Lingue e Linguaggi
Lingue Linguaggi 73 (2025), 23-42
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v73p23

ANÁLISIS COMPARATIVO DE LAS UNIDADES FRASEOLÓGICAS CON EL VERBO ECHAR EN LOS CORPUS DE APRENDIENTES *CEDEL2* Y *CAES*
Errores y competencia

MARIA ANNESE
UNIVERSITÀ DEGLI STUDI 'G. D'ANNUNZIO' CHIETI-PESCARA

Abstract - Phraseology is known to be a problematic domain for foreign language learners. However, it constitutes an important area of interlanguage, as it links it to the target culture. The problems that may arise in the process of acquiring phraseological units (PUs) concern their appropriate use, as different types of errors may arise. Against this background, the present article focuses on the occurrences of PUs containing the Spanish verb *echar* in two learner corpora available online: *CEDEL2* and *CAES*. Specifically, the aim is to determine the proficiency band (A, B or C) at which PUs containing the verb *echar* are acquired and used correctly. Next, the types of phraseological errors related to this verb are analysed and, finally, the partial phraseological richness (PPR) of the learner's language is measured in comparison with the language of native Spanish speakers (from the native *CEDEL2* control subcorpus). Potential applications of this study concern the development of corpus-based teaching activities to foster the acquisition of PUs containing both *echar* and other verbs.

Keywords: learner corpus research, phraseology, Spanish as a foreign language, error analysis, interlanguage analysis.

1. Introducción y objetivos

El presente estudio pretende analizar las unidades fraseológicas (UFS) que contienen el verbo *echar* en dos corpus de aprendientes disponibles en línea, el *CEDEL2* y el *CAES*, con el fin último de perfilarlas y promover su adquisición, intentando explicar la presencia de errores recurrentes según el nivel de competencia y promoviendo un uso variado de dichas UFS.

La investigación, en concreto, tiene un triple objetivo:

1. determinar el nivel de competencia lingüística en el que se adquieren y utilizan correctamente las distintas UFS con *echar*;
2. identificar y clasificar los errores cometidos por los aprendientes en las diferentes UFS;
3. calcular la riqueza fraseológica parcial referida al verbo *echar* y compararla con la del subcorpus nativo de control presente en el *CEDEL2*.

En los siguientes apartados, tras una exposición inicial del marco teórico en el que se basa el estudio (apartado 2), se profundizará en la metodología y se presentarán los datos preliminares relativos a los corpus de aprendientes considerados para el análisis (apartado 3). Posteriormente, en el apartado 4, se expondrá el análisis propiamente dicho, con secciones relativas a los tres objetivos descritos anteriormente. Por último, se extraerán las conclusiones del análisis y se esbozarán los desarrollos futuros y las implicaciones metodológicas del estudio (apartado 5).

2. Marco teórico

Este artículo se enmarca en el campo de estudio de la investigación sobre corpus de aprendientes (Granger *et al.* 2015). Sin embargo, para poder presentar los corpus de aprendientes de español utilizados para el análisis, es necesario introducir cuestiones clave, como el análisis de errores (AE) y la hipótesis de la existencia de la interlengua (IL), así como presentar el papel desempeñado por las UFS en la interlengua de los aprendientes de español.

2.1. El análisis de errores y UCLEE

Este trabajo se inscribe en el marco de la lingüística contrastiva, una rama de la lingüística aplicada cuyo origen se remonta a los años cuarenta con el análisis contrastivo (AC) de tipo estructuralista y conductivista. Este tipo de análisis presuponía la comparación entre una L1 y una L2 con el objetivo de identificar las áreas de dificultad en el aprendizaje y crear una gramática contrastiva que facilitara la adquisición lingüística evitando errores (Santos Gargallo 2004). Sin embargo, según esta metodología de análisis *a priori*, las incorrecciones se debían únicamente al fenómeno de la transferencia negativa de la lengua materna del aprendiente, por lo que este enfoque fue duramente criticado y tuvo que ceder el paso al más polifacético análisis de errores (AE) en los años sesenta (Santos Gargallo 2004).

Basado en los estudios de Corder (1967), inspirados en la lingüística generativo-transformacional y en los principios del cognitivismo, el AE se fundamenta en el análisis de errores *a posteriori*, ocupándose de las producciones concretas de los aprendientes de una segunda lengua o lengua extranjera.

En los años ochenta, este enfoque recibió un nuevo impulso con los estudios de Burt, Dulay y Krashen (1982), que dieron un renovado valor al concepto de error, identificado como prueba tangible del proceso de aprendizaje, pero también con un nuevo artículo de Corder (1981), que amplió el análisis de la competencia lingüística a la competencia comunicativa y subrayó la importancia de analizar también las producciones correctas de los aprendientes.

El análisis de errores, al igual que el análisis contrastivo, sufrió un declive gradual debido a una serie de deficiencias que lo caracterizaban: por ejemplo, las taxonomías utilizadas para describir los errores eran muchas y variadas y, al mismo tiempo, no lograban definir y etiquetar con precisión cada error; además, la escasez de aplicaciones didácticas hacía que este enfoque fuera bastante estéril (Rigamonti 2006, p. 21). A pesar de las críticas, se atribuye al análisis de errores el mérito de haber dado un renovado sentido positivo a los errores como elementos centrales del aprendizaje de lenguas (Santos Gargallo 2004), y de haber creado una metodología de análisis basada en tres etapas: la identificación de los errores en su contexto, el análisis y la identificación de su causa y, por último, la clasificación de los errores según la taxonomía más adecuada (Rigamonti 2006).

En épocas más recientes y paralelamente al desarrollo de las tecnologías de investigación lingüística, los principios y estudios en los que se basaba el análisis de errores han propiciado la aparición de otro enfoque: el AE asistido por ordenador (*computer-aided error analysis*, CEA), que, mediante el uso de ordenadores y especialmente de programas informáticos que pueden definirse como *error editors*, permite un análisis más detallado y exacto de las incorrecciones pero, sobre todo, posibilita la visualización de éstas en el contexto en el que aparecen y, por tanto, el desarrollo de actividades orientadas a la enseñanza centrada en el alumno (Granger 2002).

En esta dirección, en octubre de 2023 se lanzó UCLEEv2, la segunda versión del editor de errores desarrollado en el *Centre for English Corpus Linguistics* (CECL) de la Universidad Católica de Lovaina la Nueva (bajo licencia *Creative Commons*). Este editor de errores va acompañado de un Manual de etiquetado (Granger *et al.* 2022) en el que se registran las etiquetas de error según siete categorías principales: errores formales, errores gramaticales, errores léxico-gramaticales, errores léxicos, errores de puntuación, palabras redundantes/faltantes/dispuestas en orden incorrecto y, por último, errores relacionados con elecciones lingüísticas inadecuadas al contexto (*infelicities*).

Además de utilizar el sistema de etiquetas predefinido, se puede crear fácilmente un nuevo sistema de etiquetas de error, lo que lo convierte en una herramienta extremadamente flexible y adaptable a las necesidades de la investigación. UCLEEv2 también implementa una función que permite crear actividades didácticas basadas en la anotación y corrección previa de los errores; por ejemplo, es posible convertir el corpus utilizado en un texto con huecos en los que insertar la palabra o estructura lingüística que corregirá el error, con o sin la ayuda de la etiqueta que indica el tipo de error, según se desprende de la Guía del usuario (Thewissen *et al.* 2023). Como veremos, UCLEEv2 forma parte integrante de la metodología de investigación de este artículo en la medida en que, tras crear un nuevo sistema de anotación de errores (Anexo 1), permitió realizar un análisis estadístico de los mismos según su tipología (apartado 3 y 4.2).

2.2. Las UFS verbales en la interlengua

A raíz de las críticas que recibió el análisis de errores, en el campo de la lingüística contrastiva fue surgiendo un concepto que formalmente adoptó distintas denominaciones, aunque similares en su contenido (Rigamonti 2006). Nemser en 1971 hablaba de sistema aproximado, Corder en el mismo año lo definía como dialecto idiosincrásico, pero es la interlengua de Selinker (1972) el término al que se vincula la definición más acertada que ha llegado hasta nuestros días:

> [...] one would be completely justified in hypothesizing, perhaps even compelled to hypothesize, the existence of a separate linguistic system based on the observable output which results from a learner's attempted production of a TL (target language) norm. This linguistic system we will call 'interlanguage' (IL). (Selinker 1972, p. 214).

