RESLA
Revista Española de Lingüística Aplicada

https://matrix.aesla.org.es/RESLA
ISSN 0213-2028 | e-ISSN 2254-6774

Anotación de errores gramaticales en el corpus de aprendices de español COWS-L2H

Natividad Hernández Muñoz [1]*, Carmela Tomé Cornejo [1], Helena S. Belío-Apaolaza [1], Álvaro Recio Diego [1] y Sam Davidson [2]

[1]Universidad de Salamanca, España
[2]University of California, Davis, United States of America

Resumen: Los corpus de aprendices de español han despertado un creciente interés en los últimos años dada su diversidad de aplicaciones en el ámbito de la adquisición de la segunda lengua, así como en su didáctica. Dentro de los estudios metodológicos sobre la creación y el uso de corpus, uno de los aspectos menos estudiados es la anotación manual de los errores, un proceso que implica un elevado coste humano y temporal, pero que incrementa exponencialmente las posibilidades de explotación del corpus. El objetivo principal de este trabajo es presentar el diseño, el desarrollo y las pruebas de fiabilidad del etiquetado de errores gramaticales del corpus COWS-L2H y, como parte de sus aplicaciones, mostrar uno de sus usos más relevantes, el análisis de la interlengua de los aprendices a través la observación del etiquetado en una selección de textos de naturaleza pseudolongitudinal en relación con cinco fenómenos morfosintácticos propios del sintagma nominal (flexión de género y de número, uso del artículo y del pronombre personal de sujeto, y colocación del adjetivo). Los resultados no solo aportan datos sobre la progresión de los errores analizados, sino que, además, contribuyen al desarrollo metodológico de los sistemas de anotación y análisis de errores en corpus de aprendices.

Palabras clave: *corpus de aprendices de español, anotación de errores, fiabilidad, adquisición de la segunda lengua, gramática, ELE.*

Annotation of Grammatical Errors in the COWS-L2H Spanish Learner Corpus

Abstract: Spanish learner corpora have attracted increasing interest in recent years due to their diversity of applications in the field of second language acquisition, as well as in foreign language teaching. Within the methodological studies on the creation and use of corpora, one of the least explored aspects is the manual annotation of errors, a process that involves high human and time costs, but exponentially increases the possibilities of corpus exploitation. The main aim of this paper is to present the design, development and reliability tests of the grammatical error tagging in the COWS-L2H corpus and to show one of its most relevant applications, analyzing the learners' interlanguage through the analysis of the errors tags in a selection of pseudolongitudinal texts related to five morphosyntactic phenomena concerning the noun phrase (gender and number inflection, use of the article and the subject personal pronoun, and adjective placement). The results not only provide data on the progression of the errors analyzed, but also contribute to the methodological development of annotation systems and error analysis in learner corpora.

Keywords: *learner corpora, error annotation, reliability, grammar, second language acquisition, SFL.*

How to cite: Hernández Muñoz, N., Tomé Cornejo, C., Belío-Apaolaza, H. S., Recio Diego, A., & Davidson, S. (2026). Annotation of Grammatical Errors in the COWS-L2H Spanish Learner Corpus. *Revista Española de Lingüística Aplicada, 39*(2), 297-321. https://doi.org/10.58859/resla.1553

**Corresponding author:* natih@usal.es

AESLA
Asociación Española de Lingüística Aplicada

Available under the
CC BY 4.0 license. ©

1. Introducción

Los corpus de aprendices son colecciones electrónicas de textos escritos u orales producidos por estudiantes de la lengua de diferentes niveles y nacionalidades, y compilados de acuerdo con unos criterios de diseño explícitos (Granger, 2009; Granger *et al.*, 2015; Sánchez Rufat, 2015; Ferreira *et al.*, 2022). Suelen contar con una interfaz de recuperación del material más o menos completa, donde se pueden seleccionar los textos en función de las características socioculturales de los aprendientes, de forma que sirven como recurso para investigadores, profesores, diseñadores de materiales didácticos o estudiantes. En su elaboración y aprovechamiento confluyen diferentes ámbitos científicos (como la lingüística de corpus, la lingüística teórica, el procesamiento del lenguaje natural, la adquisición de segundas lenguas o la enseñanza de lenguas extranjeras), por lo que sus aplicaciones son igualmente interdisciplinares.

Dentro de las últimas disciplinas citadas, las posibilidades de uso de los corpus de aprendices han sido descritas extensamente en diferentes trabajos (Castello *et al.*, 2015; Palacios *et al.*, 2019; Le Bruyn y Paquot, 2020; Sánchez-Gutiérrez *et al.*, 2022). En ellos se presenta un amplio abanico de funcionalidades, desde enfoques centrados en el análisis y la descripción de la interlengua de los aprendices, el desarrollo de los procesos de aprendizaje o el contraste con el desempeño de hablantes nativos, hasta propuestas pedagógicas consistentes en el diseño de materiales didácticos o en la incorporación de los corpus a las hojas de ruta curriculares, a los procesos de evaluación o a la formación del profesorado. Aunque los primeros corpus de aprendices se desarrollaron en el ámbito del inglés como segunda lengua (Granger, 1998; Granger *et al.*, 2002), en las últimas décadas han experimentado un gran crecimiento en el ámbito del español como lengua extranjera (Alonso Ramos, 2016; Parodi *et al.*, 2022; Sánchez-Gutiérrez *et al.*, 2022).

Este trabajo se centra en el corpus de aprendices COWS-L2H (*Corpus of Written Spanish of L2 and Heritage Speakers*), desarrollado en la Universidad de California Davis, EE. UU. (UC Davis). Se trata de una recopilación de textos escritos por hablantes de español como segunda lengua y hablantes de herencia que combina datos longitudinales (evolución de un mismo estudiante a lo largo de su aprendizaje de español en los sucesivos cursos recibidos en la universidad) y pseudolongitudinales o transversales (datos de grupos de estudiantes de diferentes niveles), y que incorpora anotaciones de errores gramaticales a partir de una taxonomía previamente establecida (Yamada *et al.*, 2020, Sánchez-Gutiérrez *et al.*, 2024).

La lingüística de corpus recurre habitualmente a procedimientos informáticos para recopilar, estructurar y anotar los textos de los corpus, lo que permite poner a disposición de los investigadores grandes conjuntos de datos con los que se puede trabajar de manera automática. No obstante, en el tratamiento de los corpus de aprendices hay determinadas tareas que todavía no se han conseguido automatizar completamente, entre ellas, la detección y clasificación de los errores (Díaz Negrillo y Fernández Domínguez, 2006; Calero *et al.*, 2020). Así, si bien el etiquetado de las categorías gramaticales o funciones sintácticas básicas sí se realiza por medios automáticos (Van Rooy, 2015), la anotación manual de los errores sigue siendo el procedimiento más habitual, lo que implica la existencia de un equipo

humano que toma decisiones sobre cuáles son los errores, las taxonomías de desviaciones identificadas y los objetivos científicos del etiquetado (Lüdeling y Hirschmann, 2015, Díez-Bedmar, 2015 y 2021). Se trata de un proceso complejo, que requiere un amplio equipo de profesionales bilingües y que supone un gasto económico y temporal muy elevado, por lo que la mayor parte de los corpus de aprendices de español no incorporan anotaciones de errores y, si lo hacen, suelen ser conjuntos de textos menores y focalizados en fenómenos concretos (Rojo *et al.*, 2022).

En la primera parte del artículo se contextualiza el corpus que nos ocupa junto a los corpus más relevantes en el ámbito hispánico y se incluye un estado de la cuestión sobre los estudios dedicados a la corrección y anotación de los materiales (apartados 2.1 y 2.2). A continuación, presentamos el proceso de anotación de errores gramaticales llevado a cabo por el equipo de la Universidad de Salamanca (USAL) durante los años 2020-2023 para el corpus COWS-L2H en relación con las dos primeras fases descritas en Díez Bedmar (2015, 2021): preliminares, localización y corrección de errores, anotación y pilotaje para comprobar la fiabilidad del etiquetado (apartado 3). Por último, proporcionamos un primer análisis centrado en errores morfosintácticos propios del sintagma nominal como una aproximación a las posibilidades que ofrece COWS-L2H en el estudio de la interlengua, asumiendo que los errores son reflejo de esta gramática interna del aprendiz (Selinker, 1972).

2. Marco teórico

2.1. Los corpus de aprendices en español: el caso del COWS-L2H

En el contexto de la lingüística de corpus en un sentido amplio y, en concreto, dentro de la elaboración de herramientas que permitan el acceso a extensos conjuntos de textos etiquetados con información lingüística, los corpus de aprendices tienen un desarrollo relativamente reciente (Sánchez Rufat, 2015). Surgidos a finales de los años 80 y principios de los 90, los corpus de aprendices de inglés como segunda lengua constituyeron el primer eslabón dentro de la lingüística de corpus de lenguas distintas a las nativas (Granger, 1998; Granger *et al.*, 2002), pero pronto comenzaron a crearse en otros idiomas (véase la amplitud de los corpus de aprendices indexados en la herramienta Learner corpora around the world de la Universidad Católica de Lovaina). Ha sido solo en las dos últimas décadas cuando los corpus de aprendices de español han comenzado a emerger en diferentes contextos, si bien la desproporción con los de inglés sigue siendo notoria: de los 137 corpus de aprendices recogidos en Learner corpora around the world, los de español representan solo el 8 % frente al 60 % de corpus en inglés (Sánchez-Gutiérrez *et al.*, 2022).

Los corpus de aprendices de español disponibles se caracterizan por una gran heterogeneidad. Para su descripción, tomaremos como punto de partida la tipología de los textos recogidos (orales y escritos), las lenguas nativas de los estudiantes y los procesos de etiquetado, en el caso de que los posean. Para identificar los corpus, recomendamos

consultar el Indexador de Corpus de Aprendices de Español de la Universidad Complutense de Madrid (ICAE) (Díaz Sánchez *et al.*, s.f.) y, para una revisión más exhaustiva, véase Rojo *et al.* (2022).

