Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 53-63 recibido 01-04-2025 revisado 05-05-2025 aceptado 17-05-2025

GPT-3.5 para la Nivelación de ELE en un Entorno Universitario: Un Estudio Comparativo de Zero-shot Learning y Fine-tuning

GPT-3.5 for ELE Leveling in a University Environment: A Comparative Study of Zero-shot Learning and Fine-tuning

María Victoria Cantero-Romero,[1] María Estrella Vallecillo-Rodríguez,[2] Ana María Ortiz-Colón,[3] Salud María Jiménez-Zafra[2]

[1]Departamento de Filología Española, SINAI, CEATIC, Universidad de Jaén, España
[2]Departamento de Informática, SINAI, CEATIC, Universidad de Jaén, España
[3]Departamento de Pedagogía, Universidad de Jaén, España
{vcantero, mevallec,aortiz, sjzafra}@ujaen.es

Resumen: El proceso de nivelación de estudiantes extranjeros en cursos de español como lengua extranjera (ELE) es fundamental para garantizar una enseñanza acorde con sus competencias lingüísticas. Actualmente, la nivelación se realiza mediante la evaluación manual de pruebas iniciales de vocabulario, escritura, comprensión oral y comprensión escrita, lo que supone una alta carga de trabajo y retrasos en la clasificación, especialmente cuando llega una época con un gran número de estudiantes no nativos del idioma. El presente estudio explora el uso del Procesamiento del Lenguaje Natural (PLN) para automatizar la evaluación de la expresión escrita, investigando la aplicación de GPT-3.5, un modelo del lenguaje con amplio conocimiento del español, para clasificar textos según su nivel ELE. Se implementaron dos enfoques: Zero-shot Learning (ZSL), donde el modelo recibe instrucciones explícitas para identificar y justificar la clasificación de los textos, y un método supervisado, entrenando el modelo con el corpus CAES, que contiene textos nivelados desde A1 hasta C1. La evaluación se realizó con textos extraídos de clases de español para estudiantes extranjeros de una universidad, y los resultados muestran que el entrenamiento supervisado mejora significativamente la precisión del modelo, permitiéndole capturar las diferencias sutiles entre niveles. Los hallazgos destacan la necesidad de continuar investigando para optimizar los sistemas de clasificación automática en ELE y su papel en la nivelación de estudiantes no nativos.
Palabras clave: Español como Lengua Extranjera, nivelación automática de idiomas, Grandes Modelos de Lenguaje, aprendizaje supervisado.

Abstract: The process of leveling foreign students in Spanish as a Foreign Language (ELE in Spanish) courses is essential to ensure instruction that matches their linguistic competencies. Currently, this leveling is carried out through the manual evaluation of initial tests assessing vocabulary, writing, oral comprehension, and reading comprehension, which entails a high workload and delays in classification, especially with a large number of students. This study explores the use of Natural Language Processing (NLP) to automate the assessment of written expression, investigating the application of GPT-3.5, a large language model with extensive knowledge of Spanish, to classify texts according to their ELE level. Two approaches were implemented: Zero-shot Learning (ZSL), where the model receives explicit instructions to identify and justify the classification of texts, and a supervised method, where the model is trained with the CAES corpus, which contains leveled texts from A1 to C1. The evaluation was conducted using texts from a university environment (anonymized for review), and the results show that supervised training significantly improves the model's accuracy, enabling it to capture subtle differences between levels. These findings highlight the need for continued research to optimize automatic classification systems in ELE and their role in leveling non-native students.
Keywords: Spanish as a Foreign Language, automatic language leveling, Large Language Models, supervised learning.

ISSN 1135-5948 DOI 10.26342/2025-75-4

1 Introducción

La enseñanza de español como lengua extranjera (ELE) se enmarca dentro de la lingüística aplicada y se centra en la enseñanza y aprendizaje de español en hablantes no nativos. Esta enseñanza se presenta en distintos contextos y con diferentes objetivos, siendo el desarrollo de las habilidades lingüísticas el eje principal.

En el presente estudio nos centraremos en la expresión escrita de aprendices de español y más concretamente en un ámbito universitario.

Las universidades ofrecen cursos de español a estudiantes visitantes no nativos como los procedentes de becas Erasmus o programas Talentium. Dichos estudiantes pueden o no tener conocimientos previos de español y no todos cuentan con una certificación, por lo que es necesario realizar una evaluación diagnóstica inicial para determinar el nivel adecuado del estudiantado.

Sin embargo, el elevado volumen de estudiantes y el tiempo limitado para poder clasificarlos dificulta dicho proceso. Asimismo, la expresión escrita es la habilidad más laboriosa de nivelar, puesto que conlleva un análisis individual de cada texto presentado frente a la expresión oral que se nivela de manera inmediata.