La interlengua, por tanto, sería diferente tanto de la lengua meta como de la lengua materna del aprendiente, constituyendo un sistema lingüístico independiente, formado por etapas sucesivas, dinámicas y continuas (Sánchez Rufat 2015; Núñez Nogueroles 2019). El análisis de la interlengua se basa en las tres únicas categorías de datos lingüísticos que pueden observarse concretamente (Selinker 1972): las emisiones en la lengua materna del aprendiente, las emisiones en su IL y las emisiones en la lengua meta de los hablantes nativos. En analogía con la definición de interlengua, uno de los objetivos de este artículo es precisamente comprender si la producción fraseológica difiere entre aprendientes y hablantes nativos, y en qué medida. Como se ha expresado anteriormente, nos centraremos en las UFS con el verbo *echar*, entrando así en contacto con una parte específica de la interlengua de los aprendientes de español, la fraseología verbal, que se puede observar concretamente, como veremos en el siguiente apartado, gracias a la interfaz de búsqueda intuitiva de dos corpus de aprendientes empleados para el análisis, el *CEDEL2* y el *CAES*.

Como afirman Ellis *et al.* (2015), el lenguaje formulaico puede ser de gran utilidad en la adquisición eficaz de un idioma. En efecto, permite el uso de unidades lingüísticas preconfeccionadas que, si se adquieren correctamente, facilitan la producción del lenguaje,

haciéndolo más fluido. Incluso podría afirmarse, como hace el propio Ellis (2012), que la adquisición de la lengua coincide con el aprendizaje de secuencias lingüísticas, por lo que es evidente que las UFS constituyen un área fundamental de la LE y de la L2 y, en consecuencia, del desarrollo de la IL.

Para enmarcar la unidad fraseológica (UF) como ítem lingüístico, nos remitimos a la taxonomía de Corpas Pastor (1996), que identifica tres grandes dominios de unidades dentro de la fraseología española: las colocaciones, las locuciones y los enunciados fraseológicos. Cada una de las unidades identificadas se caracteriza por una serie de atributos en distintos grados, como la frecuencia de uso y la coocurrencia de las palabras constituyentes, la institucionalización y lexicalización, la idiomaticidad y la variabilidad.

Nos parecen de gran utilidad para el presente estudio las definiciones dadas por Jezek (2005) de las principales categorías de combinaciones de palabras que trataremos en el análisis del corpus examinado: las colocaciones y las locuciones o expresiones idiomáticas. Las primeras se definen como sigue (2005, p. 192):

> Una collocazione è una combinazione di parole soggetta a una restrizione lessicale, per cui la scelta di una specifica parola (il collocato) per esprimere un determinato significato, è condizionata da una seconda parola (la base) alla quale questo significato è riferito.

Una colocación que encontraremos a lo largo de nuestro estudio es, por ejemplo, *echar a la basura*, donde la base es *echar*, mientras que el colocado es *basura*. Como puede verse, aunque se trata de una UF, el nivel de idiomaticidad es bajo, cuando no nulo, ya que el significado total es composicional, es decir, inferible a partir del significado de los elementos individuales que la componen. Dentro de las colocaciones, según la taxonomía de Jezek (2005), se encuentran las construcciones con verbo de apoyo: se trata de combinaciones particulares de palabras sujetas a restricción léxica en las que el verbo está parcial o totalmente desemantizado y el sustantivo que sigue al verbo es el portador del significado. Como se verá, en la metodología del presente estudio, las construcciones con verbo de apoyo se asimilan a colocaciones, ya que se trata de un análisis realizado sobre un corpus más bien pequeño en el que distinguir entre colocaciones estándares y construcciones con verbo de apoyo no habría conducido a resultados relevantes. Sin embargo, no excluimos la posibilidad de ampliar el estudio, distinguiendo entre colocaciones y construcciones con verbo de apoyo.

Las locuciones o expresiones idiomáticas, a diferencia de las colocaciones, son UFS con un grado variable de idiomaticidad y no composicionalidad (Jezek 2005, pp. 198-199). Entre las locuciones que trataremos, como veremos, destaca la locución *echar de menos* por la frecuencia de su aparición en los corpus considerados: su significado no puede deducirse de la suma de los significados de los elementos individuales que la componen.

Según el punto de vista asumido por este estudio, la enseñanza-aprendizaje de la fraseología se configura como el nexo entre interlengua y cultura (Castillo Carballo 2002): conocer, pero sobre todo utilizar activa y eficazmente una lengua significa también saber descifrar los fragmentos lingüísticos preempaquetados que, con frecuencia y como hemos mencionado, poseen cierto grado de idiomaticidad y no composicionalidad, características que a veces los hacen opacos y difíciles de adquirir por los aprendientes. Además, el uso de UFS en el proceso de adquisición de una lengua extranjera (LE) o lengua segunda (L2) se ve obstaculizado por la falta de correspondencia (formal o semántica), en algunos casos, entre las UFS de los dos idiomas implicados, la lengua meta y la lengua materna (Martín Salcedo, 2015), lo que dificulta una producción interlingüística correcta y eficaz.

Son precisamente estas dificultades las que nos impulsan en el apartado del análisis a investigar el empleo de las UFS en los corpus de los aprendices. De hecho, en la fase de producción de un texto entran en juego diversos factores externos e internos a la tarea que pueden influir de algún modo en la producción lingüística final: como veremos con referencia a las UFS verbales que contienen *echar*, consideradas como caso de estudio, el nivel lingüístico-comunicativo y el tipo de tarea desempeñarán un papel clave no sólo en la elección, sino también en la corrección y variedad de las UFS empleadas para vehicular determinados significados.

2.2.1. Medir la riqueza fraseológica de los aprendientes

Como se verá con detalle en el apartado de metodología (apartado 3) y en el apartado del tercer objetivo de este estudio (apartado 4.3), se ha considerado oportuno adaptar un modelo estadístico utilizado por Read (2000) para calcular la riqueza léxica y retomado por Orol González y Alonso Ramos (2013) para calcular la riqueza colocacional. Si Read, por tanto, describió la riqueza léxica a partir de cuatro parámetros con el fin de evaluar la adquisición de vocabulario de los aprendientes, Orol González y Alonso Ramos adaptan su modelo al estudio de las colocaciones con el fin de comparar el uso de éstas entre nativos y aprendices en el *CEDEL2*, concretamente en un subcorpus de 100 textos de nativos y 100 de aprendientes. Lo que se desprende del estudio es que los aprendientes, aunque utilizan las colocaciones en un grado considerable, no muestran el mismo nivel de riqueza colocacional que los nativos.

Para medir la riqueza fraseológica en nuestro estudio hacemos referencia a dos de los parámetros que emplea Read para calcular la riqueza léxica: variedad y número de errores[1]. Como puede observarse, estos indicadores deben mucho al modelo de análisis del rendimiento lingüístico CAF (*complexity*, *accuracy and fluency*), descrito recientemente por Pallotti (2009) y posteriormente por Housen, Kuiken y Vedder (2022), pero empleado en el campo de la investigación sobre la adquisición de segundas lenguas y lenguas extranjeras, aunque de forma variable a lo largo del tiempo, desde los años setenta (Housen *et al.* 2022). En particular, el modelo empleado por Read, tal y como fue concebido para el análisis y la evaluación de textos escritos, se centra en la evaluación de la complejidad (en cuanto a variedad, sofisticación y densidad) y de la corrección (parámetro relativo al número de errores). La corrección es definida por Pallotti (2009) como el grado de conformidad con ciertas normas (que generalmente coinciden con la norma de los nativos de la lengua meta, no sin problemas[2]), la complejidad se caracteriza por una polisemia que dificulta su definición: nos remitimos a la definición de Pallotti (2015), en la que la complejidad es un parámetro estructural.

Como veremos, nuestro estudio se centra únicamente en la variedad fraseológica y el número de errores, combinando complejidad estructural y corrección en un modelo estadístico adaptado y simplificado, diseñado para calcular la riqueza fraseológica parcial, es decir, relativa únicamente a las UFS que contienen el verbo *echar*.

