L2: inglés	Etiquetador: 1

La artesanía chilena me parece muy elaborada y ***respendida*** [FS-lex—creado-derivación] en todo el ***país*** [omi-orto-hiato].

Cada ***region*** [EE-orto-hiato] tiene su ***propre*** [FS-lex-morf-raíz] artesanía, en el sentido que cada ***region*** [EE-orto-hiato] tiene productos típicos.

Por ***exiemplos*** [FS-lex-morf-raíz], sobre la isla de ***Chiloe*** [omi-orto-aguda], podemos encontrar muchas ropas, gorros, [adi-coma-enumerativa] o panchos para ***protegarse*** [FS-lex-morf-raíz-verbo-reg] del frio.

En el norte se **encontra** [FS-lex-morf-raíz-verbo-irreg] ***orfebreria*** [omi-orto-hiato] hecha con productos de ***extracion*** [EE-orto-hiato] de las minas. Me gustaría llevar un gorro y ***articulos*** [omi-orto-esdrújula] de cuero de Chile.

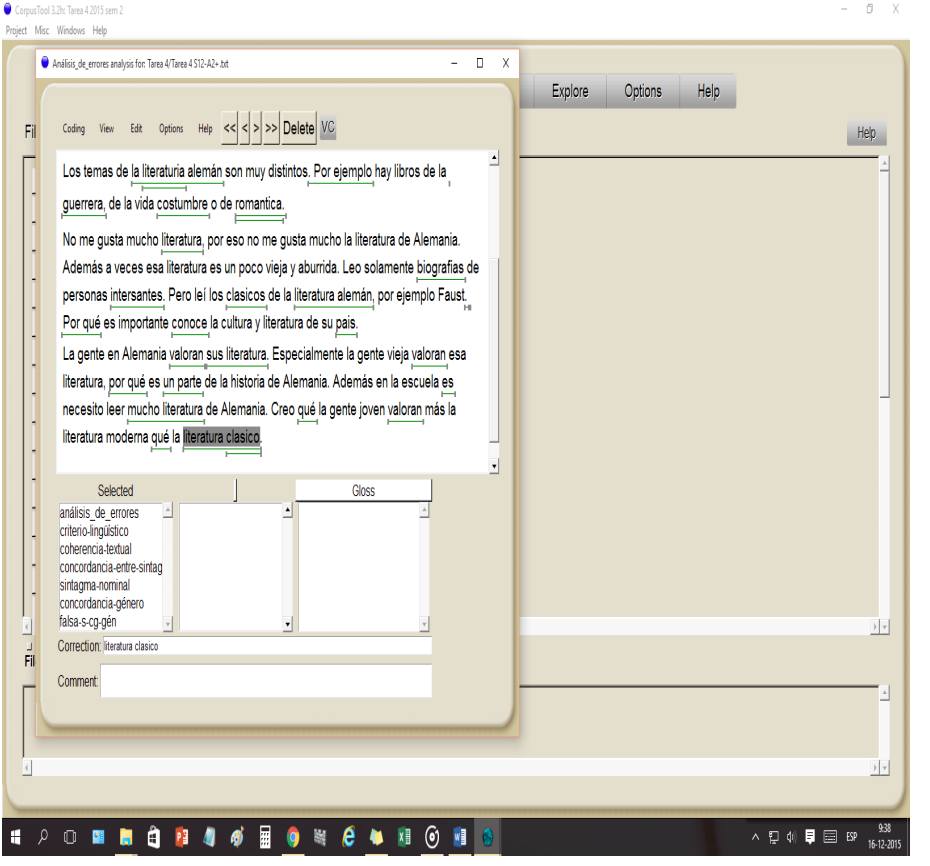
Como se puede ver en la Figura 1, los errores se marcaron con negrita y en cursiva para destacarlos en el texto. Una vez identificado el error se procedió a escribir la etiqueta entre corchetes de acuerdo con las etiquetas determinadas (FERREIRA CABRERA; ELEJALDE GÓMEZ; VINE JARA, 2014): (1) FS (falsa-selección: *FS-cg-género,* falsa selección en la concordancia gramatical de género); (2) OMI (omisión: *omi-art-def*, omisión del artículo definido); (3) ADI (adición: *adi-prep-de*, adición de la preposición de), y (4) EE (elección errónea: *omi-orto-aguda*, omisión de la tilde en la acentuación aguda).

2.3.3 Etiquetado de errores en el software Uam Corpus Tool

Para etiquetar los errores, se utilizó el programa especializado en anotación de corpus desarrollado por Mick O´Donell (2008) en el contexto del Ministerio de Educación y Ciencia de España bajo el número de licencia HUM2005-01728/FILO (The WOSLAC Project). El software es de distribución gratuita con fines específicos para la investigación y cuenta con herramientas para el etiquetado manual, semiautomático o automático. Este

software permite anotar los errores a partir de un esquema de anotación de forma manual y asignarle atributos al texto para obtener resultados acordes con las variables del estudio (lengua materna, nivel, entre otros). En la Figura 2, se puede observar un ejemplo de etiquetado de un texto.