Dada la necesidad de optimizar la nivelación de estudiantes extranjeros en cursos de ELE, resulta fundamental contar con herramientas que asistan a los evaluadores en dicho proceso. En este contexto, el Procesamiento del Lenguaje Natural (PLN) se presenta como una solución clave, ya que no solo permite evaluar la calidad sintáctica y semántica de los textos, sino también desarrollar sistemas capaces de identificar patrones característicos de cada nivel de ELE y clasificarlos automáticamente.

Hasta donde alcanza nuestro conocimiento, no existen estudios previos que aborden dicha tarea específica aplicando técnicas de PLN y centrados en la nivelación de español como lengua extranjera. Por ello, el presente trabajo propone un enfoque novedoso basado en el uso de grandes modelos de lenguaje (*Large Language Models*, LLMs), los cuales han demostrado un alto grado de comprensión del español en tareas como la generación de contranarrativas (Vallecillo et al. (2024) o la generación de textos simplificados (Espinosa Zaragoza et al. (2023). Concretamente, implementaremos dos sistemas basados en el modelo GPT-3.5: el primero empleará la técnica de *Zero-shot Learning* (ZSL), donde el modelo recibirá instrucciones explícitas para comprender y ejecutar la tarea de clasificación; el segundo se entrenará sobre un corpus de textos nivelados en ELE, lo que permitirá un aprendizaje supervisado de los distintos niveles lingüísticos. Para evaluar el desempeño de ambos sistemas, utilizaremos un conjunto de textos extraídos de clases universitarias de estudiantes extranjeros, sirviendo dicho conjunto de textos como una evaluación de nuestros sistemas en una situación real.

El resto del artículo está estructurado de la siguiente manera. En la Sección 2, se presenta una revisión del estado de la cuestión en la nivelación automática de idiomas en contextos de aprendizaje de lengua extranjera. La Sección 3 describe la metodología, los conjuntos de datos seleccionados y los experimentos realizados. En la Sección 4, se muestran los resultados obtenidos, seguidos de un análisis detallado en la Sección 5. Finalmente, en la Sección 6, se exponen las conclusiones y las líneas futuras de investigación derivadas del presente trabajo.

2 Estado de la cuestión

Con respecto al uso de los modelos de lenguaje en la enseñanza de español como lengua extranjera, podemos observar diversos estudios centrados en cómo utilizar dichos modelos en el aula, en especial ChatGPT.

En el estudio realizado por García-Peñalvo et al. (2023) se abordan los modelos como generadores de texto para dar ejemplos de lengua a los estudiantes. Asimismo, se añaden otras funciones como la traducción automática. En el presente trabajo se muestra la utilidad de los modelos de lenguaje en el aula de idiomas como un recurso para desarrollar las competencias lingüísticas.

Area-Moreira et al. (2024) apoyan la idea de utilización de los modelos de lenguaje en la clase de idiomas para generar textos y utilizarlos para que los estudiantes se relacionen con un texto próximo al real y, además, añaden que dichos modelos pueden servir para obtener explicaciones y para realizar consultas gramaticales. Asimismo, se indica que los modelos de lenguaje, pueden usarse para la elaboración de material educativo y creación de actividades de aprendizaje.

Para Hong (2023), los modelos ofrecen varias funciones en la enseñanza de idiomas: fomentan el uso auténtico del lenguaje, ya que ChatGPT puede simular las interacciones humanas y permiten a los estudiantes realizar conversaciones auténticas; también, pueden actuar como tutores personales, explicando usos gramaticales e identificando problemas lingüísticos; además, los modelos de lenguaje facilitan el aprendizaje personalizado, al poder adaptar el contenido a las necesidades individuales de cada alumno; por último, se comenta que los modelos pueden ser un apoyo al profesorado ofreciéndo sugerencias, creando planes de lecciones, generando tareas y preguntas y corrigiendo escritura de los estudiantes. Asimismo, se señala que los modelos pueden aliviar la carga de trabajo del profesorado, lo que les permitirá dedicar más tiempo a la planificación de lecciones.

Con respecto a la enseñanza de español como lengua extranjera y el uso de los modelos de lenguaje encontramos varios estudios como el realizado por Román (2023) en el que se explora la capacidad de ChatGPT para conversar con los estudiantes y comprender y resolver sus dudas.

Asimismo, también encontramos estudios más centrados en la generación de textos de comprensión lectora por modelos de lenguaje para la clase de español como lengua extranjera, como los realizados por Cantero (2024a, 2024b y 2025) en los que se analizan textos generados por ChatGPT y Llama y su adecuación con el Plan Curricular del Instituto Cervantes, en concreto, en adjetivos, tiempos verbales y léxico.