Para permitir el análisis de UFS en la interlengua de aprendientes de español, es necesario tener acceso a sus producciones, que se pueden encontrar, por ejemplo, en los corpus de aprendientes.

[1] Los otros dos parámetros empleados por Read para medir la riqueza léxica son la sofisticación y la densidad.

[2] Véase, por ejemplo, Granger (2015) para una introducción a la espinosa cuestión de la falacia comparativa, una de las principales críticas dirigidas al modelo de investigación propuesto por ella: el análisis comparativo de la interlengua (*Contrastive Interlanguage Analysis*, CIA).

2.3. Dos corpus de aprendientes del español: el CAES y el CEDEL2

En la presente subsección, introduciremos brevemente la investigación de los corpus de aprendientes, y presentaremos los corpus que se utilizarán en la sección de análisis de este artículo (apartado 4).

Los corpus de aprendientes, definidos por Granger, Gilquin y Meunier (2015, p. 1) como colecciones electrónicas de datos naturales (o casi naturales) producidos por estudiantes de lenguas extranjeras (LE) o segundas lenguas (L2) y reunidos de acuerdo con criterios de diseño explícitos, son el objeto de un campo de investigación que se desarrolló aproximadamente 30 años después de la aparición de la lingüística de corpus, con el fin de satisfacer las nuevas necesidades que surgieron en el campo de la adquisición de segundas lenguas:

> The idea of compiling learner corpora […] and applying corpus linguistic tools and methods to analyze them arose from the wish to bring to the field of Second Language Acquisition (SLA) the same kinds of benefits that corpora were providing to the linguistic field. Several linguists with a keen interest in SLA, often because they were also language teachers, concurrently but independently started to compile and analyze large electronic collections of L2 data. Their objectives in embarking on this new type of research were theoretical, i.e. they wanted to gain a better understanding of the process of learning a foreign or second language (L2), and/or practical, i.e. with a view to designing more efficient language teaching tools and methods. (Granger 2012, pp. 7-8)

Desde sus comienzos en la década de los noventa de la mano de Sylviane Granger con la fundación del CECL (*Centre for English Corpus Linguistics*) en Lovaina la Nueva, y el diseño del primer corpus de aprendientes reconocido a nivel internacional, el ICLE (*International Corpus of Learner English*) (Sánchez Rufat 2015a, 2015b), la investigación basada en corpus de aprendientes se ha desarrollado considerablemente para incluir otras lenguas además del inglés, entre ellas el español. Los corpus de aprendientes del español con los que se ha realizado el trabajo de análisis son, como ya se ha mencionado, el *CAES* (Corpus de Aprendices de Español) y el *CEDEL2* (Corpus Escrito del Español L2). A continuación, se expondrán brevemente las principales características de los dos corpus.

El proyecto *CAES* fue financiado por el Instituto Cervantes y realizado por un equipo de investigación de la Universidad de Santiago (Rojo y Palacios 2016, p. 62). Se trata de un corpus escrito que está totalmente disponible en línea (https://galvan.usc.es/caes/) y que actualmente recoge 1.045.097 unidades lingüísticas en un total de 6.591 tareas de estudiantes de once lenguas maternas (alemán, árabe, chino mandarín, francés, griego, inglés, italiano, japonés, polaco, portugués y ruso), ubicables desde el nivel A1 hasta el nivel C1 del MCER. Se excluyó el nivel C2 porque los alumnos participantes en el proyecto, al estar todavía cursando sus estudios, aún no lo habían alcanzado (Rojo y Palacios 2016, p. 63).

Además del nivel de competencia lingüístico-comunicativa del aprendiente y de su L1, los metadatos seleccionables en la interfaz de búsqueda del sitio web del *CAES* incluyen información relacionada con el alumno, como el sexo, el país de origen y la edad, e información relacionada con el texto, como la tipología textual y el tema. Adicionalmente a estos metadatos, que se pueden seleccionar en las búsquedas de concordancias de tipo KWIC (*Key Word In Context*) y frecuencias, es posible obtener más información sobre cada participante en el estudio descargando los resultados obtenidos, o sea el nivel de estudios, la edad de inicio en el estudio del español, el número de meses estudiando español y los contactos personales en países de habla hispana.

El diseño del corpus se basa en una correspondencia rigurosa entre el tipo de tarea (orientada hacia una muestra de lengua auténtica) y el nivel certificado de español, de acuerdo con las directrices establecidas por el MCER y el Instituto Cervantes para los exámenes DELE (Diploma de Español como Lengua Extranjera) (Rojo y Palacios 2016, p. 65). Cada alumno realizó 2/3 tareas seleccionadas en función de su nivel lingüístico-comunicativo en español, como se puede observar en la Tabla 1.

NIVEL	TAREA
A1	cambio trabajo, familia y nota llegar tarde
A2	persona que admira, postal vacaciones y reserva habitación
B1	carta amigo, historia graciosa y reclamación compañía aérea
B2	fumar lugares públicos, solicitud admisión
C1	reclamación compañía gas, reseña película

Tabla 1
Tareas del CAES.

El *CAES* resulta ser una herramienta de gran utilidad para los profesionales del campo de ELE, como se afirma en Parodi (2015, p. 6):

> [El *CAES*] constituye un aporte muy robusto al área, no solo por la disponibilidad de contar con un corpus diversificado en un conjunto de variables que posibilita la investigación para el español como lengua extranjera, sino también por disponer de un diseño con soportes cuidadosamente pensados y elaborados de modo muy profesional y sustentados en documentación complementaria altamente útil. El equipo de la Universidad de Santiago de Compostela con apoyo del Instituto Cervantes ha realizado un trabajo científico y tecnológico de gran calidad.

Inicialmente originado dentro del proyecto WOSLAC (*Word Order in Second Language Acquisition Corpora*) de la Universidad Autónoma de Madrid como un corpus escrito de aprendientes basado en un par de lenguas específico, a saber, L1 inglés - L2 español (Lozano y Mendikoetxea, 2013), el *CEDEL2* (http://cedel2.learnercorpora.com/) comprende ahora textos escritos y, en menor medida (alrededor del 1%) audios y transcripciones de textos orales, de aprendientes de todos los niveles y de once lenguas maternas: inglés, alemán, holandés, portugués, italiano, francés, ruso, griego, japonés, chino y árabe.

El proyecto, dirigido por Cristóbal Lozano de la Universidad de Granada desde su comienzo en 2004, pretende dar respuesta a diversas necesidades en el ámbito de la investigación sobre adquisición de segundas lenguas, con especial atención a los principios de diseño de corpus elaborados por Sinclair (2005), entre los que destacan los cinco siguientes (Lozano 2022):

1. selección del contenido del corpus basada en criterios externos y no internos, es decir, basada en las funciones comunicativas de los textos y no en su lengua;
2. los textos deben ser representativos de la lengua elegida, en este caso la lengua de los aprendientes de español;
3. los temas de las tareas también deben elegirse en función de criterios externos;
4. los subcorpus deben ser comparables, es decir, diseñados según los mismos criterios;
5. el contenido y los metadatos del corpus deben estar ampliamente documentados.

Además de esto, como afirma el propio Lozano (2022), es posible reconocer en la construcción del *CEDEL2* una adhesión sustancial a las recomendaciones en el campo de la investigación sobre corpus de aprendientes dadas por Tracy-Ventura y Paquot (2021): entre ellas, mencionamos la inclusión de una amplia gama de metadatos y la presencia de un doble corpus nativo de control, como veremos a continuación.

De forma similar al *CAES*, es posible realizar búsquedas de diversos tipos, en particular relativas a concordancias (KWIC), textos y frecuencias de palabras. La interfaz permite seleccionar una gama muy amplia de metadatos, relacionados principalmente con el alumno: L1, medio (oral o escrito), sexo, tarea, nombre del archivo, nivel de competencia, edad, edad de exposición al español, nota de la prueba de nivel, autoevaluación de la competencia, años de estudio del español y meses de estancia en el extranjero. Además de estos metadatos, como en el caso de *CAES*, encontramos una rica información adicional cuando descargamos los resultados de la búsqueda realizada, en particular metadatos relacionados con la tarea (p. ej. tiempo y lugar de realización de esta y herramientas empleadas para llevarla a cabo).