FIGURA 2: Ejemplo de etiquetado en el software *Uam Corpus Tool*



La interfaz de este programa es sencilla e intuitiva y permite al usuario interactuar entre los datos, el sistema de anotación y los procesamientos estadísticos. Una de las ventajas de este programa es la opción de etiquetado múltiple, que permite etiquetar un error en diversos niveles. Como es el caso del error encontrado en el sintagma nominal *"…*literatura *clasico…" (Sujeto 12, alemán texto No. 4),* cuya identificación de errores corresponde a

dos subniveles dentro del criterio lingüístico, estos son el sintáctico (falsa selección de la concordancia gramatical: *literatura clásica*) y el ortográfico (omisión de la tilde en palabra esdrújula: *clásica*). Al finalizar con este proceso de etiquetado, se obtuvo la frecuencia de los errores y los contextos del uso erróneo por cada categoría y criterio establecido.

2.3.4 Procesamiento de los datos: obtención de las matrices para el análisis

Pare determinar los errores más recurrentes del CAELE, se obtuvo la frecuencia de los errores en toda la muestra por cada sujeto acorde con el nivel de competencia. Para ello, el programa *Uam Corpus Tool*, ofrece la opción de procesar los datos a través de matrices en tres formatos: (1) descripción general del corpus (*describe a dataset*); (2) descripción del corpus comparando dos o más variables (*compare several datasets*), y (3) descripción de cada archivo digital del corpus (*describe each file*). Para este estudio se consideró como fuente principal la matriz generada con la descripción de los errores por cada sujeto en cada texto para el conteo de la recurrencia. En esta matriz se contabilizaron los errores repetidos por sujeto a partir de dos textos. Obsérvese un ejemplo de la matriz.

TABLA 3: Matriz por sujeto y conteo de errores por recurrencia

Análisis de recuerrencia nivel A2+ por sujeto						
Sujeto: 8 L1: alemán						
Tipo de error	**conteo errores por recurrencia**					
ARTÍCULO DEFINIDO	**Texto 1**	**Texto 2**	**Texto 3**	**Texto 4**	**Texto 5**	**Texto 6**
omi-art-def	0	**1**	0	0	**2**	**1**
falsa-s-art-def	0	0	0	0		0
adi-art-def	0	0	0	0		0
PREPOSICIONES						
omi-prep-a	0	0	0	1	1	0
omi-prep-de	0	**2**	**1**	**1**	**1**	**1**
VERBO-MODO						
falsa-s-modo-subj	0	0	**3**	**1**	**2**	0
falsa-s-modo-indic	0	**1**	0	**1**	**1**	0
ORTOGRAFÍA						
omi-orto-esdrújula	**4**	**2**	**1**	**3**	**1**	**1**
ee-orto-acen-esdrújula	0	0	0	**1**	**1**	0

La tabla 3 muestra un ejemplo de la identificación de los errores cometidos en cada texto, su frecuencia y marcación por recurrencia. Para ello, el tipo de error se marcó con sombreado gris las veces en que el sujeto incurrió en un error a través de los textos. Por ejemplo, en el error «*omi-prep-de*» el sujeto incurrió en los textos 2, 3, 4, 5, 6 (un total de 5 textos), con una frecuencia entre una o dos veces de aparición por cada texto. Esto indica una tendencia del error que se mantiene en el tiempo de manera consistente. También se puede observar que en el error «*omi-art-def*», el sujeto sólo incurrió en tres textos, sugiriendo que el error no es consistente y aparece de manera menos constante.

Para contar los errores recurrentes se procedió de acuerdo con el criterio de determinar cuántos errores de «x categoría gramatical o tipo de error» aparecían en «x textos». De esta forma, se aseguró la identificación de todas las recurrencias posibles en los textos por sujeto y la clasificación de los errores más recurrentes en toda la muestra.

Terminado este primer proceso, los datos se cuantificaron en una tabla Excel para obtener la frecuencia absoluta y los porcentajes relativos con respecto a la recurrencia por nivel de competencia. Sobre la base de estos datos, se determinó que la recurrencia se clasificaría a partir de un número mínimo de 4 sujetos por nivel.

En la Tabla 4, se ilustra un ejemplo de recurrencia de errores, se puede observar que de 26 sujetos del nivel A2, solo 6 tipos de errores fueron recurrentes en al menos 5 textos (sombreado amarillo). Por ejemplo, el error de la falsa selección de la preposición «*a*» (4 sujetos) y la omisión de la tilde diacrítica (10 sujetos) son errores recurrentes, dado que los sujetos cometieron repetidamente estos errores en al menos 5 textos. Los demás errores tuvieron menos recurrencias, tan solo uno o tres sujetos incurrieron en el mismo error, como el caso de la omisión del artículo (2 sujetos) y la falsa selección en la concordancia gramatical de número (3 sujetos). Posteriormente, se agruparon las categorías gramaticales que abarcaban de 2 a 7 textos y se registró el número de textos en los cuales era recurrente un problema lingüístico. Finalmente, se identificaron los cinco errores más recurrentes por cada nivel y se compararon con los errores más frecuentes de toda la muestra. Esto con el objeto de identificar si las tendencias de los errores más frecuentes se relacionaban con el análisis de la recurrencia.

