The EuroCALL Review
Volume 32, No. 2, 2025, 102-116
ISSN: 1695-2618
https://doi.org/10.4995/eurocall.2025.23827

RESEARCH PAPER

Intelligent (but artificial) Feedback in Spanish as a Foreign Language: Evaluation of *ChatGPT* and *Claude* as Text Correction Tools

Antoni Brosa-Rodríguez
Universitat Rovira i Virgili, Spain

antoni.brosa@urv.cat

How to cite this article:

Brosa-Rodríguez, A. (2025). Intelligent (but artificial) Feedback in Spanish as a Foreign Language: Evaluation of *ChatGPT* and *Claude* as Text Correction Tools. *The EuroCALL Review, 32*(2), 102-116. https://doi.org/10.4995/eurocall.2025.23827

Abstract

This research analyses the potential and limitations of two artificial intelligence models (ChatGPT 4 and Claude 3.7) as feedback tools for texts written by students of Spanish as a foreign language. Using selected texts from the CEDEL2 corpus, the study evaluates these models' ability to provide pedagogically appropriate corrections. The qualitative analysis focuses on four criteria: accuracy in error detection, clarity of explanations, pedagogical adequacy, and potential problems. Results reveal that both models show high precision in detecting basic errors but present significant limitations such as overcorrection and difficulties adapting explanations to the student's level. Claude stands out for its systematic structure, while ChatGPT excels in humanizing feedback. Based on

these findings, we propose an integrated model to implement these tools in the Spanish language classroom, where the teacher maintains a crucial role as mediator, and we offer practical recommendations to maximize their benefits while minimizing their risks.

Keywords

Language teaching; Computer-Assisted Language Learning, Artificial Intelligence, feedback, education technology

1. Introduction

1.1. Problem statement

The correction and feedback of written texts constitute a fundamental task in teaching Spanish as a foreign language (SFL), but it also represents one of the aspects that demands the most time and dedication from teachers (Buyse, 2014). In an educational context where personalized attention is crucial for student progress, teacher overload can compromise the quality and regularity of these corrections. This situation has been intensified in recent years by the increase in the student-to-classroom ratio and the diversification of administrative and pedagogical tasks that fall on teachers (Fernández, 2017).

In this scenario, technologies based on artificial intelligence (AI) emerge as potential tools to provide immediate and detailed feedback to students. Recent advances in large language models (LLMs) have revolutionized the possibilities of textual analysis and generation, allowing increasingly sophisticated and contextualized interactions (Kasneci et al., 2023). Among these models, ChatGPT (developed by OpenAI) and Claude (developed by Anthropic) stand out for their ability to understand and produce texts in multiple languages, including Spanish with its dialectal variants.

However, the incorporation of these technologies in the educational field is not without questions: What is their linguistic accuracy in detecting errors in SFL student texts? Are their explanations and corrections pedagogically appropriate? What limitations do they present and what risks could they introduce into the learning process?

1.2. Corrective feedback in Spanish as a foreign language learning

Corrective feedback plays an essential role in second language learning (Fernández, 2017; Bailini, 2020b). In the specific field of SFL, there has been an evolution from approaches focused on explicit correction to perspectives that consider error as an inherent part of the learning process. Based on the feedback strategies regulation scale proposed by Bailini (2020a), effective feedback can be understood as a combination of detection accuracy, explanatory clarity, pedagogical adequacy, and balance between correction and motivation.

The implementation of these principles, however, is frequently hindered by practical limitations such as the student-teacher ratio and time constraints, which has driven the search for technological alternatives that can complement teaching work without substituting the fundamental mediating role of the teacher.

1.3. Language models and Artificial Intelligence in education

Intelligent tutorial systems represent a significant advance in the automation of feedback (Coyne et al. 2023). Ferreira and Kotz (2010) have analyzed specific applications for SFL, highlighting their effectiveness for basic errors (Crossley et al. 2019), but also their limitations in contextual and pragmatic aspects.

The recent development of large language models (LLMs) such as ChatGPT and Claude has opened new possibilities thanks to features such as contextual understanding, adaptability, multilingualism, and explanatory capacity (Kasneci et al., 2023). The

applications of these models in language teaching are beginning to be explored, with studies analyzing their use in specific contexts such as preparation for the DELE exam (López Mata, 2023) or student perceptions when using them for linguistic tasks (Xiao & Zhi, 2023; Ranalli, 2021). López Mata (2023) explored the integration of ChatGPT in DELE B1 preparation classes, specifically for written tasks, presenting both the advantages and disadvantages of using this AI tool in Spanish language teaching. Her study not only provided a practical didactic sequence but also collected student impressions after implementing the pedagogical model in the classroom. This work demonstrated that while AI tools can support specific examination preparation, their effectiveness depends heavily on how they are pedagogically integrated, and the students' critical engagement with the technology. López Mata's findings on student perceptions align with our investigation of how different AI models are received and utilized in educational contexts, though her focus was narrower (examination preparation) compared to our broader analysis of corrective feedback across proficiency levels.