El *CEDEL2*, como ya se ha mencionado, cuenta también con un corpus de control nativo dual (corpus de la lengua española L1 y corpus de varias lenguas maternas de los aprendientes en desarrollo, descritos en Lozano 2022, p. 973) que permite realizar operaciones de comparación exhaustivas dirigidas a detectar posibles interferencias inter e intralingüísticas. La participación de los informantes en el proyecto es totalmente voluntaria y se concreta rellenando un formulario de Google que incluye los datos personales que convergerán en los metadatos del corpus (véase supra), una prueba de nivel ideada en 1998 en la Universidad de Wisconsin, así como una sola tarea que, a diferencia del *CAES*, no está relacionada con el nivel de competencia lingüístico-comunicativa (Lozano 2022). Las tareas son variadas y, como en el caso del *CAES*, tienen como principal objetivo la autenticidad de la muestra de lengua, con el fin de que el corpus sea lo más representativo posible de la interlengua de los aprendientes de español, como se desprende de la Tabla 2, adaptada del sitio web del corpus.

N°	Título de la tarea (en español)	Descripción
1	Región donde vives	¿Cómo es la región donde vives?
2	Persona famosa	Habla de una persona famosa.
3	Película	Resume una película que has visto recientemente.
4	Vacaciones del año pasado	¿Qué hiciste el año pasado durante las vacaciones?
5	Planes para el futuro	¿Cuáles son tus planes para el futuro?
6	Viaje reciente	Describe un viaje que hayas hecho recientemente.
7	Experiencia	Cuenta una experiencia que hayas vivido.
8	Terrorismo	Habla del problema del terrorismo en el mundo.
9	Ley anti-tabaco	¿Qué opinas de la nueva ley anti-tabaco?
10	Parejas homosexuales	¿Crees que las parejas gais tienen derecho a casarse y a adoptar niños?
11	Legalización de la marihuana	¿Crees que la marihuana se debería legalizar?
12	Inmigración	Analiza los principales aspectos de la inmigración.
13	Rana	Mira las siguientes ilustraciones. Narra una historia basada en las ilustraciones. Puedes añadir ideas nuevas o ignorar algunas que aparezcan en las ilustraciones. Por favor, comienza la historia con la frase: "Un día..." https://goo.gl/so3S6W
14	Chaplin	Mira el siguiente vídeo de Charles Chaplin (4 minutos). Haz un resumen de la historia. Puedes ver el vídeo más de una vez. https://www.youtube.com/watch?v=4QkTNJFhu-g

Tabla 2
Tareas del *CEDEL2*.

Por último, el *CEDEL2* cuenta con un corpus gemelo de aprendientes de inglés, *COREFL* (*Corpus of English as a Foreign Language*), lo que permite realizar análisis intralingüísticos, puesto que los dos corpus comparten el mismo diseño y se basan en las mismas tareas descritas anteriormente.

No trataremos en detalle las peculiaridades de la construcción e implementación de los dos corpus descritos en este apartado; nos limitaremos a decir que ambos corpus disponen de un sitio web informativo y transparente donde, además de proporcionarse información detallada acerca del diseño, el contenido y los datos numéricos del corpus, se implementan guías técnicas para el uso de la interfaz de búsqueda, que de todas formas es extremadamente intuitiva en ambos casos.

Como se verá en los siguientes apartados, partiendo de estas premisas teóricas, se elaboró una metodología de investigación que pudiera dar cuenta de los diversos factores en juego en el uso, por parte de los aprendientes, de UFS verbales que contienen el verbo *echar*, entre los que destacan, como veremos en el análisis de los datos (apartado 4), el nivel lingüístico, la complejidad sintáctica de la UF utilizada y el tipo de tarea.

3. Metodología y datos preliminares

La metodología de investigación descrita en esta sección tiene en cuenta el triple objetivo de investigación formulado en la introducción: en particular, para perfilar las UFS verbales que contienen *echar* en los dos corpus de aprendientes descritos en el apartado anterior, fue necesario evaluar diferentes aspectos para diseñar un método que fuera adecuado al tipo de análisis, aunque perfectible (apartado 5).

En primer lugar, se descargaron las ocurrencias de *echar* en formato KWIC de las plataformas en línea de *CAES* y *CEDEL2* mediante una búsqueda realizada con un

comodín (*ech**): esta búsqueda produjo un resultado de 298 casos de ech* en el corpus combinado *CEDEL2+CAES*. Posteriormente, estas listas de ocurrencias se sometieron a un proceso de selección consistente en eliminar las ocurrencias del verbo echar en usos no fraseológicos (p.ej. *El ciervo nos* ***echó*** *a mí y a mi perro al estanque debajo de la piedra grande*). También se consideraron superfluas para el estudio las apariciones incorrectas del verbo *hacer* sin h (p. ej. *Tengo* ***echos*** *mis estudios primarias y secundarias* […]) y los casos de palabras incorrectas que empiezan por ech-, como por ejemplo **echuchar* (en lugar de *escuchar*), **echografía* (*ecografía*) y **echymoses* (*equimosis*). El mismo procedimiento se utilizó para el subcorpus nativo de *CEDEL2*, que será crucial para el objetivo relativo a la medición de la riqueza fraseológica parcial (apartado 4.3). Además, se combinaron las ocurrencias resultantes en los dos corpus de aprendices y se trataron como un corpus único (*CEDEL2+CAES*), como se desprende de la Tabla 3, que recoge los datos preliminares fundamentales para el análisis.

	CAES	**CEDEL2**	**Subcorpus nativo de CEDEL2**
Tokens	1 045 097	737 398	304 211
UFS con el verbo *echar*	195		54
f. por 100,000 palabras	11		18

Tabla 3
Datos preliminares y frecuencias absolutas (aproximadas a la unidad).

En segundo lugar, se categorizaron las diferentes UFS según su estructura sintáctica, en función de las categorías de UFS con *echar* identificadas por Martín Salcedo (2015). En concreto, las ocurrencias se dividieron en cuatro tipos[3]:

1. colocaciones sin preposición (p. ej. *Pero la mujer lo ve y le echa una bronca*);
2. locuciones sin preposición (p. ej. *Pero no, ¡no era un árbol sino un ciervo que estaba echando una siesta!*);
3. colocaciones con preposición (p. ej. *se echó en la cama*);
4. locuciones con preposición (p. ej. *os echo mucho de menos*).

Posteriormente, se identificaron los errores fraseológicos en el corpus *CEDEL2+CAES* y en el subcorpus nativo de *CEDEL2*. Como referencia para identificar los errores se utilizaron diccionarios en línea (*DLE*, *DPD*, *DAMER*[4]), corpus (*esTenTen* y *CORPES XXI*[5]) y tres cuestionarios administrados a informantes nativos a través de un formulario

[3] Las colocaciones se diferencias de las locuciones porque se podrían definir como sintagmas que presentan un menor nivel de fijación e idiomaticidad (Martín Salcedo, 2015).

[4] *DLE* es el acrónimo del Diccionario de la Lengua Española de la RAE (https://dle.rae.es/), el *DPD* es el Diccionario Panhispanico de Dudas de la RAE (https://www.rae.es/dpd/) y el *DAMER* es el Diccionario de Americanismos de la Asociación de Academias de la Lengua Española (https://www.asale.org/damer/). Véase la bibliografía para más información.

[5] *EsTenTen* (Kilgariff y Renau 2013) es un amplio corpus de la lengua española que puede consultarse y analizarse en la plataforma en línea Sketch Engine. Se trata de un corpus de textos escritos recogidos en la red y representativos de las distintas variedades peninsulares y americanas de la lengua española. Según se desprende de la web de Sketch Engine (https://www.sketchengine.eu/estenten-spanish-corpus/), la versión 2018 del corpus es más amplia que la original de 2011: de hecho, alcanza un total de 20.000 millones de tokens. De forma similar a *esTenTen*, el CORPES XXI (Corpus del Español del siglo XXI) es un corpus de lengua general cuyo objetivo es presentar las características de la lengua española actual; alcanza hoy los 350 millones de formas ortográficas y consta de textos escritos y orales procedentes de España, América,

Google (llamado *Detección de errores fraseológicos (VERBO ECHAR) A, B y C*). Cada cuestionario contenía 65 frases con *echar* extraídas de los corpus de aprendientes (según el ejemplo en la Imagen 1). A los informantes se les pidió valorar si cada frase era correcta (sí) o incorrecta (no) con respecto al empleo del verbo *echar* en UFS como *echar de menos* y *echar un vistazo*, e indicar, a su discreción, el tipo de error (en la casilla *Otro*). Se les pidió detectar los errores fraseológicos, excluyendo los errores meramente gramaticales del entorno de la UF. Además, tenían la posibilidad de añadir un comentario al final del formulario.