TABLA 4: Ejemplo de recurrencia de errores presente en 5 textos en el nivel A2

Recurrencia en 5 textos de 26 sujetos			
Categoría	**Tipo de error**	**Fr. Abs**	**No. Sujetos**
Artículo	omi-art-def	18	2 de 26
Preposición	omi-prep-a	20	3 de 26
	fs-prep-a	32	4 de 26
Verbo	Fs-verbo-estar	13	1 de 26
Ortografía	omi-orto-aguda	89	6 de 26
	omi-orto-esdrújula	194	10 de 26
	omi-orto-hiato	46	4 de 26
	omi-orto-diacrítica	33	3 de 26
	ee-orto-diacrítica	10	1 de 26
	omi-orto-may-incial	14	1 de 26
Concordancia gramatica	FS-CG-gen	79	5 de 26
	FS-CG-num	27	3 de 26
	fs-cg-sujeto-predicado	50	4 de 26

3 Resultados

3.1 Frecuencias de errores en el CAELE

De acuerdo con el primer objetivo específico del estudio, se procedió a delimitar la frecuencia de errores en el CAELE a partir de una matriz general de frecuencia por año. El total de errores identificados según el criterio lingüístico fue de 8731 (véase Tabla 5), los cuales se distribuyen en cuatro categorías con sus respectivas subcategorías: (1) gramática con un 29 % (2552 errores clasificados en omisión, falsa selección y adición de categorías gramaticales); (2) la coherencia textual con un 18 % (1569 errores clasificados como falsa selección de la concordancia gramatical); (3) la ortografía con un 39 % (3448 errores, distribuidos en la omisión y colocación errónea de la acentuación), y (4) el léxico con un 13 % (1162 errores, clasificados como aspectos morfológicos y semánticos).

En el contexto de los errores gramaticales, en lo referido específicamente a las *categorías gramaticales,* se puede señalar que la mayor problemática la presentan las preposiciones, con 989 errores (véase Tabla 4), lo que

representa un 39 % del total de 2552 errores. Esto muestra que, en general, la complejidad del sistema preposicional representa un área de dificultad mayor para los aprendices de ELE, lo que corrobora los resultados de investigaciones previas (CAMPILLOS LLANOS, 2014b; FERREIRA CABRERA; LAFLEUR, 2015).

Con respecto a los errores de la subcategoría *coherencia textual*, la falsa selección de la concordancia gramatical de género es el tipo de mayor frecuencia con 968 errores, lo que corresponde a un 62 %, esto se puede explicar por el hecho de que, en muchos casos, los aprendices de ELE generalizan o desconocen las reglas de concordancia.

En la *ortografía* acentual, las palabras esdrújulas presentan un total de 823 errores, cuya problemática responde a la omisión y a la falsa selección de la tilde. En este aspecto, la ortografía puede verse influenciada por varios factores como, por ejemplo, la influencia de la lengua materna, el desconocimiento de la regla o la sobregeneralización de las normas de acentuación.

TABLA 5: Resultados generales del CAELE: criterio lingüístico

Tipo de errores				2014	2015-1	2015-2	SUBTOTAL	% =subtotal	% =8731
Criterio Lingüístico	Gramática	Categorías gramaticales	Preposición	246	293	450	989	39%	29%
			Verbo	112	219	300	631	25%	
			Artículo	73	216	206	495	19%	
			Pronombres	46	106	133	285	11%	
			Adjetivo	19	18	47	84	3%	
			Adverbio	4	20	13	37	1%	
			Sustantivo	3	17	11	31	1%	
	SUBTOTAL			503	889	1160	2552	100%	
	Coherencia textual	Concordancia sintáctica	Género	244	186	538	968	62%	18%
			Número	78	67	197	342	22%	
			Suj-pred	49	101	109	259	17%	
	SUBTOTAL			371	354	844	1569	100%	
	Ortografía	Acentual	Esdrújula	134	248	441	823	24%	39%
			Aguda	146	214	334	694	20%	
			Grave	41	85	131	257	7%	
			Sobreesdrújula	2	12	13	27	1%	
		Dierética		91	150	251	492	14%	
		Diacrítica		100	153	301	554	16%	
		Literal		86	148	367	601	17%	
	SUBTOTAL			600	1010	1838	3448	100%	
	Léxico	Morfología	Sust-adj-adv	91	166	194	451	39%	13%
			Verbo	37	73	125	235	20%	
			Fs-morf-gén	13	33	66	112	10%	
			Fs-morf-num	0	0	8	8	1%	
		Forma errónea		14	13	48	75	6%	
		Falsa selección léxica		33	103	145	281	24%	
	SUBTOTAL			188	388	586	1162	100%	
TOTAL				1662	2641	4428	8731	100%	

3.1.1 Los errores más frecuentes por subcategoría

En la Tabla 6, se evidencia la distribución de los errores más frecuentes por subcategoría, en que se observa que *la falsa selección de género* corresponde al error más frecuente de todo el corpus con un total de 968 (30 %) ocurrencias. A este le sigue *la omisión de la tilde en las esdrújulas* con un total de 791 (24 %) errores; luego *la omisión de la tilde en las agudas* con 625 (19 %) errores; *la omisión de la tilde diacrítica* con 439 errores (13 %), y *la falsa selección de uso de preposición* con 438 (13 %) errores. El total de estos errores corresponde al 37 % de la muestra, lo que equivale a 8731 errores. Dichas frecuencias sugieren que en este tipo de errores se concentran las problemáticas más importantes identificadas en el CAELE, las cuales podrían considerarse para un posterior tratamiento en el ámbito de la corrección gramatical.