García (2024) provided practical applications and recommendations for using ChatGPT in Spanish as a Foreign Language teaching, offering specific guidelines for educators on how to implement AI tools effectively in their practice. Her work contributes to the emerging framework of best practices for AI integration in language learning, emphasizing the need for teacher mediation and critical evaluation of AI-generated content.

As Ferreira and Kotz (2010) point out, computational linguistic feedback in the field of SFL must be subjected to a critical analysis that evaluates both its effectiveness and its possible adverse effects. This need for critical evaluation becomes even more relevant with the emergence of these new AI models.

1.4. Study objectives and justification

Despite these advances, the current literature presents important gaps: limited evaluation of the linguistic accuracy of LLMs in Spanish, particularly with texts by SFL students; absence of comparative studies between different models; limited attention to the pedagogical adequacy of the explanations generated; and lack of methodological proposals to integrate these tools into teaching practice.

This research addresses these questions through a comparative qualitative analysis of the feedback provided by ChatGPT and Claude on texts written by SFL students. Based on a sample from the CEDEL2 corpus (Written Corpus of Spanish as L2), it examines the ability of these systems to detect errors, explain grammatical concepts, and adapt their responses to the student's level.

This study is part of an emerging line of research on the application of AI in language teaching but it specifically focuses on the analysis of advanced models such as Claude 3.7 and ChatGPT-4, whose potential for teaching SFL has been scarcely explored to date (although in the case of EFL we have found more studies).The results obtained aim to contribute to the understanding of the strengths and limitations of these technologies, as well as to propose a framework for their effective implementation in the SFL classroom.

Next, the methodology used in this study is set out. Subsequently, the findings of the comparative analysis are presented, having been organized according to various evaluation criteria. Finally, the pedagogical implications of these results are discussed and an integrated model for the use of these tools in educational contexts is proposed, where the teacher maintains a crucial role as mediator and validator of the process.

2. Methodology

This research adopts a qualitative comparative analysis approach to evaluate the effectiveness of ChatGPT and Claude as feedback tools for texts written by SFL students. Below is a synthesis of the fundamental methodological aspects.

2.1. Corpus selection

Fifteen texts produced by SFL students were analyzed, having been selected according to the following criteria:

- Distribution by levels: 5 texts of low level (A1-A2), 5 of intermediate level (B1-B2), and 5 of advanced level (C1-C2).
- L1 diversity: The texts correspond to students with five different mother tongues (Arabic, English, Japanese, German, and Dutch).
- Text typology: Unification of narrative and descriptive texts of a cinematographic scene (fragment of a Charlie Chaplin film).
- Presence of representative errors: Texts with frequent errors at different linguistic levels (orthographic, morphological, syntactic, lexical, and pragmatic).

2.2. AI models analyzed

The study evaluates two state-of-the-art conversational AI models:

a) Claude 3.7 (Anthropic): Most recent version available at the time of the study.
b) ChatGPT-4 (OpenAI): Most advanced model in the GPT series at the time of the study.

Both were selected for representing the state of the art in generative AI and for their ability to process and generate text in Spanish with a high level of linguistic competence.

2.3. Evaluation protocol

To ensure consistency, a systematic protocol was established:

a) Standardized prompt: An identical prompt was designed for both models, requesting error analysis for pedagogical and didactic purposes. The exact text of the prompt was as follows:

I'm attaching a text written by a student who is learning Spanish as a foreign language. They had to describe what they saw in a video that was a fragment of Chaplin. I need you to analyse the text to see what errors they have made. It should serve pedagogical and didactic purposes, therefore, it would be good to have an explanation, correction proposal, etc. You should do what a teacher would do. You must take into account the students' level and their context and linguistic starting point. Send the final version to give to them.

b) Procedure: Each text was analysed by both models, collecting the 30 sets of feedback generated (15 texts × 2 models).

2.4. Analysis criteria

The qualitative analysis was based on four criteria, developed from the literature on effective feedback in SFL (Bailini, 2020a; Fernández, 2017):

a) Accuracy in error detection: Ability to correctly identify errors, considering false positives and false negatives.

b) Explanatory clarity: Quality of explanations, conceptual precision, contextualization, and exemplification.

c) Pedagogical adequacy: Adaptation to the student's level, progression, balance between correction and positive reinforcement.

d) Detected problems: Difficulties, limitations, or adverse effects introduced by the models.

For each criterion, a detailed analysis was carried out, identifying recurring patterns, representative examples, and significant differences between the models.

3. Results

The error analysis from a quantitative point of view, as shown in figures 1, 2 and 3, shows very interesting results.