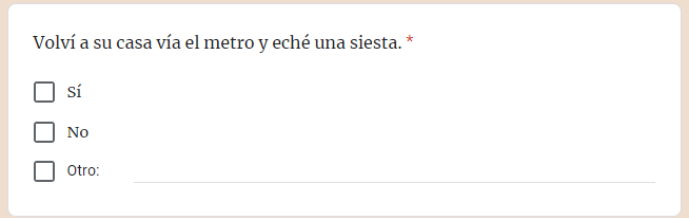

Imagen 1
Ejemplo de ocurrencia del formulario.

Tras excluir los errores gramaticales y ortográficos, como la conjugación incorrecta del verbo o la ruptura de las reglas de acentuación, las ocurrencias erróneas se anotaron en UCLEEv2 (Thewissen *et al.* 2023) y se clasificaron en las seis categorías siguientes, según un criterio lingüístico, en analogía con la taxonomía de errores de Graciela Vázquez (1991) descrita por Rigamonti (2012)[6]:

1. construcción incorrecta con respecto a la morfosintaxis (p. ej. […] **se echó un vistazo a su alrededor*);
2. posición incorrecta del modificador (p. ej. **Te echo de menos mucho*);
3. errores preposicionales, categoría que incluye la falta de una preposición necesaria, la presencia de una preposición innecesaria y la elección errónea de la preposición (p. ej. **Te echo mucho en menos*);
4. error con respecto al objeto directo (p. ej. **Echo la Georgia de menos*);
5. UF inexistente (p. ej. **Se echó la vista y se puso a fumar*);
6. contexto incorrecto desde un punto de vista pragmático (p. ej. **Espero que no te molestará mi pregunta y que no te echará de menos tu libro*).

Para lograr el tercer objetivo, se calculó la riqueza fraseológica parcial (RFP) del corpus *CEDEL2+CAES* y del subcorpus nativo del *CEDEL2* con el fin de comparar los datos y determinar si, de forma similar a la conclusión de Orol González y Alonso Ramos, los aprendices utilizan las UFS con un menor grado de corrección formal y variedad léxica que los hispanohablantes.

Paquot (2019) expresa la necesidad de estudios que se centren en el estudio de la complejidad de las unidades fraseológicas empleadas por los aprendientes: simplificando el modelo concebido por Read y adaptado por Orol González y Alonso Ramos, en línea con la visión "simple" de la complejidad de Pallotti (2015), en nuestro modelo estadístico la corrección formal calculada a través de la medición del número de errores (NE) solo se combina con el parámetro de la variedad de UFS que contienen el verbo echar (VF), tal y

Filipinas y Guinea Ecuatorial, según indica su descripción en su página web (https://www.rae.es/banco-de-datos/corpes-xxi).

[6] La anotación de errores se realizó siguiendo el *tagset* del Anexo 1, que puede utilizarse para futuros estudios.

como se expresa en el párrafo 2.2 del marco teórico. La riqueza fraseológica parcial (RFP) se obtiene restando el número de errores de la variedad fraseológica (véase la sección 4.3 para más detalles).

4. El análisis

Teniendo en cuenta las premisas teóricas de este estudio y la metodología expuesta en el apartado anterior, a continuación abordaremos los tres objetivos del estudio por separado, para centrarnos en tres cuestiones que consideramos fundamentales para sentar las bases de un análisis encaminado a trazar un perfil fraseológico, aunque sea parcial, de los aprendientes de español como lengua extranjera: la interacción entre la competencia lingüística y el uso de UFS con *echar*, la presencia de diversos tipos de estructuras y de errores y, por último, la relación entre la variedad fraseológica (vinculada, como hemos visto en el marco teórico, al concepto de complejidad) y la corrección formal (inversamente proporcional al número de errores cometidos por los aprendientes).

Estos tres nudos generan otras tantas preguntas de investigación, explicitadas al principio de cada subapartado de esta sección y acompañadas de una hipótesis preliminar. Tras el análisis propiamente dicho, extraeremos las principales conclusiones para cada objetivo, para luego integrarlas en el apartado de conclusiones, en el que se presentan las posibles futuras aplicaciones didácticas (apartado 5).

4.1. Objetivo 1: Errores Y Competencia

La primera pregunta de investigación a la que intentaremos dar respuesta puede formularse como sigue: ¿cómo influye el nivel de competencia lingüística-comunicativa en la frecuencia de uso y en el número de errores relacionados con las UFS verbales que contienen *echar*? La primera hipótesis formulada en este estudio plantearía un aumento sustancial de las ocurrencias de UFS con echar en los niveles intermedio y avanzado, ya que en éstos se concreta la enseñanza de la fraseología (García Muruais 1998); esto conlleva a una reducción progresiva del número de errores y a un aumento de su uso a medida que se desarrolla la competencia lingüística.

Para comprobar esta hipótesis de investigación con los datos de que disponemos, tras la identificación preliminar de UFS con *echar* en el corpus *CEDEL2+CAES*, fue necesario, con la ayuda de corpus de referencia (*esTenTen* y *CORPES XXI*), diccionarios (*DLE*, *DPD* y *DAMER*) y juicios de los nativos, rastrear las ocurrencias erróneas y contarlas para compararlas con las correctas, teniendo en cuenta el nivel lingüístico alcanzado por cada alumno. Con el fin de aumentar la legibilidad de los resultados, los seis niveles del MCER se agruparon en tres bandas de competencia (A = elemental, B = intermedio, C = avanzado) y se calculó la tasa de errores (número de errores en la banda de competencia en comparación con el número de los errores totales) para cada banda. Los resultados que surgieron de los cálculos y operaciones descritos se representan visualmente en el Gráfico 1.

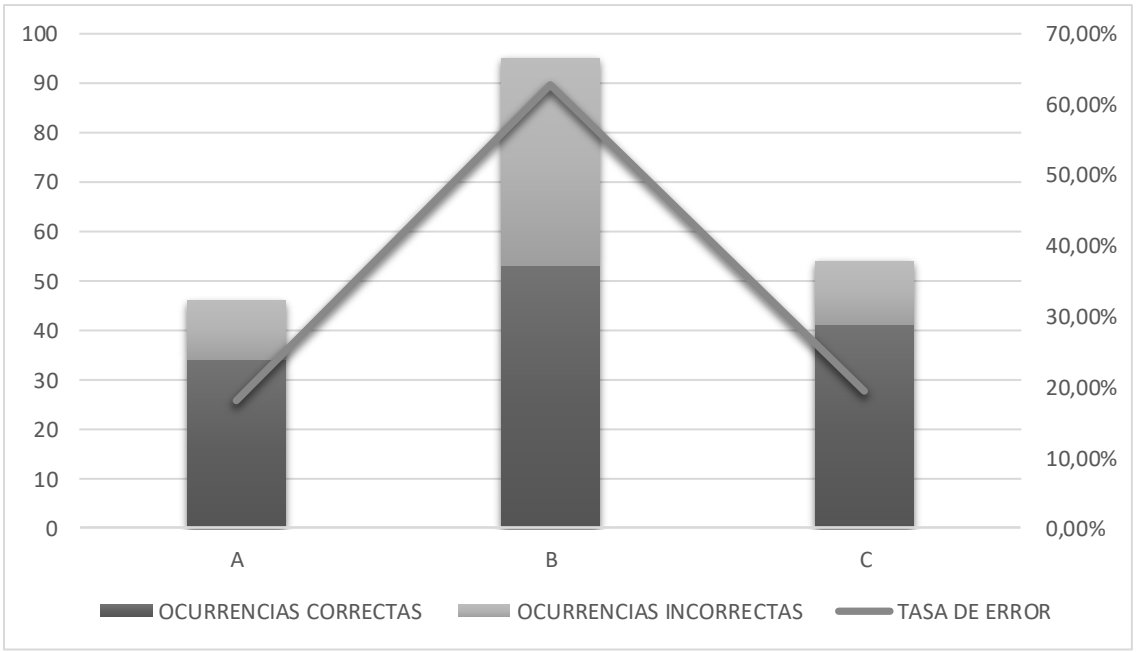

Gráfico 1
Bandas de competencia y errores.