TABLA 6: Errores más frecuentes por subcategoría

Errores más frecuentes de toda la muestra			
Nivel	**Tipo de error**	**Total**	**%=3261**
1	falsa-s-cg-gén	968	30%
2	omi-orto-esdrújula	791	24%
3	omi-orto-aguda	625	19%
4	omi-orto-diacrítica	439	13%
5	falsa-s-prep	438	13%
Total errores más frecuentes		**3261**	**100%**
Errores de la muestra: 8731			**37%**

3.2 Recurrencia de errores en el CAELE

En relación con el segundo objetivo de la investigación, se procedió a delimitar la recurrencia de errores en el CAELE. El análisis de la recurrencia se centra en el conteo de los errores observados en varios textos producidos en distintos momentos por un mismo sujeto. La recurrencia se considera a partir de un número mínimo de dos textos producidos por un mismo sujeto y corresponde a un problema sistemático sostenida en el tiempo, en que el error es persistente en el proceso de aprendizaje del estudiante. Además, en esta investigación, para delimitar la recurrencia de errores, se consideró un número mínimo de cuatro sujetos, puesto que así se cautela la representatividad de la problemática observada. El número de veces que varios estudiantes incurren en un mismo error a partir de la escritura de,

por lo menos, 2 textos puede sugerir tendencias en los resultados obtenidos, como, por ejemplo, si se trata de un error evolutivo, fosilizado u otros rasgos que pueden caracterizar la naturaleza del error y su interlengua.

Como se puede ver en el Tabla 7, el número total de errores recurrentes producidos por un mínimo de 4 sujetos corresponde a 2412 errores. Los resultados muestran que la ortografía presenta una recurrencia mayor con un 57 % del total de errores; seguida por la concordancia sintáctica con un 24 %; la preposición con un 10 %; el artículo con un 6 %, y el verbo con un 2,5 %.

Estos resultados se asemejan a algunos resultados de la frecuencia general, específicamente en lo que se refiere a los errores de ortografía acentual y de la concordancia gramatical de género. En consecuencia, estas tendencias representan las áreas más problemáticas que no sólo son frecuentes, sino que además se mantienen en el tiempo. Asimismo, la recurrencia de estos errores indica que corresponde a un tipo de error sistemático cuya interferencia en el proceso de aprendizaje dependerá, principalmente, de la gravedad del error en relación con los aspectos comunicativos (incomprensión del mensaje) y lingüísticos de la L2.

TABLA 7: Errores recurrentes de toda la muestra

Recurrencia por categoría general						
Categoría	**Tipo de error**	**Fr. Abs**	**%=2412**		**Textos**	**No. Sujetos**
Ortografía	omi-orto-esdrújula	490	20%	57%	6;5;4;3	38 de 62
	omi-orto-aguda	296	12%		6;5;4	22 de 62
	omi-orto-grave	20	1%		2	8 de 62
	ee-orto-acento-grave	64	3%		3;2	15 de 62
	omi-orto-hiato	226	9%		5;4;3;2	35 de 62
	omi-orto-diacrítica	219	9%		6;4;3;2	30 de 62
	ee-orto-diacrítica	7	0%		2	6 de 62
	ee-orto-lit-s	9	0%		2	4 de 62
	ee-orto-lit-z	9	0%		2	4 de 62
	omi-orto-mayúscula-inicial	31	1%		2	8 de 62
	fs-orto-mayús-inicial	12	0%		2	4 de 62
Concordancia sintáctica	fs-cg-gen	270	11%	24%	6;5;4;3	27 de 62
	fs-cg-gen-atributo	40	2%		2;4	12 de 62
	fs-cg-gen-epiceno	11	0%		2	4 de 62
	fs-cg-gen-anteced	9	0%		2	4 de 62
	fs-cg-núm	53	2%		4;2	13 de 62
	fs-cg-num-atributo	31	1%		2	12 de 26
	fs-cg-sujeto-pred	160	7%		5;4;3;2	28 de 62
Preposición	adi-prep-de	38	2%	10%	3;2	11 de 62
	fs-prep-de	37	2%		3;2	12 de 62
	fs-prep-a	59	2%		5;3;2	12 de 62
	adi-prep-a	16	1%		3	4 de 62
	omi-prep-a	58	2%		4;3	14 de 62
	fs-prep-para	18	1%		2	5 de 62
	omi-prep-en	7	0%		2	4 de 62
	fs-prep-por	13	1%		2	4 de 62
Artículo	omi-art-def	152	6%	6%	4;3;2	28 de 62
Verbo	fs-infinitivo	7	0%	2,5%	2	4 de 62
	fs-modo-indic	15	1%		2	6 de 62
	fs-modo-sub	22	1%		3;2	8 de 62
	fs-verbo-estar	13	1%		2	5 de 62
Total		2412	100%		7 textos	62 sujetos

3.2.1 Los errores más recurrentes por subcategoría

En la Tabla 8, se evidencia la distribución de los errores más recurrentes por subcategoría. Se puede constatar que *la omisión de la tilde en las esdrújulas* corresponde al error más recurrente de todo el CAELE con un total de 490 ocurrencias observadas en 38 sujetos. En segundo lugar, le sigue *la falsa selección de género* con un total 330 ocurrencias en 47 sujetos; luego figura *la omisión de la tilde en los hiatos* con 226 ocurrencias; *la omisión de la tilde diacrítica* con 219 ocurrencias; y *la falsa selección de uso en la preposición* con 127 errores. El total de 1392 errores corresponde al 58 % del total de los errores recurrentes (2412) del CAELE. Estas recurrencias sugieren que en estos errores debería focalizarse el tratamiento de errores a través de estrategias de Feedback Correctivo Escrito.