Figure 1

Errors detected by level.

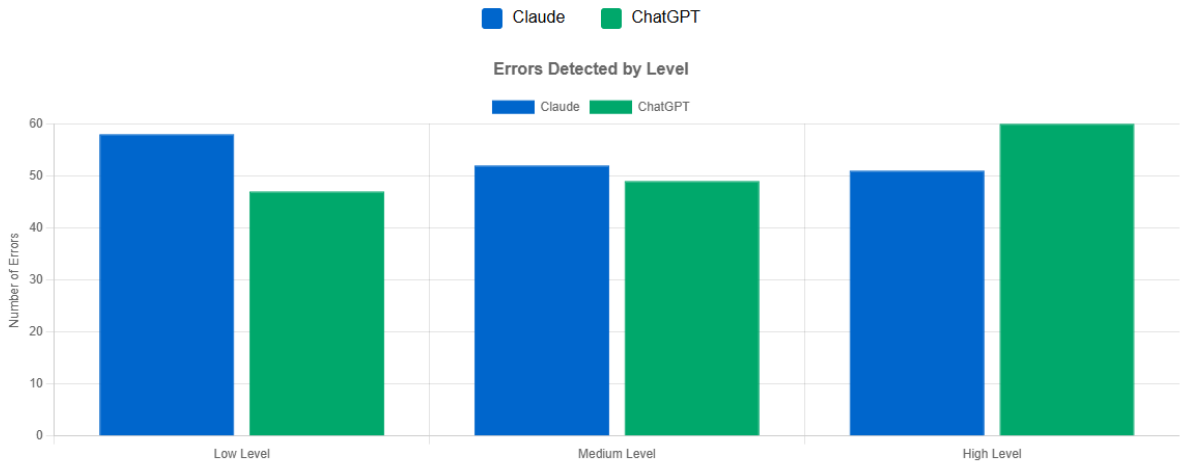


Figure 2.

Errors detected by text.

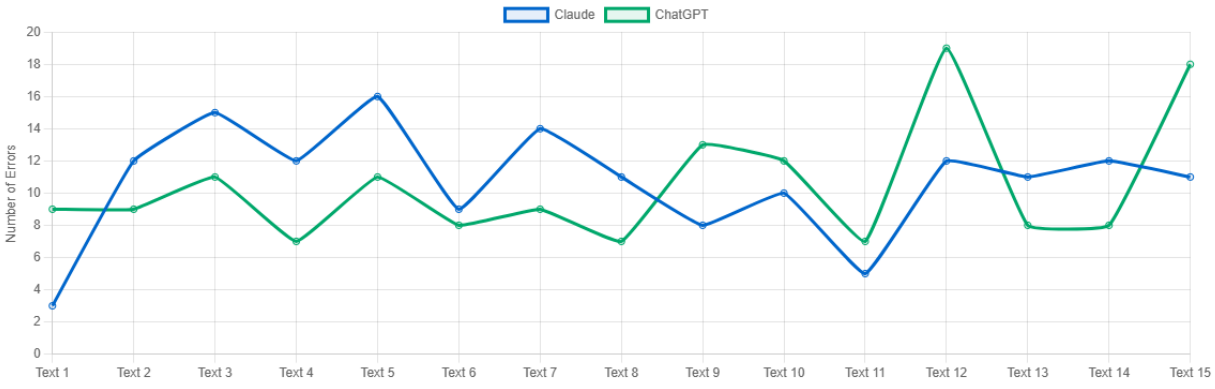
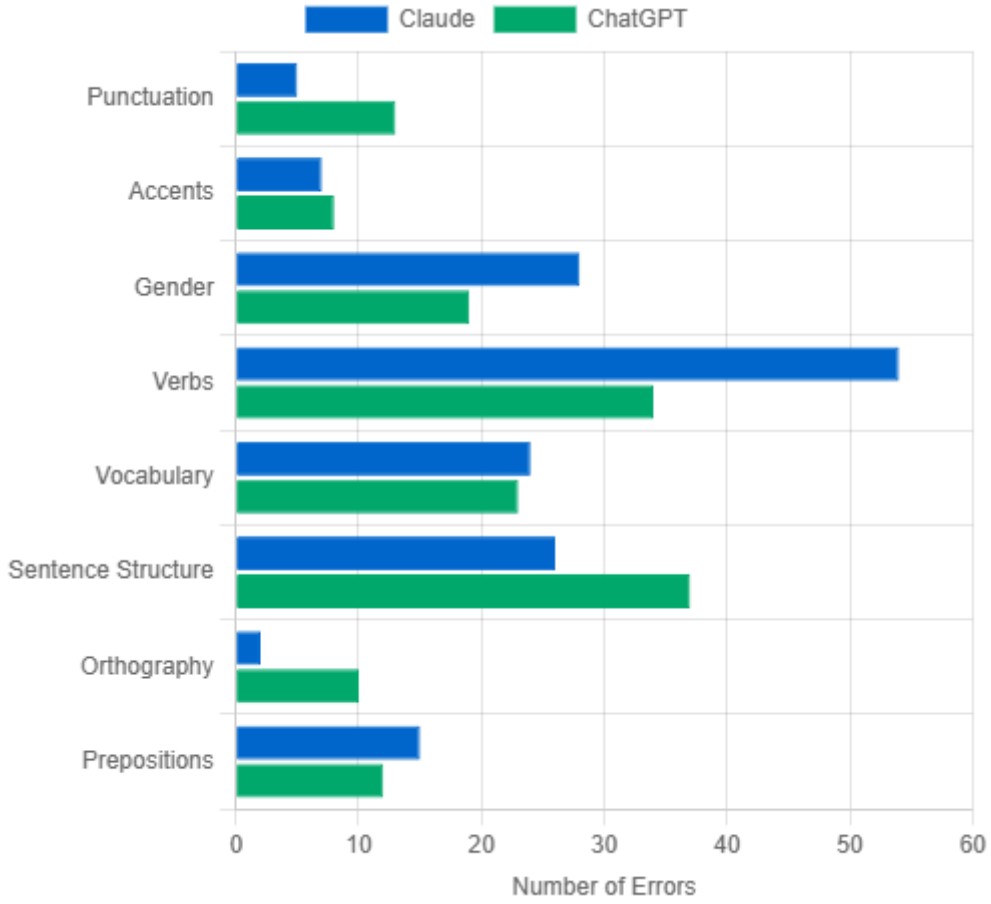


