ETIQUETAJE DE EXPRESIONES MULTIPALABRA EN ENSAYOS ESCRITOS POR NATIVOS Y NO NATIVOS DE ESPAÑOL EN UN CURSO DE DESARROLLO DE GRAMÁTICA Y COMPOSICIÓN

Miguel Da Corte

Universidade do Algarve, Faculdade de Ciências Humanas e Sociais,
Campus de Gambelas, P-8005-139 Faro, Portugal
A74072@ualg.pt

Jorge Baptista

INESC-ID Lisboa, HLT Lab, Rua Alves Redol, 9, P-1000-029 Lisboa, Portugal
jbaptis@ualg.pt

Multiword expression tagging of Spanish native and non-native speakers' written essays in a grammar and composition developmental course

Abstract: The literature on second language learning posits that there are significant differences between the use of multiword expressions (MWE) by native speakers (NS) and non-native speakers (NNS). Furthermore, it considers that levels of language proficiency can be estimated on the basis of the use of these expressions. This paper analyses the written production from a corpus of essays written by native (16 essays, 5839 words) and non-native Spanish speakers (25 essays, 7767 words) enrolled in a course focused on the development of orthographic, grammatical, lexical, semantic, and discursive skills in Spanish. This is a required course for students pursuing a certification in Translating or Interpreting (Spanish/English) in the educational setting where the study took place. The corpus was manually tagged by two linguists. The classification scheme used was inspired by other schemes found in the literature and built for similar purposes. The results show that, in general, the distribution of MWE types found in the NS and NNS partition of the corpus was not very different (Pearson correlation: 0.894). However, interesting differences were found between the categories of verbal idioms and noun constructions. Though the corpus is too small for more significant conclusions to be drawn, it is possible

to point out that different types of MWE are unevenly distributed among the native speakers' and non-native learners' written production material, and some categories may be a clearer indicator of near-native-speaker proficiency.

Keywords: multiword expressions; language proficiency; classification level; machine-learning models; developmental education courses (in Spanish)

Resumen: La literatura sobre el aprendizaje de una segunda lengua postula que existen diferencias significativas entre el uso de expresiones multipalabra (EMP) por parte de hablantes nativos (HN) y no nativos (HNN). Además, considera que los niveles de competencia lingüística pueden estimarse a partir del uso de dichas expresiones. En este trabajo se analiza la producción escrita de un corpus de ensayos escritos por hablantes nativos (16 ensayos, 5.839 palabras) y no nativos de español (25 ensayos, 7.767 palabras) matriculados en un curso centrado en el desarrollo de las habilidades ortográficas, gramaticales, léxicas, semánticas y discursivas en español. Este es un curso de matriculación obligatoria para los estudiantes que aspiran a obtener el título de traductor o intérprete (español/inglés) en el centro educativo donde se realizó el estudio. Dos expertos lingüistas etiquetaron de forma manual el corpus del estudio. El esquema de clasificación utilizado se inspiró en otros esquemas encontrados en la literatura y construido con fines similares. Los resultados mostraron que no se encontraron mayores diferencias (correlación de Pearson: 0,894) en la distribución de los tipos de EMP en el análisis del corpus de HN y HNN. Sin embargo, se observaron diferencias interesantes en algunas categorías, concretamente entre las expresiones fraseológicas verbales y los sustantivos comunes compuestos. Aunque el corpus es pequeño para llegar a conclusiones más relevantes, cabe destacar que los diferentes tipos de EMP se distribuyen de forma desigual en los textos escritos por los hablantes nativos y no nativos y que algunas categorías son un indicador más claro de un nivel de competencia más cercano al nivel de dominio del idioma materno.

Palabras clave: expresiones multipalabra; competencia lingüística; nivel de clasificación; modelos de aprendizaje automático; cursos para el desarrollo de habilidades (en español)

1. Introducción

El uso de expresiones multipalabra (EMP) ha sido motivo de continuas investigaciones debido a su idiosincrasia, complejidad y al enfoque usado para su estudio (García-Page 2008; Corpas Pastor 2017; Pasquer *et al*. 2020). A pesar de que existen diferentes criterios para la evaluación, señalización y estandarización de dichas expresiones (Farahmand *et al*. 2015), los rasgos de coocurrencia e incidencia se presentan como criterios básicos para el esclarecimiento de la función y el valor que las EMP añaden a la producción de textos escritos (Savary *et al*. 2019).

En un análisis para predecir el nivel de competencia lingüística[1] en inglés de estudiantes franceses, y tomando como referencia el Marco Común Europeo de Referencia para las Lenguas[2] (MCER), Arnold *et al*. (2018) examinaron un modelo predictivo en el que el nivel de competencia lingüística se establecía como una función dependiente del grado de complejidad de diferentes estructuras lingüísticas. Todos los datos pertenecientes a los textos no procesados fueron extraídos de la base de datos

[1] Se adopta el término *competencia lingüística* como traducción al español del concepto *proficiency*, en inglés.

[2] <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/>. Todas las páginas web en este documento se consultaron por última vez el 7 de marzo de 2023.

EFCAMDAT[3] (*EF Cambridge Open Language Database*). Se utilizó un total de 41.626 textos, correspondientes a 128 temas de redacción, de una población de 7.695 estudiantes franceses. Basándose en nueve 9 indicadores de complejidad lingüística (por ejemplo, número de palabras, número de frases, frases verbales, frases coordinadas y frases nominales complejas) y utilizando la herramienta *LE Syntactic Complexity Analyzer*[4] (L2SCA), los resultados revelaron que el número de *tokens* y el tipo de palabra desempeñan un papel fundamental en el cálculo del nivel de competencia lingüística. Concretamente, «la habilidad de generar expresiones y frases complejas y usar correctamente estructuras avanzadas, tales como oraciones relativas sin pronombre relativo, son indicadores de estudiantes con un nivel avanzado en la lengua meta» (Arnold *et al*. 2018: 4; traducción propia).

La producción de estas expresiones y frases complejas ha sido objeto de interés en el campo de Procesamiento del Lenguaje Natural (PLN) (en inglés, *Natural Language Processing*). Parte del interés reside en: (i) la relevancia de estas expresiones en la clasificación (y calificación) automática de ensayos escritos «con otras características clásicas como longitud de las EMP, recurso de clasificación (*graded resource*), vecinos ortográficos (*orthographical neighbours*), partes de la estructura gramatical de una oración, morfología, relaciones de dependencia, tiempo verbal, desarrollo del lenguaje y coherencia» (Wilkens *et al*. 2022: 62; traducción propia); (ii) el efecto que ciertas tareas o trabajos escritos tienen en la precisión de las estructuras lingüísticas usadas por aprendices de una segunda lengua y para lo cual se necesitan herramientas PLN para asegurar la exactitud de los resultados (Alexopoulou *et al*. 2017); (iii) la pedagogía que se utiliza en la enseñanza de la escritura de una segunda lengua y cómo esta influye en los índices de desarrollo de competencia lingüística y dependencia en EMP (El-Dakhs *et al*. 2022; Esfandiari & Mohammad 2022).