Con respecto a la nivelación automática de idiomas encontramos los siguientes estudios. En primer lugar, Yannakoudakis (2011) realizó una investigación con la que se pretendía clasificar textos de estudiantes de inglés como segunda lengua utilizando máquinas de soporte vectorial. En segundo lugar, un trabajo fin de máster de la Universidad del País Vasco realizado por Azurmendi (2024), en el que se evalúa si los textos analizados son de nivel C1 de euskera. Se realiza una clasificación binaria y se utilizan distintos modelos como RoBERTa.

Diversos trabajos, como los de Li et al. (2023) y Pourpanah et al. (2023), estudian distintos tipos de *prompt*, como discretos, continuos, Zero-shot Learning (ZSL) y Few-shot Learning (FSL), aplicados a diferentes tareas de PLN. Ambos estudios destacan la importancia de diseñar una buena instrucción (*prompt*) para aprovechar al máximo el potencial de los LLMs. Roumeliotis et al. (2024) exploran el ajuste de un LLM, como GPT-4, para la clasificación de correos electrónicos de spam, comparando su rendimiento con otros modelos más pequeños y tradicionales, como BERT y RoBERTa. Sus resultados muestran que ajustar un LLM a la tarea mejora la precisión en la detección de spam y reduce los falsos positivos.

A lo largo del trabajo, nos centraremos en diseñar un *prompt* que guíe al modelo en la tarea de nivelación automática de textos. Además, compararemos el rendimiento de un LLM como *gpt-3.5-turbo* utilizando las metodologías mencionadas anteriormente, que han demostrado ser efectivas en otras tareas de PLN. Concretamente, evaluaremos el *prompt* en enfoques ZSL, así como en un enfoque que requiera el ajuste del modelo a los datos específicos de la tarea.

3 Metodología

En el presente estudio, se tiene como objetivo analizar si (1) los grandes modelos de lenguaje son capaces de entender las diferencias entre los distintos niveles ELE y clasificarlos de forma automática. Además, buscamos evaluar (2) si es necesario proporcionar a los sistemas no solo herramientas o instrucciones para identificar las características de cada clase, sino también ejemplos que les ayuden a detectar y aprender las sutilezas de cada nivel o aquellas características más subjetivas asociadas a cada uno. Finalmente, pretendemos (3) identificar los principales desafíos que afrontan los sistemas al abordar la tarea, con el fin de proponer sistemas de mejora. En esta sección, se explicarán en detalle los conjuntos de datos utilizados y el sistema desarrollado para llevar a cabo los experimentos que nos permitirán responder a las preguntas que motivaron el estudio.

3.1 Corpus

En la presente investigación se han utilizado dos corpus distintos. El primero se ha usado para ajustar el modelo, mientras que el segundo se ha empleado para llevar a cabo la evaluación del sistema.

El primer conjunto de datos es el Corpus de Aprendices de Español (Cervantes, s.f.), financiado por el Instituto Cervantes y compilado por la Universidad de Santiago de

Compostela. Cuenta con un total de 3878 textos repartidos en los niveles entre A1 y C1 siguiendo el Marco Común Europeo de Referencia. Se trata de un corpus de textos de expresión escrita de estudiantes de once lenguas distintas. En el presente estudio se ha trabajado con la versión 2.1 de marzo de 2022 que amplía los textos en español de su primera versión.

A continuación, en la Tabla 1, presentamos el contenido del corpus dividido en cinco niveles.

Nivel	Tareas	Muestras	Total Muestras
A1	Email cambio de trabajo	728	2136
	Email a un familiar	703	
	Nota llegar tarde	705	
A2	Biografía persona que admiras	673	1977
	Reserva habitación de hotel	603	
	Postal de vacaciones	701	
B1	Carta a un amigo	528	1364
	Historia graciosa	382	
	Reclamación aerolínea	454	
B2	Solicitud de admisión	375	731
	Fumar en lugares públicos	356	
C1	Reseña de película	184	353
	Reclamación compañía de gas	169	
Total de muestras del corpus: 6571			

Tabla 1: Resumen de muestras corpus CAES.

El segundo corpus que se ha utilizado en el estudio es una recogida de muestras que hemos realizado de un centro de lenguas de una universidad española (en adelante CL). Dicho corpus es un reflejo real del estudiantado representado en los cursos de español como lengua extranjera. Fue compilado durante el mes de febrero de 2025 y consta de un total de 316 muestras divididas en niveles desde A1 hasta C1, según el Marco Común Europeo de Referencia. La clasificación del alumnado en los distintos niveles se realizó por los docentes del centro, teniendo en cuenta todas las habilidades lingüísticas de manera global (expresión y comprensión oral; y escrita).

En el corpus se les ha solicitado a los alumnos la redacción de tres textos de temática variada, sobre el contenido de su nivel correspondiente, siguiendo el Plan Curricular del Instituto Cervantes.

