Diagnostic Assessment of Spanish as a Foreign Language using Technology based on Natural Language Processing

María Victoria Cantero-Romero

Departamento de Filología Española, Universidad de Jaén, Jaén, Spain
SINAI, CEATIC, Universidad de Jaén, Jaén, Spain

Abstract

Leveling the written expression of students of Spanish as a foreign language can be a complex and laborious task when a large group of students is involved. If, in addition to the number of students, we add the complexity of the different language levels, the work of leveling them correctly can become an arduous task. This doctoral thesis focuses on the use of large language models (LLMs) for the automatic leveling of Spanish texts following the Common European Framework of Reference for Languages (CEFR)[1].

Keywords

Spanish as a Foreign Language, Automatic Language Leveling, Large Language Models, Natural Language Processing

1. Justification of the proposed research

Diagnostic assessment in Spanish as a Foreign Language (ELE) classrooms constitutes a key preliminary step to tailor instruction to students' linguistic proficiency. However, this process is often constrained by factors such as high enrollment in the student population and limited time at the beginning of the academic term, which may compromise the accuracy and objectivity of the assessment. These challenges are particularly evident in university programs that host international students - such as participants in Erasmus or Talentium schemes - who often present heterogeneous proficiency levels and, in many cases, lack formal language certification.

Among the skills assessed, written production poses one of the main challenges, as it requires a detailed and individualized analysis that is difficult to perform efficiently and consistently. This makes it a time-consuming and cognitively demanding task for instructors. In this context, Artificial Intelligence (AI) technologies—particularly those based on Natural Language Processing (NLP)—offer promising alternatives to automate such tasks, thus reducing the workload of educators and improving the consistency of assessment criteria. By delegating repetitive tasks to AI, teachers can redirect their efforts toward more valuable pedagogically.

Language models such as ChatGPT [2] have demonstrated considerable potential in generating and analyzing texts with syntactic and semantic precision, making them effective tools for automated assessment of written language. These technologies enable a more agile and fair classification of learners based on linguistic parameters such as grammar, cohesion, vocabulary use, and textual organization. Furthermore, their integration aligns with current pedagogical approaches that prioritize efficiency, adaptability, and personalization of learning experiences.

A growing body of research highlights the positive impact of AI in education, particularly its ability to provide real-time feedback, support differentiated instruction, and optimize instructional planning. Aligned with the principles of the Common European Framework of Reference for Languages (CEFR)[1], these tools can also improve the transparency and reliability of language assessment processes.

This thesis is motivated by the pressing need to modernize placement procedures in ELE, with a particular focus on written production. It aims to investigate the potential of LLMs as supportive tools in diagnostic assessment, with the goal of increasing efficiency, alleviating the assessment burden on

Doctoral Symposium on Natural Language Processing, 25 September 2025, Zaragoza, Spain.
vcantero@ujaen.es (M. V. Cantero-Romero)
0009-0008-7052-7322 (M. V. Cantero-Romero)

educators, and ensuring fairer, more transparent classification of non-native learners. First of all, this paper presents a selection of studies on the use of LLMs in foreign language education, with a particular focus on their application in the Spanish as a Foreign Language (ELE) classroom. It also introduces two research projects that are directly related to the core themes of this thesis. Next, the research hypothesis and main objectives of the study are outlined. This is followed by a description of the methodology employed and the experiments conducted thus far. Finally, a series of questions that have been raised at the beginning of the thesis and during the experiments will be presented.

2. Related work

Technological advances—particularly in the field of AI, have brought about profound transformations in all sectors of society, and education is no exception [3, 4]. Within this context, LLMs have emerged as one of the most innovative tools for teaching and assessment, offering new possibilities for pedagogical interaction. Trained on vast volumes of textual data [5], these models are capable of generating, analysing, and classifying texts, thereby opening up a wide range of applications in educational settings.

Bolaño-García and Duarte-Acosta [6] argue that AI can be used to personalise learning and support educators by automating routine tasks such as grading and administrative processes. They also emphasise its potential to collect detailed data on student performance, which can inform the design of more targeted instructional strategies. This perspective is shared by Fajardo et al.[7], who underscores the value of AI in higher education, particularly through personalised tutoring tailored to individual learner profiles.