El uso de expresiones y frases complejas, en su mayoría, se ve condicionado por el tema de redacción o las instrucciones proveídas, lo que la literatura advierte debe monitorearse para así evitar una distribución sesgada de dichas expresiones. Basándose en el número promedio de palabras por frase, la longitud promedio de estas y la relación de frases subordinadas con respecto al total de frases utilizadas, Alexopoulou *et al*. (2017) sostienen que la longitud de las frases aumenta con el nivel de competencia y que la complejidad lingüística se ve afectada por la complejidad de la tarea a desempeñar.

A partir de estos rasgos lingüísticos, el objetivo que nos proponemos consiste en experimentar con técnicas de PLN para realizar un análisis comparativo de textos escritos por hablantes nativos y no nativos de español. Este análisis permitirá estimar y confirmar el nivel de competencia lingüística de los hablantes no nativos, teniendo en cuenta la diversidad de EMP usadas en sus producciones escritas. Según nuestros conocimientos, ningún estudio sistemático sobre el uso de la EMP por parte de hablantes nativos se ha correlacionado con una evaluación sobre el nivel de competencia lingüística de hablantes no nativos a nivel académico. Esta investigación

[3] <https://philarion.mml.cam.ac.uk/>.
[4] <https://sites.psu.edu/xxl13/l2sca/>.

representa un primer paso hacia ese objetivo. Al mismo tiempo, procuramos identificar los rasgos lingüísticos que son más predictivos del nivel de competencia escrita de los hablantes no nativos versus los hablantes nativos y si ciertas EMP –tipos y *tokens* con combinaciones de dos o más palabras– son usadas por uno de los grupos con mayor frecuencia.

La necesidad concreta de contrastar y analizar el uso de EMP entre hablantes nativos y no nativos surge del hecho de que el dominio de estas expresiones, ya catalogado como pertinente para los niveles superiores de competencia según el MCER (Consejo de Europa 2002), se asocia, según estudios previos (Dahunsi & Ewata 2022; Hinkle 2023), con altos niveles de competencia lingüística. Apoyando esta afirmación, en un análisis sobre el uso de EMP (Da Corte & Baptista 2022a), realizado en el mismo contexto que aquí se presenta, se analizó la producción escrita de hablantes nativos de inglés (L1) que participaron en un curso de dos niveles (nivel 1 y 2) de desarrollo de gramática y composición como requisito indispensable para ser admitidos en un programa académico. Los experimentos realizados compararon el efecto sobre un conjunto de clasificadores de un corpus sin anotaciones y en una versión del mismo corpus anotada con EMP. Los resultados mostraron que la identificación de las EMP, como características léxicas, mejoró la precisión de la clasificación de textos en un 8,1 %, en cuanto a los niveles (nivel 1 y 2) del curso de gramática y composición, por encima del experimento base (sin anotación) y consecuentemente la ubicación en el nivel correspondiente en dicho curso. Ya que el contexto es el mismo y la expectativa general se ajusta al marco teórico de este estudio, se seleccionó como indicador principal de este estudio las EMP.

En cuanto a la organización de este documento, primero, se muestra, en la sección de métodos, una descripción detallada del contexto institucional y los participantes de este estudio (Sección 2.1; 2.2). Se incluye también información sobre la constitución del corpus (2.3), así como los tipos de EMP seleccionadas para el etiquetaje del corpus y el esquema de anotación (2.4). Los experimentos con herramientas de aprendizaje automático se presentan de manera resumida (2.5) con los resultados y algunas reflexiones (3). Las conclusiones y trabajo futuro se presentan en la última sección (4).

2. Métodos

En esta sección se incluye: el contexto institucional, los participantes con el nivel de competencia lingüística correspondiente, el corpus utilizado, los tipos de EMP y esquema de anotación empleados y los respectivos experimentos con herramientas de aprendizaje automático.

2.1. Contexto institucional y participantes

Un total de veintidós estudiantes de un instituto de educación superior (Oklahoma, Estados Unidos) participaron en este estudio. Estos estudiantes estaban matriculados en un programa (bilingüe) de entrenamiento intensivo de traducción e interpretación (inglés-español; español-inglés). De estos veintidós estudiantes, quince manifestaron

tener el inglés como lengua materna y siete el español, siendo esta la lengua que usan con más frecuencia.

Ambos grupos participaron en un curso de dieciséis semanas de español como segunda lengua que tenía como objetivo principal el desarrollo de aspectos ortográficos, gramaticales, semánticos y discursivos. Este curso es denominado Estudio Intermedio de Gramática y Composición en Español y se imparte completamente en español, la lengua meta. Para los hablantes nativos de español, la participación en este curso les permite afianzar el conocimiento en cuanto a su lengua materna o simplemente confirmar sus niveles de competencia. Este curso es uno de los requisitos en el plan de estudios de Traducción e Interpretación, además de ser un elemento indispensable para ser técnico superior (primer paso en la obtención de una licenciatura). Para los hablantes no nativos de español, este curso permite corregir y desarrollar algunos patrones léxico-sintácticos y discursivos que llevan a un aumento del nivel de competencia lingüística en una segunda lengua. Cabe destacar que la participación en este curso, para los hablantes no nativos, es posible después de completar satisfactoriamente tres semestres de español – Niveles 1 al 3.

2.2. Nivel de competencia de los participantes

Según la tabla de convalidaciones del MCER (Consejo de Europa 2002) y el Consejo Americano para la Enseñanza de Idiomas Extranjeros[5] (*ACTFL*, por sus siglas en inglés), en relación con la escala de competencia lingüística, los estudiantes no nativos muestran un nivel de competencia equivalente al B1/B2. Este nivel de competencia se estima empleando dos sistemas de evaluación: uno formativo y otro sumativo. La evaluación formativa incluye la redacción de un párrafo sobre un tema en particular y un ejercicio de comprensión oral, con un fragmento de texto pregrabado, en la que los participantes responden a 5 preguntas de selección múltiple o verdadero/falso. La evaluación formativa se aplica al final de cada lección (6 en total), en cada nivel. La evaluación sumativa consiste en una prueba oral, con preguntas y respuestas, usando ciertos temas de discusión que únicamente son compartidos en el momento de la prueba (que se aplica solo al final del semestre). La escala de evaluación oral se enfoca en cuatro aspectos principales: número de funciones lingüísticas, contexto y contenido, precisión y tipo de discurso (ACTFL 2016); (Alpine Testing Solutions 2020).

Como parte del programa de estudio, los hablantes nativos y los no nativos demuestran y desarrollan sus habilidades de producción escrita según cuatro enfoques discursivos: descripción, narración, información y opinión. Cada ensayo tiene un tema abstracto en particular, por ejemplo: (i) relaciones personales en la sociedad global de hoy; (ii) experiencias personales en ciudades grandes, cosmopolitas en comparación con ciudades más pequeñas y poco desarrolladas; (iii) impacto de la tecnología en la vida cotidiana; y (iv) la convivencia en familia. Estos temas, indicados aquí a modo de ejemplo, fueron propuestos a los participantes y sobre ellos se realizaron los ensayos que constituyen el corpus objeto de este estudio. Las producciones escritas de ambos grupos se caracterizan por el uso y dominio de diferentes tiempos verbales, así como por otros rasgos lingüísticos de interés, particularmente el uso de EMP.