A continuación, en la Tabla 2, podemos ver el resumen de las muestras recogidas.

Nivel	Tareas	Muestras	Total Muestras
A1	Email a un familiar	29	97
	Invitación a una fiesta	32	
	Escribir un texto sobre ti	36	
A2	Postal de vacaciones	28	85
	Biografía de un familiar/amigo	28	
	¿Qué has hecho esta semana?	29	
B1	Email a un amigo	24	71
	Opinión de una película	24	
	¿Qué hiciste en vacaciones?	23	
B2	Solicitud de admisión a curso de verano	18	54
	Carta para ser publicada en periódico	18	
	Ventajas y desventajas del transporte público	18	
C1	Reclamación compañía de gas	3	9
	Reseña de una película	3	
	Defensa del transporte público	3	
Total de muestras del corpus: 316			

Tabla 2: Resumen de muestras Corpus CL.

3.2 Sistema

Para la realización de nuestros experimentos hemos utilizado el modelo *gpt-3.5-turbo* de OpenAI (OpenAI, 2023). Elegimos dicho modelo por varias razones. En primer lugar, es un modelo de lenguaje grande con un entrenamiento extenso en textos de diversos idiomas y temáticas, lo que garantiza una comprensión avanzada del lenguaje. Además, su API permite ejecutar experimentos de manera eficiente sin los requisitos de hardware que demandan modelos abiertos de tamaño similar, como LLaMA o Mistral. El coste total de ejecución de los experimentos realizados ha sido de 25 euros.

Dado que, hasta donde alcanza nuestro conocimiento, el estudio aquí presentado es el

primero sobre clasificación automática del nivel ELE, decidimos comenzar con un modelo privado de alto rendimiento para establecer una línea base sólida. En futuros trabajos, planeamos comparar su rendimiento con modelos abiertos más pequeños que requieren menos recursos para evaluar su viabilidad y explorar posibles mejoras en nuestro sistema.

3.3 Experimentos propuestos

En el estudio se plantean dos experimentos que consideramos clave para poder responder a las preguntas que motivaron la investigación:

1. **Experimento basado en prompting Zero-shot Learning (ZSL).** El método se basa en la elaboración de una instrucción clara (*prompt*) que permita al modelo comprender la tarea a realizar. En nuestro estudio, una lingüista experta en la nivelación del español para hablantes no nativos diseñó la instrucción, en las que se especificaron las características de cada nivel ELE. Se le pidió al modelo clasificar los textos según estos criterios y justificar su decisión. El *prompt* completo se encuentra en el Anexo 1. Para este experimento, utilizamos el modelo *gpt-3.5-turbo* con su configuración predeterminada
2. **Experimento basado en el entrenamiento o *fine-tuning* del modelo (FT).** En este caso, buscamos ajustar el modelo *gpt-3.5-turbo* para la tarea de nivelación automática ELE. El modelo se entrenó con su configuración predeterminada utilizando el corpus CAES, estableciendo una única época por ser el valor que mejores resultados nos dió en el conjunto de validación. El conjunto de datos CAES se dividió en tres particiones estratificadas por niveles ELE: 90% para entrenamiento, 5% para validación y 5% para prueba. Los resultados de evaluación del modelo en el conjunto de prueba del corpus CAES mostraron un desempeño con una precisión total de 0,9885, un recall total de 0,9857 y un F1-score total de 0,9869. Sin embargo, en el experimento buscamos evaluar el sistema en un entorno real en lugar de en datos procedentes del corpus de entrenamiento, donde, como hemos podido comprobar, el rendimiento del modelo suele ser especialmente alto. Por ello, para analizar su capacidad de generalización, el modelo fue aplicado al corpus CL, objeto de este estudio.

3.4 Evaluación

Para evaluar los distintos sistemas propuestos, es necesario definir las métricas que permitirán comparar los resultados y medir el rendimiento de los modelos. En este caso, utilizaremos las métricas más comunes en tareas de clasificación de textos (Sebastiani, 2002): Macro-Precisión (P), Macro-Recall (R) y Macro-F1 (F1). Las mencionadas métricas se calcularán tanto de forma global como por clase, con el fin de identificar las dificultades en la clasificación de cada nivel.

Adicionalmente, se realizará una evaluación manual por parte del personal docente del centro universitario (manteniendo el anonimato de qué centro se trata en la revisión) a los resultados obtenidos en el experimento FT para obtener una evaluación cualitativa del desempeño del modelo.

4 Resultados

En la presente sección se presentan los resultados obtenidos en los experimentos realizados. Tal como mencionamos en la Sección 3.2, para evaluar nuestros sistemas se ha utilizado el corpus CL, recopilado de distintas clases y niveles de un centro universitario. La Tabla 3 muestra los valores de las métricas macro-precisión, macro-recall y macro-F1 para cada nivel ELE y para el conjunto de todas las clases.