A growing body of research [8, 9, 7, 6] has highlighted the impact of AI on educational assessment, shifting the focus from traditional testing methods to more individualised, learner-centred approaches. One of the key benefits of AI systems is their ability to provide real-time feedback, enabling students to identify areas for improvement immediately—an aspect that has been linked to increased motivation and improved academic outcomes [6]. Moreover, as Area-Moreira et al. [10]and García Peñalvo et al. [11] point out, tools like ChatGPT facilitate automated grading, content creation, plagiarism detection, and the development of exams and questionnaires, thereby reducing teachers' workload and optimising institutional resources.

Recent studies have also begun to explore the use of LLMs specifically in the field of language education. García Peñalvo et al. [11] propose their use as text generators and machine translators, supporting the development of language competencies in the classroom. Likewise, Area-Moreira et al. [10] highlight their potential for producing authentic learning materials and resolving grammatical queries. Hong [12] adds that LLMs can function as virtual tutors—simulating conversations, providing linguistic explanations, generating customised content, and assisting educators in planning and assessment tasks.

In the specific context of Spanish as a Foreign Language (ELE), several studies have examined the pedagogical potential of LLMs. Román Mendoza [13] explores the use of ChatGPT as a conversational agent to address learner queries. Based on this line of research, previous works of my doctoral thesis [14, 15, 16] evaluate how texts generated by ChatGPT and LLaMA align with the proficiency levels outlined in the Instituto Cervantes' Curriculum Plan (PCIC) [17], focusing on linguistic variables such as adjective use, verb tense distribution, and lexical selection.

A key challenge in ELE instruction lies in accurately placing learners at the appropriate proficiency level at the start of a course. Prior research [18] indicates that traditional placement tests often rely on closed-ended items and neglect writing tasks, largely due to the difficulty of manual assessment. The automation of this process through language modelling therefore represents a major innovation.

Along these lines, earlier studies have explored the automatic classification of texts in various languages. Yannakoudakis et al.[19] employed support vector machines to evaluate English writing, while Azurmendi Arrue[20] used models such as RoBERTa to analyse Basque texts at the C1 level. However, there are no studies, to date, which deal with the leveling of written expression texts in Spanish.

Recent advances in prompt engineering have further emphasised the importance of designing effective

instructions to guide LLM performance. Studies by Li[21] and Pourpanah et al.[22] highlight the benefits of Zero-shot and Few-shot Learning strategies in specialised tasks. Meanwhile, Roumeliotis et al.[23] demonstrate that task-specific fine-tuning of advanced models like GPT-4 can outperform smaller pre-trained models such as BERT or RoBERTa in similar contexts.

3. Hypothesis and objectives

The main hypothesis of this work is the use of AI and specifically LLMs to facilitate the teaching task when leveling students of Spanish as a foreign language. Therefore, we set the following objectives:

- Conduct a preliminary review of the state of the art on language leveling using LLMs, focusing on Spanish and other languages.
- Evaluate and compare the performance of current LLMs in the analysis of written texts.
- Study and assess the effectiveness of general-purpose LLMs in the context of language leveling of ELE learner texts.
- Fine-tune and implement a custom language model specifically adapted to the task of language leveling in ELE.
- Develop a dedicated corpus of ELE learner texts annotated with proficiency levels, to be used for training and evaluating the proposed LLM.

4. Methodology and proposed experiments

The methodology proposed for the completion of this doctoral thesis is as follows:

- **Review of the state of the art.** A comprehensive review of existing literature will be conducted, focusing on text leveling and the application of LLMs to language learning and assessment. This phase will also include the identification and analysis of existing learner corpora for ELE, which may serve as reference or training data in later stages.
- **Analysis of LLMs' knowledge of proficiency frameworks.** The extent to which LLMs internalise and reflect the descriptors of the Common European Framework of Reference for Languages (CEFR) [1], specifically in the context of ELE, will be assessed.
- **Application of LLMs to text leveling.** Several models, both closed-source (e.g., GPT-4) and open-source (e.g., LLaMA), will be applied to the task of ELE text leveling.
- **Development of a dedicated learner corpus.** A custom corpus of ELE learner texts will be compiled and annotated to support further experimentation and research.
- **Experimental validation and evaluation.** The resources already available, together with those generated during the research will be used to conduct experiments assessing the accuracy and reliability of the selected LLMs in leveling tasks.
- **Prototype development.** Based on the insights and results obtained from the previous phases, a working prototype will be designed and implemented to automate ELE text leveling using LLMs.