Lo que se desprende del Gráfico 1 no es totalmente coherente con la primera hipótesis: si bien es cierto que a la franja elemental le corresponde la menor incidencia de UFS con *echar* (46 de 195 ocurrencias totales, con una tasa de error del 26,1%), en los niveles intermedios se observa un aumento de UFS con *echar,* pero también la mayor incidencia de errores en el corpus (95 ocurrencias y una tasa de error del 44,2%). Por último, de acuerdo con la hipótesis planteada, aunque se observa una menor incidencia de ocurrencias en C que en B, las ocurrencias de los aprendientes de nivel avanzado son globalmente más correctas (54 ocurrencias y una tasa de error del 24,1%).

La hipótesis, por tanto, sólo se confirma parcialmente: no se produce el aumento de usos esperado en la banda C y la tasa de errores muestra una tendencia no lineal, o sea no descendiente, sino con un pico de errores en la banda intermedia. Estos resultados pueden estar motivados por la altísima incidencia (129 casos, el 66% de las ocurrencias totales) de la locución *echar de menos*, rastreable sobre todo en la tarea "carta amigo" asociada al nivel B1 en el *CAES*: como es obvio, en una carta a un amigo es frecuente expresar sentimientos de nostalgia; como veremos en el objetivo siguiente, esto conlleva varios errores en el corpus considerado.

4.2. Objetivo 2: Análisis De Errores

La pregunta de investigación que guía el segundo objetivo se refiere a la conexión entre los tipos de UFS con *echar* y los distintos errores cometidos por los aprendientes: ¿cómo pueden relacionarse ambas taxonomías? La hipótesis formulada para responder a esta pregunta retoma el concepto de complejidad introducido en el marco teórico: es posible, por tanto, que las UFS estructuralmente más complejas por estar compuestas simplemente por más palabras (y, por tanto, tener un mayor margen de error) presenten más desviaciones de la norma.

Para categorizar las UFS con *echar*, se utilizó aquí la taxonomía establecida por Martín Salcedo en su estudio sobre la traducibilidad de la fraseología con *echar* en

portugués (2015). Véase la Tabla 4 para un recuento detallado de las UFS con *echar* según su categoría: dicha tabla no sólo muestra las ocurrencias correctas e incorrectas, sino también la tasa de error (en este caso relativa al número total de casos erróneos) y un ejemplo de estructura para cada categoría, con el fin de ofrecer una visión lo más completa posible de los datos analizados.

CATEGORÍAS DE UFS	OCURRENCIAS INCORRECTAS / OCURRENCIAS TOTALES	TASA DE ERROR	Ejemplos
Colocaciones con preposición	4 / 19	5,97%	*echarse en la cama*
Locuciones con preposición	47 / 129	70,15%	*echar de menos*
Colocaciones sin preposición	3 / 11	4,48%	*echar la culpa*
Locuciones sin preposición	13 / 36	19,40%	*echar una mano*

Tabla 4
Taxonomía de las UFS con *echar* en el corpus *CEDEL2+CAES*: errores y ejemplos.

Como puede observarse, si consideramos las construcciones estructuralmente (en este caso, sintácticamente) más complejas en su conjunto (Pallotti 2015), notamos que se caracterizan por una mayor tasa de error que las estructuras sin preposición, especialmente en el caso de las locuciones. Al mismo tiempo, considerando las locuciones en su conjunto, dado que se caracterizan por la idiomaticidad y la no composicionalidad (apartado 2.2), parece claro que son más difíciles de asimilar para los estudiantes de español como lengua extranjera. Un último dato bastante interesante es la presencia preponderante de locuciones con preposición que, además de ser las más difíciles, como hemos señalado, son también las más utilizadas por los aprendientes. En breve profundizaremos en este aspecto.

La segunda taxonomía que introduciremos en este apartado se refiere a los errores encontrados en el corpus *CEDEL2+CAES*. En la tabla 5 se reporta el cómputo de los errores, junto con la tasa de frecuencia del error con respecto al total de los errores y un ejemplo para cada categoría[7].

	1 CONSTRUCCIÓN INCORRECTA	2 ERROR MODIFICADOR	3 ERROR PREPOSICIONAL	4 ERROR OBJ.DIRECTO	5 UF INEXISTENTE	6 CONTEXTO INCORRECTO
NÚMERO DE ERRORES	33	15	15	5	2	2
TASA de frecuencia del error	46%	21%	21%	7%	3%	3%
EJEMPLOS	**[…] se echó un vistazo a su alrededor.*	**[…] te echo de menos mucho.*	**[…] han echado de perder.*	**Echo de menos a ti.*	**Se echó la vista y se puso a fumar.*	**[…] no te echará de menos tu libro.*

Tabla 5
Taxonomía de los errores y ejemplos.

[7] Además, es necesario distinguir entre errores y ocurrencias erróneas: algunas ocurrencias erróneas se caracterizan por la presencia de dos errores combinados, de modo que el número de errores es ligeramente superior al número de ocurrencias erróneas presentado anteriormente.

A continuación, se triangularon los datos de la Tabla 4 con los de la Tabla 5: más del 70% de los errores de tipo 1 (24 errores de un total de 33) se cometieron en estructuras complejas, es decir, las que contienen preposición, lo que corrobora la hipótesis inicial de que las UFS con preposición son las más críticas para los aprendientes. Por tanto, la suposición inicial también se ve confirmada por otras dos categorías de error: la alta frecuencia de estructuras que llevan preposición justifica toda la categoría de errores de tipo 2 y, por último, los errores de tipo 3 se cometen solo en la locución *echar de menos*. Las categorías de error 4, 5 y 6 resultan ser menos relevantes a nivel global: sin embargo, en un corpus pequeño como el considerado, desempeñan un papel bastante importante y nos permiten aumentar nuestro conocimiento sobre las áreas de dificultad de los aprendientes, para luego permitir el desarrollo de actividades didácticas *ad hoc*.

4.3. Objetivo 3: Riqueza Fraseológica Parcial

A continuación, se presenta la pregunta de investigación que motiva el tercer y último objetivo del presente estudio. Como hemos visto en la Tabla 3 que recoge los datos preliminares, los hablantes nativos de la muestra usan más UFS con *echar* que los aprendientes, pero ¿en qué medida y de qué manera difieren ambas categorías de hablantes en el uso de estas UFS? Se supone que, conforme a lo expuesto por Orol González y Alonso Ramos (2013) para las colocaciones, los aprendices no alcanzan la riqueza fraseológica de los hablantes nativos y, al mismo tiempo, se supone que su riqueza fraseológica guarda una relación de proporcionalidad directa con el dominio.

Utilizando los dos parámetros que hemos discutido en la sección sobre metodología, a saber, la variedad fraseológica y el número de errores, se calcula la riqueza fraseológica parcial en el corpus *CEDEL2+CAES* y en el subcorpus de control nativo de *CEDEL2*, utilizando las siguientes fórmulas:

VARIEDAD FRASEOLÓGICA (VF) = NÚMERO DE UFS LEMA[8] / NÚMERO TOTAL DE UFS
NÚMERO DE ERRORES (NE) = NÚMERO DE ERRORES / NÚMERO TOTAL DE UFS
RIQUEZA FRASEOLÓGICA PARCIAL (RFP) = VF – NE

	CEDEL2 NATIVOS	CEDEL2+CAES APRENDIENTES	A1	A2	B1	B2	C1	C2
VF	0,33	0,08	0,13	0,08	0,09	0,56	0,38	0,32
NE	0,04	0,37	0,5	0,24	0,48	0,56	0,27	0,21
RFP	0,29	-0,29	-0,37	-0,16	-0,39	0	0,11	0,11

Tabla 6
Cálculo de la riqueza fraseológica parcial.