TABLA 8: Errores más recurrentes por subcategoría

Errores más recurrentes de toda la muestra					
Nivel	**Tipo de error**	**Recurrencia**	**%=1392**	**Textos**	**No. Sujetos**
1	omi-orto-esdrújula	490	35%	6;5;4;3	38 de 62
2	falsa-cg-gen	330	24%	6;5;4;3	47 de 62
3	omi-orto-hiato	226	16%	5;4;3;2	35 de 62
4	omi-orto-diacrítica	219	16%	5;4;3;2	30 de 62
5	falsa-s-prep	127	9%	6;4;3;2	33 de 62
Total		**1392**	**100%**	**%=2412**	**58%**

3.3 Recurrencia de errores por Nivel de Competencia en el CAELE

De acuerdo con el tercer objetivo de este estudio, se realizó un análisis de los datos para determinar los errores recurrentes por nivel de competencia.

3.3.1 Recurrencia de errores en el nivel A2

En el nivel A2, se determinó un número total de 1372 errores recurrentes (56 %) de 2412 errores recurrentes de todo el CAELE (véase Tabla 9). Dichos errores se distribuyen en 5 categorías: (1) la *ortografía* con un 57,9 %; (2) *la concordancia sintáctica* con un 22,5 %; (3) *la preposición* con un 8,8 %; (4) *el artículo* con un 6,6 %, y (5) *el verbo* con un 4,2 %.

Los errores recurrentes identificados condicen con el proceso de desarrollo de interlengua de niveles preintermedios como en el caso del A2. Los errores en el uso de las categorías gramaticales, la ortografía y la concordancia gramatical, persisten a lo largo del aprendizaje. En ese sentido, el conocimiento lingüístico que obtenga el sujeto en niveles posteriores, dependerá del éxito con que logre asimilar el sistema lingüístico del español y emplearlo en diferentes contextos comunicativos.

TABLA 9: Errores recurrentes nivel A2

Recurrencia resumida de 26 sujetos nivel a2						
Error	**Tipo**	**Fr.abs**	**%=1372**		**Textos**	**No. de sujetos**
Ortografía	omi-orto-aguda	194	14,1%	57,9%	5;6	11 de 26
	omi-orto-grave	10	0,7%		2	4 de 26
	omi-orto-esdrújula	266	19,4%		5;4	16 de 26
	omi-orto-hiato	156	11,4%		5;4;3;2	19 de 26
	omi-orto-diacrítica	80	5,8%		4	8 de 26
	ee-orto-acento-grave	20	1,5%		2	6 de 26
	ee-orto-diacrítica	7	0,5%		2	6 de 26
	ee-orto-lit-s	9	0,7%		2	4 de 26
	ee-orto-lit-z	9	0,7%		2	4 de 26
	omi-orto-mayúscula-inicial	31	2,3%		2	8 de 26
	fs-orto-mayús-inicial	12	0,9%		2	4 de 26
Concordancia sintáctica	fs-cg-num	37	2,7%	22,5%	4;2	9 de 26
	fs-cg-n-atributo	31	2,3%		2	12 de 26
	fs-cg-gen	79	5,8%		5	5 de 26
	fs-cg-gen-anteced	9	0,7%		2	4 de 26
	fs-cg-g-atrib	18	1,3%		2	8 de 26
	fs-cg-gen-epiceno	11	0,8%		2	4 de 26
	fs-cg-sujeto-pred	124	9,0%		5;4;3;2	19 de 26
Preposición	adi-prep-de	15	1,1%	8,8%	2	6 de 26
	fs-prep-a	42	3,1%		5;2	8 de 26
	fs-prep-de	14	1,0%		2	6 de 26
	fs-prep-para	18	1,3%		2	5 de 26
	fs-prep-por	13	0,9%		2	4 de 26
	omi-prep-a	12	0,9%		3	4 de 26
	omi-prep-en	7	0,5%		2	4 de 26
Artículo	omi-art-def	91	6,6%	6,6%	4;3;2	14 de 26
Verbo	fs-infinitivo	7	0,5%	4,2%	2	4 de 26
	fs-modo-indic	15	1,1%		2	6 de 26
	fs-modo-sub	22	1,6%		3;2	8 de 26
	fs-verbo-estar	13	0,9%		2	5 de 26
Total		1372	100%	100%		26

3.3.2 Errores más recurrentes en el nivel A2

A partir de la recurrencia observada en el nivel A2, se puede señalar que 5 son los errores más recurrentes (Tabla 10), los cuales suman un total de 831 errores (el 61 % del total de 1372 errores recurrentes del A2). Cuatro son los errores más recurrentes, a saber: (1) *la omisión de esdrújula* (32 %); (2) la *omisión de la aguda* (23 %); (3) *la omisión del hiato* (19 %); (4) *la falsa selección concordancia sintáctica sujeto-predicado* (15 %), y (5) *la omisión de artículo definido* (11 %).