Sin duda, uno de los hitos más relevantes en el avance de los corpus de aprendices de español es la propuesta y elaboración del CAES (Corpus de aprendices de español como lengua extranjera), un proyecto promovido por el Instituto Cervantes y desarrollado por la Universidad de Santiago de Compostela. En su versión 2.1 de marzo de 2022, se compone de 6561 textos escritos por 2544 aprendices de 11 lenguas maternas y de distintos niveles de competencia (desde A1 hasta C1). Las tareas producidas por los aprendices se anotan morfosintácticamente de forma automática y a continuación se desambiguan manualmente.

Otra de las compilaciones más amplias es el Corpus Escrito del Español como L2 (CEDEL2), recogido en la Universidad de Granada (Lozano, 2022) y compuesto por más de un millón de palabras en su versión 2.0 de septiembre de 2020. CEDEL2 incorpora las producciones escritas y orales de aprendices de once lenguas maternas y distintos niveles de competencia, así como datos de hablantes nativos de diferentes orígenes a fin de posibilitar análisis contrastivos. Como el CAES, CEDEL2 posee etiquetas relativas a las clases de palabras en el subcorpus de aprendices de español, el subcorpus de nativos de español y el de nativos de inglés. En este caso, las categorías gramaticales se anotan automáticamente mediante el etiquetador Freeling sin un proceso de desambiguación posterior. Incluye también textos escritos y orales el corpus LANGSNAP, integrado por dos subcorpus paralelos de 56 anglófonos aprendices de español y de francés en contexto de inmersión, con un diseño longitudinal muy extenso. En la misma línea, el Corpus de Aprendices Taiwaneses de Español (CATE) en su versión 2.0 integra dos subcorpus: el Corpus Escrito de Aprendices de Español (CEATE, 2005-2011) y el Corpus Oral de Aprendices de Español (COATE, 2013-2017).

Dentro de los corpus escritos, en la universidad de Alcalá se compiló el Corpus para el Análisis de Errores de Aprendices de E/LE, el CORANE (Cestero y Penadés, 2009), de naturaleza longitudinal y con un total de 1091 composiciones elaboradas por 321 aprendices de más de veinte lenguas maternas diferentes y de niveles de competencia elemental, intermedio, avanzado y superior. Integrado también por composiciones escritas de estudiantes de distintas lenguas maternas, se encuentra el Corpus de Aprendientes de Español como Lengua Extranjera (CAELE). Este corpus de corte longitudinal se compone de 1217 textos producidos por 201 estudiantes de intercambio de nivel A2 y B1, en el contexto universitario chileno (Ferreira *et al.*, 2022). CAELE incluye anotación de errores, en un primer momento, basada en una taxonomía que contemplaba errores léxicos, gramaticales y ortográficos de distintos tipos (Ferreira y Elejalde, 2017) y que posteriormente se actualizó con un criterio etiológico (Ferreira y Elejalde, 2020, Ferreira *et al.*, 2022). Con una denominación similar, pero sin relación directa, el Corpus de Aprendientes de Español como Lengua Extranjera y Segunda Lengua (CÁELE/2) es un proyecto del Instituto Caro y Cuervo que se propone “diseñar, recopilar, sistematizar y construir un corpus escrito y

oral de muestras de aprendientes de español, extranjeros, sordos e indígenas en Colombia" (Hincapié, 2018). Para terminar este grupo, a partir de los textos producidos para la obtención de los Diplomas de Español como Lengua Extranjera (DELE) de nivel B2, se desarrolló en la Universidad de Salamanca el corpus ESPALEX (Bustos Gisbert y Sánchez Iglesias, 2011), con más de dos millones de palabras y la participación de 6288 informantes de 85 países diferentes.

En cuanto a los corpus escritos que solo poseen registros de aprendices de una lengua materna destacan el Corpus de ELE en Japón, CELEN (Valverde, 2020), compuesto en su versión 1.2 de abril de 2023 por 6196 textos escritos por 1035 aprendices japoneses; el CINEAS, Corpus de Interlengua Española de Aprendices Sinohablantes, coordinado desde la Universidad de Lleida, con un total de 4339 tareas producidas por estudiantes de español de origen chino con distintos niveles de competencia (desde A1 hasta C1) en contexto de inmersión y de no inmersión (Calero *et al.*, 2020); el Aprescrilov I y II (Aprender a Escribir en Lovaina), con datos transversales (Aprescrilov I) y longitudinales (Aprescrilov II) de aprendientes de español como L3 de nivel universitario (Buyse et al., 2016); el Corpus del español de los italianos (CORESPI y Bailini, 2018), con 474 textos y más de 120 000 palabras; el SAELE, el Corpus de Aprendientes Suecos de ELE (Pino Rodríguez, 2012), integrado por textos argumentativos de aprendientes de nivel A2-B1, o The Tartu Learner Corpus of Spanish as a L3+, con 885000 palabras de muestras de escritura académica de estudiantes estonios.

Por otro lado, entre los corpus que recogen textos orales de estudiantes de diferentes nacionalidades destacan el Corpus Oral de Español como Lengua Extranjera (CORELE), que compila las producciones de 40 aprendices de niveles A2 y B1 de los cursos de español de la Universidad Autónoma de Madrid y de la Universidad Complutense de Madrid; y el Corpus Fono.ele, centrado en el desarrollo del componente fónico y compuesto por muestras de habla de aprendientes de español de niveles A2, B1, B2 y C1 de seis nacionalidades distintas (Blanco Canales, 2011).

Finalmente, de los corpus orales centrados en aprendientes de español de la misma lengua materna, cabe señalar el Spanish Learner Language Oral Corpus (SPLLOC 1 y 2), con muestras de 120 participantes británicos de distintos niveles, además de un subcorpus adicional de hablantes nativos; el SPT (Spanish corpus Proficiency level Training), que proporciona 327 grabaciones de 38 informantes angloparlantes con diferentes niveles de español, entre los que se incluyen 17 hablantes de herencia (Koike y Witte, 2016); el Spanish in Texas Corpus, que ofrece entrevistas y conversaciones de 97 hablantes bilingües inglés-español que viven en Texas tanto en archivos de vídeo como de audio, además de textos con las transcripciones completas y anotaciones referidas a las clases de palabras (Bullock y Toribio, 2013); el Corpus Oral de Interlengua Español-Italiano (CORINÉI), en el que estudiantes de italiano llevan a cabo interacciones orales entre pares con nativos (Chiapello y González Royo, 2020), o el corpus COPINE, Corpus Oral de Peticiones en Interacciones

Naturalizadas en Español (Marsily, 2022), con 120 participantes, 60 estudiantes de ELE de origen belga con el francés como lengua materna y 60 hablantes nativos de español de España.

En este contexto descrito, se integra el Corpus of Written Spanish of L2 and Heritage Speakers (COWS-L2H), que reúne una extensa colección de composiciones escritas por estudiantes de español de la UC Davis. Con 5383 textos, este corpus proporciona datos transversales y longitudinales de aprendices de español como lengua extranjera, así como de hablantes de herencia, lo que supone un elemento diferenciador con respecto a la mayoría de los corpus disponibles.

La recogida de datos se inició en 2017 y se llevó a cabo en todos los cursos de español que se ofrecen en UC Davis, tanto del programa de español para extranjeros como del específico para hablantes de herencia. El hecho de que los datos se recopilen en una sola institución permite trabajar con una muestra homogénea en cuanto al contexto de aprendizaje. Dado que todos siguen el mismo programa de español, comparten objetivos, contenidos, metodología y materiales didácticos, lo que supone una ventaja a la hora de analizar el desarrollo de la interlengua y posibilita análisis relativos al impacto de tales factores institucionales en el desarrollo de la competencia escrita (Yamada *et al.* 2020). Cada trimestre se pide a los estudiantes que elaboren dos redacciones y se les ofrece un crédito extra por su participación. Por tanto, un alumno que completa todo el programa (obligatorio para los que cursan el grado en español) puede aportar al corpus un total de 14 textos, o bien de 6 en el caso de los hablantes de herencia, de forma que, además de datos transversales, se obtienen datos longitudinales, lo que supone un nuevo elemento diferenciador de COWS-L2H (Sánchez-Gutiérrez y Fernández Mira, 2022).

Hasta la fecha, las redacciones solicitadas se corresponden con los siguientes temas: “Una persona famosa”, “Unas vacaciones perfectas”, “Una persona especial en tu vida”, “Una historia terrible”, “Tú mismo”, “Una experiencia bonita”, “Chaplin” y “Un lugar que te disguste” (Yamada *et al.*, 2020). Todos estos temas cuentan con textos anotados de acuerdo con una taxonomía de errores gramaticales previamente definida (véase 3) y se han proporcionado versiones corregidas que, junto con las originales, se han empleado para entrenar sistemas de procesamiento del lenguaje natural para la corrección de errores gramaticales (Davidson *et al.*, 2020).

En cuanto a su disponibilidad, COWS-L2H se encuentra alojado en un repositorio de GitHub bajo licencia Creative Commons, a plena disposición de la comunidad investigadora. En este sentido, entre los trabajos que ya han hecho uso de los datos del corpus, se distinguen al menos tres líneas: investigaciones en procesamiento del lenguaje natural (PLN), en adquisición de la L2 y en didáctica de la lengua (véase Lozano y Fernández Mira (2022) para una propuesta de los tipos de investigaciones que posibilita el corpus en relación con las decisiones que subyacen a su diseño).

En la primera línea de investigación encontramos el trabajo ya citado de Davidson *et al.* (2020), así como el de Miaschi *et al.* (2020), en el que se emplean datos longitudinales de COWS-L2H para extraer los rasgos lingüísticos que caracterizan la escritura de los estudiantes en los distintos niveles de competencia y desarrollar un sistema automático de clasificación, además de explorar la relación entre los rasgos más predictivos y el plan de estudios seguido por los alumnos.