La Tabla 6 – que recoge los datos estadísticos resultantes de los cálculos mencionados, divididos según el nivel lingüístico y el corpus al que se refieren – muestra que los aprendientes presentan en general niveles bajos de RFP, especialmente en los niveles iniciales en los que aparecen valores negativos (A1, A2, B1): sorprendentemente, el nivel en el que los alumnos alcanzan el valor más bajo es el nivel intermedio B1. Además, se

[8] La expresión *UFS lema* se refiere al recuento de los distintos tipos de UFS: p. ej., en el corpus considerado se encuentran distintas formas de la UF lema *echar de menos*, como *echo de menos*, *echaba de menos* y *eché de menos*; en el calculo de la variedad fraseológica se divide el número de UFS lema por el número total de UFS, que incluye las diferentes formas lingüísticas en las que se concretan las UFS lema.

observa un grado 0 de RFP en el nivel B2 y niveles positivos de RFP en los niveles C1 y C2, los que más se aproximan al valor de RFP presente en el subcorpus nativo.

En conjunto, podemos afirmar que la hipótesis relativa a este tercer objetivo ha quedado solo en parte corroborada: si bien es cierto que los aprendientes no alcanzan la RFP mostrada por los nativos, también lo es que la tendencia de este valor no asciende de forma constante y progresiva al aumentar el nivel de competencia lingüística y, por el contrario, presenta un descenso en el nivel intermedio B1; el aumento de la RFP que se registra en los niveles más altos, en cambio, era previsible. Esta conclusión, como veremos en el apartado 5, conlleva implicaciones didácticas considerables.

5. Conclusiones

A la luz de los resultados, es posible trazar un perfil fraseológico parcial de los aprendientes de español de la muestra analizada: en general, el nivel B1 parece ser el más crítico, tanto en términos de NE como de RFP. Además, el error más frecuente cometido por los aprendientes es la construcción incorrecta.

El tipo de UF más difícil y, al mismo tiempo, el más utilizado por los alumnos resulta ser la estructura con preposición, en concreto la locución *echar de menos*, adquirida precisamente en el nivel B1 según Penadés Martínez (2002). Estos resultados están en parte relacionados con el concepto de *opportunity of use* (Caines y Buttery 2017), ya que el emparejamiento del nivel B1 con la tarea "carta amigo" en el *CAES* tiene un efecto determinista en la selección de UFS por parte del aprendiente, que exterioriza sus emociones a través de dicha locución.

En los niveles superiores, el número de errores disminuye significativamente, lo que hace que la RFP aumente, demostrando que el aumento del dominio se corresponde con una mayor competencia fraseológica. En general, los aprendientes utilizan UFS con *echar*, pero no alcanzan el grado de variedad y corrección de los hablantes nativos, como concluyen Orol González y Alonso Ramos (2013) con respecto a las colocaciones.

El presente estudio, partiendo de la parcialidad que lo caracteriza al ser un análisis de UFS que contienen un solo verbo, es decir, *echar*, se propone como una investigación exploratoria de un método que también puede aplicarse a UFS que contengan otros verbos productivos desde el punto de vista fraseológico en la lengua española, como *tener*, *hacer*, *ser* y *estar*. Además, es una aproximación a las áreas de dificultad de los aprendices de español por lo que concierne a la competencia fraseológica, por lo que sería útil centrarse no sólo en los niveles intermedios y avanzados, sino también en las etapas más tempranas del aprendizaje de la lengua.

Entendiendo la fraseología como un facilitador en la memorización y uso de fragmentos lingüísticos preconfeccionados, el objetivo último de este estudio es subrayar la necesidad de facilitar la adquisición de estructuras fraseológicas: aunque la bibliografía cuenta con numerosos estudios válidos relativos a la enseñanza de UFS (Mendizábal de la Cruz, Sastre Ruano 2017; Peramos Soler *et al*. 2010; Timofeeva Timofeev 2013), es posible explotar los corpus de aprendientes para el diseño de actividades basadas en corpus que promuevan la competencia fraseológica en diferentes niveles. Según lo que hemos visto, en los niveles intermedios el énfasis debe ponerse en la corrección (ya que el nivel B1 puede considerarse una etapa crítica en el proceso de aprendizaje de UFS), mientras que en los niveles avanzados el énfasis debe ponerse en la variedad, para que los aprendientes puedan alcanzar un nivel de RF más alto. Aprovechando al máximo el

potencial de UCLEEv2, según esta perspectiva, es posible crear actividades didácticas a partir de un corpus de aprendientes, como se menciona en el apartado 2.1.

Por último, se considera oportuno en estudios futuros centrarse en producciones de aprendientes que comparten una misma L1[9] para evaluar el papel de la interferencia interlingüística en el cálculo global de errores. La lengua materna del alumno, sin embargo, no resulta ser el único metadato que se puede explotar; como hemos visto en el apartado 2.3, tanto *CAES* como *CEDEL2* ofrecen un abanico de metadatos con los que se podría llevar a cabo un análisis de errores más en profundidad vinculado a variables sociales o a las características propias de la tarea considerada.

Nota biográfica: Maria Annese es doctoranda en Lenguas, Literaturas y Culturas en Contacto en la Universidad 'G. d'Annunzio' de Chieti-Pescara, con un proyecto titulado "Análisis de la interlengua italiano-español en corpus de aprendices". Estudiante en la Universidad 'G. d'Annunzio' de Chieti-Pescara, en 2021 obtiene el Máster en Lenguas, Literaturas y Culturas Modernas con una tesis titulada "Análisis comparativo del lenguaje femenino y masculino en español: lengua e identidad de género", que ha dado lugar a dos publicaciones (Annese 2022 y Annese 2023). Desde abril de 2022 hasta febrero de 2023, es becaria de investigación en la Universidad de Pescara. Participa en el Proyecto de Interés Nacional DISBIOCOM (20227FWEZ5), financiado por el Ministerio de Educación italiano y la Unión Europea NextGenerationEU. Sus intereses de investigación se centran en los corpus de aprendices, la fraseología, el español como lengua extranjera y el análisis del discurso.

Dirección de correo electrónico del autor: mariaannese95@gmail.com

[9] Véase Bailini (2016), obra en la que se plantea el estado de la cuestión del estudio de la interlengua L1 italiano-L2 español y viceversa, y se presenta el análisis de dos corpus de aprendientes recopilados por la misma autora (CORESPI y CORITE), prestando particular atención al papel de la interferencia interlingüística entre las dos lenguas que se definen *afines*, por tener una distancia lingüística escasa.

Bibliografía

Asociación de Academias de la Lengua Española: Diccionario de americanismos (DAMER) [en línea], https://www.asale.org/damer/, [Consulta: 21/02/2024].

Bailini S. 2016, *La interlengua de lenguas afines. El español de los italianos, el italiano de los españoles*, Milano, LED.

Caines A. y Buttery P. 2017, *The effect of task and topic on opportunity of use in learner corpora*, en Flowerdew L. y Brezina V. (ed.), *Learner corpus research: New perspectives and applications*, Londres, Bloomsbury Publishing Academic, pp. 5–27.

Castillo Carballo M.A. 2002, *Conocimiento cultural en la adquisición de la L2: la fraseología*, en "El Español, Lengua del Mestizaje y la Interculturalidad, Actas del XIII Congreso Internacional de la ASELE", Murcia.

Corder S.P. 1967, *The Significance of Learners' Errors*, en "International Review of Applied Linguistics in Language Teaching" 5, pp. 161-170.

Corder S.P. 1981, *Error Analysis and Interlanguage*, Oxford, Oxford University Press.

Corpas Pastor G. 1996, *Manual De Fraseología Española*, Madrid, Gredos.

Dulay H., Burt M. y Krashen S. 1982, *Language Two*, Oxford, Oxford University Press.

Ellis N. 2012, *Formulaic language and second language acquisition: Zipf and the phrasal teddy bear*, en "Annual Review of Applied Linguistics" 32, pp. 17-44.

Ellis N., Simpson-Vlach R., Römer U, O'Donnell M.B. y Wulff S. 2015, *Learner corpora and formulaic language in second language acquisition research*, en Granger S., Gilquin G. y Meunier F., (ed.) *The Cambridge Handbook of Learner Corpus Research, Cambridge Handbooks in Language and Linguistics*, Cambridge, Cambridge University Press.