2.3. Constitución del corpus

El corpus utilizado para este análisis tiene un total de 81.466 caracteres que provienen de la colección de textos escritos producidos por los hablantes nativos y no nativos de español en el curso de gramática y composición. Se obtuvo una muestra de dos ensayos por estudiante durante dos períodos claves en el semestre en el que se desarrolló el curso: uno al principio del semestre (antes de recibir formación sobre contenidos gramaticales y ortográficos) y el segunda después de cinco semanas. En cuanto al contenido de las muestras, los ensayos de los hablantes nativos contienen aproximadamente 2.140 caracteres, con una desviación media de 1.203; en comparación, los ensayos de los no nativos contienen un 9 % menos en el promedio de caracteres usados (1.889) y una desviación media de 964. El tamaño de los ensayos escritos por parte de los hablantes nativos oscila entre los 383 y 4.711 caracteres, mientras que el número de caracteres en los ensayos de los hablantes no nativos oscila entre 691 y 4.647.

Estas cifras, que se presentan de manera resumida en la Tabla 1, demuestran que el tamaño de los ensayos de los hablantes nativos y no nativos es comparable, pero confirman la necesidad de tener un corpus equilibrado en cuanto al número de caracteres por unidad de muestreo, por nivel y temas presentados en los ensayos.

Indicador	HN	HNN	Corpus
Número de caracteres (mínimo)	383	691	383
Número de caracteres (máximo)	4.711	4.647	4.711
Número total de caracteres	**34.247**	**47.219**	**81.466**
Media	2.140	1.889	1.987
Desviación media	1.203	964	1.056

Tabla 1. Número de caracteres, media y desviación media de los ensayos escritos por los hablantes nativos (HN) y hablantes no nativos (HNN)

Dos lingüistas, con experiencia en tareas de anotación de corpus y dominio en los conceptos de EMP utilizados en este estudio, etiquetaron manualmente el corpus. Dado el pequeño tamaño del corpus, la anotación se llevó a cabo de forma independiente y las discrepancias existentes se discutieron para llegar a una delimitación y clasificación consensuadas.[5] Los elementos de cada una de las EMP se enlazaron con un carácter de subrayado o barra baja ('_'), para así ser considerados y tratados como una sola palabra gráfica (*token*, en inglés) en la herramienta de aprendizaje automático, Orange[6]. Esta herramienta se utiliza en experimentos de clasificación automática y en procesos de minería de datos. Las etiquetas siguen los lineamientos diseñados y aplicados con anterioridad a un corpus en inglés de estudiantes anglohablantes que desarrollan sus habilidades en su lengua materna (Da Corte & Baptista 2022a).

[5] Dadas las condiciones descritas, el cálculo del acuerdo entre anotadores no se llevó a cabo.
[6] <https://orangedatamining.com/>.

Las categorías de EMP (con sus respectivos códigos) delimitadas y empleadas en el etiquetaje manual del corpus fueron las siguientes: adverbios (ADV); locuciones conjuntivas (CONJ); sustantivos comunes compuestos (N); entidades nombradas (EN); preposiciones (PREP); pronombres (PRON); construcciones con verbo soporte (SVC); y, por último, expresiones fraseológicas verbales (VIDIOM).

2.4. Tipos de EMP y esquema de anotación

Las etiquetas aplicadas al corpus de los hablantes nativos y no nativos, en preparación para el análisis sobre la incidencia de EMP, se agrupan bajo la categoría de patrones lexicales y sintácticos. Estos patrones contribuyen con la construcción de frases y secuencias de palabras que, al combinarse con otras frases o palabras, definen o moldean el nivel discursivo del escritor y, por consiguiente, su nivel de competencia. El uso e incidencia de las EMP contribuyen a un análisis más holístico de ensayos escritos (Siyanova-Chanturia & Spina 2020), así como también resaltan, con mayor precisión, el desarrollo de las habilidades escritas en cuanto al nivel de competencia.

Debido a la variedad de las EMP, se utilizó como marco referencial para la clasificación de dichas expresiones una combinación de criterios formales, sintácticos y semánticos ya establecidos por otros autores como: Laporte (2018), quien presenta sugerencias en cuanto a la clasificación de construcciones con verbo soporte; Pasquer *et al*. (2020), quienes mencionan las expresiones fraseológicas verbales y describen su variabilidad; Nam & Park (2020) y Kochmar *et al*. (2020), quienes: (i) incluyen preposiciones, entidades nombradas, locuciones conjuntivas y sustantivos comunes compuestos en su lista de tipos de EMP; (ii) presentan sugerencias en cuanto a la construcción de modelos de aprendizaje automático para la clasificación de textos (por nivel), teniendo en cuenta la proporción e incidencia de las EMP que estos modelos reconocen en las fases de entrenamiento y prueba.

Los tipos de EMP, con sus respectivas definiciones, etiquetas, enlaces ('_') y algunos ejemplos[7] (incluyendo EMP que son parcialmente fijas), se presentan a continuación:

- Adverbio (ADV): EMP que funciona como adverbio simple. Ejemplos de adverbios empleados por los hablantes nativos incluyen: *casi_siempre*; *con_más_gusto_que_nunca*; *con_el_pasar_del_tiempo*; *al_contrario*; *en_solo_unas_palabras*; *en_la_actualidad*; *desde_mi_perspectiva*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_el_pasado*; *por_primera_vez*; *en_cualquier_lugar*; *por_supuesto*; *en_el_mundo_de_hoy*; *cara_a_cara*; *a_menudo*; *con_mucho_gusto*.
- Locución conjuntiva (CONJ): EMP usada como conjunción enlazada con un verbo. Ejemplos de conjunciones empleadas por los hablantes nativos incluyen: *a_medida_que*; *si_no*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_vez_de*; *antes_de_que*; *debido_a*; *así_que; como_si*.
- Sustantivo común compuesto (N): EMP usada como sustantivo. Ejemplos de sustantivos comunes empleados por los hablantes nativos incluyen: *roles_de_género*;

[7] La lista completa de EMP sin registros duplicados puede consultarse en Da Corte & Baptista (2022b).

nuevas_generaciones; *redes_sociales*; *señal_de_afecto*; *salud_mental*; *pista_de_hielo*; *pensamiento_crítico*; *lenguas_nativas*; *programas_de_noticias*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *cultura_de_la_soltería*; *relaciones_personales*; *habilidades_sociales*; *tiempo_libre*; *riesgo_financiero*; *palabras_sabias*; *almas_gemelas*; *casas_embrujadas*; *medios_de_interacción_social*.