Dentro de los estudios de adquisición se situaría el de Fernández Mira *et al.* (2021), que analiza la influencia del tema de la redacción en las puntuaciones de diversidad léxica del texto en relación con dos variables, la valencia positiva o negativa del tema, y su cercanía o cotidianeidad. En todos los niveles, los textos con mayor diversidad léxica se corresponden con los que abordan una experiencia positiva y con los que describen a un personaje famoso. Posteriormente, Sánchez-Gutiérrez y Fernández Mira (2022) combinan datos transversales y longitudinales de COWS-L2H para analizar la progresión de los estudiantes en relación con la diversidad léxica que muestran sus composiciones escritas. La adopción de este enfoque mixto les permite obtener una visión más completa del objeto de estudio: mientras que el análisis transversal proporciona tendencias evolutivas generales, el longitudinal permite agrupar a los estudiantes en función del tipo de trayectoria que muestran. Se emplean también datos longitudinales y transversales en Minnillo *et al.* (2022), en este caso, para evaluar las predicciones de la Hipótesis del aspecto léxico (LAH), la Hipótesis del tiempo pasado por defecto (DPTH) y la Hipótesis de la distribución sesgada (DBH) en relación con la adquisición del pretérito perfecto simple y del imperfecto en español como L2. Y en este mismo grupo se inscribe el trabajo de Hidalgo (2023) sobre los errores relativos al género gramatical cometidos por aprendientes de ELE y hablantes de herencia, y sobre su relación con el sexo de los participantes, así como el de Carando *et al.* (2023), donde se compara la evolución de la complejidad estructural de los textos de los aprendientes de ELE y los hablantes de herencia.

Por último, si bien Lozano y Fernández Mira (2022) ofrecen sugerencias sobre posibles usos pedagógicos tanto de COWS-L2H como CEDEL2, el trabajo de Sampedro (2022) es el primero que desarrolla una propuesta didáctica basada en el uso de corpus de aprendices de ELE entre los que se encuentra COWS-L2H. En concreto, Sampedro (2022) diseña una unidad para aprendientes de nivel B1 en torno a la función comunicativa de dar consejos en la que combina el enfoque inductivo con el de PPP (presentación, práctica y producción).

2.2 El etiquetado de errores en los corpus de aprendices

Dentro de las posibilidades de análisis de los materiales del corpus, ha habido dos enfoques predominantes, el *Análisis contrastivo de la interlengua —Contrastive Interlanguage Analysis* (Granger, 1996)— y el *Análisis de errores asistido por ordenador —Computer-aided error analysis* (término introducido por Dagneaux *et al.*, 1998)—. El primero compara las actuaciones lingüísticas recogidas en los corpus con otros corpus de nativos o entre grupos de aprendices de la lengua, mientras que el segundo tiene como objetivo principal etiquetar y analizar los errores cometidos por los aprendices. Aunque la presencia de corpus de

aprendices con etiquetado de errores está creciendo (Thewissen, 2015), la complejidad que entraña el proceso de anotación y el coste humano y temporal que conlleva hacen que este segundo enfoque siga siendo el menos frecuente. En efecto, los corpus con etiquetado de errores son minoritarios, según se ha señalado para diversas lenguas (español, Campillos Llanos, 2014, 2016; Alonso Ramos, 2016, y Rojo *et al.*, 2022; inglés, Díez-Bedmar, 2015, 2021 y Fuchs y Werner, 2018; alemán Lüdeling *et al.*, 2005, etc.). Además, estas anotaciones no suelen estar integradas en el corpus general, sino que forman parte normalmente de tratamientos de datos independientes (Van Rooy, 2015).

Como describe Díez Bedmar (2021), la mayor parte de los procesos de etiquetado para el análisis de la interlengua se llevan a cabo para errores concretos y se realizan sobre un conjunto de textos limitado. Por ejemplo, en el nivel gramatical, uno de los fenómenos del ámbito nominal más estudiados es la presencia o ausencia del sujeto pronominal (*null/overt pronouns*) en diferentes contextos sintácticos y pragmáticos. Un trabajo reciente de Domínguez y Arche (2022) utiliza el corpus oral SPLLOC, referenciado en 2.1, para analizar la evolución de este fenómeno en 60 estudiantes británicos, repartidos en tres niveles de español, a través de dos tareas (120 textos) y concluye, como la mayoría de los estudios previos, que, aunque los hablantes anglófonos tienen dificultades para interpretar los sujetos en algunos contextos, especialmente en los niveles iniciales, se produce una clara evolución con el aumento del grado de competencia. La ventaja de los corpus extensos de acceso abierto es que permiten al investigador ampliar fácilmente el espectro de estudio, los perfiles de los estudiantes o las tipologías textuales si la investigación lo requiere (como las mismas autoras hacen a partir de un estudio de Domínguez, 2013), además de posibilitar la replicabilidad del trabajo y la consulta de casos concretos.

Tal y como indicamos en la introducción, este estudio persigue mostrar las posibilidades de análisis que ofrece el etiquetado de errores del corpus COWS-L2H. Dicho enfoque está claramente influido por la trayectoria del análisis de errores tradicional llevado a cabo en la lingüística aplicada desde los años 70 (Corder, 1967), pero se centra en textos producidos por los aprendices en las condiciones más naturalísticas posibles.

Quizá la mayor crítica que ha recibido este enfoque en la aplicación a la lingüística de corpus es la necesidad de establecer un estándar prescriptivo como punto de partida para realizar el etiquetado y determinar el error (Lüdeling y Hirshmann, 2015), lo que plantea serias dificultades en lenguas pluricéntricas como es el español incluso a la hora de delimitar el corpus (Rojo *et al.*, 2022). Otro de los problemas señalados tiene que ver con la conceptualización del error. Dado que el error se entiende como una desviación de la norma, algunos autores prefieren evitar ese término y emplear expresiones como *formas no codificadas* o *formas no convencionales*. En cualquier caso, siempre es necesario partir de una hipótesis de uso que se considera como la práctica deseable por parte de los alumnos —la denominada *target hypothesis* en Lüdeling y Hirchmann (2015, pp. 142-146)—. En nuestro etiquetado hemos considerado los usos recogidos en las obras normativas de las Academias (RAE y ASALE 2005, 2009), las cuales incorporan, hasta cierto punto, las diferentes variedades estándar de la lengua española.

Antes de continuar, debemos hacer hincapié en los diferentes tipos de anotaciones incorporadas a los corpus de aprendices: los etiquetados morfosintácticos, que distinguen categorías gramaticales, fundamentalmente, y permiten búsquedas de elementos lingüísticos concretos, y los etiquetados de identificación y corrección de errores, que marcan las características de la interlengua de los estudiantes comparándolas con las soluciones nativas. Los primeros suelen llevarse a cabo con una herramienta automática de marcación de las formas lingüísticas junto con una revisión y ajuste por parte del investigador (Van Rooy, 2015; Campillos Llanos, 2016). Este aparato de búsqueda puede servir para localizar clases de palabras en su contexto (por ejemplo, sustantivos o sustantivos masculinos), ciertos patrones combinatorios (por ejemplo, «sustantivo + adjetivo») o incluso funciones sintácticas simples. Como ejemplo para el español, dentro del enfoque del análisis contrastivo de la interlengua (Rojo *et al.*, 2022), puede verse el caso del recuento de uso del pretérito imperfecto de subjuntivo por diferentes lenguas de origen en Rojo y Palacios (2022, pp. 82-83).

Por otro lado, las anotaciones que marcan errores están concebidas para facilitar al investigador la identificación de las desviaciones que, en fenómenos concretos, muestra la interlengua de los aprendices. En estos casos, se asigna un código de error a la secuencia específica que se juzga como una desviación y, en algunos casos como el nuestro, se proporciona también la solución correcta (Lüdeling y Hirchmann, 2015; Díez-Bedmar, 2021). Esto requiere, en primer lugar, una selección de los errores que se van a etiquetar (no pueden marcarse todos los que identifican la interlengua del aprendiente) y un diseño, testeo y aplicación de las etiquetas seleccionadas por parte del investigador. Se trata, por tanto, de anotaciones orientadas a problemas concretos (Leech, 1997, p. 15).

A partir de corpus con etiquetado automático se han llevado a cabo análisis de errores extrayendo una serie de ejemplos y anotándolos posteriormente de acuerdo con los fenómenos que se desea investigar (véase, por ejemplo, la propuesta de Palacios *et al.* (2019), donde se analizan manualmente ejemplos extraídos mediante la interfaz de búsqueda que incorpora el CAES para examinar la confusión entre *ser* y *estar,* y las preposiciones *por* y *para*). Este procedimiento semiautomático de análisis de errores se enmarca en el denominado *Análisis de Errores Asistido por Computador* (Dagneaux *et al.*, 1998).

Ejemplos de este mecanismo de etiquetado son dos estudios recientes sobre la anáfora. Lozano y Quesada (2023) realizan anotaciones manuales en una selección de 75 textos del CEDEL con etiquetas implementadas en la *UAM Corpus Tool* versión 6.2 para analizar la resolución de las anáforas con antecedente ambiguo, un fenómeno ampliamente investigado desde modelos experimentales, pero no desde la lingüística de corpus. Con este objetivo, diseñan un sistema detallado de etiquetado específico, donde se incluyen el número de potenciales antecedentes de cada anáfora y la posición de cada uno de ellos. Para el mismo fenómeno, Bruscato y Baptista (2022) recogen dos microcorpus específicos de estudiantes de origen brasileño, uno de ellos en español (15 textos) y el otro en inglés (30 textos). Para la incorporación manual de las anotaciones utilizan el software libre *Recogito annotation tool*, que permite establecer relaciones orientadas entre anáforas y

antecedentes, y diseñan un sistema de etiquetado con el que se indica la forma, la función, el segmento oracional, la ambigüedad y la relación anafórica o catafórica (Bruscato y Baptista, 2022, p. 9). En cualquier caso, este tipo de análisis de errores que utiliza el apoyo de software específico (véase Calero *et al.* (2020) para una revisión de estas herramientas aplicadas a textos de aprendices de español), no permite en ningún caso la automatización completa del trabajo. Según se ha señalado previamente, el diseño de las etiquetas y la decisión de qué fragmentos van a ser anotados recaen necesariamente sobre un equipo humano.