Figure 3.

Error types detected by model.



First of all, the number of errors detected is very similar in both models, and their number broken down by levels is also very similar. There are 161 errors detected by Claude and 156 by ChatGPT. However, an inverse pattern in detection is observed: Claude detects more at low levels and less at high levels, while ChatGPT detects less at low levels and more at high levels. As we cannot really know the whole internal process of the models, the only thing we can point out is that Claude probably chooses to do a complete search for errors, while ChatGPT, perhaps, focuses on the main errors at a low level and only goes into detail at advanced levels.

Secondly, it can be observed that the specific error detection behaviour per text is similar. In general, the peaks behave similarly, and Claude always detects more errors than ChatGPT. The only observable change in trend is precisely in the advanced level group, where Claude's behaviour is consistent with previous analyses, but ChatGPT shows two detection peaks.

Thirdly, figure 3 shows an interesting disparity in some categories. The disparity in correcting verbs, spelling and gender problems is striking, as these are supposed to be objective facts with little interpretation. The disparity in sentence structure is more understandable given that one can be more demanding or more precise with this requirement depending on the criteria being added or the level of learners. The tie in vocabulary, which behaves rather similarly to sentence structure, is surprising. Perhaps these results are due to the fact that lower-level learners make more fundamental errors, and more advanced learners make more errors with sentence structure. Therefore, we understand that the disparity in the detection of objective elements is due to the fact that ChatGPT chooses not to point out all the errors (because it is not able to detect them or

because it does not want to overwhelm the low-level students) or because ChatGPT is detecting errors in verb forms or gender agreement which, in reality, do not exist and it is inventing them. Thus, we consider this type of quantitative analysis to be incomplete, because error detection is not equivalent to correct error detection (Almusharraf & Alotaibi 2022; Ranalli et al., 2017). This is why we consider it relevant to carry out a detailed qualitative analysis of the behaviour of the two models based on the outputs generated from the axes that we consider fundamental (qualitatively) of an error correction analysis.

3.1. Accuracy in error detection

3.1.1. Claude 3.7

Claude showed high general precision in identifying errors in texts by SFL students, with special effectiveness in detecting:

a) Orthographic errors, including accent marks ("encontro → encontró", "despues → después").

b) Morphosyntactic errors, such as agreement problems and incorrect use of prepositions.

c) Expression errors and communicative adequacy, beyond mere ungrammaticality, such as the addition of connectives, for example.

However, some relevant limitations were also observed:

- Over-precision: In texts of initial levels, Claude tended to point out errors related to advanced grammatical structures (such as the use of the subjunctive) that exceed the expectations for that level. For example, in an A1-A2 level text, it pointed out constructions that, although improvable, would be appropriate for a later learning stage.

- Imprecise characterization of errors: Occasionally, Claude incorrectly labelled the type of error detected. A representative case was the classification of "Un bebe → Un bebé" as a gender agreement error, when in reality it is an accent error.

- Hypercorrection of acceptable forms: There were cases in which Claude marked as erroneous constructions that, although less frequent or preferable in certain contexts, are grammatically acceptable. Or, for example: "enciende su cigarrillo → enciende su cigarrillo", where there was no error to correct.