Al observar la Tabla 3, podemos ver que, en general, si calculamos las métricas agrupando todas las clases, el experimento FT supera ampliamente al experimento ZSL (por ejemplo, en la métrica macro-F1: 0,560 vs. 0,291). Esto sugiere que entrenar el modelo proporcionando ejemplos de cada clase mejora su rendimiento, ya que le permite conocer los límites de cada nivel a partir de los datos de entrenamiento.

Si analizamos los resultados por niveles, podemos concluir que cada uno presenta características particulares. En la clasificación del nivel A1, el experimento ZSL obtiene una mayor precisión (0,732 vs. 0,672), mientras que FT alcanza un mejor F1-score (0,757 vs. 0,670). Esto indica que ZSL suele clasificar

correctamente los textos de este nivel cuando los predice, aunque FT logra identificar un mayor número de textos A1 de manera correcta. Para el nivel A2, el experimento ZSL presenta un recall alto (0,765) pero una baja precisión (0,369), lo que significa que detecta muchos casos pero con un alto número de falsos positivos. En contraste, el experimento FT mejora en precisión (0,559) y recall (0,671), logrando un mejor equilibrio. En el caso del nivel B1, ambos experimentos muestran un rendimiento bajo, aunque el experimento FT sigue obteniendo mejores resultados en las métricas evaluadas. Consideramos que puede deberse a que los textos de nivel B1 estén clasificados globalmente en este nivel por otra habilidad lingüística o a que comparten características con los niveles fronterizos (A2 y B2) y quizás las diferencias no sean tan claras como en otros niveles que tienen mayor o menor calidad en el lenguaje utilizado. Esto sugiere que debemos invertir esfuerzos para definir mejor las características de dicha clase y proporcionarle un mayor número de ejemplos al modelo durante el entrenamiento. Para el nivel B2, el experimento FT mejora drásticamente respecto al experimento ZSL (0,783 vs. 0,351 en la métrica F1), lo que indica que entrenar al modelo con ejemplos reales es clave para capturar mejor las características y patrones de este nivel. Finalmente, analizados los textos correspondientes a un nivel C1, observamos que el experimento ZSL no logra clasificar correctamente ningún texto de dicho nivel, mientras que el experimento FT alcanza un valor de F1 de 0,286. Este resultado indica que el rendimiento del experimento FT en nivel mencionado es bajo, lo cual era esperado debido a que es la clase minoritaria en nuestro conjunto de datos, por lo que el modelo cuenta con pocos ejemplos de entrenamiento para identificarla correctamente. Aun así, vemos que, incluso con un conjunto de datos reducido, el entrenamiento ya proporciona mejoras.

Nivel ELE	ZSL			FT		
	P	**R**	**F1**	**P**	**R**	**F1**
A1	**0,732**	0,619	**0,670**	0,672	**0,866**	0,757
A2	0,369	**0,765**	0,498	0,559	0,671	0,610
B1	0,342	0,197	0,250	0,513	0,281	0,364
B2	0,333	0,185	0,351	**0,947**	0,667	**0,783**
C1	0,000	0,000	0,000	0,250	0,333	0,286
Total	0,355	0,320	0,291	0,588	0,564	0,560

Tabla 3: Resultados de los experimentos según cada nivel de ELE y agrupando todos los niveles. Los experimentos se han evaluado con el corpus CL en las métricas Macro-Precisión, Macro-Recall y Macro-F1. En negrita se marcan los valores más altos obtenidos en cada métrica. ZSL: experimento de Zero-shot Learning, FT: experimento en el que se ajusta el modelo mediante fine-tuning.

Adicionalmente, y dado que no es lo mismo que un modelo se equivoque al clasificar un texto de nivel B1 como A2 que como C1, hemos decidido calcular el coeficiente de acuerdo de los experimentos realizados con el *gold_label* utilizando la métrica *Weighted Kappa*. Esta métrica la calculamos aplicando dos tipos de penalización: una penalización lineal, proporcional a la distancia entre categorías, y una penalización cuadrática, que penaliza en mayor medida los desacuerdos más grandes. Los resultados de la métrica pueden verse en la Tabla 4. Es importante tener en cuenta que los valores más cercanos a 1 indican un mayor acuerdo entre las predicciones del modelo y las etiquetas de referencia. En concreto, se observa que el experimento FT obtiene coeficientes de acuerdo significativamente más altos que el experimento ZSL, tanto con penalización lineal como cuadrática. Los resultados refuerzan la idea de que adaptar un modelo a la tarea, hace que aprenda mejor los límites entre las clases y comete errores menos severos que cuando no tiene conocimiento previo.