TABLA 10: Errores más recurrentes del nivel A2

Errores más recurrentes del nivel A2					
Nivel	**Tipo de error**	**Recurrencia**	**%=831**	**Textos**	**No. Sujetos**
1	omi-orto-esdrújula	266	32%	5;4	16 de 26
2	omi-orto-aguda	194	23%	5;6	11 de 26
3	omi-orto-hiato	156	19%	6;4;3	19 de 26
4	Falsa-s-cg-suj-pred	124	15%	6;4	19 de 26
5	omi-art-def	91	11%	2;3	14 de 26
Total		**831**	**100%**	**%=1372**	**61%**

De estos errores, la omisión del artículo se puede explicar como una posible interferencia de la lengua materna de algunos sujetos, dado que lenguas como inglés y alemán, prescinden del uso de artículo en los casos donde se nombran objetos o eventos generales. Con respecto a la concordancia entre sujeto y predicado, muchos de los errores corresponden al desconocimiento de la regla de sustantivos colectivos que en español corresponden al singular, como por ejemplo, «…*la gente es*…» que en algunas lenguas concuerda en plural dado que es un sustantivo plural «…*people are*…». También se observa el desconocimiento de la regla de concordancia con los verbos gustar y encantar, en que los sujetos eligen incorrectamente la concordancia, ya que la relacionan con la persona gramatical que expresa el gusto por algo «…*me gusta* las películas romáticas*…».

3.3.3 Recurrencia de errores en el nivel B1

Del total de 2412 errores recurrentes en el CAELE, se delimitó un número de 1040 errores recurrentes en el nivel B1 que corresponden a un 43 %. Dichos errores se distribuyen, de manera general, en problemáticas relativas a la *ortografía* (acentual, dierética y diacrítica), *la concordancia sintáctica* (género y número) y las *categorías gramaticales de la preposición y el artículo*. Como se observa en la Tabla 11, el mayor porcentaje de errores se concentra en la ortografía, de los cuales, la omisión de la esdrújula uno de las más recurrentes.

La segunda problemática con mayor recurrencia se refiere a la concordancia gramatical con un 25 %, en que se observa una tendencia a la falsa selección del género gramatical con un 18 %, seguida de la falsa selección de la concordancia entre sujeto predicado con una recurrencia de 9 sujetos en tres textos. Con respecto al uso incorrecto de las categorías gramaticales, en la preposición se identificó la omisión de la preposición «a», lo que representa un 4 % de 46 errores con una recurrencia de 10 sujetos en 4 y 3 textos.

TABLA 11: Errores recurrentes del nivel B1

Recurrencia por categoría nivel B1						
Categoría	**Tipo de error**	**Fr. Abs**	**%=1040**		**Textos**	**No. Sujetos**
Ortografía	omi-orto-esdrújula	224	22%	57%	6;4;3	22 de 36
	omi-orto-aguda	102	10%		6;4	11 de 36
	ee-orto-acen-grave	44	4%		3;2	9 de 36
	omi-orto-grave	10	1%		2	4 de 36
	omi-orto-diacrítica	139	13%		6;4;3	16 de 36
	omi-orto-hiato	70	7%		2;3	16 de 36
Concordancia sintáctica	falsa-s-cg-gén	191	18%	25%	6;4;3	22 de 36
	falsa-s-cg-núm	16	2%		4	4 de 36
	falsa-s-cg-suj-pred	36	3%		3	9 de 36
	fs-cg-gen-atributo	22	2%		4	4 de 36
Preposición	omi-prep-a	46	4%	12%	4;3	10 de 36
	falsa-s-prep-de	23	2%		3	6 de 36
	adi-prep-de	23	2%		3	5 de 36
	falsa-s-prep-a	17	2%		3	4 de 36
	adi-prep-a	16	2%		3	4 de 36
Artículo	omi-art-def	61	6%	6%	3;2	14 de 36
Total		1040	100%		6 textos	36 sujetos

Estos resultados muestran que, en el nivel B1, los sujetos presentan una mejor precisión lingüística en comparación con los del nivel A2 y se asemejan a los resultados de estudios previos realizados por Ferreira Cabrera, Elejalde Gómez y Vine Jara (2014), Vázquez (2009) y Quiñones (2009a). En estos estudios se evidencian principalmente errores ortográficos, el uso incorrecto de las preposiciones, el artículo y la concordancia gramatical persisten en la transición de un nivel preintermedio (A2) al intermedio (B1).

No obstante, las problemáticas relacionadas con la concordancia sintáctica persisten a lo largo del aprendizaje, observándose así recurrencia en niveles preintermedios e intermedios. Asimismo, se observa una tendencia al uso incorrecto del sistema ortográfico, lo que podría interpretarse como el desconocimiento de las reglas de acentuación general, dierética y diacrítica independiente de la lengua materna de los sujetos (SÁNCHEZ JIMÉNEZ, 2009). En esta última, cabe destacar que el conocimiento y procesamiento de la ortografía es un aspecto complejo para cualquier aprendiz de lengua. Por esta razón, el tratamiento de los errores de ortografía debe focalizarse considerando el aprendizaje de aspectos prosódicos, fonológicos y lingüísticos que atañen a la lengua objeto de estudio.

3.3.4 Errores más recurrentes en el nivel B1

A partir de la recurrencia general observada en el nivel B1, se pueden delimitar los 5 errores más recurrentes de dicho nivel (Tabla 12). Estos suman un total de 726 errores (70 % del total de errores recurrentes del B1, 1040 errores), los cuales se distribuyen de la siguiente forma: (1) *omisión de esdrújula* con una recurrencia de un 31 %; (2) *concordancia sintáctica incorrecta de género gramatical* con un 26 %; (3) *omisión en la tilde diacrítica* con un 19 %; (4) *omisión de la tilde en palabras agudas* con un 14 %, y (5) *omisión de la tilde en hiatos* con un 10 %.