- Entidad nombrada (EN): EMP que designa una entidad extralingüística y hace referencia a personas, organizaciones, lugares y eventos (Erdmann *et al.* 2019). Ejemplos de entidades nombradas empleadas por los hablantes nativos incluyen: *Guglielmo_Marconi* (persona); *Cristo_Rey* (lugar); *Pew_Research_Center* (organización); *Tulsa_Race_Massacre* (evento). En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *Tacos_Don_Francisco* (organización); *Acción_de_Gracias* (evento); *Nueva_York* (lugar).
- Preposición (PREP): EMP que funciona como una preposición enlazada con un sustantivo. Ejemplos de preposiciones empleadas por los hablantes nativos incluyen: *a_diferencia_de*; *por_culpa_de*: *en_frente_de*; *a_través_de*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *en_lo_que_respecta_a*; *en_torno_a*; *al_otro_lado_de*; *en_lugar_de*; *después_de*.
- Pronombre (PRON): EMP que funciona como pronombre. Entre los ejemplos de los pronombres empleados por los hablantes nativos y los no nativos, destacan los siguientes: *todo_el_mundo* (indefinido); *el_uno_al_otro* (complemento de eco); *sí_mismas*; *entre_sí* (recíprocos).
- Construcción con verbo soporte (SVC) o verbos leves (en inglés, *light verb constructions*): EMP que resulta de la combinación de un verbo con un sustantivo actuando como un predicado semántico, de la misma manera que un verbo simple o adjetivo, independientemente de la existencia de una nominalización (sustantivo-adjetivo; sustantivo-verbo). Este sustantivo predicativo selecciona argumentos y determina las propiedades sintácticas (estructurales) y semánticas distribucionales de sus oraciones base, así como las transformaciones que permite. En este trabajo, se siguió la perspectiva del Léxico-Gramática (Gross 1996) presentada recientemente por Fotopoulou *et al*. (2021) y Baptista *et al*. (2022) para la definición de las construcciones con verbo soporte. Ejemplos de construcciones con verbo soporte empleadas por los hablantes nativos incluyen: *juega_un_papel*; *tener_cuidado*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *asumir_el_compromiso*; *tiene_celos*; *dar_apoyo*; *hacer_diligencias*; *ofrecer_paz*; *guardo_gratos_recuerdos*; *pasar_#_tiempo*; *harás_conexiones*.[8]
- Expresión fraseológica verbal, frecuentemente idiomática (VIDIOM): EMP que resulta de la combinación de un verbo con por lo menos un argumento (sujeto; objeto) y que juntos tienen fuertes restricciones combinatorias. Estas expresiones suelen ser idiomáticas, es decir, su significado global no es composicional. Ejemplos de expresiones fraseológicas verbales empleadas por los hablantes

[8] Nótese que las EMP con verbo soporte se reproducen como aparecen en el corpus y no han sido lematizadas (en la mayoría de los casos, el infinitivo se destaca). Esta misma observación se aplica a las expresiones fraseológicas verbales que se detallan a continuación.

nativos incluyen: *irse_de_la_casa_de_sus_padres*; *navegar_la_sociedad_de_hoy*; *dieron_las_herramientas*; *aprender_de_los_errores_que_cometemos*; *dar_lo_mejor_de_sí*; *no_todo_es_color_de_rosa*. En cuanto a los ejemplos de los hablantes no nativos, destacan los siguientes: *caminando_de_puntillas*; *viven_en_el_momento*; *pasar_un_buen_momento*; *ponerse_pesado*; *elegido_caminos_diferentes*; *viene_a_la_mente*; *interponen_en_el_camino*.

Nótese que en algunos casos la EMP es solo parcialmente fija. Este es el caso de muchas EMP SVC en el que el determinante puede variar libremente (*tener_ mucho/poco_cuidado_de_algo*) o de EMP VIDIOM en las que no solo el sujeto, sino también la distribución del complemento es libre en algunas instancias (algo *viene_a_la_mente* de alguien).

En la Tabla 2, se muestra el total de EMP (y los tipos de EMP) identificadas en el corpus con la distribución de frecuencia normalizada por mil palabras.

EMP	**HN**	**Frec. mil**	**HNN**	**Frec. mil**
ADV	107	18,325	101	13,003
CONJ	1	171	9	1,159
N	102	17,469	191	24,591
NE	33	5,652	33	4,249
PREP	7	1,199	21	2,704
PRON	2	343	13	1,674
SVC	43	7,364	70	9,012
VIDIOM	27	4,624	19	2,446
Total	322	55,146	457	58,839
Número total de palabras	5.839		7.767	

Tabla 2. Distribución y frecuencia normalizada por mil palabras (Frec. mil) de los tipos de expresiones multipalabra (EMP) por categorías y grupo de participantes: hablantes nativos (HN) y hablantes no nativos (HNN)

Un total de 779 EMP fueron identificadas en el corpus de textos escritos por parte de los hablantes nativos y los no nativos (número total de palabras: 13.606). En la tabla presentada arriba se observa que los hablantes nativos contribuyen con el 41 % de las EMP identificadas mientras que los no nativos contribuyen con el 59 %. Según la incidencia de EMP y su distribución de frecuencia normalizada por mil palabras (55,146), los hablantes nativos presentan un mayor uso de adverbios (ADV, 107; 18,325), sustantivos comunes compuestos (N, 102; 17,469), construcciones con verbo soporte (SVC, 43; 7,364) y expresiones fraseológicas verbales (VIDIOM, 27; 4,624). En el corpus de los hablantes no nativos (Frec. mil: 58,839), se observan con mayor incidencia los adverbios (ADV, 101; 13,003), sustantivos comunes compuestos (N, 191; 24,591) seguido por construcciones con verbo (SVC, 70; 9,012).

Dos muestras de textos escritos acompañan la Figura 1, para resaltar algunas de las EMP (enlazadas y debidamente etiquetadas) ya explicadas:

Muestra de ensayo escrito por un Hablante Nativo (HN): Lo que me hace preguntarme, ¿realmente estamos listos para ser padres? O incluso ¿qué es el ser un buen padre? *Desde_mi_perspectiva/ADV*, todo padre de familia ha *dado_lo_mejor_de_sí/VIDIOM* pero aun así han afectado a sus hijos a través de estas creencias heredadas. Un claro ejemplo es la creencia de que el padre no debe llorar o mostrar alguna *señal_de_afecto/N*, causando una herida de abandono o creando dureza en un niño. Es por eso por lo que cuando veo jóvenes parejas que construyen familias desde una edad muytemprana, lo primero que me *viene_a_la_mente/VIDIOM* es "aún no estamos listos, ¿porque queremos correr?"

Muestra de ensayo escrito por un Hablante No Nativo (HNN): *Por_lo_tanto/ADV*, aprovechar las características de los lugares donde vive la gente puede ser una gran manera de evitar la dependencia en la tecnología para conocer gente nueva. Cuando se pueden utilizar otros *medios_de_interacción_social/N*, la dinámica de las interacciones de los pueblos cambia. Las personas pueden comportarse de forma más natural entre sí porque no hay *pantallas_de_computadora/N* ni aplicaciones que se *interponen_en_el_camino/VIDIOM*. Cuando tienes una conversación *por_Internet/ADV* no estás cada vez más *cerca_de/PREP* la otra persona.
Por_eso/ADV, te recomiendo explorar el rascacielos que siempre habías pensado que era hermoso con alguien, y *harás_conexiones/SVC* reales con la gente.

