Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa

Measuring writing skills in Spanish as a foreign language with generative artificial intelligence

María-Victoria Cantero Romero - *Universidad de Jaén, UJA (España)*
María-Teresa Martín-Valdivia - *Universidad de Jaén, UJA (España)*
Ana María Ortiz-Colón - *Universidad de Jaén, UJA (España)*
Salud María Jiménez-Zafra - *Universidad de Jaén, UJA (España)*

RESUMEN

La irrupción de la Inteligencia Artificial Generativa (IAG), y en particular de los Modelos de Lenguaje de gran tamaño (LLMs) como ChatGPT, está transformando el ámbito educativo, especialmente en la enseñanza de lenguas extranjeras. Este artículo analiza el potencial de estas tecnologías para automatizar la evaluación de la competencia escrita en español como lengua extranjera (ELE), una tarea especialmente laboriosa al inicio de los cursos universitarios dirigidos a estudiantes Erasmus. La metodología se basa en tres experimentos con el Corpus de Aprendices de Español del Instituto Cervantes. En el primero, se utilizó la técnica de *zero-shot learning*, proporcionando al modelo un *prompt* con los descriptores del Plan Curricular del Instituto Cervantes. En el segundo y tercer experimentos, se ajustó el modelo mediante *fine-tuning* con el 90 % y el 80 % del corpus, respectivamente, reservando el resto para validación y prueba. Los resultados muestran que los modelos ajustados son capaces de identificar el nivel de competencia escrita con una precisión significativamente superior al enfoque sin entrenamiento previo. Estos hallazgos evidencian que los LLMs pueden emplearse para agilizar procesos de evaluación inicial en cursos de ELE, reduciendo la carga docente y mejorando la eficiencia. Se concluye que la IAG representa una herramienta complementaria valiosa en contextos educativos multiculturales y multilingües, siempre que su uso esté guiado por criterios pedagógicos sólidos.

Palabras clave: enseñanza de lenguas; expresión escrita; inteligencia artificial; innovación pedagógica relevante.

ABSTRACT

The emergence of Generative Artificial Intelligence (GAI)—particularly Large Language Models (LLMs) such as ChatGPT—is transforming the educational landscape, especially in the field of foreign language instruction. This article explores the potential of these technologies to automate the assessment of writing proficiency in Spanish as a Foreign Language (SFL), a task that is especially time-consuming at the beginning of university-level courses for Erasmus students. The study is based on three experiments conducted using the Spanish Learner Corpus compiled by the Instituto Cervantes. The first experiment applied a zero-shot learning approach by prompting the model with level descriptors from the Instituto Cervantes's Curriculum Plan. In the second and third experiments, the model was adjusted through fine-tuning using 90% and 80% of the corpus, respectively, with the remaining data reserved for testing and validation. The results indicate that the fine-tuned models significantly outperform the zero-shot configuration in identifying the correct proficiency levels of learner texts. These findings demonstrate that LLMs can be effectively employed to streamline the initial placement process in SFL courses, thus reducing the workload of instructors and improving efficiency. The study concludes that GAI can serve as a valuable complementary tool in multilingual and multicultural educational settings, provided its use is guided by sound pedagogical principles.

Keywords: language instruction; writing; artificial intelligence; teaching method innovations.

INTRODUCCIÓN

La evolución de la Inteligencia Artificial ha supuesto un punto de inflexión en todos los ámbitos y, en especial, en el educativo (Aparicio Gómez, 2023; Hernández-León y Rodríguez-Conde, 2024; Zambrano Campozano, 2025). El avance de esta disciplina abre un campo de nuevas investigaciones y trabajos con los que poder implantar la Inteligencia Artificial, en concreto, la generativa, en las aulas. Como ya nos señalan Bolaño-García y Duarte-Acosta (2024), la Inteligencia Artificial generativa ha ganado atención en educación, puesto que mejora la personalización del aprendizaje y la retroalimentación en tiempo real. Zapata Ros (2024) apoya esta idea de personalización además de la disponibilidad de información. Asimismo, Fajardo et al. (2023) remarcan la utilización de estas herramientas para la personalización del aprendizaje en la educación universitaria, adaptándolo a las preferencias y necesidades de cada estudiante mediante tutorías guiadas y virtuales. García-Peñalvo et al. (2024) señalan la importancia de preparar tanto a docentes como a discentes en el uso de la Inteligencia Artificial generativa, puesto que estará presente en todos los aspectos de la vida. Además, como nos indica Moreno (2019) hay que destacar el potencial de la Inteligencia Artificial generativa para transformar la educación al crear entornos de aprendizaje adaptativos, ajustados al rendimiento de los estudiantes, como pueden ser para estudiantes con necesidades educativas especiales. En su estudio sobre herramientas de Inteligencia Artificial generativa, Area-Moreira et al. (2024), además de las funciones ya mencionadas, indican que estas herramientas pueden ser utilizadas para la automatización de tareas como apoyo al profesorado en la labor docente e incluso como sistemas antiplagios. En consonancia con lo anterior, Chan y Tsi (2023) señalan el uso de la Inteligencia Artificial generativa como herramienta auxiliar para los profesores y no como una sustitución de estos. Moreno (2019) también señala tres enfoques con los que trabajar en educación: con la Inteligencia Artificial generativa, con robótica educativa y con las plataformas de autoaprendizaje. Es en el primero de ellos en el que nos vamos a centrar en este estudio. Tanto Barroso-Osuna y Cabero-Almenara (2025) como Owan et al. (2023) identifican una serie de beneficios del uso de la Inteligencia Artificial, en concreto, la generativa, en educación, entre los que destacan la optimización de tiempo docente y la evaluación automatizada y precisa. En este sentido, Crespo Mendoza et al. (2024) señalan que puede mejorar la precisión como la fiabilidad de las evaluaciones.

Como se ha mencionado anteriormente, el enfoque principal de este trabajo es la Inteligencia Artificial generativa, concretamente los Modelos de Lenguaje (LLMs, por sus siglas en inglés), que han supuesto una nueva herramienta con la que poder llevar a cabo estudios enfocados a mejorar las tareas docentes. Los LLMs son herramientas de generación de lenguaje natural entrenadas con una gran cantidad de textos (Wang, 2024). García-Peñalvo et al. (2024) señalan distintas funciones que pueden llevar a cabo los LLMs, como el apoyo a la investigación, la creación de contenido educativo, la generación de chatbots para interactuar con el alumnado ofreciendo

Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), pp. 353-379. https://doi.org/10.5944/ried.45486

una retroalimentación autodirigida, el uso como complemento de los motores de búsqueda, el parafraseado de texto, la enseñanza de idiomas o la generación de exámenes y cuestionarios. Relacionado con estas dos últimas funciones es donde se enmarca el presente estudio, la nivelación en la enseñanza de idiomas y, en concreto, en el español como lengua extranjera.

La Inteligencia Artificial generativa presenta la oportunidad de poder realizar evaluaciones adaptativas a cada estudiante con una retroalimentación inmediata y específica, sugiriendo posibles soluciones (Barroso-Osuna y Cabero-Almenara, 2025). Asimismo, los Modelos de Lenguaje pueden ser usados para una automatización (García-Peñalvo, 2024) de la corrección tanto en pruebas de opción múltiple como en respuestas de texto abierto (Moreno, 2019). Como ventaja de usar los LLMs para evaluación, se pueden destacar la eficiencia (Area-Moreira et al., 2024) y la detección del plagio (García-Peñalvo et al., 2024). Asimismo, García-Peñalvo et al. (2024) señalan que los LLMs mejoran la productividad del profesorado.

Con respecto al uso de los Modelos de Lenguaje y la enseñanza de idiomas, diversos estudios (Baskara y Mukarto, 2023; Salguero Romero, 2023; Wang, 2024) han demostrado su utilidad. Baskara y Mukarto (2023) señalan cómo ChatGPT es capaz de generar textos realistas que acerquen a los estudiantes a la realidad de la lengua. Por otro lado, Hong (2023) señala como ventaja poder utilizar estas herramientas para agilizar la corrección de exámenes, liberando a los docentes de carga, dándoles oportunidad de centrarse en la preparación de las clases.

Uno de los retos de la enseñanza de idiomas y, en este caso, de la enseñanza de español como lengua extranjera es la adecuada nivelación de los estudiantes al iniciar los cursos de español. Que un estudiante se encuentre en un nivel adecuado de aprendizaje de español es esencial para una correcta progresión del aprendizaje, puesto que, si se encuentra en un nivel superior, el alumno puede frustrarse, y a la inversa, si se encuentra en un nivel inferior, el alumno puede no encontrar motivación. Es por ello que, la nivelación, es clave en la iniciación de los cursos de idiomas.