- Bias towards peninsular varieties: Claude showed a tendency to privilege forms from peninsular Spanish, marking as errors correct uses in other varieties. A clear example was the correction "ella se enoja → ella se enfada", indicating that "enfada" is preferable, even though "enoja" is perfectly valid in many varieties of Spanish.

3.1.2. ChatGPT-4

ChatGPT also demonstrated high precision in error detection, although with a different pattern of strengths and limitations:

- Good general identification of orthographic, lexical, and basic agreement errors.

- Difficulty identifying causes: A problematic aspect was the incorrect causal attribution of certain errors. For example, in a case where the student wrote "se sentió" instead of "se sintió", ChatGPT-4 indicated: "In Spanish we use the preterite to tell actions of the past", when in reality it was not an error in the selection of the verb tense, but the ignorance of the change of the lexeme in an irregular verb.

- Hypercorrection: Like Claude, ChatGPT-4 tended to mark grammatically acceptable constructions as erroneous. A significant example was marking "quien está caminando" as an error because "que está caminando" is more frequent, considering the second option more common, when the first one is actually more precise and appropriate for the task.

- Imprecision in contextual explanation: ChatGPT-4 appropriately marked errors such as the student's expression "negro y blanco" instead of "blanco y negro" when describing the type of movie. However, the mere marking of the error without contextual explanation can lead to the student being confused that this order of appearance of colours is like this in any other context, something that is not real, although it may seem so due to the lack of information in the correction feedback.

- Dialectal misinformation: Cases of erroneous information about dialectal variants were observed, such as when it stated: "'Cigarillo' is a more common term in some countries, but in standard Spanish, 'cigarro' is more frequent". In this case, he is explaining that the frequent and standard Spanish form is 'cigarro', while 'cigarrillo' is a less frequent and non-standard form. The frequency criterion is not supported by any data or reality. Moreover, if one consults the academic normative (Real Academia Española de la Lengua) one will observe that the preferred form is 'cigarrillo'. Moreover, the dialectal references confuse students more than they help.

3.2. Clarity of explanations

3.2.1. Claude 3.7

Regarding the explanatory quality of the feedback, Claude presented the following characteristics:

- Contextualized positive reinforcement: Claude systematically highlighted positive aspects of the analysed texts, helping to strengthen knowledge and confirm learning. However, concrete exemplification was sometimes lacking. For example, it indicated "You have good vocabulary to express the sequence of events" without specifying what those words were.

- Structured format: Feedback was generally presented in list format, facilitating the identification of different elements. However, this format sometimes resulted in excessive compartmentalization that made it difficult to perceive interrelated problems.

- Brief grammatical explanations: Claude did not limit itself to pointing out errors and offering corrections but frequently provided concise grammatical explanations. However, this practice was not consistent; for example, when correcting "vio a una carta → vio una carta", it failed to explain the rules for using the preposition "a" with direct objects.

- Memory-based approach for certain aspects: There was a tendency to offer memory-based solutions for phenomena that require understanding of rules, particularly in the case of accent marks, where the underlying principles were not explained.

3.2.2. ChatGPT-4

The analysis of the explanations provided by ChatGPT revealed:

- General guidelines without development: ChatGPT-4 tended to offer generic guidelines ("it is better to write shorter and clearer sentences") without providing concrete indications on how to achieve this or specific examples.

- Conceptual confusions: Potentially confusing or imprecise explanations were detected. A representative example was: "le amo" → "lo quiera" (correct subjunctive, and we use "lo" because it is a masculine direct object)", where it did not explain why the verb was changed or clarify that "amo" is not subjunctive.

- Generic positive feedback: The positive observations provided by ChatGPT-4 were quite similar across different texts, regardless of their specific characteristics, which could affect the credibility perceived by students if results were compared.

- Incomplete explanations: Like Claude, ChatGPT-4 tended to offer corrected versions without explaining the rationale for the changes or the context of the expressions. For example, when correcting dialectal uses like "carriola" to "cochecito", it did not adequately contextualize the variation (when, in fact, it is not appropriate to mark this as an error).

3.3. Pedagogical adequacy

3.3.1. Claude 3.7

The adequacy of Claude's feedback to the educational context of SFL showed the following traits:

- Adaptation to level: In general, Claude adequately adjusted its explanations to the presumed level of the student, albeit with the exceptions already mentioned regarding over-precision.

- Effective motivation: The feedback systematically incorporated motivating elements and recognition of achievements, potentially contributing to a positive learning environment.

- Excessive technical terminology: At times, especially with texts at beginner levels, Claude used grammatical terminology that was too technical, which could make comprehension difficult for students.

- Decontextualization of reformulations: When suggesting sentence reformulations, there was occasionally insufficient context for the student to fully understand the proposal.