En este sentido, Díez-Bedmar (2015, pp. 41-50) presenta un detallado análisis de las fases que debe seguir la creación y la puesta en marcha del proceso de etiquetado de errores. Este será el modelo que utilicemos en el apartado siguiente para describir nuestra propuesta. En la primera fase, considerada preliminar, se decide la taxonomía de errores que se van a etiquetar y el número y el grado de especificidad de las etiquetas. La segunda fase integra diferentes pasos: detectar los errores, realizar un pilotaje para clasificarlos y describirlos a partir de un perfilado del sistema de etiquetado (*fine-tuning the error tagging system*) y poner en marcha los sistemas de fiabilidad. Como resultado de esta segunda fase se confecciona un manual de etiquetado con un conjunto de etiquetas fácilmente comprensibles para los anotadores (Granger, 2003). La tercera fase comprende los procesos posteriores al etiquetado de postedición y preparación del corpus para su consulta.

Según Granger *et al.* (2007), el sistema de etiquetado debe ser jerárquico en su estructura y, al mismo tiempo, flexible, de forma que admita la incorporación de nuevas etiquetas durante y después del proceso de anotación, esto es, que permita su sofisticación si el investigador lo considera oportuno. Asimismo, el diseño de esta taxonomía debe estar guiada por un objetivo de investigación y su aplicación, mediada por la subjetividad del anotador, debe ser rigurosa y lo más sistemática posible. Para ello, se recomienda que los anotadores sean nativos de la variedad en la que se están realizando las correcciones (españoles en nuestro caso), pero que dominen también la lengua materna de los estudiantes (inglés para nosotros) (Granger, 2003; Díaz-Negrillo y Fernández-Domínguez, 2006). Además, para asegurar la consistencia, la validez y la eficacia del procedimiento, han de incorporarse medidas de fiabilidad tanto del sistema de etiquetado como de las correcciones realizadas por los anotadores (Dahlmeier *et al.*, 2013; Lüdeling y Hirchmann, 2015). Se aconseja que estas últimas sean realizadas por dos anotadores independientes (Granger, 2009; Lüdeling y Hirchmann, 2015) y que haya un tercero para desambiguar los casos dudosos. Se trata, en definitiva, de un proceso largo y laborioso, que requiere un equipo amplio de investigadores que cumplan con los requisitos expuestos y trabajen en coordinación (Díez Biedmar, 2021).

3. Descripción del proceso de etiquetado del COWSL2-H

La anotación de errores de COWS-L2H es el resultado de una colaboración entre la Universidad de California Davis, responsable del corpus, y la Universidad de Salamanca, en el marco de un proyecto de investigación de tres años de duración (desde 2020 a 2023). Con anterioridad a dicha colaboración, el equipo de UC Davis había iniciado ya el proceso de etiquetado con la definición del esquema de anotación, integrado en el texto (*in-text annotation*), y el diseño de las primeras etiquetas —en concreto, las relativas a la flexión de género y número, y a la *a* de complemento directo (Davidson *et al.*, 2020), si bien de esta última se prescindió en la taxonomía posterior—. La labor del equipo salmantino, integrado por cuatro coordinadores y doce investigadores, ha consistido en la adaptación y ampliación de la taxonomía de errores (fase 1), y en la localización, corrección y anotación de los errores definidos, con el correspondiente proceso de pilotaje para garantizar su fiabilidad (fase 2). Junto con este trabajo de anotación manual, los investigadores de la USAL llevaron a cabo un proceso paralelo de corrección por pares de los textos con el fin de que tales correcciones holísticas sirvieran para desarrollar una herramienta de corrección automática (Davidson *et al.*, 2020).

3.1 Fase 1

Frente a otros etiquetados más generales (Díaz Negrillo y Fernández Domínguez, 2006), el de COWS-L2H se centra en el nivel gramatical y comprende algunos de los fenómenos morfosintácticos, tanto del ámbito nominal como verbal, que mayor atención reciben en la gramática aplicada y en su didáctica (Gutiérrez Araus, 2022), y que son causa de interferencia con el inglés como L1 (Bueso y Casamián, 2001; Rothman *et al.*, 2018). En concreto, las anotaciones de COWS-L2H atienden a las siguientes categorías: flexión de género y número, pronombres personales (específicamente de sujeto), artículos, preposiciones (hasta la fecha, *por* y *para*), verbos copulativos (*ser* y *estar*), adjetivos y morfología verbal.

En cada una de estas categorías los tipos de errores que se contemplan se refieren a problemas de concordancia de género y/o de número, atribuciones indebidas del género o del número, presencia o ausencia indebida del pronombre personal de sujeto o del artículo, intercambio de las preposiciones *por* y *para* o de los verbos *ser* y *estar*, errores de colocación del adjetivo en el sintagma nominal y selección de la forma verbal incorrecta (en cuanto a la flexión de tiempo, modo, persona o número, o por la elección de la raíz —irregular que se regulariza o viceversa—). De esta forma, a las etiquetas iniciales para los errores relacionados con la flexión nominal se suma toda una serie de etiquetas relativas a la flexión verbal, lo que permite analizar la adquisición de la morfología flexiva en su conjunto y posibilita la investigación de diversos fenómenos morfosintácticos, como los valores y las relaciones temporales, la selección modal, los procesos de concordancia sujeto-verbo o el uso de las formas no personales del verbo, entre otros. El resto de etiquetas presenta un alcance más específico: atienden a comportamientos de determinadas unidades en distintas clases de palabras que caracterizan la evolución de la interlengua

de los estudiantes de español L2 (Rothman *et al.*, 2018). Se trata de temas clásicos de la gramática ELE, recientemente abordados también desde la lingüística de corpus, como el uso de las preposiciones *por* y *para* y de los verbos copulativos, los sujetos nulos, los sintagmas nominales escuetos o la posición del adjetivo (véase, por ejemplo, Palacios *et al.*, 2019; Guijarro-Fuentes *et al.*, 2023; Domínguez y Arche, 2022; Yonekawa y San Mateo Valdehíta, 2022).

Una vez localizados, estos errores se corrigen y se anotan dentro del texto (*in-text annotation*) de acuerdo con el siguiente esquema:

[error del estudiante]{forma corregida}<etiquetas de error>

En las etiquetas de error se distinguen dos niveles: primero se identifica el tipo de error (concordancia, atribución, presencia o ausencia indebida, intercambio, colocación, etc.) y a continuación se indica la clase de palabras afectada. Véanse los ejemplos siguientes:

(1) (McAdams es muy simpática con [su]{sus}<na:det> amigos.

(Error de concordancia de número —*number agreement*— en un determinante)

(2) Johnny Depp es un actor famoso. [Él]{}<pr:su:pron> nació en 1963 en Kentucky.

(Error de presencia indebida de un sujeto pronominal)

(3) []{El}<ab:det:art:def> Señor Sandler es muy feliz.

(Error de omisión del artículo definido)

(4) Había un perro bebé color marrón que [era]{estaba}<conf:v:seta> triste porque nadie lo quería y no tenía casa.

(Error de confusión de los verbos copulativos *ser* – *estar*)

(5) “The Rock” [lucha]{luchó}<fs:st:t> en la WWE hace mucho tiempo.

(Error de selección de una forma verbal simple en relación con el tiempo)

Se define, por tanto, una taxonomía bidimensional en la que se combinan, en buena medida, las dos sugeridas por Dulay *et al.* (1982) —la referida a las categorías lingüísticas afectadas por el error y la que indica la forma en que el aprendiz ha modificado la estructura superficial, es decir, la estrategia que ha puesto en marcha (omisión, adición, intercambio, colocación, etc.)—, en línea con lo que se defiende, por ejemplo, en Granger (2003) o Díaz Negrillo y Fernández Domínguez (2006). La taxonomía completa puede consultarse en el Anexo I.

En su diseño, se asumen los principios de consistencia, informatividad, flexibilidad y reutilización definidos en Granger (2003, p. 467) para la creación de sistemas de anotación de errores eficaces y se establece una taxonomía precisa y jerárquica (que aporta sistematicidad en la aplicación de las etiquetas, pero que deja margen para su combinación, ampliación o profundidad) y un conjunto de etiquetas manejable, pero

informativo. En este sentido, el hecho de proporcionar una clasificación del error y la categoría gramatical afectada, además de la forma corregida, hace que el etiquetado de COWS-L2H sirva para: 1) realizar una identificación más rápida y eficaz de los fenómenos problemáticos para su análisis o incorporación a propuestas didácticas; 2) realizar un análisis longitudinal o transversal de los fenómenos etiquetados; 3) identificar elementos léxicos o contextos gramaticales donde se dan determinados tipos de errores, y 4) realizar análisis contrastivos.

3.2 Fase 2

Según se ha señalado, esta fase comprende el proceso de localización, corrección y anotación de los errores, así como el análisis de su fiabilidad (Díez Bedmar, 2015, 2021). Tras una formación específica para familiarizarse con el corpus y con el sistema de etiquetado, el equipo de la USAL, integrado por hablantes nativos de español con dominio del inglés como L2 y especializados en la enseñanza de español como lengua extranjera, ha sido el encargado de anotar manualmente los textos seleccionados por el equipo de UC Davis. Para ello, se ha seguido un sistema de anotación por pares con la posibilidad de consultar dudas específicas a los dos coordinadores responsables del diseño de la taxonomía de errores. De esta forma, el manual de etiquetado de COWS-L2H se ha ido completando a lo largo del proceso con una sección de dudas resueltas. En los tres años del proyecto (2020-2023), el equipo salmantino ha anotado 2948 textos —lo que supone el 54,8 % del corpus— y ha aportado un total de 9425 versiones (recuérdese que el proyecto incluía también una parte de corrección de textos para el entrenamiento de un modelo de procesamiento del lenguaje natural).