The methodology outlined was implemented through a series of experiments designed to assess the ability of LLMs in the task of ELE text leveling.

As mentioned above, our initial objective was to determine whether LLM, specifically ChatGPT [2] and LLaMa [24], possess knowledge of the principal reference frameworks used in language teaching, namely the Common European Reference Framework for Languages (CEFR) [1] and the Cervantes Institute Curriculum Plan (PCIC) [17]. The experiments conducted to date have addressed this objective through the following phases. In the first phase, a set of experiments was conducted, consisting of the design of prompts instructing both models to generate reading comprehension texts at different proficiency levels. Then, these texts were analyzed from several linguistic perspectives. First, a lexical analysis was conducted to assess whether the selection and diversity of vocabulary aligned with the

expected lexical difficulty at each CEFR level. Second, an analysis of verb tenses was performed to evaluate the grammatical suitability relative to the assigned level. Finally, the use of adjectives was examined, focusing on their frequency, complexity, and discursive function within the text. In a second phase, experiments were conducted to explore the ability of ChatGPT [2] and LLaMa [24] to identify the linguistic level of written texts. A targeted study was carried out using ChatGPT to level texts from the CAES corpus (Corpus de Aprendientes de Español como Lengua Extranjera) [25] compiled by the Instituto Cervantes. Using carefully crafted prompts, the model's ability to classify authentic learner texts according to CEFR levels was evaluated. This analysis enabled a review of the consistency of the model in identifying linguistic characteristics typical of each level and its potential as a support tool for the automated classification of linguistic corpora.

5. Research Elements Proposed for Discussion

This study is based on the following primary research questions:

- To what extent do LLMs encode knowledge of the Spanish proficiency levels defined by the Common European Framework of Reference for Languages (CEFR)?
- Can LLMs accurately identify the level of linguistic proficiency of a written text when asked appropriately?
- What specific features should prompts include to ensure reliable and valid level classification?
- Which LLM demonstrates the highest performance in the task of automatic text leveling using carefully engineered prompts?
- Could alternative approaches, such as model fine-tuning, yield more accurate results than prompt-based methods?

Based on the preliminary findings, a set of complementary research questions is also formulated:

- Following fine-tuning with the CAES corpus (Corpus de Aprendientes de Español como Lengua Extranjera), is the model capable of generalising to accurately level texts from other learner corpora?
- What is the minimum data requirement for effective fine-tuning, and how should such data be partitioned to optimise model performance?
- After fine-tuning, does the model assign CEFR levels based primarily on the recognition of salient linguistic features, or does it rely instead on thematic or topical cues?

Acknowledgments

My sincere thanks to my thesis tutors Salud María Jiménez Zafra and Ana María Ortíz Colón for guiding me along this process, to the doctoral program of my beloved University of Jaén

Declaration on Generative AI

During the preparation of this work, the author used GPT-4 in order to: Grammar and spelling check.

References

[1] Council of Europe, Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume, Council of Europe Publishing, Strasbourg, 2020. URL: https://www.coe.int/lang-cefr.

[2] OpenAI, Chatgpt (versión gpt-4), https://chat.openai.com, 2025.