- Proposals for complementary activities: A noteworthy aspect was the consistent inclusion of suggestions for additional activities adapted to the detected problems, closing each feedback with a complementary question or exercise.

3.3.2. ChatGPT-4

The evaluation of the pedagogical adequacy of ChatGPT revealed:

- Premature presentation of corrected versions: ChatGPT-4 tended to present a complete version of the corrected text at the beginning of its feedback, which could hinder the process of discovery and gradual construction of learning.

- Initial positive reinforcement: Like Claude, ChatGPT-4 typically began with recognition of positive aspects, although as mentioned, these tended to be less personalized.

- Effective use of conversational attenuators: A noteworthy aspect was the use of natural attenuators ("a bit", "perhaps") that humanized the correction and potentially made it less intimidating.

- Distortion of corrective priorities: There were cases where the same weight was given to errors of different communicative relevance. For example, marking as an error "pegaba a él" instead of "le pegó" without distinguishing between stylistic preference and ungrammaticality.

- Incomplete explanations of complex phenomena: In cases such as "uno peceno niño, in Spanish we say 'un niño pequeño'", important rules such as adjective position were not explained, missing learning opportunities.

- Little systematic organization: The lists of errors were presented without categorization by type, making organized learning and the connection between revision and underlying grammatical principles difficult.

3.4. Detected problems

3.4.1. Claude 3.7

The analysis identified several potential problems in Claude's feedback:

- Internal contradictions: Inconsistencies were observed within the same feedback, such as offering a corrected version with sentence restructuring and then proposing as a complementary activity to perform that same restructuring.

- Ambiguous signalling: At times, Claude pointed out correct phrases by explicitly indicating them as such, without clarifying why they were included in the errors section, generating potential confusion.

- *Horror vacui*: A tendency to look for improvable aspects, even in low-level texts where certain imperfections could be considered admissible for the learning stage, was detected, reflecting an apparent need to always point out areas for improvement.

3.4.2. ChatGPT-4

Among the problems identified in ChatGPT's feedback are:

- *Horror vacui*: Like Claude, ChatGPT-4 showed a tendency to "fill the void" with corrections, even when they were not necessary. For example, in an A2 level text, it suggested changing "deshacerse del niño" to "no sabía qué hacer con él" because, as it indicated, the former "sounds a bit harsh", despite the fact that both expressions are correct and the first is perfectly understandable.

- Lack of dialectal contextualization: ChatGPT-4 tended to mark terms specific to non-peninsular varieties as errors without adequately contextualizing. A clear example was marking "carriola" as an error, indicating that "in Spain we say 'cochecito' or 'carrito de bebé'", without recognizing the validity of the term in other variants of Spanish.

- Pragmatic corrections without justification: Cases were observed where changes were suggested based on subjective perceptions of pragmatic adequacy without offering a clear explanation of why. This type of correction could generate insecurity in the student regarding constructions that are actually grammatically valid.

- Inconsistency in correction criteria: ChatGPT-4 did not maintain uniform criteria between various similar texts, suggesting a certain arbitrariness in its corrective decisions.

4. Discussion

The results obtained reveal significant strengths and limitations of the analysed models as feedback tools in SFL. Below, we discuss the most relevant implications of these findings.

4.1. Interpretation of results

The high general accuracy in the detection of basic errors shown by both models coincides with the observations of Ferreira and Kotz (2010) on the potential of computational systems in linguistic feedback. However, problems such as over-precision, incorrect characterization of errors, and hypercorrection show important limitations in their understanding of the learning process.

The *horror vacui* detected in both models contradicts the principles of selective and adapted feedback proposed by Bailini (2020b). Likewise, the bias towards peninsular varieties suggests the presence of imbalances in the training data, as pointed out by Xiao and Zhi (2023) in their studies on the use of ChatGPT in learning contexts.

Regarding the differences between models, it is significant that Claude 3.7 offers a more systematic structure with complementary activities (i.e. it asks students to do new activities related to the topic or to expand their writing with a focus on other aspects), while ChatGPT-4 stands out in the humanization of feedback through conversational attenuators. This distinction has important implications for the affective reception of feedback, an aspect that Bailini (2020a) identifies as crucial in the learning process.

4.2. Didactic implications

These technologies may contribute to reducing educational inequalities in contexts where personalized attention is limited due to large class sizes or resource constraints. By providing consistent, detailed feedback to all students regardless of their circumstances, AI tools could help democratize aspects of language learning traditionally dependent on high levels of teacher availability. This potential benefit, however, depends on implementing these tools with appropriate safeguards and awareness of their limitations.

Our findings suggest several practical recommendations for effectively implementing AI models in SFL classrooms, tailored to different educational stakeholders. Teachers should approach AI-generated feedback as a starting point rather than a finished product. Prior to sharing AI feedback with students, educators would benefit from reviewing the content to identify and correct potential errors or inaccuracies, ensuring students receive reliable guidance. This review process need not be exhaustive but should focus on areas where AI models have demonstrated limitations.