Hasta ahora, esta nivelación se ha realizado a través de pruebas de opción múltiple o entrevista (Biedma Torrecillas et al., 2012). Estas pruebas, principalmente, se centran en obtener el nivel que el alumno tiene con respecto a la expresión y comprensión tanto oral como escrita. Las pruebas de nivel de español existentes, como la prueba de nivel del Instituto Cervantes, consisten en contestar una serie de ítems escritos de respuesta cerrada (verdadero-falso; emparejar u ordenamiento) de dificultad ascendente (Centro Virtual Cervantes, s. f.). Esta misma tipología se observa actualmente en contextos internacionales, como en las pruebas de nivel de español de Columbia University y de University of Wisconsin–Madison, que también se basan en ejercicios de opción múltiple sin incluir producción escrita libre (Columbia University, s. f.; University of Wisconsin–Madison, s. f.). No obstante, las pruebas de nivel de español no se han centrado en las pruebas de expresión escrita de manera directa con escritura de textos completos, sino de manera indirecta, a

través de la respuesta de ítems, debido a las limitaciones que conlleva, como pueden ser la falta de inmediatez o la compleja nivelación cuando se trata de grandes grupos.

Para solucionar este problema, el presente estudio se centrará en la nivelación de la prueba de expresión escrita. Como se ha mencionado anteriormente, los Modelos de Lenguaje son capaces de automatizar la calificación de trabajos escritos (García-Peñalvo et al., 2024), además de manera rápida, ahorrando tiempo a los docentes (Area-Moreira et al., 2024), siendo este el principal motivo por el que no se ha incluido la prueba de expresión escrita en las pruebas de nivel actuales.

En este contexto, resulta necesario establecer una base conceptual que permita comprender los avances y limitaciones de la evaluación automatizada de la escritura y, de forma particular, los fundamentos lingüísticos que sustentan la clasificación de niveles A1–C1. En el siguiente apartado se desarrolla este marco teórico, que servirá de apoyo para la propuesta metodológica del presente estudio.

El presente artículo contiene la siguiente estructura. Tras la introducción y el marco teórico, se presenta la fundamentación tecnológica y elección del modelo, los objetivos del estudio y la metodología. A continuación, se exponen los resultados, seguidos de un apartado sobre la pertinencia pedagógica del modelo. Finalmente, el artículo concluye con la discusión y conclusiones, donde se ha añadido un subapartado sobre aspectos éticos y licencias de uso.

MARCO TEÓRICO

El presente apartado expone, en primer lugar, una revisión del estado del arte en torno a la evaluación automatizada de la escritura y los fundamentos lingüísticos que sustentan la clasificación por niveles del Marco Común Europeo de Referencia (MCER) y el Plan Curricular del Instituto Cervantes (PCIC). Esta base conceptual servirá para contextualizar la propuesta del estudio y, finalmente, para presentar los objetivos que guían la investigación.

Evaluación automatizada de la escritura: evolución y enfoques actuales

La clasificación automática de textos de aprendientes mediante modelos de lenguaje se inscribe en una tradición más amplia de evaluación automática de la escritura (AWE), cuyo desarrollo ha dado lugar a múltiples herramientas y sistemas que conviene tener en cuenta para contextualizar el presente trabajo. La AWE ha evolucionado considerablemente en las últimas décadas, convirtiéndose en una herramienta cada vez más común en contextos educativos. Sistemas pioneros como *e-rater* (Burstein et al., 2003), desarrollado por ETS, han sido empleados ampliamente en pruebas estandarizadas, valiéndose de métricas lingüísticas, gramaticales y discursivas para estimar la calidad textual. Por su parte, *Coh-Metrix* (McNamara et al., 2014) permite realizar análisis detallados de la cohesión, la complejidad sintáctica y la legibilidad, proporcionando una aproximación multifactorial al discurso escrito.

Estudios recientes, como el de Zhang (2021), ofrecen revisiones sistemáticas de estos sistemas, destacando su transición desde enfoques basados en reglas hacia modelos impulsados por el aprendizaje automático y el procesamiento del lenguaje natural. En esta misma línea, Wang et al. (2022) analizan enfoques actuales orientados a la evaluación de textos argumentativos, centrados en componentes discursivos como la estructura del razonamiento, la evidencia o la organización. Otras propuestas, como *Writing Mentor* (Burstein et al., 2018), integran la evaluación automática con retroalimentación formativa, favoreciendo procesos de autorregulación en la escritura académica. Asimismo, herramientas como *Write & Improve*, desarrollada por la Universidad de Cambridge, ejemplifican cómo es posible proporcionar feedback automático inmediato sobre textos producidos por estudiantes de lenguas extranjeras, facilitando el aprendizaje autónomo y guiado (Cambridge English, s. f.).

Fundamentos lingüísticos de la clasificación por niveles A1–C1

La clasificación de los textos producidos por aprendices de español como lengua extranjera en niveles A1–C1 se fundamenta en los descriptores establecidos por el Marco Común Europeo de Referencia para las Lenguas (Consejo de Europa, 2002) y el Plan Curricular del Instituto Cervantes (Instituto Cervantes, 2006). Estos documentos definen de manera detallada las competencias lingüísticas, pragmáticas y sociolingüísticas asociadas a cada nivel, proporcionando una base sólida para la evaluación.

En el nivel A1, el repertorio léxico es muy limitado y se restringe a transacciones básicas y expresiones cotidianas. Los textos son muy breves y sencillos, con oraciones simples y un promedio reducido de palabras por enunciado. Predomina el uso de las formas regulares del presente de indicativo y se emplea un repertorio gramatical elemental (Consejo de Europa, 2002; Instituto Cervantes, 2006).

En el nivel A2, el aprendiz puede producir textos breves que transmiten información sencilla sobre temas familiares. Se observa un mayor repertorio léxico y estructuras ligeramente más complejas, incorporando tiempos verbales del pasado de indicativo (pretérito perfecto, imperfecto e indefinido) y algunas formas irregulares de presente. También aparece el uso del imperativo afirmativo (Consejo de Europa, 2002; Instituto Cervantes, 2006).

En el nivel B1, el repertorio léxico es más amplio y permite elaborar textos que cumplen una tarea comunicativa concreta, manteniendo una estructura coherente. Gramaticalmente, se manejan con cierta soltura tiempos como el futuro simple, el condicional simple y el pretérito pluscuamperfecto, además de introducir el presente de subjuntivo y el imperativo negativo. El discurso muestra una mayor cohesión y variedad de recursos conectores (Consejo de Europa, 2002; Instituto Cervantes, 2006).

En el nivel B2, el usuario dispone de un repertorio lingüístico amplio y preciso, capaz de sostener argumentaciones complejas y descripciones detalladas. Se

manejan con fluidez oraciones compuestas y subordinadas, así como un uso seguro de los tiempos verbales de indicativo (presente, pasados, futuros y condicionales) y de subjuntivo (presente, imperfecto, perfecto y pluscuamperfecto). La cohesión textual es consistente y se emplean matices léxicos adecuados a diferentes registros (Consejo de Europa, 2002; Instituto Cervantes, 2006).

En el nivel C1, el repertorio lingüístico y no lingüístico es lo suficientemente amplio y flexible para abordar cualquier tipo de transacción o interacción comunicativa, incluso en contextos académicos o profesionales exigentes. El aprendiz es capaz de producir textos extensos y complejos con una estructura clara y bien organizada, utilizando con precisión todos los tiempos verbales y un amplio rango de recursos sintácticos y léxicos (Consejo de Europa, 2002; Instituto Cervantes, 2006).

FUNDAMENTACIÓN TECNOLÓGICA Y ELECCIÓN DEL MODELO

En el presente estudio se ha utilizado el modelo de lenguaje ChatGPT, desarrollado por OpenAI, en su versión 3.5, lanzada en noviembre de 2022 (OpenAI, 2022). A pesar de no ser un modelo de acceso en abierto, esta versión permite realizar un *fine-tuning* a través de la API de OpenAI. La selección de la versión 3.5 frente a la versión posterior se justifica por la posibilidad de realizar este proceso, lo cual no es posible en la última versión de ChatGPT.

Para esta investigación se optó por emplear GPT-3.5 en lugar de modelos más recientes o de código abierto debido a una combinación de factores técnicos, económicos y metodológicos. Al tratarse de una tarea novedosa —la nivelación automática de la competencia escrita en español como lengua extranjera—, se consideró pertinente evaluar el rendimiento de un modelo generalista ampliamente probado, como GPT-3.5, tanto en su configuración base como mediante *fine-tuning*. Este modelo, accesible a través de la API de OpenAI, no requiere infraestructuras computacionales avanzadas y ofrece una arquitectura sólida con buena cobertura del español (Li, 2023; Pourpanah et al., 2023). Además, presenta una relación adecuada entre rendimiento, coste y velocidad de respuesta, factores clave para validar la viabilidad del enfoque en esta fase exploratoria (Roumeliotis et al., 2024). Por contraste, en el momento de los experimentos, modelos como GPT-4 suponían un coste económico considerablemente mayor y tiempos de inferencia más prolongados, lo que reforzó la decisión de utilizar GPT-3.5 como referencia inicial.