TABLA 12: Errores más recurrentes del nivel B1

Errores más recurrentes del nivel B1					
Nivel	**Tipo de error**	**Recurrencia**	**%=726**	**Textos**	**No. Sujetos**
1	omi-orto-esdrújula	224	31%	6;4;3	22 de 36
2	falsa-s-cg-gén	191	26%	6;4;3	22 de 36
3	omi-orto-diacrítica	139	19%	6;4;3	16 de 36
4	omi-orto-aguda	102	14%	6;4	11 de 36
5	omi-orto-hiato	70	10%	2;3	16 de 36
Total		**726**	**100%**	**%=1040**	**70%**

4 Conclusiones

Este artículo ha centrado su atención en la temática de la frecuencia y recurrencia de errores en ELE basada en un análisis de errores asistido por computador. Los principales objetivos han sido delimitar los errores lingüísticos más frecuentes y recurrentes en el CAELE.

El análisis de recurrencia entendido como una forma de identificar los errores frecuentes persistentes en un período determinado, permite dilucidar la sistematicidad, consistencia o gravedad de los errores. En dicho contexto, los errores pueden clasificarse según su frecuencia y sistematicidad, permitiendo así un análisis que contribuye en los procesos de corrección y mejoramiento de la habilidad escrita en ELE. En efecto, un aporte relevante que hace este tipo de análisis al área de la adquisición del español como lengua extranjera es poder determinar con mayor precisión los errores que deberían tratarse y focalizarse con tratamientos efectivos para su corrección, como los sugeridos en el ámbito del FCE.

Los resultados indican que los errores más *frecuentes* están relacionados con *falsa selección de género gramatical*, uso incorrecto de la *acentuación de las palabras esdrújulas, agudas y diacríticas y falsa selección de preposición*. En esta misma línea, los errores más *recurrentes* se registran en la *acentuación de las palabras esdrújulas y falsa selección de género gramatical* (aunque en orden inverso). Luego, a diferencia de las frecuencias, en tercer lugar figura la *problemática acentual en los hiatos*, que no coincide con los errores más frecuentes. Finalmente, los dos últimos errores coinciden también con los resultados obtenidos en las frecuencias generales, esto es: *omisión de la tilde en los diacríticos*, y *falsa selección de preposición*.

En cuanto al nivel de competencia, en ambos niveles el error más recurrente corresponde a *la omisión de la tilde en las palabras esdrújulas*. No obstante, luego se observan algunas diferencias como *omisión de la tilde en las palabras agudas* e *hiato* en el nivel A2 y la *falsa selección de género gramatical* y *omisión de tildes en diacríticos* en el B1. También se observan diferencias entre los niveles en los dos últimos errores. En el nivel A2 se observan *la falsa selección de concordancia gramatical entre sujeto y predicado* y *la omisión del artículo definido*. En cambio, en el nivel B1, destacan *la omisión de la tilde en palabras agudas* y *hiatos*. Todos estos resultados indican que, si bien los errores más recurrentes son similares en ambos niveles, es relevante considerar las diferencias en cuanto a la recurrencia obtenida, puesto que estaría arrojando una tendencia importante para el orden en el tratamiento de los errores según el nivel de competencia.

Los hallazgos encontrados en el estadio actual de esta investigación en materia de errores lingüísticos nos permiten avanzar en la delimitación de las relaciones entre tipos de errores, niveles de lengua y FCE. Consecuente con ello, actualmente estamos implementando un modelo de tratamiento de errores diferenciado para aprendices de nivel A2 y de nivel B1 en cuanto a los errores tratados y al tipo de estrategias de FCE más adecuados tanto para el nivel de lengua como para el tipo de error enfocado. La meta final es poder apoyar de manera más efectiva la problemática de los errores y contribuir a la mejora de la precisión escrita de los aprendices de ELE.

Como trabajo futuro también planificamos realizar estudios de tipo cualitativo que aporten con explicaciones de los errores aquí encontrados y sus relaciones con variables individuales como el nivel de lengua y la lengua materna de los sujetos. Sería interesante, además, indagar los errores cometidos por los estudiantes en su interlengua considerando para ello, tanto su L1 como el dominio de otras lenguas (L2, L3).

Finalmente, en lo que respecta al diseño e implementación del CAELE, consideramos que dicha colección de textos constituye un aporte valioso en el ámbito de los corpus de textos escritos en formato electrónico de ELE. Esto posibilita la investigación y el análisis de producciones escritas reales de ELE y sus problemáticas lingüísticas, enriqueciendo así también el ámbito de la adquisición y enseñanza del español como lengua extranjera. El CAELE se está incrementando semestralmente con la recolección de nuevos textos en formato electrónico de aprendientes de ELE.

Referencias

ALEXOPOULOU, A. Los criterios descriptivo y etiológico en la clasificación de los errores del hablante no nativo: una nueva perspectiva. *Porta Linguarum*, Granada, n. 5, p. 17-35, 2006.

CAMPILLOS LLANOS, L. Errores léxicos en el español oral no nativo: análisis de la interlengua basado en corpus. *Revista ELUA*, Alicante, v. 28, 2014a.