Figura 1. Muestras balanceadas de ensayos escritos por un hablante nativo (HN) y un hablante no nativo (HNN).

Ya presentada la descripción del corpus, la selección de los tipos de EMP y el esquema de anotación seleccionado, discutido y aceptado por ambos lingüistas, se procede a explicar los experimentos realizados con la herramienta de minería de datos Orange.

2.5. Experimentos de aprendizaje automático para la clasificación de textos

Para determinar el nivel de competencia de los hablantes nativos y los no nativos nos centramos en un problema de clasificación con dos niveles. En este trabajo, se adoptó un enfoque en herramientas de aprendizaje automático para investigar el impacto de EMP en el proceso de clasificación, usando como datos la información del corpus descrito en la Sección 2.3.

Dado que la composición de los ensayos de ambos grupos, en cuanto al número de caracteres, varía ampliamente, de 383 a 4.711 caracteres, con una media de 1.987 caracteres y una desviación media de 1.056, según datos presentados en la Tabla 1, la preparación de datos consistió en la división de cada ensayo en varios segmentos de tamaño similar, dejando intactas las frases completas. Los segmentos tienen entre 600 y 700 caracteres, con un promedio de 636 caracteres y un máximo de 985 caracteres.

Para los experimentos de aprendizaje automático, usamos la herramienta de aprendizaje automática Orange (Demšar *et al.* 2013). Este sistema es completo, fácil de usar y contiene los elementos básicos requeridos para el preprocesamiento y gestión de datos, así como también los algoritmos de aprendizaje utilizados más comúnmente, incluyendo uno de redes neuronales (*Neural Network;* NN por sus siglas en inglés). Orange dispone de varias herramientas de visualización de datos que permiten un análisis más detallado en cuanto al impacto de las EMP en la clasificación de textos de estudiantes nativos y no nativos. El flujo de trabajo básico adoptado para cada uno de los experimentos se muestra a continuación en la Figura 2.

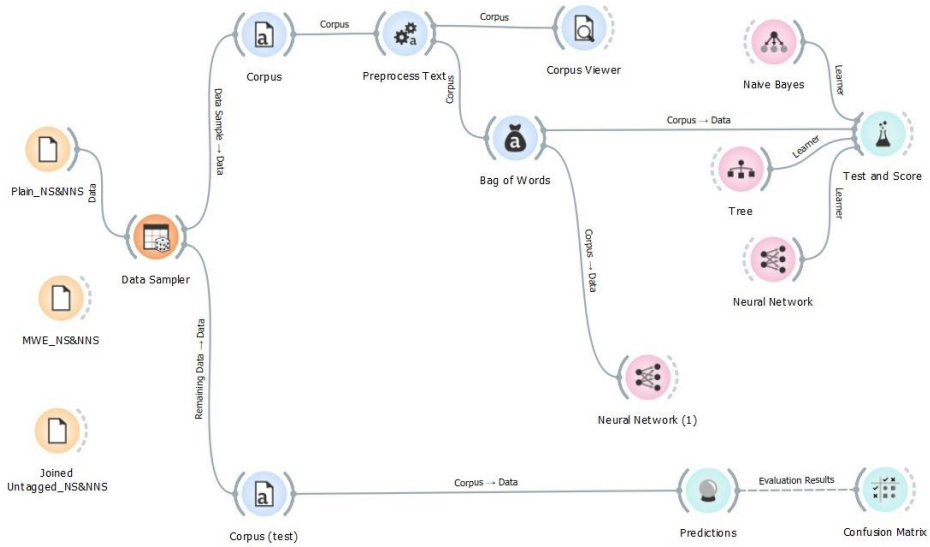

Figura 2. Configuración del flujo de trabajo en Orange para las fases de entrenamiento y prueba del modelo experimental

La Figura 2 muestra (a la izquierda) las tres versiones del corpus utilizadas en este estudio: (i) sin etiquetaje; (ii) con etiquetaje y enlace de las EMP como un solo *token* con el símbolo de barra baja ('_'); (iii) sin etiquetaje, pero con enlace de la EMP y que corresponden a los tres experimentos básicos explicados a continuación. Le sigue un *widget* de muestreo de datos (*Data Sampler*, en inglés, tal como lo muestra la Figura 2) que divide aleatoriamente el corpus en dos particiones: 75 % para la fase de entrenamiento y 25 % para la fase de prueba.

Los tres experimentos efectuados inicialmente con el fin de estimar, de manera general, el efecto de las EMP en el criterio de clasificación por nivel de los hablantes nativos y los no nativos fueron:

- **Experimento 1**, con el corpus de ensayos escritos por los hablantes nativos y los no nativos, sin etiquetaje ni enlace de las EMP.
- **Experimento 2**, con el corpus de ensayos con etiquetaje y enlace de todas las EMP (779). Con base en este diseño, se efectuaron experimentos adicionales que

incluyeron la eliminación, por separado, de los tipos EMP y otros con la eliminación de dos tipos de EMP, con sus respectivos enlaces. Estos experimentos (Exp 4 al Exp 18) se detallan en la Tabla 3.

- **Experimento 3**, con el corpus de ensayos escritos por los hablantes nativos y los no nativos, sin etiquetaje, pero con las EMP enlazadas.

La Figura 2 muestra también el conjunto de modelos de aprendizaje utilizados: *Naïve Bayes* (NB), *Neural Network* (NN), (*Decision*) *Tree*. El *widget* de prueba y puntuación (en inglés, *Test & Score*) se utilizó para entrenar y probar los modelos, solo con los datos del corpus de entrenamiento (partición con el 75 % del corpus), con el fin de seleccionar aquel con el mejor desempeño basado en la precisión de clasificación. Tras varios experimentos, se seleccionó el modelo NN, ya que mostró sistemáticamente mejores resultados. Da Corte & Baptista (2022a) también observaron el mismo desempeño del modelo NN en experimentos similares en un contexto parecido, pero con estudiantes en un curso de desarrollo de gramática en inglés. Se utilizó el *widget* de predicciones (en inglés, *Predictions*) para aplicar al modelo NN los datos de prueba no vistos previamente (partición con el 25 % del corpus) y evaluar la clasificación de los hablantes nativos y los no nativos en un escenario realista. Se utilizó el *widget* matriz de confusión (en inglés, *Confusion Matrix*) para inspeccionar en profundidad los resultados obtenidos.