Experimentos	Lineal	Cuadrático
Gold - ZSL (CAES)	0.2120	0.3481
Gold - FT (CAES)	**0.9833**	**0.9900**
Gold - ZSL (CEALM)	0.3820	0.5127
Gold - FT (CEALM)	**0.5888**	**0.6889**

Tabla 4: Nivel de acuerdo entre los experimentos realizados. *Gold* hace referencia a las etiquetas tomadas como gold standard en los experimentos. En negrita se señalan los casos en los que hay un mayor grado de acuerdo para cada conjunto de datos probado.

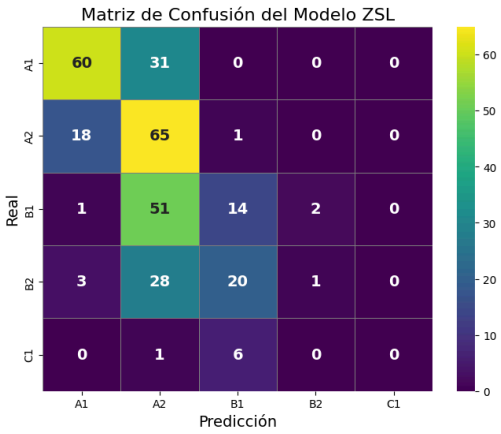

Figura 1: Matriz de confusión para el experimento *Zero-shot learning* evaluado en el corpus CL (experimento ZSL).

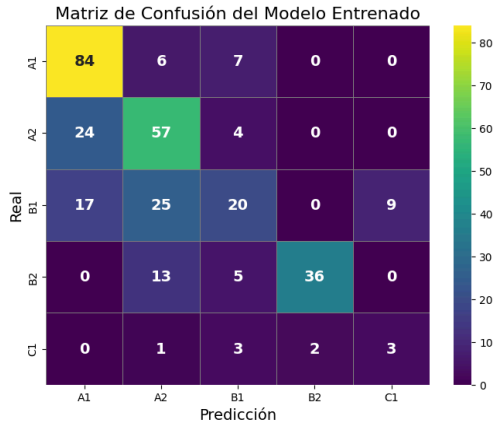

Figura 2: Matriz de confusión para el experimento en el que se realiza el entrenamiento del modelo con el corpus CAES y posteriormente es evaluado con el corpus CL (experimento FT).

5 *Análisis de errores*

A raíz de los resultados obtenidos, se realiza un análisis manual, por los docentes del centro universitario, del experimento FT, por ser el que muestra un rendimiento superior. En el análisis encontramos diferentes tipos de errores resumidos en la Tabla 5. Por un lado, tenemos textos clasificados erróneamente por el modelo y, por otro lado, textos anotados en un determinado nivel y que el modelo ha nivelado correctamente en un nivel distinto. En el último caso, el nivel inicial en el que se ha clasificado se debe a un nivel global del alumnado, no sólo a su nivel de expresión escrita, como se ha mencionado anteriormente en el apartado 3.1.

El análisis llevado a cabo por los docentes se ha realizado siguiendo las directrices del Plan Curricular del Instituto Cervantes (Cervantes, s.f.). En dicho análisis, se ha observado que los textos del nivel A1, que eran muy cortos (en total 7 muestras), el modelo los ha clasificado en un nivel superior por encontrarse bien escritos, sin tener en cuenta el número de palabras, se puede visualizar en el *Texto-1-EM* de la Tabla 6 del Anexo 2.

El análisis llevado a cabo en el nivel A2 muestra que el modelo ha nivelado con precisión varias muestras (en total 5) que no estaban clasificadas en el corpus original en dicho nivel, un ejemplo de ello se puede ver en el *Texto-1-NC* de la Tabla 7 del Anexo 3. Lo mencionado refleja que el modelo ha sido capaz de clasificar correctamente tras el aprendizaje del patrón de dicho nivel del entrenamiento.

Con respecto al nivel B1, al igual que ocurre en el nivel A2, en la revisión manual, se ha comprobado que el modelo nivela de manera correcta 30 muestras que no se encontraban clasificadas en dicho nivel, lo cual se puede observar en el *Texto-2-NC* de la Tabla 6 del Anexo 2. Puede deberse a varios motivos. Por un lado, el nivel B1, como se ha mencionado en la sección 4, es un nivel intermedio y los límites entre sus niveles próximos pueden ser difusos. Por otro lado, como se ha comentado, al ser una clasificación inicial en la que se han tenido en cuenta otras habilidades lingüísticas a parte de la expresión escrita, es probable que los textos hayan sido anotados en dicho nivel por una habilidad lingüística distinta a la expresión escrita. Asimismo, se ha observado que la tarea 2 del nivel B1 del corpus CL: "*Opinión sobre una película"*, es similar a la tarea 1 del nivel C1 en el corpus CAES: "*Reseña de película*", y que hay 9 textos que han sido clasificados por temática, por lo que se puede señalar que el modelo, en determinadas ocasiones, clasifica siguiendo las temáticas con las que fue entrenado (*Texto-2-EM*, Anexo 2). Por otro lado, se ha detectado que el modelo clasifica en el nivel A1 a los textos con errores de puntuación (*Texto-3- EM*, Anexo 2), en total 8 textos.