García Muruais M.T. 1998, *Propuestas para la enseñanza de unidades fraseológicas en la clase de E/LE*, en "El español como lengua extranjera. Del pasado al futuro: actas del VIII Congreso Internacional de ASELE", Alcalá de Henares.

Granger S. 2002, *A Bird's-Eye View of Learner Corpus Research*, en Granger S., Hung J. y Petch-Tyson S. (ed.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*, Amsterdam, John Benjiamins, pp. 3-33.

Granger S. 2012, *How to use foreign and second language learner corpora*, en Mackey A. y Gass S. (ed.) *Research Methods in Second Language Acquisition: A Practical Guide*, Oxford, Blackwell, pp. 7-29.

Granger S. 2015, *Contrastive interlanguage analysis: A reappraisal*, en "International Journal of Learner Corpus Research*"* 1, pp. 7-24.

Granger S. Gilquin G., Meunier F. 2015, *Introduction: learner corpus research – past, present and future*, in "The Cambridge Handbook of Learner Corpus Research", Cambridge, Cambridge University Press.

Granger S., Swallow H., Thewissen J. 2022, *The Louvain Error tagging Manual. Version 2.0*, en "CECL Papers 4", Louvain-la-Neuve, Centre for English Corpus Linguistics/Université catholique de Louvain.

Housen A., Kuiken F., Vedder I. 2012, *Complexity, accuracy and fluency: Definitions, measurement and research*, en "Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA", pp. 1-20.

Jezek E. 2005, *Lessico. Classi di parole, strutture, combinazioni*, Bologna, Il Mulino.

Kilgarriff A. y Renau I. 2013, *esTenTen, a vast web corpus of Peninsular and American Spanish*, en "Procedia-Social and Behavioral Sciences" 95, pp. 12-19.

Lozano C. 2022, *CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research*, en "Second Language Research", pp. 965-983.

Lozano C., Mendikoetxea A. 2013, *Learner corpora and Second Language Acquisition: The design and collection of CEDEL2*, en Díaz-Negrillo A., Ballier N. y Thompson P. (ed.), *Automatic Treatment and Analysis of Learner Corpus Data*, Amsterdam, John Benjamins, pp. 65-100.

Martín Salcedo J. 2015, *Échale ganas que te echamos una mano. Fraseología con el verbo echar*, en "V Congresso Nordestino de Professores de Espanhol, 2014, Teresina, Anais do V Congreso Nordestino de Profesores de Espanhol", Brasilia, Ministerio de Educación, Cultura y Deporte, pp. 170-176.

Mendizábal de la Cruz N. Y Sastre Ruano M.Á. 2017, *Problemas de las unidades fraseológicas verbales y su aplicación a la enseñanza del español como lengua extranjera*, en del Barrio de la Rosa F. (ed.), *Palabras Vocabulario Léxico: La lexicología aplicada a la didáctica y a la diacronía*, pp. 49-62.

Núñez Nogueroles E.E. 2019, *Pasado, presente y futuro de los corpus de aprendices de ELE, Una revisión bibliográfica*, en "ReiDoCrea - Monográfico sobre Perspectivas transnacionales en la enseñanza de lenguas" 8 [3], pp. 170-190.

Orol González A. y Alonso Ramos M. 2013, *A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish*, en "Procedia - Social and Behavioral Sciences" 95, pp. 563-570.
Pallotti G. 2009, *CAF: Defining, Refining and Differentiating Constructs*, en "Applied Linguistics" 30 [4], pp. 590–601.
Pallotti G. 2015, *A simple view of linguistic complexity*, en "Second Language Research" 31 [1], pp. 117-134.
Parodi G. 2015, *Corpus de aprendices de español (CAES)*, en "Journal of Spanish Language Teaching" 2 [2], pp. 194-200.
Penadés Martínez I. 2002, *Diccionario de locuciones verbales para la enseñanza del español*, Madrid, Arco/Libros.
Peramos Soler N., Leontaridi E. y Ruiz Morales M. 2009, *Las unidades fraseológicas del español: su enseñanza y adquisición en la clase de ELE*, en Barrio Barrio J.F. (coord.), *Actas de las Jornadas de Formación del Profesorado en la Enseñanza de L2/ELE y la Literatura Española Contemporánea*, Sofía, Ministerio de Educación de España y Universidad de Sofía "San Clemente de Ojrid", pp. 185-204.
Read J. 2000, *Assessing Vocabulary (Cambridge Language Assessment)*, Cambridge, Cambridge University Press.
Real Academia Española y Asociación de Academias de la Lengua Española: Diccionario panhispánico de dudas (DPD) [en línea], https://www.rae.es/dpd/, 2.ª edición (versión provisional). [Consulta: 21/02/2024].
Real Academia Española: Banco de datos (Corpes XXI) [en línea]. Corpus del Español del Siglo XXI (CORPES). http://www.rae.es, [Consulta: 21/02/2024].
Real Academia Española: Diccionario de la lengua española, 23.ª ed., [versión 23.7 en línea]. https://dle.rae.es, [Consulta: 21/02/2024].
Rigamonti D. 2012, *Problemas de lingüística de la adquisición y enseñanza del e/le a italófonos*, Milán, LED Edizioni Universitarie.
Rojo G. y Palacios Martínez I. 2016, *Learner Spanish on computer: The CAES 'Corpus de Aprendices de Español' project*, en Alonso Ramos M. (ed.), *Spanish learner corpus research: Current trends and future perspectives*, Amsterdam, John Benjamins, pp. 55–87.
Sánchez Rufat A. 2015a, *Análisis contrastivo de interlengua y corpus de aprendientes: precisiones metodológicas*, en "Pragmalingüística" 23, pp. 191-210.
Sánchez Rufat A. 2015b, *La investigación de corpus de aprendientes y el desarrollo de los estudios de la interlengua del español*, en "Language Design" 17, pp. 57-84.
Santos Gargallo I. 2004, *El análisis de errores en la interlengua del hablante no nativo*, en *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/ lengua extranjera (LE)*, Madrid, SGEL, pp. 391-410.
Selinker L. 1972, *Interlanguage*, en "IRAL - International Review of Applied Linguistics in Language Teaching" 10, pp. 209-231.
Sinclair J. 2005, *How to build a corpus*, en Wynne M. (ed.) *Developing Linguistic Corpora: A Guide to Good Practice*, Oxford, Oxbow books, pp. 79–83.
Thewissen J., Granger S. y Swallow H. 2023*, The UCLouvain Error Editor User guide - version 2.0*, en *CECL Papers 6*, Louvain-la-Neuve, Centre for English Corpus Linguistics/Université catholique de Louvain.
Timofeeva Timofeev, L. 2013, *La fraseología en la clase de lengua extranjera: ¿misión imposible?*, en "Onomázein" 28, dicembre, Santiago del Chile, Pontificia Universidad Católica de Chile, pp. 320-336.
Tracy-Ventura N., Paquot M. 2021, *The future of corpora in SLA*, en Tracy-Ventura N. and Paquot M., (ed.), *The Routledge handbook of second language acquisition and corpora*, Abingdon, Routledge.
Vázquez G. 1991, *Análisis de errores y aprendizaje de español/lengua extranjera*, Frankfurt, Peter Lang.

Anexo 1

Este apéndice contiene el *tagset* utilizado en UCLEEv2 para anotar los errores. Para utilizarlo, es necesario seguir las instrucciones del manual del usuario del Error Editor (Thewissen 2023). El orden de los errores corresponde al presentado en el artículo:

1)CI indica una construcción incorrecta con respecto a la morfosintaxis;
2)PMOD indica la posición incorrecta del modificador;
3)EP es la etiqueta utilizada para subrayar los errores preposicionales, categoría que incluye la falta de una preposición necesaria (PF), la presencia de una preposición innecesaria (PI) y la elección errónea de la preposición (PE);
4)OD señala un error con respecto al objeto directo;
5)UFI indica una UF inexistente;
6)CONT designa una UF empleada en un contexto incorrecto desde un punto de vista pragmático.

```
[CI
  CI
]
[PMOD
  PMOD
]
[EP
  PF
  PI
  PE
]
[OD
  OD
]
[UFI
  UFI
]
[CONT
  CONT
]
```