Dado que la anotación manual de errores es un proceso largo y laborioso, que depende de la actuación de un equipo humano, es necesario comprobar su consistencia como requisito previo para considerar la validez del etiquetado y, en última instancia, la calidad de cualquier análisis basado en las anotaciones del corpus (Dobrić, 2023). Las medidas principales para garantizar la fiabilidad del etiquetado analizan el grado de coincidencia entre pares de anotadores (*inter-annotator agreement*) y también entre los propios anotadores (*intra-annotator agreement*), con el etiquetado de los mismos textos al cabo de un tiempo. Esta coincidencia interanotador e intraanotador se expresa generalmente mediante índices de acuerdo, como el coeficiente kappa de Cohen, que es el que se emplea aquí (Cohen, 1960; Carletta, 1996; Artstein y Poesio, 2008; Dahlmeier *et al.*, 2013; Díez Bedmar, 2015). En este caso, se ha tenido en cuenta como *agreement* tanto las coincidencias en el etiquetado como en los ítems seleccionados para ser anotados, es decir, que un ítem no haya sido etiquetado por ambos correctores también se considera acuerdo.

De los 2948 textos anotados, se seleccionaron aleatoriamente 490 (el 20 %) para los análisis de fiabilidad. Para el acuerdo interanotador, se establecieron 20 pares de correctores y se llevó a cabo una comparación de 30 textos por cada par. En el caso del acuerdo intraanotador, se solicitó el reetiquetado de 30 textos aleatorios de diferentes años. Según se ha señalado, la fórmula estadística seleccionada fue el kappa de Cohen y,

para la preparación y alineación de los textos, se empleó un código *ad hoc* programado con Python. Los resultados muestran un índice kappa superior a 0.6 en la comparación *inter-* y a 0.7 para el caso de *intra-*, lo que supone una buena estimación del grado de acuerdo (Tabla 1).

Tabla 1. Kappa para la estimación del grado de acuerdo.

EKappa	Acuerdo
<0	No acuerdo
0.0-0.2	Insignificante
0.2-0.4	Bajo
0.4-0.6	Moderado
0.6-0.8	Bueno
0.8-1.0	Muy bueno

4. Evolución de errores en el sintagma nominal

4.1 Metodología

Para este estudio se ha tomado un conjunto de 905 textos de COWS-L2H (aproximadamente un 17 % del corpus total y un 30 % de los textos anotados), con una media de 251 palabras por texto y un total de 227 236 *tokens*. De las diversas propuestas temáticas del corpus (véase 2.2), para este estudio se han seleccionado dos a fin de comprobar el impacto del tipo de texto en los errores producidos y en su evolución: "Tú mismo" (*Yourself*) (491 textos) y "Una experiencia bonita" (*Beautiful*) (414 textos).

Se propone un estudio de naturaleza transversal, para el que se consideran textos escritos por estudiantes de español como lengua extranjera (no por hablantes de herencia) de distintos niveles de los cursos de español general. En concreto, se toman 644 textos de nivel Introductorio (correspondientes a tres asignaturas de primer año y equivalentes a un A1-A2) y 261 textos del curso de Composición (dos asignaturas de tercer año equivalentes a un B1+ / B2). Esta clasificación en niveles es un reflejo del universo de estudio: en términos generales, el número de cursos de español introductorio que se ofertan cada trimestre en UC Davis es dos o tres veces mayor que el de los cursos de Composición (Yamada *et al.*, 2020, p. 23). Por transparencia, en el análisis de resultados nos referiremos a estos dos grupos como nivel inicial y avanzado.

De los errores etiquetados (véase 3.1), se seleccionan aquí los que afectan al sintagma nominal, esto es, aquellos que tienen que ver con la flexión de género y de número, con el uso del pronombre personal de sujeto y del artículo, y con la colocación del adjetivo. El objetivo de este análisis es determinar cuáles de dichos aspectos gramaticales del ámbito nominal presentan mayores dificultades para los aprendices norteamericanos y cómo es

su evolución, planteamiento que se aproxima a los estudios exploratorios enmarcados en el análisis de errores que persiguen la identificación de fenómenos lingüísticos problemáticos (Lüdeling y Hirchmann, 2015).

4.2 Resultados y discusión

Los 905 textos analizados contienen un total de 9463 errores relativos a las categorías anotadas en el sintagma nominal. Tal y como se refleja en la Figura 1, el error más frecuente se corresponde con la presencia indebida de un pronombre personal de sujeto (3051 casos, 32 %), seguido por los errores de concordancia de género (2235 casos, 24 %) y la ausencia de artículo (1695 casos, 18 %). Cuando consideramos los distintos errores que afectan a estas categorías (presencia o ausencia, concordancia o atribución, etc.), se comprueba que los pronombres personales de sujeto siguen siendo la categoría más problemática, seguidos de la flexión de género y los artículos (Figura 2).

Esta distribución de los errores se explica en buena medida por interferencias con la L1: frente al inglés, el español es una lengua de sujeto nulo con flexión nominal de género y, aunque en ambas se emplean artículos como marcadores de definitud, sus usos menos prototípicos difieren significativamente. En este sentido, es razonable que la flexión de número ofrezca menos dificultades a los aprendientes, mientras que la colocación del adjetivo, distinta también en ambas lenguas, sería, en principio (puesto que se trata de un fenómeno más restringido), menos problemática.

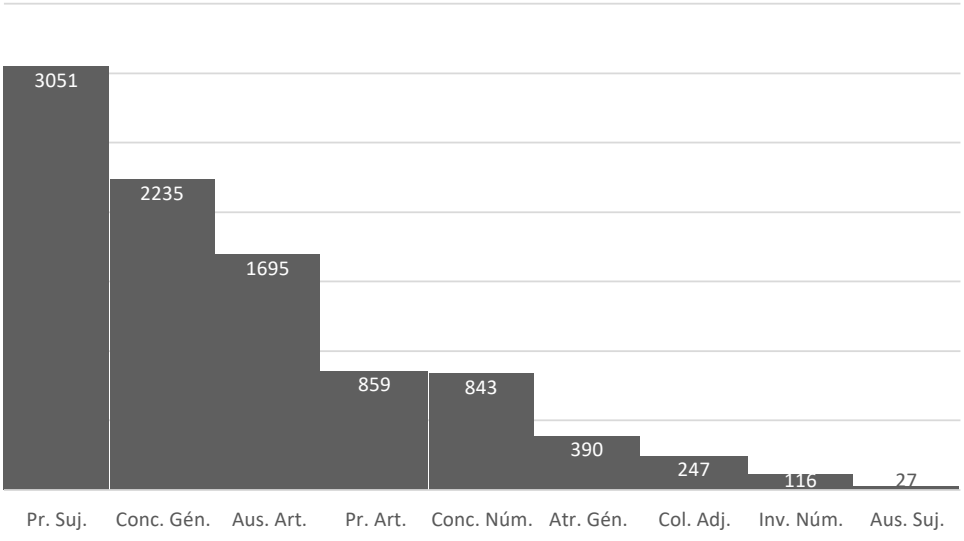


Figura 1. Tipos de errores en el SN por su frecuencia absoluta.

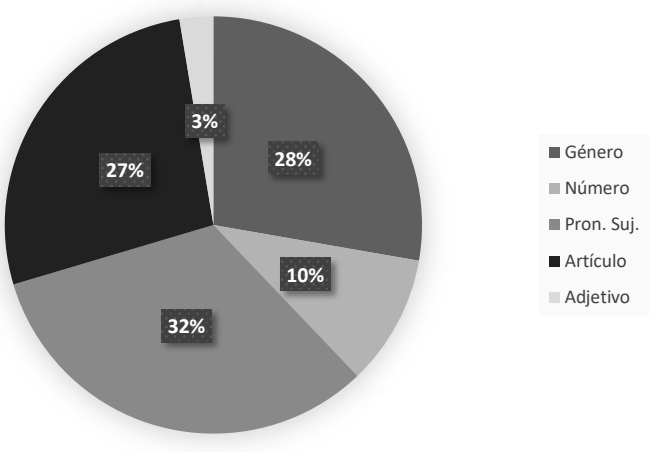


Figura 2. Distribución de los errores en cada categoría.

Por otro lado, si bien la presencia indebida de un sujeto pronominal es el error más frecuente, es también el que experimenta un descenso más pronunciado en la comparación por niveles de competencia (Figura 3), tal y como concluyen las investigaciones previas sobre el fenómeno (Domínguez, 2013; Domínguez y Arche, 2022). La progresión es también favorable en los errores relativos a la concordancia de número y a la ausencia de artículo, mientras que se registra un leve aumento (de 0,02 puntos) en la media de errores correspondientes a la atribución incorrecta de género y la presencia de artículo. No obstante, en términos generales, todas las categorías examinadas muestran un menor número de errores en las composiciones de nivel avanzado que en las de nivel introductorio (Figura 4).