No obstante, conviene señalar que GPT-3.5 presenta limitaciones frente a modelos más recientes, como GPT-4, que ofrecen una comprensión contextual más precisa, un entrenamiento con conjuntos de datos más amplios y heterogéneos y una mayor capacidad de razonamiento. Estas características los convierten en candidatos especialmente adecuados para investigaciones futuras, tanto en la evaluación automatizada de la competencia escrita alineada con los niveles del MCER como en la generación de corpus sintéticos de alta calidad que permitan entrenar sistemas especializados.

Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), pp. 353-379. https://doi.org/10.5944/ried.45486

OBJETIVO DEL ESTUDIO

Para llevar a cabo el presente estudio se han adaptado los descriptores establecidos por el Marco Común Europeo de Referencia para las Lenguas (Consejo de Europa, 2002) y el Plan Curricular del Instituto Cervantes con el fin de establecer criterios claros y operativos que permitan a un modelo de lenguaje, en este caso ChatGPT, clasificar de forma automática las producciones escritas de aprendices de español según su nivel de competencia.

A pesar de los avances mencionados en la evaluación automática de la escritura, pocos estudios se han centrado en la nivelación de textos escritos producidos por aprendices de lenguas extranjeras y, en concreto, en el caso específico de la enseñanza de español como lengua extranjera. Las investigaciones existentes suelen abordar la retroalimentación o corrección automática, pero no la clasificación de producciones escritas según niveles del MCER o del Plan Curricular del Instituto Cervantes, especialmente en tareas de evaluación inicial. Esta laguna es especialmente relevante en contextos como el universitario, donde la heterogeneidad del alumnado internacional, como en los programas Erasmus, exige herramientas eficientes para la nivelación. Este estudio propone una solución innovadora basada en Modelos de Lenguaje generativos (ChatGPT), que permite clasificar automáticamente textos escritos de aprendices de enseñanza de español como lengua extranjera (ELE) según su nivel de competencia, reduciendo la carga docente y mejorando la gestión del proceso de inicio de curso. Así, el trabajo no solo complementa investigaciones previas centradas en la mejora textual, sino que amplía el alcance de la evaluación automatizada hacia tareas de diagnóstico inicial en contextos de enseñanza de segundas lenguas.

Es por ello que, como objetivo general del estudio, se ha planteado evaluar la eficacia del modelo de lenguaje ChatGPT como herramienta innovadora para nivelar la expresión escrita de los estudiantes de español. Asimismo, se espera conseguir una serie de objetivos específicos: i) conocer si ChatGPT con su entrenamiento previo es capaz de nivelar de manera adecuada; ii) comprobar si ChatGPT es capaz de nivelar si es ajustado con un corpus nivelado con el Plan Curricular del Instituto Cervantes y el Marco Común Europeo de Referencia para las lenguas; y iii) determinar el impacto de ChatGPT en la eficiencia del proceso de nivelación del español como lengua extranjera.

METODOLOGÍA

Para llevar a cabo la presente investigación se ha utilizado el Corpus de Aprendices de Español (en adelante CAES) (Palacios Martínez et al., 2019), elaborado por el Instituto Cervantes en colaboración con la Universidad de Santiago de Compostela. Asimismo, se ha empleado el modelo de lenguaje ChatGPT en su versión 3.5, desarrollado por OpenAI (OpenAI, 2022).

En este estudio se han realizado tres tipos de pruebas experimentales: una de ellas utilizando la técnica de *zero-shot learning*, y las otras dos mediante *fine-tuning* del modelo.

El procedimiento de *fine-tuning* se realizó con una única época de entrenamiento, manteniendo el resto de parámetros por defecto según la API de OpenAI. Para garantizar la reproducibilidad, se empleó una semilla aleatoria fija (valor 42), es decir, un valor de referencia que permite que los experimentos puedan reproducirse siempre con los mismos resultados. No se aplicaron técnicas de balanceo de clases, puesto que se optó por conservar la distribución real del corpus, reflejando así condiciones auténticas de nivelación en el aula. De este modo, los resultados del sistema se ajustan mejor a los desafíos propios de escenarios educativos reales, sin introducir modificaciones artificiales en la representación de los niveles. No obstante, se reconoce que en investigaciones futuras podrían aplicarse estrategias de compensación para comparar su efecto en la equidad y la robustez del modelo. Tampoco se empleó validación cruzada, en coherencia con el carácter inicial y experimental del estudio.

En relación con los parámetros técnicos del entrenamiento, se mantuvieron los valores predeterminados de la API de OpenAI en aspectos como el tamaño de lote (*batch size,* es decir, el número de ejemplos procesados a la vez), la tasa de aprendizaje (*learning rate,* que indica la velocidad con la que el modelo ajusta sus parámetros durante el entrenamiento) y las funciones de pérdida (métricas que miden la diferencia entre la predicción del modelo y el resultado esperado). No se aplicaron técnicas adicionales de regularización ni estrategias de *early stopping* (interrupción anticipada del entrenamiento para evitar sobreajuste), puesto que el objetivo principal era validar la viabilidad del enfoque más que optimizar el modelo al máximo rendimiento. Esta decisión se enmarca en el carácter exploratorio de la investigación, orientado a comprobar la aplicabilidad del modelo a la tarea de nivelación automática de textos.

Corpus CAES

El corpus CAES fue recogido por la Universidad de Santiago de Compostela y financiado por el Instituto Cervantes. Se recopilaron datos informatizados desde octubre de 2011 hasta diciembre de 2020 de centros, en su mayoría universitarios, de distintos países como España, Estados Unidos, Brasil, Egipto, Irlanda o Portugal. Los estudiantes que formaron parte del proyecto procedían de once lenguas de origen distintas (inglés, chino mandarín, portugués, árabe, ruso, alemán, francés, griego, italiano, japonés y polaco).

En este estudio se ha trabajado con la versión 2.1 del corpus CAES de marzo de 2022, con la que actualizaron la primera recogida de datos, que contaba con un número menor de datos, con un total de 1423 participantes, frente a los 2544 participantes de la actualización de 2022.

En el corpus se encuentran ejemplos de distintos niveles de español según el Marco Común Europeo de Referencia para las Lenguas, desde A1 hasta C1. En los niveles A1, A2 y B1 se recopilan textos pertenecientes a tres tareas de distinta tipología que los estudiantes debían escribir con una extensión de entre 30 y 200 palabras. Los niveles B2 y C1 cuentan con una muestra de dos tareas por nivel y con una extensión de entre 275 y 500 palabras.

Los temas que se han identificado en cada nivel son los siguientes: En A1, la primera tarea consiste en un correo electrónico sobre cambiar de trabajo, con un total de 728 ejemplos; la segunda tarea consiste en un correo electrónico sobre su familia, con 703 ejemplos; la última tarea consiste en una nota acerca de llegar tarde, con una muestra de 705. En el nivel A2, la primera tarea que encontramos es una biografía, con 673 ejemplos; la segunda tarea consiste en la reserva de una habitación de hotel, con una muestra de 603 textos; y la última consiste en escribir una postal sobre sus vacaciones, que cuenta con una muestra de 701 ejemplos. Se puede comprobar que las muestras recogidas de los niveles iniciales rondan los 2 000 ejemplos por nivel, dando una muestra significativa de dichos niveles. Asimismo, la temática de los textos se encuentra relacionada con las funciones y los productos textuales correspondientes de su nivel según el Plan Curricular del Instituto Cervantes (Instituto Cervantes, 2006).

En el nivel B1 se encuentran, igualmente, tres tareas distintas. La primera, escribir una carta a un amigo, con una muestra de 528 textos; la segunda tarea, escribir un correo electrónico sobre una reclamación a una compañía aérea, con una muestra de 454; y la última tarea, una narración de una historia, con una muestra de 382. Hay que señalar que estas tareas, al igual que las de nivel A1 y A2, se encuentran relacionadas con las funciones que se describen en el Plan Curricular.

En los niveles B2 y C1, las tareas se reducen a dos. En el nivel B2, en la primera tarea deben redactar una solicitud de admisión, con un total de 375 muestras; mientras que la segunda tarea del nivel B2 trata sobre un texto donde deben argumentar el tema de fumar en lugares públicos, con una muestra de 356.

Con respecto al nivel C1, la primera tarea consiste en escribir una reclamación a una compañía de gas, con una muestra de 169; y en la segunda tarea, deben escribir una reseña de una película, con una muestra de 184 textos. Se puede ver cómo la muestra se reduce significativamente en estos niveles al disminuir el número de tareas. Asimismo, es notable, cómo en el nivel C1 las muestras son menores, no llegando a 400 muestras en este nivel.

Al igual que en los niveles anteriores, se puede observar que las tareas de los niveles B2 y C1 corresponden igualmente a las funciones descritas en el Plan Curricular.