Effective implementation requires strategic rather than comprehensive teacher intervention. Rather than contextualizing every piece of AI feedback, teachers can focus their efforts on areas where our analysis showed AI limitations are most problematic—such as complex grammatical explanations, dialectal variations, and level-appropriate corrections. This selective approach maintains educational quality while still reducing overall teacher workload compared to traditional manual correction. Given the varying reliability of AI feedback across different linguistic domains, selective use represents another valuable strategy. Teachers might employ these tools primarily for first-round reviews or for aspects where they have demonstrated greater reliability, such as detection of basic orthographic errors, while maintaining more direct intervention for complex grammatical, pragmatic, or cultural issues where AI systems show inconsistent performance.

Innovative educators can transform the very limitations of these models into pedagogical opportunities. By inviting students of higher proficiency levels to critically evaluate AI-generated feedback, teachers can develop students' metalinguistic awareness and analytical skills. This approach, similar to that proposed by López Mata (2023), positions students as active evaluators rather than passive recipients of feedback.

Finally, teachers can significantly improve AI performance through careful design of specific prompts. By developing optimized instructions that guide the models towards appropriate feedback for particular contexts, assignments, or student levels, educators can mitigate some of the systems' tendencies for inappropriate correction styles.

Students engaging with AI feedback tools would benefit from developing a critical perspective with regard to the guidance they receive. Rather than accepting AI corrections as authoritative, learners should approach this feedback with healthy scepticism, comparing it with other sources when uncertainties arise. This critical engagement itself constitutes a valuable learning process that strengthens linguistic judgment.

Understanding the complementary nature of these tools is equally important. Students should view AI feedback as supplementing, not replacing, teacher guidance and conventional learning resources. Particularly for complex or nuanced aspects of language learning, human expertise and traditional explanatory materials often provide clarity that AI systems currently cannot match.

4.3. Integrated model for the SFL classroom

Based on our findings, we propose a three-phase model for the implementation of these tools in the SFL classroom:

a) The teacher as mediator: Selecting appropriate tools, validating the feedback generated by AI, and contextualizing it pedagogically.

b) The student as critical user: Developing digital competences and evaluative capacity to interact effectively with these systems, in line with the autonomy approach proposed by Benson (2006) and García Pujals and Lasagabaster (2019).

c) AI as a complementary tool: Providing immediate and systematic feedback that can complement, never substitute, teacher intervention.

This model recognizes both the transformative potential of AI and its inherent limitations, maximizing benefits and minimizing risks through teacher mediation and the development of critical thinking in students.

4.4. Limitations and future lines

Among the limitations of this study are the small sample size, the constant evolution of AI models, and the absence of evaluation with real users. As future lines of research, it would be valuable to develop longitudinal studies on the real impact of these tools, optimize specific instructions for SFL, and explore the perceptions of students and teachers, following the line of works such as those of Slamet (2024) and Feng Teng (2024).

5. Conclusions

Our analysis reveals that both Claude and ChatGPT demonstrate considerable potential as feedback tools for SFL students, particularly excelling in the detection of basic grammatical errors and provision of corrections. However, this potential is tempered by notable limitations including tendencies toward overprecision, incorrect characterization of errors, inadequate adaptation to student levels, and insufficient contextualization of explanations. These limitations highlight the complexity of fully automating the nuanced work of language instruction.

Claude 3.7 distinguishes itself through systematic structuring and consistent inclusion of complementary activities, creating a more comprehensive learning experience. In contrast, ChatGPT-4 excels at humanizing feedback through conversational attenuators, potentially making corrections less intimidating for learners. Despite these differences, both systems share problematic tendencies including *horror vacui* (the compulsion to correct even correct constructions), bias favoring peninsular Spanish variants over equally valid alternatives, and providing insufficient explanations for complex linguistic phenomena.

The contradictions, inconsistencies, and excessive generalizations detected in both systems' feedback firmly establish that these AI models cannot function autonomously as substitutes for teacher feedback. Their maximum utility emerges when they operate as

complementary tools under the supervision and mediation of experienced educators who can contextualize, supplement, and occasionally correct the AI-generated feedback. This finding aligns with broader educational research suggesting that technology functions best as an augmentation rather than replacement of human expertise.

Despite these limitations, both models offer significant possibilities for fostering student autonomy by providing opportunities for immediate feedback outside the classroom setting. This immediate response capability can complement teacher attention and support the development of self-regulation capacity – which, as Benson (2006) and Bailini (2020b) emphasize, constitutes an essential component of effective language learning. The accessibility of these tools allows students to receive feedback at times and places convenient to them, potentially accelerating the learning process.