CAMPILLOS LLANOS, L. Las preposiciones en el habla no nativa de nivel intermedio: análisis de la interlengua basado en corpus. *Revista de Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, Madrid, v. 16, 2014b.

CORDER, S. P. Error analysis, interlanguage and second language acquisition. *Language Teaching*, Cambridge, v. 8, n. 4, p. 201-218, 1975. https://doi.org/10.1017/S0261444800002822

CRUZ PIÑOL, M. *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: ArcoLibros, 2012. 192 p.

DAGNEAUX, E.; DENNESS, S.; GRANGER, S. Computer-aided error analysis. *System*, Amsterdam, v. 26, n. 2, p. 163-174, 1998. https://doi.org/10.1016/S0346-251X(98)00001-3

ELLIS, R. *Second language acquisition*. Oxford: Oxford University Press, 1997.

FERREIRA CABRERA, A. *Proyecto Fondecyt de investigación nº 1110812*: Un sistema tutorial inteligente para la focalización en la forma en la enseñanza del español como lengua extranjera. 2011.

FERREIRA CABRERA, A. *Proyecto Fondecyt de investigación nº 1140651*: El feedback correctivo escrito directo e indirecto en la adquisición y aprendizaje del español como lengua extranjera. 2014.

FERREIRA CABRERA, A.; ELEJALDE GÓMEZ, J.; VINE JARA, A. Análisis de errores asistido por computador basado en un corpus de aprendientes de español como lengua extranjera. *Revista signos*, Valparaíso, v. 47 n. 86, p. 385-411, 2014.

FERREIRA CABRERA, A.; LAFLEUR, N. Description and analysis of the most common prepositional errors in Spanish as L2. *Lingüística y Literatura*, Medelín, n. 68, p. 57-79, 2015.

FERREIRA CABRERA, A.; VINE JARA, A.; ELEJALDE GÓMEZ, J. Hacia una prueba de nivel en español como lengua extranjera. *RLA – Revista de Lingüística Teórica y Aplicada*, Concepción, v. 51, n. 2, p. 73-103, 2013.

GONZÁLEZ ROYO, C. Skype y la interacción oral nativo/no nativo: funciones y rutinas conversacionales en Corinei, un corpus de interlengua español e italiano. In: HERNÁNDEZ, C.; CARRASCO, A.; ÁLVEREZ, E. (Eds.). *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera.* Valladolid: Universidad de Valladolid, p. 283-293, 2012

GRANGER, S. Computer learner corpus research: current status and future prospects. *In Applied Corpus Linguistics. A Multidimensional Perspective, U. Connor & T.A. Upton (Eds). Amsterdan:Rodopi,* p. 123-145, 2004. https://doi.org/10.1163/9789004333772_008

GRANGER, S. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, Alexandria, v. 37, n. 3, p. 538-546. 2003. https://doi.org/10.1080/23247797.2015.1084685

LONG, M. H.; LARSEN-FREEMAN, D. *An introduction to second language acquisition research*. London: Longman, 1991.

LOZANO, C.; MENDIKOETXEA, A. Learner corpora and second language acquisition. *Automatic treatment and analysis of learner corpus data*, Granada, v. 59, 2013.

MACKEY, A.; GASS, S. M. *Second language research: Methodology and design*. 2. ed. Abingdon: Routledge, 2015.

PINO RODRIGUEZ, A. Palabras en interacción: un corpus de aprendices suecos de E/LE. *A survey of corpus-based research*, 2009. p. 470-487. Disponible en: <https://goo.gl/XAHPR3>. Acceso en: 29 jun. 2017.

PINO RODRIGUEZ, A. Saele, un corpus de aprendices suecos de E/LE. In: INTERNATIONAL CONFERENCE ON CORPUS LINGUISTICS, 4., 2012, Jaén. *Anais*… Murcia: Alienco, 2012.

QUIÑONES, V. A. Análisis de errores ortográficos en aprendices: alemanes de español como lengua extranjera. *Philologia Hispalensis*, Sevilla, v. 23, n. 1-2, p. 17-35, 2009a.

QUIÑONES, V. A. El análisis de errores en el campo de ELE. Algunas cuestiones metodológicas. *Revista Nebrija de Lingüística aplicada a la Enseñanza de las Lenguas*, Madrid, v. 5, n. 1, 2009b.

ROJO, G.; PALACIOS MARTÍNEZ, I. M. Corpus de aprendices de español (CAES). *Journal of Spanish Language Teaching*, Oxford, v. 2, n.2, 194-200, 2015. https://doi.org/10.1080/23247797.2015.1084685

SÁNCHEZ JIMÉNEZ, D. Una aproximación a la didáctica de la ortografía en la clase de ELE. *Marco ELE – Revista de didáctica español como lengua extranjera*, Valencia, v. 9, 1-22, 2009.

TORIJANO, J. A. Los que nos enseñan los errores. *Signum: Estudos da Linguagem*, Londrina, v. 9, n. 1, p. 141-206, 2006. https://doi.org/10.5433/2237-4876.2006v9n1p141

VÁZQUEZ, G. Análisis de errores, el concepto de corrección y el desarrollo de la autonomía. *Revista Nebrija de Lingüística aplicada a la Enseñanza de las Lenguas*, Madrid, v. 5, n. 10, 2009.

Data de submissão: 02/08/2016. Data de aprovação: 08/02/2017.