La anotación del corpus fue realizada por especialistas en enseñanza de español como lengua extranjera de la Universidad de Santiago de Compostela. Cada texto fue clasificado en uno de los niveles del MCER siguiendo los criterios establecidos en el Plan Curricular del Instituto Cervantes y las pautas definidas en el propio

proyecto (Palacios Martínez et al., 2019; Universidad de Santiago de Compostela, s. f.). Para garantizar la fiabilidad de la clasificación, los textos fueron evaluados de forma independiente por varios anotadores y, posteriormente, revisados de manera conjunta hasta alcanzar un consenso. Este procedimiento asegura que la asignación de niveles sea coherente y consistente, lo que convierte al corpus en un recurso sólido para la investigación y la evaluación automática de la producción escrita.

En la Tabla 1, se presenta un resumen de los niveles con respecto a las tareas abordadas y al número total de muestras recogidas.

Asimismo, Cantero (2024) realizó un estudio de este corpus en el que se pueden sacar los siguientes resultados. En el nivel A1, el promedio de palabra por oración se encuentra entre 10.9 y 11.7 y las palabras más frecuentes son conectores simples y un léxico limitado. En el nivel A2, el promedio de palabras por oración es mayor, tratándose entre 12.5 a 14.6, con un léxico frecuente simple. En el siguiente nivel, B1, se puede observar que aumenta a 12.0 - 16.5 el promedio de palabras por oración, asimismo, el vocabulario es más amplio y complejo que el de niveles anteriores. Con respecto al nivel B2, el promedio de palabras por oración es aún más complejo, llegando a encontrarse entre 17.7 y 21.5, y en relación con el léxico se utilizan conectores más complejos y un vocabulario más especializado. Por último, en el nivel C1, el promedio de palabras por oración se encuentra entre 20.7 y 23.3 con un vocabulario complejo y variado.

Tabla 1
Resumen de muestras

Nivel	Tareas	Muestras
A1	Correo electrónico cambio de trabajo	728
	Correo electrónico familia	703
	Nota llegar tarde	705
A2	Biografía persona que admiras	673
	Reserva habitación de hotel	603
	Postal de vacaciones	701
B1	Carta a un amigo	528
	Historia graciosa	382
	Reclamación compañía aérea	454
B2	Solicitud admisión	375
	Fumar en lugares públicos	356
C1	Reseña película	184
	Reclamación compañía de gas	169

Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), pp. 353-379. https://doi.org/10.5944/ried.45486

Prompt utilizado

Una vez analizado el corpus, se ha elaborado un *prompt* específico para realizar las pruebas con el modelo. Como señala Morales-Chan (2023), un buen *prompt* puede garantizar el éxito de la tarea. Por ello, es importante definir el objetivo y proporcionar un contexto suficiente.

El siguiente *prompt* fue utilizado en las tres pruebas realizadas en este estudio —*zero-shot learning* (Prueba 1) y *fine-tuning* (Pruebas 2 y 3)— con el fin de mantener una coherencia metodológica en los criterios de evaluación. El diseño del *prompt* se basó en los descriptores lingüísticos del Plan Curricular del Instituto Cervantes y del Marco Común Europeo de Referencia para las Lenguas.

> *Tú eres un experto lingüista especializado en enseñanza de español como lengua extranjera. Tu tarea es indicar el nivel de español como lengua extranjera de los textos siguiendo el Plan Curricular del Instituto Cervantes.*
>
> *Aquí tienes una descripción de los distintos niveles.*
>
> *Niveles A1 y A2 Transacciones básicas relacionadas con su entorno.*
>
> *A1: Repertorio limitado de léxico, textos muy breves y sencillos, un promedio de 10 palabras por oración. Formas regulares del presente de indicativo.*
>
> *A2: Textos breves con información sencilla, un promedio de 12 palabras por oración. Tiempos verbales del pasado de indicativo: Pretérito perfecto, imperfecto e indefinido. Formas irregulares de presente de indicativo. Imperativo afirmativo.*
>
> *Niveles B1 y B2 desenvolverse con textos sobre temas de su interés gustos y preferencias.*
>
> *B1: Vocabulario amplio pero sencillo, realizar textos con una tarea concreta. Presente de indicativo, pretérito perfecto, imperfecto e indefinido de indicativo, futuro simple, condicional simple, pretérito pluscuamperfecto de indicativo, presente de subjuntivo. Imperativo negativo.*
>
> *B2: Repertorio lingüístico amplio, oraciones subordinadas. Tiempos verbales de indicativo: presente, pretérito perfecto, imperfecto, indefinido, futuro simple y compuesto, condicional simple y compuesto, pretérito pluscuamperfecto. Tiempos verbales de subjuntivo: presente, pretérito imperfecto, pretérito perfecto y pluscuamperfecto.*
>
> *C1 transacciones de todo tipo. Disponen de un repertorio de recursos lingüísticos y no lingüísticos lo suficientemente amplio y rico. Pueden enfrentarse a una amplia serie de textos extensos y complejos. Todos los tiempos verbales de indicativo y de subjuntivo el presente, pretérito perfecto, imperfecto y pluscuamperfecto.*

> *Ahora vas a recibir un TEXTO y teniendo en cuenta lo explicado anteriormente y los errores gramaticales indica al final de tu respuesta con la etiqueta 'NIVEL:' el nivel del TEXTO (A1, A2, B1, B2 o C1).*
> *TEXTO: "..."*

Con este *prompt* se ha añadido información descriptiva breve de cada nivel, siguiendo el análisis del corpus comentado anteriormente. Asimismo, siguiendo el Plan Curricular del Instituto Cervantes (Instituto Cervantes, 2006), se ha incluido una descripción de los tiempos verbales utilizados en cada uno de los niveles y los tipos de textos siguiendo las indicaciones de los productos textuales. De esta forma, se le proporciona al modelo un contexto más amplio para que su respuesta sea más precisa y ajustada a las necesidades que se le solicitan.

Pruebas con el Modelo de Lenguaje ChatGPT

En el estudio se han realizado un total de tres pruebas distintas para evaluar la capacidad del modelo al nivelar los textos del corpus.

La primera de las pruebas fue un *zero-shot learning*. En este proceso, el modelo no recibe ejemplos específicos, sino que se basa en el conocimiento previo del mismo. Para realizar esta prueba, se ha utilizado únicamente el *prompt* que se ha mencionado en el apartado anterior.

En las pruebas dos y tres se ha realizado *fine-tuning*. Este proceso consiste en realizar una especialización de un modelo preentrenado para que realice una tarea específica, adaptándolo a un conjunto específico de datos que se le proporciona. Este conjunto de datos lo obtuvimos del corpus CAES, considerándolo un conjunto de ejemplos claros para el modelo. Para realizar el *fine-tuning*, se le proporciona al modelo una entrada-salida especificando la entrada y el tipo de salida que queremos que nos proporcione. En este caso, el *prompt* es el que se ha mencionado en el apartado anterior. Por otro lado, la salida que se le ha solicitado ha sido el nivel de español. Asimismo, para realizar el *fine-tuning*, se ha reservado una parte del conjunto de datos para verificar la respuesta.

Asimismo, las pruebas 2 y 3 se diferencian entre sí por la división que se ha realizado del corpus. Para la segunda prueba, se realizó una división del corpus en 90 %, 5 % y 5 %. El 90 % del corpus se utilizó para entrenar el modelo, un 5 % para validarlo y el 5 % restante para testearlo. En la tercera prueba, se realizó una división de 80 % - 20 %, utilizando el 80 % del corpus para el entrenamiento y el 20 % para el testeo.

Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), pp. 353-379. https://doi.org/10.5944/ried.45486

RESULTADOS

Para analizar los resultados, se han empleado tres medidas de evaluación ampliamente utilizadas en tareas de clasificación:

- **Precisión** (*precision*): indica el porcentaje de ejemplos que el modelo clasificó en un determinado nivel y que realmente pertenecen a ese nivel.
- **Cobertura** (*recall*): señala la proporción de ejemplos de un nivel específico que el modelo identificó correctamente. Por ejemplo, el porcentaje de textos de nivel A1 detectados como A1 respecto al total de textos A1 existentes en el corpus.
- **F1-score**: valor único que combina precisión y cobertura mediante su media armónica, proporcionando una medida equilibrada del rendimiento. Esta métrica es especialmente útil cuando es importante que el modelo no solo acierte, sino que también detecte todos los casos posibles de cada categoría.

A continuación, se presentan los resultados de las tres experimentaciones:

Experimentación *zero-shot learning*

Como se ha mencionado anteriormente, en esta técnica, al modelo no se le proporciona ningún ejemplo, se realiza mediante el *prompt* elaborado. En este caso los resultados que se han obtenido son los siguientes:

Tabla 2
Resultados experimentación ZSL

Etiquetas	Precisión	Cobertura	F1-score
A1	0.9375	0.1402	0.2439
A2	0.3333	0.7677	0.4648
B1	0.2400	0.2609	0.2500
B2	0.4286	0.8110	0.1364
C1	0.0000	0.0000	0.0000

La Tabla 2 muestra que, en el nivel A1, la precisión es alta, es decir, los textos que clasifica como nivel A1 tienen una probabilidad alta de ser de este nivel. Sin embargo, la cobertura es bastante baja, el modelo ha detectado como A1 un porcentaje bajo de textos del corpus en general. Por tanto, el F1-score da un resultado bajo.