En cualquier caso, según se defiende en Sánchez-Gutiérrez y Fernández Mira (2022), este análisis transversal (y exploratorio) debería completarse con datos longitudinales que permitieran matizar dichas tendencias generales en función de los tipos de trayectorias evolutivas que mostraran los participantes. Frente a la visión tradicional de la adquisición

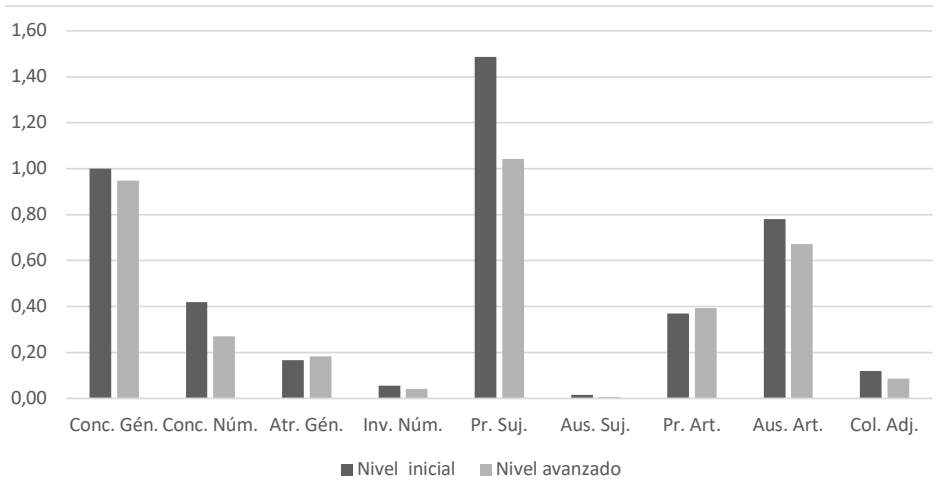


Figura 3. Progresión de cada tipo de error.

de la L2, se asume que el desarrollo de la competencia gramatical no es un proceso lineal sino dinámico, de forma que los contenidos morfosintácticos supuestamente adquiridos no se mantienen estables en la interlengua de los aprendientes, sino que son susceptibles de degradarse, perderse, volver a recuperarse, etc. (Sánchez-Gutiérrez y Fernández Mira, 2022, p. 375).

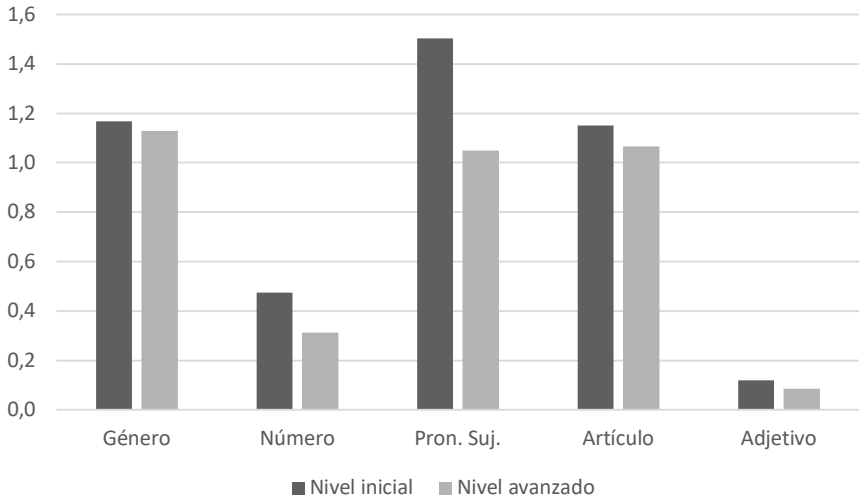


Figura 4. Progresión de los errores en cada categoría.

En cuanto al tipo de texto, los aprendientes cometen más errores en las redacciones en las que hablan de sí mismos que en los textos que describen una experiencia bonita, especialmente en los niveles iniciales (Figura 5). Esta tendencia se aprecia en todos los errores analizados (salvo en la ausencia de sujeto pronominal, de forma muy poco representativa), pero es especialmente acusada en los errores de ausencia de artículo, de presencia de sujeto pronominal, concordancia de género y presencia de artículo (Figura 6). De hecho, al considerar los errores en función de la categoría afectada, el artículo presenta un mayor impacto del cambio textual, seguido por el sujeto y la flexión de género (Figura 7).

Estos datos ponen de manifiesto la necesidad de completar los análisis cuantitativos de los corpus con un enfoque cualitativo para conseguir una correcta interpretación de los resultados. En este caso, se podría comprobar si el hecho de que haya más usos incorrectos del artículo en las redacciones en las que los aprendientes hablan de sí mismos se relaciona con la presencia de sustantivos referidos a partes del cuerpo o prendas de vestir, por ejemplo, en los que el uso del artículo es distinto en ambas lenguas (*Tengo los ojos azules – I have blue eyes*). En la misma línea, cabe suponer que un tema como "Tú mismo" habrá propiciado más ejemplos de pronombre personal de sujeto innecesario y, en consonancia con los resultados de Hidalgo (2023), habría que determinar si las alumnas, al describirse a sí mismas, tienden a usar formas masculinas.

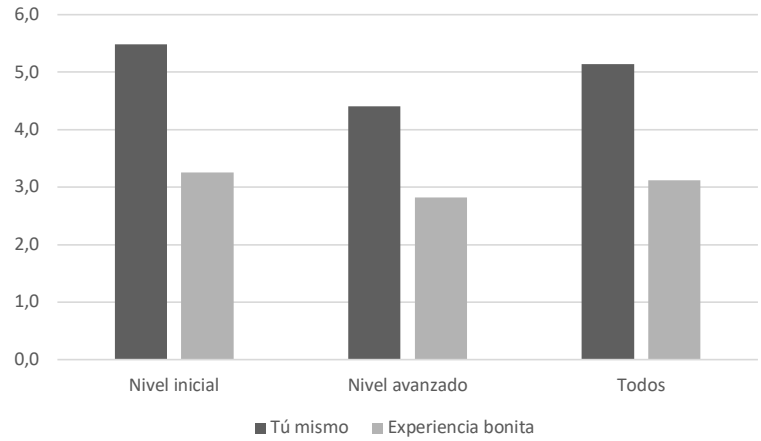


Figura 5. Media de errores por tipo de texto.

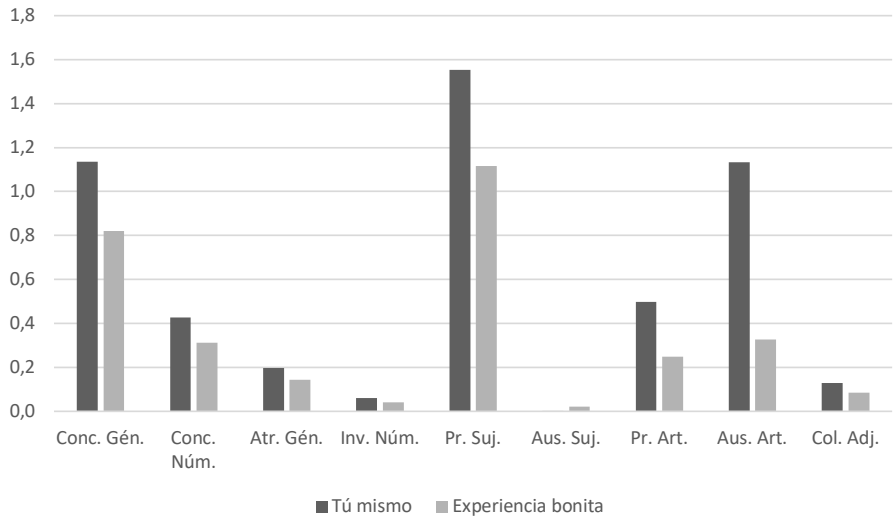


Figura 6. Tipo de error en cada tipo de texto.

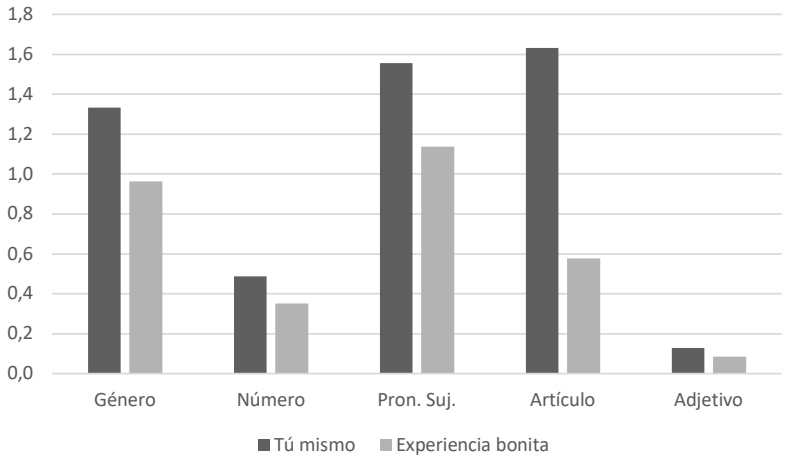


Figura 7. Errores por texto en cada categoría.

4. Conclusiones

En el marco de los estudios de corpus de aprendices de español, este trabajo ha contribuido a la caracterización del Corpus of Written Spanish of L2 and Heritage Speakers (COWS-L2H) desde una doble perspectiva: metodológica e investigadora.

Desde una perspectiva metodológica, se ha presentado el proceso de anotación de errores gramaticales que se ha llevado a cabo, desde el diseño de la taxonomía empleada hasta las medidas de fiabilidad calculadas (inter e intraanotadores) a fin de garantizar la validez de los datos. A pesar de la inversión de tiempo y esfuerzo que requiere este proceso, coincidimos con Granger (2003) o Van Rooy (2015), entre otros muchos, en que este tipo de anotaciones incrementa enormemente el valor del corpus como herramienta de investigación y facilita su explotación en aplicaciones diversas, en nuestro caso, en el ámbito de la adquisición de la gramática del español como lengua adicional o de su didáctica, pero también en el marco de la gramática teórica (aunque los datos de corpus han sido menos frecuentes en la tradición generativa, en realidad, pueden servir como contribución empírica para cualquier escuela).