En el modelo de aprendizaje diseñado para los tres experimentos, utilizando el *widget* de muestreo, los datos se dividieron en tres particiones para ejecutar una validación cruzada, dejando 2/3 para el entrenamiento y 1/3 para las pruebas de predicción de clasificación por nivel. Las tareas de preprocesamiento de datos consistieron en: (i) la *tokenización*, manteniendo tanto las palabras como los signos de puntuación como *tokens*, ya que se estima que la puntuación es un buen indicador del nivel de desarrollo de las habilidades escritas a nivel académico; (ii) la selección del *Averaged Perceptron Tagger* como etiquetador de partes del discurso (*Part-of-Speech Tagge*r; PoS por sus siglas en inglés). No se utilizó ninguna transformación de palabras, ya que la distinción entre mayúsculas y minúsculas puede ser relevante en la clasificación por nivel entre los hablantes nativos y los no nativos en tareas de redacción con fines académicos.

Como paso adicional en la etapa de preprocesamiento, se verificó y comprobó si el filtrado de palabras reservadas (en inglés, *stopwords*) repercutió en los resultados. Para ello, se utilizó un listado estándar de estas palabras en español disponible en línea[9]. En la mayoría de las configuraciones experimentales, la eliminación de las palabras reservadas, tal como se demuestra en la Tabla 3 (Exp 1bis y Exp 3bis), no mejoró ni empeoró los resultados, confirmando así que las palabras reservadas no influyen significativamente en la información semántica del texto en el que aparecen (HaCohen-Kerner *et al*. 2021). Sin embargo, utilizando el corpus etiquetado (Exp 2bis) se alcanzó el mayor nivel de precisión en la fase de entrenamiento (0,892), aunque en la fase de prueba, se observó uno de los resultados más débiles (0,711). Este último nivel de precisión es significativamente inferior al del experimento que

[9] <https://countwordsfree.com/stopwords/spanish>.

retuvo las palabras reservadas (Exp 2: 0,816). Por lo tanto, se decidió omitir el filtrado de las mismas en el resto de los experimentos.

Se compararon también dos métodos de representación de datos que se encuentran disponibles en la caja de herramientas Orange: la bolsa de palabras (*Bag-of-Words* o BoW, por sus siglas en inglés) y la integración de documentos (*Document Embeddings*, o DE, por sus siglas en inglés). En el método BoW, solo se tiene en cuenta la frecuencia de las palabras del texto, ignorando su orden relativo. En el método DE, se tienen en cuenta tanto la frecuencia como la información de co-ocurrencia de las palabras, ya que el texto se convierte en un vector de datos. Se confirmó que el método de la bolsa de palabras produce sistemáticamente mejores resultados. Con resultados similares obtenidos por HaCohen-Kerner *et al*. (2021), se observa, por ejemplo, en la Tabla 3, los resultados del Exp 2tri que corresponden a la utilización del método de representación de datos por medio de integración de documentos y que son significativamente peores que el Exp 2: 0,662 en la fase de entrenamiento y 0,5 en la fase de prueba. Consecuentemente, se descartó la opción de integración de documentos como método de representación de datos y se mantuvo solo la opción de bolsa de palabras (sin palabras reservadas) en el resto de los experimentos.

Corpus (Ensayos segmentados)	**Exp**	**T&S**	**Pred #**	**EMP**	
Sin etiquetaje	**Exp 1**	0,797	**0,842**	0	
Sin etiquetaje (con pr)	Exp 1bis	0,878	**0,816**	0	
Con etiquetaje y unión	**Exp 2**	**0,824**	0,816	779	
Con etiquetaje y unión (sin pr)	Exp 2bis	**0,892**	0,711	779	
Con etiquetaje y unión (con id)	Exp 2tri	0,662	0,500	779	
Sin etiquetaje con unión	**Exp 3**	0,784	**0,842**	779	
Sin etiquetaje con unión (sin pr)	Exp 3bis	0,770	**0,816**	779	%EMP
~ sin SVC	Exp 4	0,824	0,789	90	0,12
~ sin VIDIOM	Exp 5	**0,770**	0,789	41	0,05
~ sin ADV	Exp 6	0,811	0,789	183	0,23
~ sin PREP	Exp 7	0,797	0,789	24	0,03
~ sin CONJ	Exp 8	0,824	**0,763**	7	0,01
~ sin NE	Exp 9	0,838	0,816	52	0,07
~ sin N	Exp 10	**0,770**	0,816	289	0,37
~ sin PRON	Exp 11	0,838	0,816	13	0,02
~ sin SVC&VIDIOM	Exp 12	0,824	0,842	131	0,17
~ sin CONJ&VIDIOM	Exp 13	**0,784**	0,789	48	0,06
~ sin CONJ&SVC	Exp 14	0,851	0,816	97	0,12
~ sin PREP&ADV	Exp 15	0,811	0,763	207	0,27
~ sin PREP&CONJ	Exp 16	0,851	0,816	31	0,04
~ sin ADV&CONJ	Exp 17	0,838	0,737	190	0,24
~ sin N&VIDIOM	Exp 18	**0,784**	0,816	330	0,40

Tabla 3. Resultados obtenidos en las fases de entrenamiento (T&S) y prueba (Pred) en los tres diseños experimentales (Exp 1, Exp 2 y Exp 3) y experimentos complementarios (~ bis, ~ tri) seguidos de los experimentos derivados del Exp 2 y con la eliminación sucesiva (una a la vez) decada tipo EMP y de diferentes combinaciones de tipos de EMP[10]

[10] Leyenda: pr = palabras reservadas, id = integración de documentos.

3. Análisis de los resultados

En los dos primeros experimentos (corpus sin etiquetaje y corpus con etiquetaje y enlace), la adición de las etiquetas de EMP sugiere una mejoría en el nivel de rendimiento del modelo simple establecido como base, ya que pasa, en la fase de entrenamiento, de una precisión de clasificación de 0,797 a 0,824. Sin embargo, en la fase de prueba, la precisión del modelo con el corpus sin etiquetar aumenta ligeramente en comparación al corpus con el etiquetado. Esto puede deberse a una combinación de factores: (i) el hecho de utilizar todas las etiquetas juntas pudo haber degradado la precisión de los resultados; (ii) el pequeño tamaño del corpus y el muestreo aleatorio de la parte del corpus no etiquetada utilizada para las pruebas pudo haber sesgado los resultados. Además, dado que el corpus usado en los experimentos incluía las etiquetas de las EMP, que luego se utilizaron en la bolsa de palabras en la etapa de preprocesamiento como *tokens*, no se descarta que su presencia haya influido en los resultados.

Para entender el impacto que tienen las etiquetas en la clasificación y corroborar las suposiciones mencionadas, se llevó a cabo el Experimento 3, manteniendo como un único *token* la EMP (enlazada) y eliminando todas las etiquetas (8) seleccionadas. En este experimento, el modelo reconoce la EMP presente en el corpus, mas no considera las etiquetas con los tipos de expresiones. Los resultados de este experimento muestran una ligera disminución en el rendimiento del modelo en la fase de aprendizaje (0,784), por debajo de la base, pero coincide con esta en la fase de prueba (0,842).