En el nivel A2, ocurre el fenómeno opuesto al descrito en el nivel A1. La precisión es baja, por lo que acierta en menor medida, sin embargo, la cobertura es alta. Por lo tanto, podemos decir que, en este nivel, aunque detecta en mayor medida los textos de A2, tiene una baja precisión a la hora de detectar el nivel correcto.

En relación con el nivel B1, ambos factores, precisión y cobertura, son bajos. En este nivel, el modelo tiene problemas tanto para detectar los textos de nivel B1 como para identificarlos en su nivel correcto.

Con respecto al nivel B2, se observa que la cobertura es alta, detecta de la mayoría de los textos de nivel B2, no obstante, la precisión es moderada, no acierta en la mayoría de las clasificaciones.

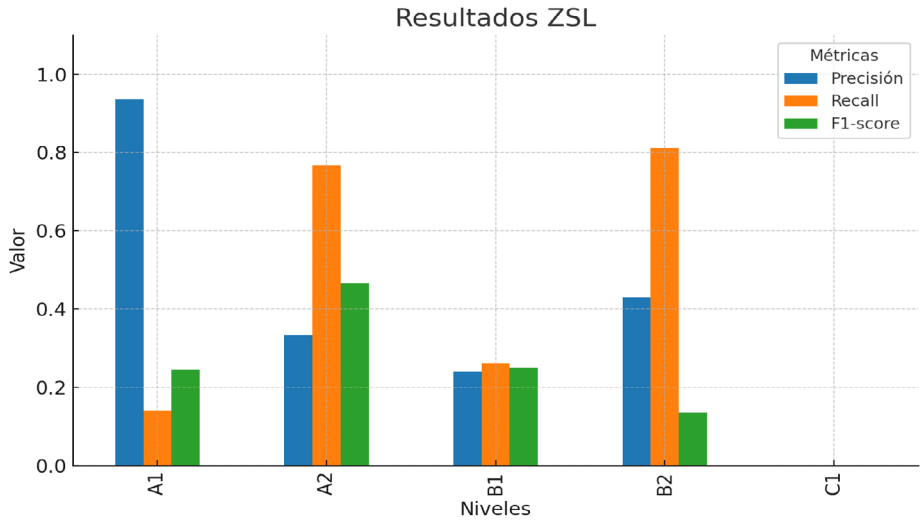
Por último, es llamativo el caso del nivel C1, ya que tanto precisión como cobertura son 0, no detecta ningún texto de dicho nivel.

Asimismo, al tratarse de una experimentación a través de un *prompt*, el modelo no solo dio en sus respuestas el nivel del texto, sino que también añadió comentarios en cada una sobre los errores encontrados. Ejemplos completos de estas respuestas, que incluyen los textos originales y las correcciones propuestas por el modelo, se presentan en el Anexo 1.

La Figura 1 complementa esta información mostrando de forma comparativa los valores de precisión, cobertura y F1-score para cada nivel evaluado en la configuración *zero-Shot Learning*.

Figura 1

Resultados por nivel en la experimentación zero-shot learning



Cantero Romero, M.-V., Martín-Valdivia, M.-T., Ortiz-Colón, A. M., & Jiménez-Zafra, S. M. (2026). Measuring writing skills in Spanish as a foreign language with generative artificial intelligence [Medición de la habilidad escrita en español como lengua extranjera con inteligencia artificial generativa]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), pp. 353-379. https://doi.org/10.5944/ried.45486

Experimentación *fine-tuning 90-5-5*

En esta segunda experimentación, como ya se ha mencionado anteriormente, se ha realizado un *fine-tuning* del corpus y se ha dividido en 90 %, 5 % y 5 % para el entrenamiento, validación y prueba del modelo. A continuación, se pueden observar los resultados de esta experimentación:

Tabla 3
Experimentación fine-tuning 90-5-5

Etiquetas	Precisión	Cobertura	F1-score
A1	0.9905	0.9720	0.9811
A2	0.9519	1.0000	0.9754
B1	1.0000	0.9565	0.9778
B2	1.0000	1.0000	1.0000
C1	1.0000	1.0000	1.0000

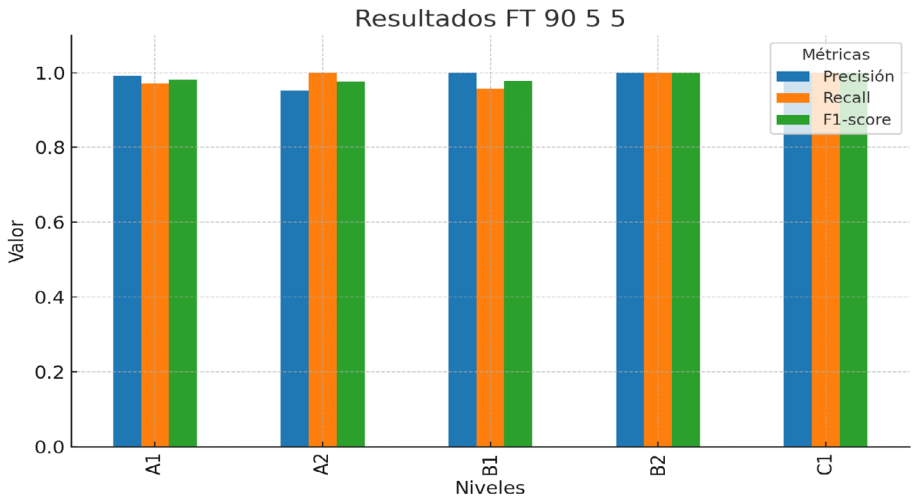
La Tabla 3 evidencia que los resultados presentan niveles de precisión y cobertura más altos que los realizados con el *zero-shot learning*. En el nivel A1, se puede ver que el modelo casi no comete errores y que detecta casi todos los textos del nivel A1.

En el nivel A2, como podemos apreciar, el modelo detecta todos los textos del nivel A2, con una precisión muy alta.

Con respecto al nivel B1, el modelo predice de manera correcta todos los textos, asimismo, cuenta con una cobertura alta, con un pequeño porcentaje no detectado.

Por último, con los niveles B2 y C1, los resultados muestran que el modelo detecta todos los textos de estos niveles y acierta en todas las ocasiones. No obstante, los resultados de este experimento demuestran que el modelo puede predecir correctamente todos los niveles con valores cercanos o iguales a 1. La Figura 2 complementa esta información mostrando de forma comparativa los valores de precisión, cobertura y F1-score para cada nivel evaluado en la configuración *fine-tuning 90-5-5*.

Figura 2
Resultados por nivel en la experimentación fine-tuning 90-5-5



Experimentación *fine-tuning 80-20*

En la tercera experimentación, se realizó un *fine-tuning* dividiendo el corpus en dos porcentajes, un 80 % para el entrenamiento y un 20 % para la validación del modelo. A continuación, se pueden observar los resultados obtenidos.

En esta experimentación, como refleja la Tabla 4, en el nivel A1, tanto la precisión como la cobertura muestran un nivel elevado de aciertos y de detección de textos.

Los niveles A2 y B1 muestran un resultado similar al de A1, aun teniendo un resultado ligeramente inferior en la precisión en el nivel A2 y en la cobertura en el nivel B1.

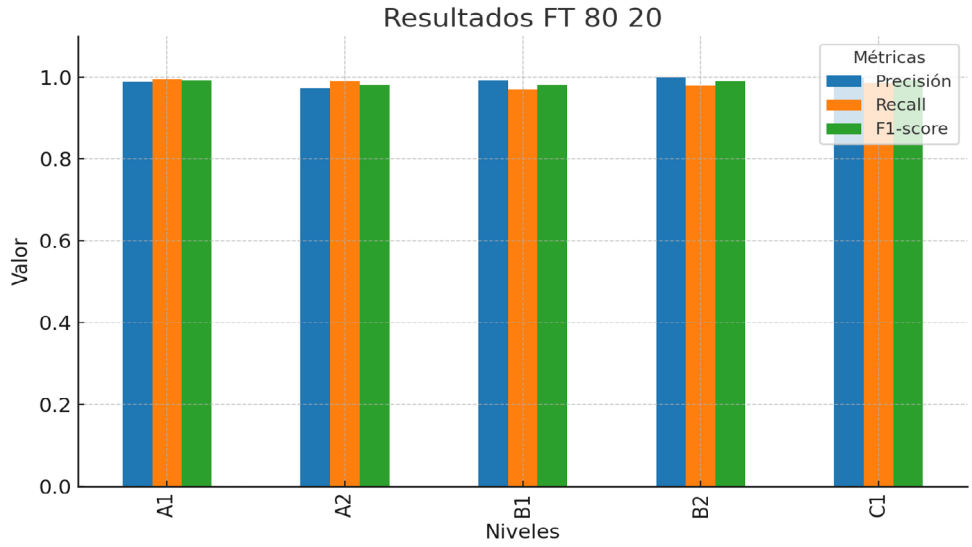
Con respecto a los niveles B2 y C1, observamos que la precisión del modelo es excelente, puesto que predice de manera correcta estos niveles, aunque la cobertura sea moderadamente menor.