El sistema de etiquetado debe responder a una reflexión previa sobre la finalidad del corpus. Así, la propuesta implementada en el COWS-L2H parte de un conjunto de etiquetas que comprende algunos de los fenómenos morfosintácticos más representativos de la interlengua de los aprendices angloparlantes, de forma que, a partir de una estructura jerárquica (Granger, 2009), además de aportar información sobre fenómenos amplios, puede usarse como punto de partida para la anotación de errores gramaticales más específicos, permitiendo afinar en distintos niveles de anotación (*fine-tuned*, Díez Bedmar, 2021). Se trata de un etiquetado muy especializado, que no habría sido posible sin la intervención de un equipo humano con conocimiento de ambas lenguas y formación específica en gramática aplicada. En efecto, a pesar de los avances en cuanto a la automatización del etiquetado, estas prácticas siguen presentando muchas limitaciones frente a la anotación manual.

Desde una perspectiva investigadora, se ha mostrado el potencial que ofrece el etiquetado para el análisis del desarrollo de la competencia gramatical de los aprendientes a través de un estudio exploratorio centrado en errores propios del sintagma nominal. Los resultados obtenidos nos han permitido confirmar los fenómenos más problemáticos y las tendencias generales en cuanto a su evolución, así como la influencia que ejerce el tipo de texto solicitado. Asimismo, este análisis preliminar nos ha servido para comprobar la conveniencia de emprender investigaciones que combinen datos transversales y longitudinales, y la necesidad de completar los análisis cuantitativos con los cualitativos para poder interpretar adecuadamente los resultados.

En definitiva, COWS-L2H brinda amplias posibilidades como recurso de investigación en español como lengua adicional. La incorporación de anotaciones precisas sobre aspectos gramaticales altamente representativos de la interlengua de aprendices angloparlantes posibilita la realización de estudios tanto generales como centrados en fenómenos muy

CC BY

específicos, como, por ejemplo, la concordancia de persona en el verbo. Asimismo, se trata de uno de los pocos corpus de aprendices de español que proporciona datos transversales y longitudinales, y que incluye una submuestra de hablantes de herencia, además de hablantes de ELE. Por último, el hecho de que todos los participantes pertenezcan a la misma institución y respondan a las mismas tareas permite, por un lado, un amplio control de las condiciones de aprendizaje y, por otro, un análisis sobre el impacto del modelo textual o del tema de la redacción.

CRediT Author contribution / *Contribución de los autores*
Conceptualization / *Conceptualización:* N.H.M. y C.T.C.
Formal Analysis / *Análisis formal:* N.H.M. y C.T.C.
Methodology / *Metodología:* N.H.M., C.T.C., H.S.B y A.R.D.
Writing / *Redacción:* N.H.M., C.T.C., H.S.B y A.R.D.
Research dataset / *Datos de investigación:* N.H.M., C.T.C., H.S.B y A.R.D. y S.D.

Funding, data availability, and copyright / *Financiación, disponibilidad de datos y derechos de autoría*
Funding / *Financiación:* Contrato de investigación "Corrección y anotación del corpus de aprendices de español COWS-L2H" (2020/00260/001), Universidad de California Davis (EEUU).
Image use consent / *Consentimiento para el uso de imágenes:* Propiedad de los autores.
Conflict of interest / *Conflicto de intereses:* No existe conflicto de intereses.
Data availability statement/ *Declaración de disponibilidad de datos:* https://github.com/ucdaviscl/cowsl2h.
License/ *Licencia:* This article is published under the CC BY 4.0 License / *Este artículo se publica bajo la Licencia CC BY 4.0*.

Editorial history / *Fechas del proceso editorial*
Received / *Recibido:* 30/06/2023
Accepted / *Aceptado:* 30/04/2024
Published / *Publicado:* 15/11/2026

Referencias

Alonso Ramos, M. (2016). Spanish learner corpus research: Achievements and challenges. En M. Alonso Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 3–31). John Benjamins. https://doi.org/10.1075/scl.78

Artstein, R., & Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Bailini, S. (2018). *Corpus del Español de los Italianos*. https://corespiycorite.altervista.org/?doing_wp_cron=1671141855.4830360412597656250000

Blanco Canales, A. (2011). Fono.Ele, una herramienta web para la investigación de la competencia fónica y la formación de profesores. En C. Hernández González, A. Carrasco Santana, & E. Álvarez Ramos (Eds.), *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera* (pp. 129–140). ASELE.

Bueso Fernández, I., & Casamián Sorrosa, P. (2001). *Diferencias de usos gramaticales entre el español y el inglés. Gramática contrastiva*. Edinumen.

Bullock, B. E., & Toribio, A. J. (2013). *The Spanish in Texas corpus project*. Center for open education resources and language learning (COERLL).

Bustos Gisbert, J. M., & Sánchez Iglesias, J. J. (2011). ESPALEX: Un corpus para el estudio de la adquisición del español como lengua extranjera. En C. Hernández González, A. Carrasco Santana y E. Álvarez Ramos (Eds.), *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera* (pp. 149–159). ASELE.

Buyse, C., Fernández Pereda, L., & Verveckken, K. (2016). The Aprescrilov corpus, or broadening the horizon of Spanish language learning in Flanders. En M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 143–168). John Benjamins. https://doi.org/10.1075/scl.78.06buy

Calero Fernández, M. A., Serrano Zapata, M., & Gómez Devís, M. B. (2020). Codificación y etiquetado en los corpus de aprendices y su aplicación didáctica: La propuesta del Corpus de Interlegua Española de Aprendices Sinohablantes (CINEAS). *E-Aesla. Revista digital de lingüística aplicada, 6*, 206–222.

Campillos Llanos, L. (2014). Análisis de errores pragmático-discursivos en un corpus oral de español como lengua extranjera. *Círculo de Lingüística Aplicada a la Comunicación, 58*, 23–59. https://doi.org/10.5209/rev_CLAC.2014.v58.45469

Campillos Llanos, L. (2016). PoS-tagging a Spanish oral learner corpus: Criteria, procedure, and a sample análisis. En M. Alonso Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 89–116). John Benjamins. https://doi.org/10.1075/scl.78.04cam

Carando, A., Minnillo, S., Fernández-Mira, P., Davidson, S., Sagae, K., & Sánchez-Gutiérrez, C. (2023). Writing development in Spanish as a second and heritage language: A corpus study on complexity. *Journal of Spanish Language Teaching, 10*(1), 59–71. https://doi.org/10.1080/23247797.2023.2201989

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Castello, E., Arckeley, K., & Coccetta, F. (2015). *Studies in learner corpus linguistics research and applications for foreign language teaching and assessment*. Peter Lang.

Cestero, A. M., & Penadés, A. I. (2009). *Corpus de textos escritos para el análisis de errores de aprendices de E/LE (CORANE)*. Ediciones Universidad de Alcalá de Henares.

Chiapello, S., & González Royo, C. (2020). Oral narrative in Italian as a foreign language at an intermediate level (B1). En R. Roig-Vila (Ed.), *La docencia en la Enseñanza Superior. Nuevas aportaciones desde la investigación e innovación educativas* (pp. 1152–1161). Octaedro.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching, 5*, 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System, 26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS Corpus of Learner English. En *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22–31). Association for Computational Linguistics. Atlanta, Georgia.

Davidson, S., Yamada, A., Fernández Mira, P., Carando, A., Sánchez-Gutiérrez, C. H., & Sagae, K. (2020). Developing NLP tools with a new corpus of learner Spanish. En *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 7238–7243).

Díaz Negrillo, A., & Fernández Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista española de lingüística aplicada, RESLA, 19*, 83–102.

Díaz Sánchez, P., Fernández-Pampillón, A. M., & Amenós Pons, J. (s. f.). *Indexador de Corpus de Aprendices de Español*. http://repositorios.fdi.ucm.es/corpus_aprendices_espa%c3%b1ol/view/paginas/view_paginas.php?id=1

Díez Bedmar, M. B. (2015). Dealing with errors in learner corpora to describe, teach and assess EFL writing: Focus on article use. En E. Castello, K. Arckeley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics research and applications for foreign language teaching and assessment* (pp. 37–70). Peter Lang.

Díez Bedmar, M. B. (2021). Error analysis. En N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 90–104). Routledge. https://doi.org/10.4324/9781351137904-9

Dobrić, N. (2023). Identifying errors in a learner corpus – the two stages of error location vs. error description and consequences for measuring and reporting inter-annotator agreement. *Applied Corpus Linguistics, 3*. https://doi.org/10.1016/j.acorp.2022.100039

Domínguez, L. (2013). *Understanding interfaces: Second language acquisition and first language attrition of Spanish subject realization and word order variation*. John Benjamins. https://doi.org/10.1075/lald.55

Domínguez, L., & Arche, M. J. (2022). Early use od null and overt subjects in L2 Spanish. Evidence from two oral tasks. En K. McManus & M. S. Schmid, *How special are early birds? Foreign language teaching and learning* (pp. 189–225). Freie Universität Berlin/ Languace Science Press.

Dulay, H., Burt, M., & Krashen, S. D. (1982). *Language two*. Rowley.

Fernández-Mira, P., Morgan, E., Davidson, S., Yamada, A., Carando, A., Sagae, K., & Sánchez-Gutiérrez, C. H. (2021). Lexical diversity in an L2 Spanish learner corpus. The effect of topic-related variables. *International Journal of Learner Corpus Research, 7*(2), 230–258. https://doi.org/10.1075/ijlcr.20017.fer

Ferreira, A., & Elejalde, J. (2017). Análisis de errores recurrentes en un corpus de aprendices de español como lengua extranjera (Corpus CAELE). *Revista Brasileira de Linguística Aplicada, 17*(3), 509–538. https://doi.org/10.1590/1984-6398201710927

Ferreira, A., & 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://doi.org/10.15446/fyf.v33n1.84182

Ferreira Cabrera, A. 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*, 137–155. https://doi.org/10.5209/clac.71174

Fuchs, R., & Werner, V. (2018). The use of statives progresives by school learners of English and the importance of the variable context. En Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and crosslinguistic perspective in learner corpus research* (pp. 195–224). John Benjamins. https://doi.org/10.1075/ijlcr.17010.fuc

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual an learner corpora. En K. Aijmer, N. Altengerg, & M. Johansson (Eds.), *Langeuges in Contrast. Text-based Cross-linguistic Studies* (pp. 37–51). Lund University Press.