AI models such as Claude and ChatGPT represent a significant advance in the possibilities of automated feedback for SFL learning. Their ability to analyze complex texts, detect a wide range of errors, and provide structured explanations far exceeds the capabilities of previous tutorial systems. However, their current limitations—particularly in deeply understanding the learning process and responding appropriately to specific contexts—confirm that their optimal role remains that of complementary tools supporting, rather than replacing, teacher mediation.

The future of SFL teaching will likely evolve toward integrated models where human expertise and AI capabilities complement each other to create more personalized, immediate, and effective learning experiences. As Xiao and Zhi (2023) and Feng Teng (2024) suggest, AI can function as a collaborative companion in the teaching-learning process rather than either an adversary or replacement for human educators.

In this emerging educational landscape, parallel advancement of both research and practice becomes crucial. These domains must inform each other iteratively to develop approaches that maximize the potential of technologies' while minimizing associated risks. Through critical and reflective integration of AI in this field, we can meaningfully contribute to our fundamental objective: facilitating the development of students' communicative competence in an increasingly interconnected and multilingual world.

Ethical statement

We declare that there are no conflicts of interest or ethical concerns. The AI tools ChatGPT and Claude have been used as the object of study. Claude has also been used as a proofreader.

References

Almusharraf, H. & Alotaibi, H. (2022). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning*, *28*. https://doi.org/10.1007/s10758-022-09592-z

Bailini, S. (2020a). El feedback como herramienta didáctica para el desarrollo de la autonomía en la adquisición de lenguas extranjeras. *Philologia Hispalensis*, *34*, 25-39. https://dx.doi.org/10.12795/PH.2020.v34.i01.02

Bailini, S. (2020b). El feedback interactivo y la adquisición del español como lengua extranjera. Mimesis.

Benson, P. (2006). Autonomy in language teaching and learning. *Language Teaching*, *40*, 21-40. https://doi.org/10.1017/S0261444806003958

Buyse, K. (2014). Una hoja de ruta para integrar las TIC en el desarrollo de la expresión escrita: recursos y resultados. *Journal of Spanish Language Teaching*, *1*(1), 101-115. https://doi.org/10.1080/23247797.2014.898516

Coyne, S., Sakaguchi, K., Galvan-Sosa, D., Zock, M. & Inui, K. (2023). Analyzing the performance of GPT-3.5 and GPT-4 in grammatical error correction. *arXiv:2303.14342*. https://doi.org/10.48550/arXiv.2303.14342

Crossley, S. A., Bradfield, F. & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270. https://doi.org/10.17239/jowr-2019.11.02.01

Feng Teng, M. (2024). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*. https://doi.org/10.1016/j.caeai.2024.100270

Fernández, S. (2017). Evaluación y aprendizaje. *MarcoELE: Revista de Didáctica Español Lengua Extranjera*, *24*, 1-43. http://marcoele.com/descargas/24/fernandez-evaluacion_aprendizaje.pdf

Ferreira, A., & Kotz, G. (2010). ELE-Tutor Inteligente: Un analizador computacional para el tratamiento de errores gramaticales en Español como Lengua Extranjera. *Revista signos*, 43(73), 211-236. https://dx.doi.org/10.4067/S0718-09342010000200002

García, M. (2024). ChatGPT: posibles aplicaciones y recomendaciones de uso en ELE. In *ELEUK ampliando horizontes: propuestas didácticas y avances en investigación* (pp. 121-139). Instituto Cervantes.

García Pujals, A., & Lasagabaster, D. (2019). El efecto de la evaluación y la retroalimentación en la autonomía, la motivación y el aprendizaje del español como L3. *Revista Española de Lingüística Aplicada*, 32(2), 455-485. https://doi.org/10.1075/resla.17050.gar

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*. https://doi.org/10.1016/j.lindif.2023.102274

López Mata, D. (2023). ChatGPT en la clase de preparación al DELE. Propuesta didáctica e impresiones de los estudiantes de ELE. *Revista Nebrija De Lingüística Aplicada a La Enseñanza De Lenguas*, 17(35). https://doi.org/10.26378/rnlael1735533

Ranalli, J. (2021) L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52.* https://doi.org/10.1016/j.jslw.2021.100816

Ranalli, J., Link, S. & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8-25. https://doi.org/10.1080/01443410.2015.1136407

Slamet, J. (2024). Potential of ChatGPT as a digital language learning assistant: EFL teachers' and students' perceptions. *Discoveries in Artificial Intelligence*, 4. https://doi.org/10.1007/s44163-024-00143-2

Xiao, Y., & Zhi, Y. (2023). An Exploratory Study of EFL Learners' Use of ChatGPT for Language Learning Tasks: Experience and Perceptions. *Languages*, 8(3). https://doi.org/10.3390/languages8030212