Tabla 4
Experimentación fine-tuning 80-20

Etiquetas	Precisión	Cobertura	F1-score
A1	0.9884	0.9953	0.9918
A2	0.9727	0.9899	0.9812
B1	0.9925	0.9707	0.9815
B2	1.0000	0.9795	0.9896
C1	1.0000	0.9859	0.9929

La Figura 3 complementa esta información mostrando de forma comparativa los valores de precisión, cobertura y F1-score para cada nivel evaluado en la configuración *fine-tuning 80-20*.

Figura 3
Resultados por nivel en la experimentación fine-tuning 80-20



Análisis estadístico

Para evaluar la significación estadística de los resultados, se estimaron intervalos de confianza al 95 % (IC95 %) de los valores Macro-F1 mediante **bootstrap**, una técnica de remuestreo con reemplazo que permite calcular márgenes de confianza sin necesidad de asumir una distribución estadística específica. Se realizaron 1 000 réplicas para obtener estos intervalos, los cuales indican el rango dentro del cual se espera que se encuentre el valor verdadero del Macro-F1 con un 95 % de probabilidad. Adicionalmente, se llevaron a cabo pruebas de hipótesis para contrastar los resultados. El experimento *zero-shot learning* obtuvo un valor Macro-F1 de 0,2190 con un intervalo de confianza IC95 % = [0,1785, 0,2617], mientras que los experimentos con *fine-tuning* alcanzaron valores cercanos a la perfección: un valor Macro-F1 de 0,9869 con un intervalo de confianza IC95 % = [0,9748, 0,9959] en la partición 90-5-5 y un valor Macro-F1 de 0,9874 (IC95 % = [0,9812, 0,9933]) en la partición 80-20. Las comparaciones se realizaron mediante pruebas de hipótesis no paramétricas basadas en bootstrap —es decir, contrastes estadísticos que no requieren asumir una distribución concreta de los datos y que se apoyan en múltiples remuestreos aleatorios—. Estos análisis mostraron que ambos experimentos con *fine-tuning* superan de forma estadísticamente significativa al experimento *zero-shot* ($\Delta^1 \approx 0{,}77$; $p < 0{,}001$ en ambos casos). Por el contrario, no se observaron diferencias significativas entre las dos experimentaciones de *fine-tuning* ($\Delta = -0{,}0005$; IC95 % = [−0,0133, 0,0113]; $p = 0{,}964$). Estos resultados confirman que el *fine-tuning* mejora sustancialmente la capacidad del modelo para evaluar la competencia escrita en ELE.

PERTINENCIA PEDAGÓGICA DEL MODELO Y APLICACIÓN TANTO EN CONTEXTOS PRESENCIALES COMO A DISTANCIA

Los resultados obtenidos muestran que un modelo de lenguaje ajustado con un corpus especializado puede alcanzar un rendimiento muy alto en la clasificación automática de textos escritos según los niveles del MCER. Esta capacidad tiene un claro potencial para facilitar la labor docente, especialmente en la etapa inicial de colocación del estudiantado, constituyendo una innovación pedagógica en el uso de inteligencia artificial para la enseñanza de lenguas.

En un entorno de aprendizaje real, el sistema podría integrarse en un *Learning Management System* (LMS) como módulo de evaluación diagnóstica. El flujo de uso sería sencillo: el estudiante envía un texto, el modelo lo clasifica automáticamente por nivel y presenta el resultado al docente mediante una rúbrica alineada con los descriptores del Plan Curricular del Instituto Cervantes (PCIC). De esta forma, el profesorado podría asignar al estudiante el curso correspondiente a su nivel real, optimizando el ajuste de los grupos y evitando desfases que puedan afectar al progreso de aprendizaje.

En implementaciones futuras, y a partir de análisis más exhaustivos, el sistema podría generar informes más detallados basados en el PCIC, identificando áreas lingüísticas específicas a reforzar (por ejemplo, uso de tiempos verbales, cohesión textual o repertorio léxico). Esta información permitiría que el curso no solo se ajuste al nivel del alumno, sino también a sus principales carencias, orientando la programación didáctica hacia la mejora de dichos aspectos.

En contextos de educación a distancia y aprendizaje en línea, la integración en un LMS tendría la misma función diagnóstica, clasificando automáticamente al alumnado desde el primer acceso. Esto permitiría mantener la eficiencia en la organización de grupos incluso en entornos sin contacto presencial, algo especialmente relevante en cursos masivos o programas virtuales de matrícula continua.

En comparación con herramientas como *Write & Improve* (Cambridge English, s. f.), centradas principalmente en retroalimentación correctiva y evaluación formativa, la propuesta aquí presentada aporta un enfoque complementario: la clasificación automática por niveles MCER. Esta función, combinada con la posibilidad de integración directa en plataformas educativas, refuerza su valor como herramienta de apoyo docente para optimizar la colocación inicial y agilizar la labor de planificación de los cursos.

En todos los casos, el sistema se concibe como apoyo docente y no como sustituto, con el objetivo de agilizar las tareas de evaluación inicial y liberar tiempo para actividades de mayor valor añadido, como la retroalimentación individualizada o el seguimiento del progreso.

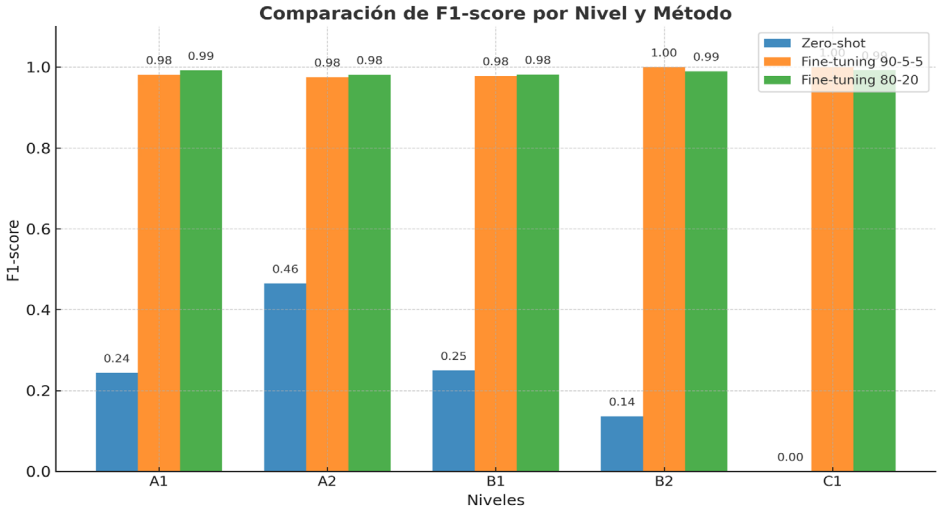
Entre las posibles mejoras futuras, se plantea la ampliación del corpus a más géneros y temáticas, la inclusión de descriptores léxicos y pragmáticos adicionales, la experimentación con otros modelos de lenguaje y la adaptación del sistema a otros idiomas y niveles educativos. También resultaría pertinente evaluar su impacto real en el rendimiento y la motivación del alumnado, así como su integración fluida en distintos LMS y contextos educativos.

DISCUSIÓN Y CONCLUSIONES

Una vez realizadas las pruebas y obtenidos los resultados, se observa un contraste claro entre el desempeño del modelo en la configuración *zero-shot learning* y en las experimentaciones con *fine-tuning*. En la primera, el modelo no logra una precisión adecuada en ninguno de los niveles, destacando únicamente un valor relativamente alto en A1, aunque con una cobertura muy baja. En los niveles A2 y B2, la cobertura es medio-alta (0,76 y 0,81), lo que indica que el modelo detecta la mayoría de los textos de esos niveles, pero con baja precisión. El nivel C1 es especialmente problemático, con valores nulos tanto en precisión como en cobertura. Estos datos confirman que, con su preentrenamiento original, el modelo no es capaz de nivelar textos de ELE de forma fiable siguiendo el Plan Curricular del Instituto Cervantes.

En contraste, las dos pruebas de *fine-tuning* (90-5-5 y 80-20) proporcionaron valores de precisión y cobertura muy próximos a 1 en todos los niveles, con especial solidez en B2 y C1. La comparación entre ambos experimentos sugiere que un mayor volumen de datos de entrenamiento puede favorecer el rendimiento, aunque incluso con menos datos el modelo mantiene una alta capacidad de predicción. Este comportamiento coincide con lo observado en trabajos previos sobre adaptación de modelos para tareas específicas de procesamiento del lenguaje natural (García-Peñalvo, 2024; García-Peñalvo et al., 2024), donde la personalización del sistema incrementa notablemente su eficacia.