En cuanto al nivel B2, se han encontrado 8 muestras en las que el nivel indicado por el modelo cuenta con una mayor precisión que el nivel inicial.

Por último, en el nivel C1 se ha podido comprobar que 5 de las muestras recogidas se encuentran en un nivel más cercano al propuesto por el modelo, ya que se trata de textos menos complejos que los correspondientes a un nivel C1 (*Texto-3-NC*, Anexo 3).

Tipología de errores	Muestras
Textos cortos para el nivel	7
Textos clasificados correctamente por el modelo	43
Tarea con temática similar	9
Textos con errores de puntuación	8
Textos complejidad distinta al nivel	5

Tabla 5: Resumen de los errores encontrados.

6 Conclusiones

Los resultados del presente estudio han dado respuesta a los objetivos que se plantearon en la Sección 3 relativa a la metodología seguida. Con respecto al objetivo de (1) si los modelos eran capaces de entender las diferencias de niveles en ELE y clasificarlos de manera automática, se puede observar que sí son capaces. No obstante, como se ha demostrado en la comparativa entre los experimentos ZSL y FT, es necesario proporcionar instrucciones claves al modelo y ajustar dicho modelo con ejemplos representativos de cada clase para mejorar la detección de nivel, dando así respuesta a nuestro objetivo (2).

Asimismo, se han podido identificar (3) las dificultades que el modelo ha presentado a la hora de nivelar de manera correcta los textos escritos, como los errores de puntuación, la longitud de los textos o la asociación del nivel según la similitud entre las tareas descritas en los textos.

Sin embargo, se ha comprobado que el modelo puede clasificar el nivel de español del texto de manera aún más precisa que la asignación inicial. Puede deberse a que los alumnos se encuentran en un determinado nivel por sus conocimientos globales de las habilidades lingüísticas, por lo que es probable que su expresión escrita se encuentre más cerca de la propuesta por el modelo que del nivel que se les ha asignado de manera global. Como resultado, nuestra propuesta ayudará a determinar el nivel de expresión escrita del alumnado, facilitando así a los docentes en su clasificación y adaptación del curso a su alumnado.

Como trabajo futuro, se planea continuar con la implementación de sistemas de nivelación ELE automáticos utilizando modelos abiertos, como LLaMA o Salamandra, este último es un modelo entrenado con una proporción de textos en español superior a los existentes actualmente. Esta elección permitirá tener un control total sobre el sistema implementado y evitar depender de las posibles modificaciones que una empresa privada pueda realizar en su modelo. Además, se pretende elaborar un *prompt* más detallado para identificar con claridad las características de cada clase, reduciendo ambigüedades y asegurando que pequeños fallos comunes entre los hablantes del español en la escritura de los textos a clasificar no generen nivelaciones erróneas. También se explorarán otras estrategias que permitan incorporar características sintácticas y semánticas de los textos, ayudando al modelo a clasificarlos correctamente. Finalmente, se trabajará en la generación de más conjuntos de datos de calidad para mejorar la precisión de la tarea.

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 es parte del Proyecto CONSENSO (PID2021-122263OB-C21), del Proyecto MODERATES (TED2021-130145B-I00), y del Proyecto SocialTox (PDC2022-133146-C21) financiados por MCIN/AEI/10. 13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR, y parte del Proyecto FedDAP (PID2020-116118GA-I00) y del Proyecto Trust-ReDaS (PID2020-119478GB-I00) financiados por MICINN/AEI/10.13039/501100011033. 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+.

Bibliografía

Area-Moreira, M., A. Del Prete, A- L. Sanabria-Mesa y M. B. Sannicolás-Santos. 2024. No todas las herramientas de IA son iguales. Análisis de aplicaciones inteligentes para la enseñanza universitaria. *Digital Education Review*, 45, 141-149.

Azurmendi Arrue, E. 2024. Euskarazko lehen C1 ebaluatzaile automatikoa [Trabajo de Fin de Máster, Universidad del País Vasco (UPV/EHU)].

Cantero Romero, M. V. 2024a. Modelos de lenguaje y ELE. Uso de los adjetivos. En *Innovación en el aula: nuevas estrategias didácticas en humanidades*. pp. 843-860.

Cantero Romero, M. V. 2024b. Modelos de lenguaje y ELE. Uso de los tiempos verbales. En *Avances en los estudios de lingüística hispánica: perspectivas teóricas y aplicadas entre lengua y sociedad*. pp. 203-220.