Granger, S. (Ed.). (1998). *Learner English on computer*. Longman.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480. https://doi.org/10.1558/cj.v20i3.465-480

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. En K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–32). John Benjamins. https://doi.org/10.1075/scl.33.04gra

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: Learner corpus research-past, present and future. En S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.001

Granger, S., Hung, J., & Pech-Tyson, S. (2002). *Computer learner corpora, second language acquisition, and foreign language teaching*. John Benjamins. https://doi.org/10.1075/lllt.6

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL, 19*(3), 252–268. https://doi.org/10.1017/S0958344007000237

Guijarro-Fuentes, P., Amenos-Pons, J., & Ahern, A. (2023). Acquisition of estar + adjective in L2 Spanish by L1 French and Portuguese speakers. *Spanish in Context, 20*(2), 282–312. https://doi.org/10.1075/SIC.00085.GUI

Gutiérrez Araus, M. L. (2022). *Problemas fundamentales de la gramática del español 2/L* (8.ª edición). Arco/ Libros.

Hidalgo, M. (2023). *Estudio del género gramatical en el COWS-L2H Aspectos lingüísticos y extralingüísticos* [Trabajo final de máster, Universidad de Salamanca].

Hincapié, D. (2018). Corpus de aprendientes de español como lengua extranjera y segunda lengua (CÁELE/2): El componente escrito. *Forma y Función, 31*(2), 129–143. https://doi.org/10.15446/fyf.v31n2.74659

Koike, D., & Witte, J. (2016). Spanish Corpus Proficiency Level Training Website and Corpus: An open-source, online resource for corpus linguistics studies. En M. Alonso Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives* (pp. 169–196). John Benjamins. https://doi.org/10.1075/scl.78.07koi

Le Bruyn, B., & Paquot, M. (2020). *Learner corpus research and second language acquisition: An attempt at bridging the gap*. Cambridge University Press. https://doi.org/10.1017/9781108674577.002

Leech, G. (1997). Teaching and language corpora: A convergence. En A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). Longman. https://doi.org/10.4324/9781315842677-1

Lozano, C. (2022) CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research, *Second Language Research, 38*(4), 965-983. https://doi.org/10.1177/02676583211050522

Lozano, C., & Fernández Mira, M. (2022). Designing, compiling and interrogating corpora in L2 Spanish acquisition research. *Journal of Spanish Language Teaching, 9*, 190–206. https://doi.org/10.1080/23247797.2022.2157086

Lozano, C., & Quesada T. (2023). What corpus data reveal about the position of antecedent strategy: Anaphora resolution in Spanish monolinguals and L1 English-L2 Spanish bilinguals. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1246710

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. En S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–158). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.007

Lüdeling, A., Walter, M., Kroymann, E., & Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of corpus linguistics, 1*, 14–17.

Maraschin Bruscato, A., & Baptista, J. (2022). The synchronous and asynchronous learning of anaphora: A corpus-based analysis with learners of English and Spanish. *Research in Corpus Linguistics, 11*(1), 1–28. https://doi.org/10.32714/ricl.11.01.01

Marsily, A. (2022). La conciencia pragmática de los estudiantes nativos y no nativos del español: El caso de las peticiones hacia profesores. *Revista Signos. Estudios De Lingüística, 55*(108), 338–362. https://doi.org/10.4067/S0718-09342022000100338

Miaschi, A., Davidson, S., Brunato, D., Dell'Orletta, F., Sagae, K., Sanchez-Gutierrez, C. H., & Venturi, J. (2020). Tracking the evolution of written language competence in L2 Spanish learners. En *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 92–101). https://doi.org/10.18653/v1/2020.bea-1.9

Minnillo, S., Sánchez-Gutiérrez, C. H., Carando, A., Davidson, S., Fernández Mira, P., & Sagae, K. (2022). Preterit-imperfect acquisition in L2 Spanish writing: Moving beyond lexical aspect. *Research in Corpus Linguistics, 10*(1), 156–184. https://doi.org/10.32714/ricl.10.01.08

Palacios, N., Barcala, A., & Rojo, V. (2019). El Corpus de Aprendices de Español (CAES) y sus aplicaciones para la enseñanza/aprendizaje del español como lengua extranjera. En M. Blanco, H. Olbertz, & V. Vázquez Rozas (Eds.), *Corpus y construcciones. Perspectivas hispánicas* (pp. 273–303). Universidade de Santiago de Compostela.

Parodi, G., Cantos-Gómez, P., & Howe, C. (2022). *Lingüística de corpus en español / The Routledge handbook of Spanish corpus linguistics*. Routledge. https://doi.org/10.4324/9780429329296

Pino Rodríguez, A. (2012). El uso de combinaciones de palabras con que en un corpus de aprendices suecos de español como lengua extranjera. *Anuari de Filologia. Estudis de Lingüística, 2*, 211–212.

RAE & ASALE. (2005). *Diccionario panhispánico de dudas* [en línea]. https://www.rae.es/dpd/

RAE & ASALE. (2009). *Nueva gramática de la lengua española*. Espasa.

Rojo, G., & Palacios, I. (2022). Los corpus de aprendices de español como L2. En G. Parodi & P. Cantos Gómez, *Lingüística de corpus en español. Árbol académico/Chad Howe* (pp. 74–88). https://doi.org/10.4324/9780429329296-7

Rojo, G., Palacios, I., Sampedro, M., & Marsily, A. (2022). Los corpus de aprendices de español LE/L2: Panorama actual y perspectivas futuras. *Journal of Spanish Language Teaching, 9*(2), 174–189. https://doi.org/10.1080/23247797.2022.2157085

Rothman, J., González Alonso, J., & Miller, D. (2018). The acquisition of second language Spanish morphosyntax. En K. L. Geeslin (Ed.), *The Cambridge handbook of Spanish linguistics* (pp. 689–715). Cambridge University Press. https://doi.org/10.1017/9781316779194.032

Sampedro Mella, M. (2022). Corpus de aprendices y cuestiones de discurso en la interlengua: Un nuevo acercamiento aplicado a la enseñanza. *Foro de profesores de E/LE, 18*, 163–182. https://doi.org/10.7203/foroele.18.24965

Sánchez-Gutiérrez, C., & Fernández Mira, P. (2022). Datos longitudinales en corpus de aprendientes de español. En G. Parodi *et al.* (Eds.), *Lingüística de corpus en español, The Routledge handbook of Spanish corpus linguistics* (pp. 374–387). Routledge. https://doi.org/10.4324/9780429329296-28

Sánchez-Gutiérrez, C., De Cock, B., & Tracy-Ventura, N. (2022). Spanish corpora and their pedagogical uses: Challenges and opportunities. *Journal of Spanish Language Teaching, 9*(2), 105–115. https://doi.org/10.1080/23247797.2022.2157078

Sánchez Rufat, A. (2015). Análisis contrastivo de interlengua y corpus de aprendientes: Precisiones metodológicas. *Pragmalingüística, 23*, 191–210. https://doi.org/10.25267/Pragmalinguistica.2015.i23.11

Selinker, L. (1972). Interlanguage. Product Information. *International Review of Applied Linguistics in Language Teaching, 10*, 209–241. https://doi.org/10.1515/iral.1972.10.1-4.209

Thewissen, J. (2015). *Accuracy accross proficiency levels. A learner corpus aproach*. Presses Universitaires de Louvain.

Valverde, P. (2020). Diseño y creación de un corpus de aprendices de ELE en Japón (CELEN). E-AESLA. *Revista Digital de Lingüística Aplicada, 6*, 223–240. https://cvc.cervantes.es/lengua/eaesla/pdf/06/16.pdf

van Rooy, B. (2015). Annotating learner corpora. En S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 79–105). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.005

Yamada, A., Davidson, S., Fernández-Mira, P., Carando, A., Sagae, K., & Sánchez-Gutiérrez, C. (2020). COWS-L2H: A corpus of Spanish learner writing. *Research in Corpus Linguistics, 8*(1), 17–32. https://doi.org/10.32714/ricl.08.01.02

Yonekawa, K., & San Mateo Valdehíta, A. (2022). Análisis de errores en el uso del artículo en textos de aprendientes japoneses de español L2. Estudio de corpus. *Cuadernos CANELA, 33*, 71–89. https://doi.org/10.2107/canela.33.0_71

Anexo I

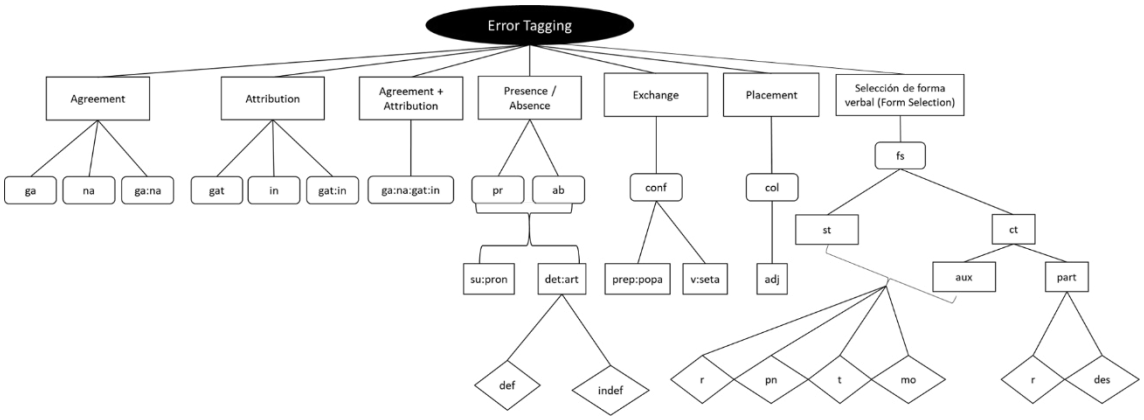


Figura 8. Listado completo de etiquetas utilizadas.