Figura 4
Comparación resultados de los experimentos



Si se compara con herramientas como *Write & Improve* (Cambridge English, s. f.), centradas en retroalimentación formativa y detección de errores, el presente estudio aporta un enfoque complementario: la clasificación automática por niveles MCER, que podría integrarse en entornos de aprendizaje para optimizar la colocación inicial de estudiantes y agilizar la labor docente. Asimismo, estudios como los de Burstein et al. (2003, *e-rater*) o McNamara et al. (2014, *Coh-Metrix*) ya habían mostrado la utilidad de combinar características lingüísticas y métricas automáticas para evaluar la escritura; nuestros resultados refuerzan esta línea, evidenciando que un modelo ajustado con corpus específicos alcanza niveles de rendimiento muy altos.

Por otro lado, en la experimentación *zero-shot learning*, el modelo, además de indicar el nivel, generó comentarios sobre errores y sugerencias de corrección

de los textos. Esta funcionalidad, aunque no relevante para el objetivo central de este estudio, podría explorarse en futuras investigaciones como recurso de apoyo didáctico para que los estudiantes identifiquen y corrijan sus errores.

En cuanto a las limitaciones, el estudio se ha realizado sobre un único corpus debido a la escasez de corpus especializados y nivelados de ELE que cumplan criterios homogéneos y controlados por expertos. Asimismo, la clasificación se ha basado en descriptores gramaticales presentes en el *prompt* (por ejemplo, repertorio verbal, número promedio de palabras por oración, uso de ciertos tiempos y modos), sin integrar de forma sistemática estructuras sintácticas complejas, recursos léxicos específicos o aspectos pragmáticos como la adecuación o la coherencia discursiva, que resultan especialmente relevantes en niveles intermedios y avanzados.

Como líneas de trabajo futuro, se propone:

1. Diseñar y validar nuevos corpus de ELE nivelados por expertos, con mayor diversidad temática y representatividad de niveles.
2. Explorar la aplicación del método con otros modelos de lenguaje para contrastar su rendimiento.
3. Ampliar el conjunto de descriptores lingüísticos, incorporando indicadores sintácticos, léxicos y pragmáticos.
4. Experimentar con variaciones controladas del *prompt* para evaluar su impacto en la clasificación.
5. Validar el sistema con textos auténticos de estudiantes en contextos educativos reales e integrar la herramienta en plataformas de gestión del aprendizaje (LMS).

En conjunto, los resultados alcanzados confirman que la inteligencia artificial generativa y, en particular, los modelos de lenguaje ajustados con datos específicos pueden convertirse en un recurso eficaz para agilizar la nivelación inicial del estudiantado de ELE, apoyando la labor docente y optimizando los procesos de enseñanza-aprendizaje.

Dentro de esta reflexión resulta pertinente abordar también los aspectos éticos y legales implicados en el uso de corpus de aprendientes, que se desarrollan en el subapartado siguiente.

Aspectos éticos y de licencia de uso

En el presente estudio se ha utilizado el Corpus de Aprendices de Español (CAES), que se encuentra disponible en línea para fines académicos y de investigación (Universidad de Santiago de Compostela, s. f.). Sus textos han sido anonimizados previamente, de modo que no incluyen datos personales identificables, y fueron recopilados en contextos educativos regulados y con validación experta, lo que garantiza un tratamiento adecuado de la información.

Desde un punto de vista ético, resulta necesario considerar que la ampliación o combinación del corpus con otros recursos debe atender tanto a la claridad en las licencias de uso como a la prevención de sesgos asociados a la lengua materna o al contexto sociocultural. Estos factores son esenciales para asegurar una clasificación justa y reproducible de los textos y, en última instancia, para garantizar un uso responsable de la inteligencia artificial en entornos educativos.

Agradecimientos

Esta publicación del proyecto Desarrollo de Modelos ALIA está financiada por el Ministerio para la Transformación Digital y de la Función Pública y por el Plan de Recuperación, Transformación y Resiliencia –Financiado por la Unión Europea– NextGenerationEU. Este trabajo también ha sido parcialmente financiado por el Proyecto CONSENSO (PID2021-122263OB-C21), el Proyecto MODERATES (TED2021-130145B-I00) y el Proyecto SocialTox (PDC2022-133146-C21), financiados por MCIN/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR, el proyecto ROMANET (CERV-2024-CHAR-LITI-101215052), financiado por la Unión Europea en el marco del programa Ciudadanos, Igualdad, Derechos y Valores, y el proyecto HEART-NLP-UJA (PID2024-156263OB-C21), financiado por MICIU/AEI/10.13039/501100011033 y por FEDER/UE. El trabajo de investigación realizado por Salud María Jiménez-Zafra es parte de la ayuda RYC2023-044481-I, financiada por MICIU/AEI/10.13039/501100011033 y por el FSE+.

NOTAS

1. Para cuantificar las diferencias entre modelos, se utilizó el símbolo Δ (delta), que en estadística se emplea convencionalmente para indicar la diferencia entre dos valores. Así, Δ=Macro-F1(fine-tuning) – Macro-F1(zero-shot) ≈ 0,77 significa que la diferencia de Macro-F1 entre los experimentos comparados es aproximadamente 0,77.

REFERENCIAS

Aparicio Gómez, W. O. (2023). La inteligencia artificial y su incidencia en la educación: transformando el aprendizaje para el siglo XXI. *Revista Internacional de Pedagogía e Innovación Educativa*, *3*(2), 217-229. https://dialnet.unirioja.es/servlet/articulo?codigo=9624350

Area-Moreira, M., Del Prete, A., Sanabria-Mesa, A. L. y Sannicolás-Santos, M. B. (2024). No todas las herramientas de IA son iguales: análisis de aplicaciones inteligentes para la enseñanza universitaria. *Digital Education Review*, *45*, 141-149. https://doi.org/10.1344/der.2024.45.141-149

Barroso-Osuna, J. y Cabero-Almenara, J. (2025). Potencialidades de la inteligencia artificial en la personalización de la educación. En P. Román-Graván, J. Barroso-Osuna, J. Cabero-Almenara y C. Llorente-Cejudo (Eds.), *Visiones sobre la integración educativa de la inteligencia*

artificial (1.ª ed.). Dykinson. https://doi.org/10.14679/4177

Baskara, R. y Mukarto, M. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *7*(2), 343-358. https://doi.org/10.21093/ijeltal.v7i2.1387

Biedma Torrecillas, A., Chamorro Guerrero, M. D., Lozano, G. y Sánchez Cuadrado, A. (2012). Diseño y validación de las pruebas de nivel del CLM de la Universidad de Granada. En *Actas del VII Congreso ACLES: Multilingüismo en los centros de lengua universitarios: evaluación, acreditación, calidad y política lingüística* (pp. 26-37). ACLES. https://dialnet.unirioja.es/servlet/libro?codigo=501925

Bolaño-García, M. y Duarte-Acosta, N. (2024). Una revisión sistemática del uso de la inteligencia artificial en la educación. *Revista Colombiana de Cirugía, 39*(1), 51-63. https://doi.org/10.30944/20117582.2365

Burstein, J., Elliot, N., Beigman Klebanov, B., Madnani, N., Napolitano, D., Schwartz, M., Houghton, P. y Molloy, H. (2018). Writing mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, *2*, 285-313. https://doi.org/10.37514/JWA-J.2018.2.1.12

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L. y Harris, M. D. (2003). E-rater as a diagnostic tool for writing instruction. En *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations* (pp. 79-81). Association for Computational Linguistics.

Cambridge English. (s. f.). *Write & Improve*. https://writeandimprove.com

Cantero, M. V. (2024). Aproximación a un posible uso de ChatGPT para nivelar la expresión escrita en ELE. En F. M. Sirignano, R. Martínez Roig y A. López Padrón (Eds.), *Enseñanza y aprendizaje en la era digital desde la investigación y la innovación* (pp. 55-64). Octaedro.

Centro Virtual Cervantes. (s. f.). *Ítem de respuesta cerrada*. https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/itemrespuestacerrada.htm

Chan, C. K. Y. y Tsi, L. H. Y. (2023). *The AI revolution in education: Will AI replace or assist teachers in higher education* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2305.01185

Columbia University, Department of Latin American and Iberian Cultures. (s. f.). *Spanish placement exam*. Recuperado el 25 de julio de 2025, de https://laic.columbia.edu/content/spanish-second-language-placement-exam

Consejo de Europa. (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Instituto Cervantes; Ministerio de Educación, Cultura y Deporte. https://cvc.cervantes.es/ensenanza/biblioteca_ele/marco/cvc_mer.pdf

Crespo Mendoza, R., Rodríguez López, W., Montenegro Patrel, M. y Tomalá Tomalá, G. (2024). IA: una herramienta para asistir a los docentes en la evaluación de los estudiantes. *Conocimiento Global, 9*(2), 305-323. https://doi.org/10.70165/cglobal.v9i2.423

Fajardo, G. M., Ayala, D. C., Arroba, E. M. y López, M. (2023). Inteligencia artificial y la educación universitaria: una revisión sistemática. *Magazine de las Ciencias: Revista de Investigación e Innovación, 8*(1), 109-131. https://doi.org/10.33262/rmc.v8i1.2935

García-Peñalvo, F. J. (2024). Cómo afecta la inteligencia artificial generativa a los procesos de evaluación. *Cuadernos de Pedagogía, (549)*.