Como parte de este estudio, se realizaron ocho experimentos adicionales (Experimentos 4-11) con el fin de evaluar el impacto de cada tipo de EMP en los resultados generales descritos anteriormente. Para ello, adoptamos una estrategia de *one-out* o 'una a la vez', eliminando una etiqueta de EMP a la vez, por ejemplo, VIDIOM, ejecutando el modelo de aprendizaje con la misma configuración y manteniendo las otras siete EMP (etiquetadas y enlazadas). De este modo, el modelo reconoce la misma información asociada con el Experimento 2, excepto la información relativa al tipo de EMP que se está poniendo a prueba. Dado que cada tipo de EMP está distribuido de forma diferente en el corpus, presentamos en las columnas de la izquierda, en la Tabla 3, el número y el porcentaje de EMP eliminadas en cada experimento. El argumento para interpretar los resultados es el siguiente: cuanto mayor sea la caída o disminución de la precisión del criterio de clasificación en los resultados generales de cada experimento, mayor es la contribución de ese tipo de EMP (en particular) al rendimiento del modelo.

Estos resultados fueron comparados con el número de EMP que quedaron en el corpus en cada experimento, al utilizar la estrategia *one-out*. Se calculó y encontró un coeficiente de correlación de Pearson débil y positivo de 0,259 entre la precisión de los modelos en la fase de entrenamiento y el número de EMP. Al comparar los resultados de la fase de prueba con el número de EMP en el corpus, se encuentra un coeficiente de Pearson más interesante, pero aun así moderado y negativo de -0,349. Este coeficiente parece indicar que en la fase de entrenamiento, como se esperaba, cuanto mayor es el número de EMP disponibles, mejores son los resultados de precisión que produce el modelo. Sin embargo, probablemente debido a que el muestreo

se realizó con un corpus pequeño, se observa el efecto contrario en la fase de prueba, lo que limita llegar a conclusiones más definitivas al respecto.

Los resultados muestran que la eliminación de los tipos de expresiones fraseológicas verbales (VIDIOM) y sustantivos comunes compuestos (N) producen la mayor disminución en la precisión (0,054) de los modelos en la fase de aprendizaje. Las expresiones VIDIOM son usadas por los hablantes nativos y los no nativos, en cuanto a frecuencia, de manera comparable: 27 incidencias por parte de los nativos y 19 por los no nativos. Esta mínima diferencia alude al nivel similar (y avanzado) de competencia entre los dos grupos. En cuanto al uso de sustantivos compuestos, los no nativos muestran un mayor uso de estas EMP, con 191 incidencias, en comparación con 102 por parte de los nativos.

En cuanto a la precisión de clasificación, le siguen inmediatamente los resultados obtenidos con la eliminación de las preposiciones (PREP) (0,027). Los experimentos con la eliminación de los tipos de EMP construcciones con verbo soporte (SVC) y locuciones conjuntivas (CONJ) no modificaron la precisión (0,824) del modelo base (Experimento 2). Por último, los experimentos con nombres de entidades (NE) y pronombres (PRON) mostraron una diferencia positiva (0,838) con respecto al modelo base, sugiriendo que la presencia de estas expresiones puede dificultar la evaluación de las destrezas escritas de los hablantes nativos y los no nativos. Sin embargo, en la fase de pruebas, es la eliminación de CONJ la que produce la mayor disminución en el rendimiento del modelo (0,053), situándose los tipos de EMP SVC, VIDIOM, ADV y PREP en segundo lugar, ex aequo (0,027) y sin ningún cambio en las categorías restantes NE, N y PRON.

Dado que la combinación de EMP puede tener una mayor incidencia en la precisión de la clasificación que una sola categoría, se realizaron siete experimentos adicionales (Exp 12-18, Tabla 3). Estos experimentos siguen las ideas expuestas por Siyanova-Chanturia & Spina (2020), en un estudio a gran escala de EMP en ensayos escritos por aprendices de una segunda lengua (italiano). Estos autores afirmaron que diferentes combinaciones de patrones de EMP «pueden captar una descripción más compleja y detallada de los patrones de desarrollos de hablantes no nativos» (Siyanova-Chanturia & Spina 2020: 453, traducción propia). Se procedió a adoptar una estrategia de *two-out* o 'dos a la vez': eliminando dos etiquetas de EMP a la vez con sus respectivos enlaces y ejecutando el modelo de aprendizaje con la misma configuración y manteniendo etiquetadas las EMP restantes. Las combinaciones de CONJ&VIDIOM (Exp 13) y N&VIDIOM (Exp 18) fueron las que reportaron la mayor reducción en el nivel de clasificación en la fase de entrenamiento (0,784), siendo N&VIDIOM la que mayor peso tuvo en cuanto al número de EMP (40 %), en relación con el total de EMP en todo el corpus.

4. Conclusiones y futuros estudios

Los tipos de EMP VIDIOM (expresiones fraseológicas verbales) y N (sustantivos comunes compuestos) tienen un mayor impacto en el modelo de aprendizaje en la fase de entrenamiento, mientras que, a pesar del pequeño número de elementos

presentes en el corpus utilizado, el tipo de EMP CONJ (locuciones conjuntivas) parece tener el mayor impacto en la fase de prueba. Cuando se combinaron y eliminaron dos tipos de EMP a la vez, las combinaciones CONJ&VIDIOM y N&VIDIOM fueron las que generaron la mayor reducción en el nivel de clasificación en la fase de entrenamiento, siendo N&VIDIOM la combinación que mayor peso tuvo en cuanto al número de EMP presentes en todo el corpus.

Estos resultados, aunque todavía provisionales y extraídos de un corpus pequeño, pueden indicar áreas de desarrollo léxico (y tipos de EMP) que deberían tomarse en cuenta para enriquecer las estrategias pedagógicas y de competencia lingüística en los programas de estudios orientados a la enseñanza del español como lengua extranjera. Aunque muchos estudios se ocupan de esta temática, estos suelen centrarse en las destrezas de lectoescritura en una segunda lengua. En este estudio también destacamos la importancia de estas habilidades para quienes tienen el español como lengua materna.

En el contexto de Oklahoma (4.019.800 habitantes),[11] existe una gran demanda de intérpretes y traductores calificados para desempeñar ambas tareas en español e inglés, particularmente, en el ámbito médico y legal (en este artículo no abordamos terminología técnica). Esta demanda se debe al aumento de hispanohablantes (actualmente, 11,8 % de la población) que tienen un nivel bajo de competencia en inglés (36,6 %)[12]. Por consiguiente, es necesario constatar que los participantes en los programas de certificación de traducción e interpretación: primero, estén en el nivel correcto y reciban la formación adecuada en cuanto al español como lengua extranjera; y, segundo, puedan comunicarse debidamente en dicha lengua como lengua meta o de destino (en inglés, *target language*).

En este análisis se tomó en cuenta un curso de desarrollo enfocado en la gramática y composición escrita de textos en español. Nos proponemos realizar un estudio de cursos de desarrollo en otros idiomas, como, por ejemplo, el inglés, para investigar si la eliminación o combinación de tipos de EMP arrojan resultados más significativos en cuanto a la clasificación del nivel de competencia lingüística de los participantes. Adicionalmente, sería de interés identificar si la disminución o aumento del uso de EMP específicas por parte de los hablantes nativos versus no nativos se debe al reconocimiento de dichas palabras o si su uso se ve limitado por los temas sugeridos para la elaboración de ensayos.