[3] R. F. Zambrano Campozano, Impacto de la inteligencia artificial en la transformación digital de la educación, Serie Científica De La Universidad De Las Ciencias Informáticas 18 (2025) 177–192.
[4] W. O. Aparicio Gómez, La inteligencia artificial y su incidencia en la educación: Transformando el aprendizaje para el siglo xxi, Revista internacional de pedagogía e innovación educativa 3 (2023) 217–230.
[5] L. Wang, El uso de chatgpt y gemini para la preparación de actividades de comprensión oral de cle: un estudio comparativo, https://ddd.uab.cat/record/300353, 2024. Dipòsit Digital de Documents de la UAB.
[6] M. Bolaño-García, N. Duarte-Acosta, Una revisión sistemática del uso de la inteligencia artificial en la educación, Revista Colombiana de Cirugía (2023) 51–63. doi:10.30944/20117582.2365.
[7] G. M. Fajardo, D. C. Ayala, E. M. Arroba, M. López, Inteligencia artificial y la educación universitaria: Una revisión sistemática, Magazine de las Ciencias Revista de Investigación E Innovación 8 (2023) 109–131. doi:10.33262/rmc.v8i1.2935.
[8] R. D. Moreno, La llegada de la inteligencia artificial a la educación, Revista de Investigación En Tecnologías de la Información 7 (2019) 260–270. doi:10.36825/riti.07.14.022.
[9] V. R. García-Peña, A. B. Mora-Marcillo, J. A. Ávila Ramírez, La inteligencia artificial en la educación, Dominio de las Ciencias 6 (2020) 648–666. doi:10.23857/dc.v6i3.1421.
[10] M. Area-Moreira, A. Del Prete, A. L. Sanabria-Mesa, M. B. Sannicolás-Santos, No todas las herramientas de ia son iguales. análisis de aplicaciones inteligentes para la enseñanza universitaria, Digital Education Review 45 (2024) 141–149. doi:10.1344/der.2024.45.141-149.
[11] F. J. García Peñalvo, F. Llorens-Largo, J. Vidal, La nueva realidad de la educación ante los avances de la inteligencia artificial generativa, RIED Revista Iberoamericana de Educación a Distancia 27 (2023) 9–39. doi:10.5944/ried.27.1.37716.
[12] W. C. H. Hong, The impact of chatgpt on foreign language teaching and learning: Opportunities in education and research, Journal of Educational Technology and Innovation 5 (2023) 38–53.
[13] E. Román Mendoza, Formular preguntas para comprender las respuestas: Chatgpt como agente conversacional en el aprendizaje de español como segunda lengua, 2023. Publicado en marcoELE.
[14] M. V. Cantero Romero, Modelos de lenguaje y ele. uso de los adjetivos, in: Innovación en el aula: nuevas estrategias didácticas en humanidades, Editorial Académica Española, Madrid, 2024, pp. 843–860.
[15] M. V. Cantero Romero, Modelos de lenguaje y ele. uso de los tiempos verbales, in: Avances en los estudios de lingüística hispánica: perspectivas teóricas y aplicadas entre lengua y sociedad, Editorial Síntesis, Madrid, 2024, pp. 203–220.
[16] M. V. Cantero Romero, El léxico ele en los modelos de lenguaje, RILEX. Revista sobre investigaciones léxicas 8 (2025) 155–185.
[17] Instituto Cervantes, Plan curricular del Instituto Cervantes. Niveles de referencia para el español, Biblioteca nueva, Madrid, 2006.
[18] A. Biedma Torrecillas, M. D. C. Chamorro Guerrero, G. Lozano, A. Sánchez Cuadrado, Diseño y validación de las pruebas de nivel del clm de la universidad de granada.(1)teoría, in: Actas del VII Congreso ACLES, 2012, pp. 26–37.
[19] H. Yannakoudakis, T. Briscoe, B. Medlock, A new dataset and method for automatically grading esol texts, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, University of Cambridge, 2011, pp. 180–189.
[20] E. Azurmendi Arrue, Euskarazko lehen C1 ebaluatzaile automatikoa, Master's thesis, Universidad del País Vasco (UPV/EHU), 2024. Trabajo de Fin de Máster.
[21] Y. Li, A practical survey on zero-shot prompt design for in-context learning, 2023. URL: https://doi.org/10.26615/978-954-452-092-2_069. arXiv:2309.13205, arXiv preprint.
[22] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, A review of generalized zero-shot learning methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 4051–4070.
[23] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, Next-generation spam filtering: Comparative fine-tuning of llms, nlps, and cnn models for email spam classification, Electronics 13 (2024) 2034.

doi:10.3390/electronics13112034.
[24] Meta AI, Llama 2: Large language model meta ai, https://ai.meta.com/llama, 2023. Modelo de lenguaje de código abierto.
[25] Instituto Cervantes and Universidad de Santiago de Compostela, Caes, https://galvan.usc.es/caes, 2021.