García-Peñalvo, F. J., Llorens-Largo, F. y Vidal, J. (2024). La nueva realidad de la educación ante los avances de la

inteligencia artificial generativa. *RIED-Revista Iberoamericana de Educación a Distancia, 27*(1), 9-39. https://doi.org/10.5944/ried.27.1.37716

Hernández-León, N. y Rodríguez-Conde, M. J. (2024). Inteligencia artificial aplicada a la educación y la evaluación educativa en la universidad: Introducción de sistemas de tutorización inteligentes, sistemas de reconocimiento y otras tendencias futuras. *Revista de Educación a Distancia (RED), 24*(78), Artículo 6. https://doi.org/10.6018/red.594651

Hong, W. C. H. (2023). *The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. Journal of Educational Technology and Innovation, 5*(1), 38-53. https://doi.org/10.61414/jeti.v5i1.103

Instituto Cervantes. (2006). *Plan curricular del Instituto Cervantes: Niveles de referencia para el español* (3 vols.). Biblioteca Nueva. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Li, Y. (2023). *A practical survey on zero-shot prompt design for in-context learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.13205

McNamara, D. S., Graesser, A. C., McCarthy, P. M. y Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Morales-Chan, M. A. (2023). *Explorando el potencial de ChatGPT: Una clasificación de prompts efectivos para la enseñanza*. Universidad Galileo. https://biblioteca.galileo.edu/tesario/handle/123456789/1348

Moreno, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación en Tecnologías de la Información, 7*(14), 260-270. https://doi.org/10.36825/riti.07.14.022

OpenAI. (2022). *ChatGPT (versión 3.5) [Modelo de lenguaje de inteligencia artificial]*. https://openai.com

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O. y Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(8), em2307. https://doi.org/10.29333/ejmste/13428

Palacios Martínez, I., Barcala Rodríguez, F. M. y Rojo, G. (2019). El corpus de aprendices de español (CAES) y sus aplicaciones para la enseñanza y aprendizaje del español como lengua extranjera. En M. Blanco, H. Olbertz y V. Vázquez Rozas (Eds.), *Corpus y construcciones: Perspectivas hispánicas* (pp. 273-301). Universidade de Santiago de Compostela (Verba, Anexo 79). https://doi.org/10.15304/9788417595876

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R. y Lim, C. P. (2023). A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(4), 4051-4070. https://doi.org/10.1109/TPAMI.2022.3182926

Roumeliotis, K. I., Tselikas, N. D. y Nasiopoulos, D. K. (2024). Next-generation spam filtering: Comparative fine-tuning of LLMs, NLPs, and CNN models for email spam classification. *Electronics, 13*(11), 2034. https://doi.org/10.3390/electronics13112034

Salguero Romero, P. (2023). *La traducció pedagògica i l'ús de ChatGPT-3 a classes d'anglès com a segona llengua per a nens i nenes* [Trabajo de fin de grado, Universitat Autònoma de Barcelona]. Repositorio UAB. https://ddd.uab.cat/record/279383

Universidad de Santiago de Compostela. (s. f.). *Corpus de aprendices de español (CAES)*. https://galvan.usc.es/caes

University of Wisconsin–Madison, Testing and Evaluation Services. (s. f.). *Spanish*

placement test. University of Wisconsin–Madison. https://testing.wisc.edu/centerpages/spanishtest.html

Wang, L. (2024). *El uso de ChatGPT y Gemini para la preparación de actividades de comprensión oral de CLE: Un estudio comparativo* [Trabajo de fin de máster, Universitat Autònoma de Barcelona]. Repositorio UAB. https://ddd.uab.cat/record/300353

Wang, X., Lee, Y. y Park, J. (2022). *Automated evaluation for student argumentative writing: A survey* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.04083

Zambrano Campozano, R. F. (2025). Impacto de la inteligencia artificial en la transformación digital de la educación. *Serie Científica de la Universidad de las Ciencias Informáticas, 18*(1), 177-192. https://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2306-24952025000100177

Zhang, S. (2021). Review of automated writing evaluation systems. *Journal of China Computer-Assisted Language Learning, 1*(1), 170-176. https://doi.org/10.1515/jccall-2021-2007

Zapata Ros, M. (2024). IA generativa y ChatGPT en educación: Un reto para la evaluación y ¿una nueva pedagogía? *Revista Paraguaya de Educación a Distancia (REPED), 5*(1), 12-44. https://doi.org/10.56152/reped2024-vol5num1-art2

ANEXOS

Anexo 1. Ejemplos completos de respuestas del modelo

Texto del corpus:

"Hola Eva.
Que hace dos meses, hemos estudiado espanol en madrid.
He vuelt a Japon la semana pasada.
Glacias por tu ayuda.
Yo escribo mi familia, mi mujer, hija y hijo.
Mi mujer trabaya en Idioma.
Hija trabaya en una tienda de palo, hijo en un enterprise electrolico.
Vivimos en yokohama cerca de tokyo.
Vamos a ir a madrid que mes viene
Por favor, vernos en madrid.
Hasta plonto." (CAES, 2022)

Respuesta del modelo:

Errores encontrados en el texto:

- "Que hace dos meses, hemos estudiado espanol en madrid." La forma correcta sería "Hace dos meses, estudiamos español en Madrid."
- "He vuelt a Japon la semana pasada." La forma correcta sería "He vuelto a Japón la semana pasada."
- "Glacias por tu ayuda." Debería ser "Gracias por tu ayuda."
- "Yo escribo mi familia, mi mujer, hija y hijo." Se debería decir "Te escribo sobre mi familia: mi mujer, mi hija y mi hijo."

- “Mi mujer trabaya en Idioma.” La forma correcta sería “Mi mujer trabaja en Idiomas.”
- “Hija trabaya en una tienda de palo, hijo en un enterprise electrolico.” Debería ser “Mi hija trabaja en una tienda de ropa, y mi hijo en una empresa electrónica.”
- “Vamos a ir a madrid que mes viene.” La forma correcta sería “Vamos a ir a Madrid el mes que viene.”
- “Hasta plonto.” Debería ser “Hasta pronto.”

NIVEL: A2 (OpenAI, 2022)

—

Texto del corpus:

“El 24 de Noviembre

Estimados señores

El 22 de noviembre pasado, viajé con Tonta Airlines de Badaros a Madrid con el vuelo TA 222.

Estoy aquí en Madrid hasta el 27 para algunas citas profesionales.

Cuando embarqué en Badaros, la azafata me dijo que mi maleta estaba demasiado grande y que tenia que facturarla.

La maleta esta de marca Samsonite, de modelo XYZ1234, color negra, con mis nombre y dirección muy visibles.

Cuando llegamos a Madrid, la maleta no fue encontrada.

Hace a dos dias que estoy esperando pero nadie de su compañía me ha llamado con información.

Evidentemente, no puedo quedar sin cambiarme o lavarme me.

Entonces compré algunas cosas y me parece que Tonta Airlines tenga que indemnizarme.

Yo incluido la lista de mis compras y les pide que se pongan en contacto para decirme si sean de acuerdo de pagar el 50% des estas compras.

Se pueden contactarme por correo electronico o llamarme en el movil 660 12 34 45 y evidentemente, en cualquier caso que Tonta Airlines encuentre mi maleta, me daría mucha alegria.

Un saludo” (CAES, 2022)

Respuesta del modelo:

El texto presenta varios errores gramaticales y algunos elementos propios de un nivel de español como lengua extranjera B1:

- Uso incorrecto de tiempos verbales: “tenia” en lugar de “tenía”.
- Errores en el uso de los artículos: “la maleta esta” en lugar de “la maleta está”.
- Uso incorrecto de pronombres: “Hace a dos días” en lugar de “Hace dos días”.
- Errores de concordancia de género y número: “me parece que Tonta Airlines tenga que indemnizarme” debería ser “me parece que Tonta Airlines debe indemnizarme”.

NIVEL: B1 (OpenAI, 2022)

Fecha de recepción del artículo: 1 de junio de 2025
Fecha de aceptación del artículo: 9 de septiembre de 2025
Fecha de aprobación para maquetación: 8 de octubre de 2025
Fecha de publicación en OnlineFirst: 28 de octubre de 2025
Fecha de publicación: 1 de enero de 2026