Referencias bibliográficas

ACTFL (2016), *Assigning CEFR ratings to ACTFL assessments* [disponible en <https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Ass essments.pdf>, 7/3/2023].

Alexopoulou, Theodora – Michel, Marije – Murakami, Akira – Meurers, Detmar (2017), «Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques», *Language Learning* 67(S1), 180-208.

[11] <https://www.census.gov/quickfacts/OK>.

[12] <https://www.migrationpolicy.org/data/state-profiles/state/language/OK>.

Alpine Testing Solutions (2020), *Examination of the ACTFL Writing Proficiency Test (WPT) in English, Russian, and Spanish for the ACE Review – Part B: Statistical Analysis & Evidence of Validity*, Orem, UT: Alpine Testing Solutions.

Arnold, Taylor – Ballier, Nicolas – Gaillat, Thomas – Lissòn, Paula (2018), «Predicting CEFRL levels in learner English on the basis of metrics and full texts», *Proceedings of the 20th Conférence Sur l'Apprentissage Automatique*. INSA de Rouen, 20-22 June 2018, *ArXiv:1806.11099*.

Baptista, Jorge – Mamede, Nuno – Reis, Sonia (2022), «Support Verb Constructions across the Ocean Sea», *Proceedings of the 18th Workshop on Multiword Expressions* @ LREC2022, Marseille, France. European Language Resources Association, 26-36.

Consejo de Europa (2002/2020), *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, Enseñanza, Evaluación,* [disponible en <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf>; <https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco_complementario/mcer_volumen-complementario.pdf>, 7/3/2023].

Corpas Pastor, Gloria (2017), «Collocational constructions in translated Spanish: what corpora reveal», en Mitkov, R. (ed.), *Computational and Corpus-Based Phraseology. Second International Conference, Europhras 2017,* Londres: Springer, 29-40.

Da Corte, Miguel – Baptista, Jorge (2022a), «A Phraseology Approach in Developmental Education Placement», en Corpas Pastor, G. – Mitkov, R. – Kunilovskaya, M. – Caro Quintana, R. (eds.) *Computational and Corpus-based Phraseology*, Proceedings of EUROPHRAS 2022, Malaga, September 28-30, 2022, Londres: Springer, 79-86.

Da Corte, Miguel – Baptista, Jorge (2022b), «Lista de expresiones multipalabra detectadas en ensayos escritos en un curso de desarrollo de gramática y composición» [disponible en <https://doi.org/10.13140/RG.2.2.35177.57443>, 7/3/2023].

Dahunsi, Toyese Najeem – Ewata, Thompson Olusegun (2022), «An exploration of the structural and colligational characteristics of lexical bundles in L1–L2 corpora for English language teaching», *Language Teaching Research* 1-17 [disponible en <https://doi.org/10.1177/13621688211066572>, 7/3/2023].

Demšar, Janez – Curk, Tomaž – Erjavec, Aleš – Gorup, Črt – Hočevar, Tomaž – Milutinovič, Mitar – Možina, Martin – Polajnar, Matija – Toplak, Marko – Starič, Anže – Stajdohar, Miha – Umek, Lan – Žagar, Lan – Žbontar, Jure – Žitnik, Marinka – Zupan, Blaž (2013), «Orange: data mining toolbox in Python», *The Journal of machine Learning research* 14(1), 2349-2353.

El-Dakhs, Dina Abel Salam – Khan, Shazia Khalid – Al-Khodair, Maram (2022), «Do foreign language learners mine input texts for multiword expressions? The case of writing story retellings», *Ampersand* 9, 100080.

Erdmann, Alexander – Wrisley, David Joseph – Brown, Christopher – Cohen-Bodénès, Sophie – Elsner, Micha – Feng, Yukun – Brian, Joseph – Joyeux-Prunel, Béatrice – de Marneffe, Marie-Catherine (2019), «Practical, efficient, and customizable active learning for named entity recognition in the digital humanities», *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, 2223-2234.

Esfandiari, Rajab – Ahmadi, Mohammad (2022), «Phraseological Complexity and Academic Writing Proficiency in Abstracts Authored by Student and Expert Writers», *English Teaching & Learning*, 1-20.

Farahmand, Meghdad – Smith, Aaron – Nivre, Joakim (2015), «A multiword expression data set: annotating non-compositionality and conventionalization for English noun compounds», *Proceedings of the 11th Workshop on Multiword Expressions,* 29-33.

Fotopoulou, Aggeliki – Laporte, Éric – Nakamura, Takuya (2021), «Where Do Aspectual Variants of Light Verb Constructions Belong?», *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021),* 2-12.

García-Page, Mario (2008), *Introducción a la fraseología española: estudio de las locuciones*, Barcelona: Anthropos.

Gross, Maurice (1996), «Lexicon-grammar», en Brown, K. – Miller, J. (eds.), *Concise Encyclopedia of Syntactic Theories*, Cambridge: Pergamon, 244-259.

HaCohen-Kerner, Yaakov – Miller, Daniel – Yigal, Yair (2020), «The influence of preprocessing on text classification using a bag-of-words representation». *PloS ONE* 15(5), e0232525 [disponible en <https://doi.org/10.1371/journal.pone.0232525>, 7/3/2023].

Hernández, Mireia – Costa, Alber – Arnon, Inbal (2016), «More than words: multiword frequency effects in non-native speakers», *Language, Cognition and Neuroscience* 31(6), 785-800.

Hinkel, Eli (2023). «Teaching and Learning Multiword Expressions», *Handbook of Practical Second Language Teaching and Learning,* Nueva York: Routledge, 435-448.

Kochmar, Ekaterina – Gooding, Sian – Shardlow, Matthew (2020), «Detecting multiword expression type helps lexical complexity assessment», *LREC 2020: Proceedings of the 12th Conference on Language Resources and Evaluation,* The European Language Resources Association (ELRA), 4426-4435.

Laporte, Éric (2018), *Choosing features for classifying multiword expressions,* en Sailer, M. – Markantonatou, S. (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, Berlín: Language Science Press, 143-186.

Nam, Daeyeon – Park, Kwanghyun (2020), «I will write about: Investigating multiword expressions in prospective students' argumentative writing», *Plos one* 15(12), e0242843.

Pasquer, Caroline – Savary, Agata – Ramisch, Carlos – Antoine, Jean-Yves (2020), «Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?», *Proceedings of the 28th International Conference on Computational Linguistics,* 3333-3345.

Savary, Agata – Cordeiro, Silvio Ricardo – Ramisch, Carlos (2019), «Without lexicons, multiword expression identification will never fly: A position statement», *Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019).* Association for Computational Linguistics, 79-91.

Siyanova-Chanturia, Anna – Spina, Stefania (2020), «Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study», *Language Learning* 70(2), 420-463.

Wilkens, Rodrigo – Seibert, Daiane – Wang, Xiaoou – François, Thomas (2022), «MWE for Essay Scoring English as a Foreign Language», *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference,* 62-69.