Cantero Romero, M. V. 2025. El léxico ELE en los modelos de lenguaje.*RILEX. Revista sobre investigaciones léxicas*, 8/I. pp. 155-185

Cervantes, C. C. V. (s. f.-b). CVC. Plan Curricular del Instituto Cervantes. Niveles de referencia para el español. https://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/

Espinosa Zaragoza, I., Abreu Salas, J., Moreda, P., y Palomar, M. 2023. Automatic Text Simplification for People with Cognitive Disabilities: Resource Creation within the ClearText Project. *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 68–77, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

García Peñalvo, F. J., F. Llorens Largo y J. Vidal. 2023. 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.

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.

Instituto Cervantes; Universidad de Santiago de Compostela. (s. f.). CAES. https://galvan.usc.es/caes

Li, Y. (2023). A Practical Survey on Zero-Shot Prompt Design for In-Context Learning. abs/2309.13205. https://doi.org/10.26615/978-954-452-092-2_069

OpenAI. (2023). OpenAI GPT-3.5 turbo API. https://platform.openai.com/docs/models/gpt-3.5-turbo

Pourpanah, F., M. Abdar, Y. Luo, X. Zhou, R. Wang, y C. P. Lim. 2023. A Review of Generalized Zero-Shot Learning Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4051-4070

Román Mendoza, E. 2023. Formular preguntas para comprender las respuestas: ChatGPT como agente conversacional en el aprendizaje de español como segunda lengua | marcoELE. marcoELE.

Roumeliotis, K. I., N. D. Tselikas y D. K. Nasiopoulos (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

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1), 1-47.

Vallecillo Rodríguez, M. E., M. V. Cantero Romero, I. Cabrera De Castro, A. Montejo Ráez y M. T. Martín Valdivia. 2024. CONAN-MT-SP: A Spanish Corpus for Counternarrative Using GPT Models. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688, Torino, Italia. ELRA and ICCL.

Yannakoudakis, H., T. Briscoe y B. Medlock. 2011. A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 180–189. University of Cambridge.

A Anexo 1: Prompt utilizado

Prompt

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: Texto a evaluar.

Figura 3: *Prompt* utilizado para la realización de nuestros experimentos.

B Anexo 2: Tabla errores del modelo FT

Id	Texto	Nivel CL	Nivel FT
Texto-1-EM	*quiero mi dos hermanos*	A1	B1
Texto-2-EM	*Veo Gladiator II no hay poco tiempo y estoy un poco decepcionado. Me gusta la trama pero no esta muy diferente del primero. Me gusta globalmente la película pero me gustaría ver un poco más los emperadores, ve solo el gladiator y su historia no era diferente de aquella del gladiator en la primera película. El personaje del gladiator era un poco aburrido...*	B1	C1
Texto-3-EM	*Hola, mi amiga! Estas bien? Ahora estoy estudiando en Jaén, España. Jaén es el lugar mu bien para vivir. Todos los personas son amable, y cosas para comprar es barato! Pero una cosa que es difícil para mi es casi nada persona local habla inglés. Siento nerviosa todos los veces cuando necesito hablar en español en por ejemplo las tiendas. Pero este es oportunidad buena tan bien para aprender español. Qué estas haciendo en to destinación de estudiar extranjero? Quiero escuchar to historia tan bien! Adiós!*	B1	A1

Tabla 6: Ejemplos de los textos del corpus CL que han sido clasificados erróneamente por el modelo del experimento basado en el ajuste a la tarea de nivelación automática ELE (experimento FT).

C Anexo 3: Tabla aciertos del modelo

Id	Texto	Nivel CL	Nivel FT
Texto-1-NC	*Fui a bar y comi un waffle garande con mi amiga. Tambien, bebi un zervesa y vino. Mi amiga habla sobre su familia.*	A2	A1
Texto-2-NC	*Me encantan la película como esta en la que la trama es muy complicada y con muchos culpos de escena*	B1	A2
Texto-3-NC	Hola. Buenos días! Escribo otra vez a ustedes porque tengo algunos problemas con el gas de mi piso. Ya he pagado y el hombre todavía no estuvo aquí para hacer que el gas vuelva a funcionar, también ya he enviado mensaje y llamado el teléfono de la compañía pero nadie mi ha contestado. Envío otra vez una reclamación para intentar finalmente lograr y que mi contesten pronto. Gracias, estoy esperando la respuesta.	C1	B1

Tabla 7: Ejemplos de los textos del corpus CL que han sido clasificados correctamente por el modelo del experimento basado en el ajuste a la tarea de nivelación automática ELE (experimento FT), pese a que la etiqueta inicial ELE del corpus era errónea.