ShAnEL-2: A Multilingual Benchmarking Dataset for Short-Answer Language Learning Exercises

Jasper Degraeuwe, Thomas Moerman
Language and Translation Technology Team (LT^3), Ghent University
Ghent, Belgium
jasper.degraeuwe@ugent.be, thomas.moerman@ugent.be

Abstract

Before using GenAI models as EdTech tools, their pedagogical suitability should be corroborated. In this paper, we present `ShAnEL-2`, a novel multilingual dataset comprising 1,185 student responses to short-answer language learning exercises corrected by teachers. We use `ShAnEL-2` to establish an initial benchmark of (1) "off-the-shelf" GenAI models and (2) retrieval-augmented generation (RAG) techniques for the automated correction of this exercise type. With an overall accuracy of 90% and recall of 95%, few-shot RAG (which adds previously corrected responses to the prompt) outperforms the off-the-shelf baseline and textbook RAG setup (which adds coursebook materials) by up to 7 (accuracy) and 5 (recall) percentage points. These results confirm that LLMs learn better from examples than from analysing context and highlight GenAI's particular potential as a correction assistant for teachers.

Keywords: short-answer exercises, second language acquisition, educational technology (EdTech), automated exercise correction, generative artificial intelligence (GenAI), retrieval-augmented generation

1. Introduction

Recently, generative artificial intelligence (GenAI) tools and their underlying large language models (LLMs) have revolutionised the field of education (Bozkurt, 2023). Since OpenAI's release of ChatGPT in 2022, GenAI models have been widely used in an off-the-shelf manner by students (e.g., as writing assistants or advanced search engines), teachers (e.g., to provide corrective feedback), and coursebook designers alike (e.g., to create infographics). In a similar vein, the domain of educational technology (EdTech) has witnessed the integration of GenAI and LLMs becoming common practice (Concannon et al., 2023). EdTech tools can be subdivided into three main domains (Stošić, 2015): (1) technology as a tutor, (2) technology as a teaching tool, and (3) technology as a learning tool.

However, since GenAI models are not specifically designed for educational purposes, we should exercise caution when using them as EdTech tools. When GenAI is used as a tutor, for example, it should be avoided at all costs that correct student responses are flagged as incorrect (and vice versa). In other words, before implementing GenAI models in any of the three abovementioned EdTech domains, we should first scientifically corroborate that they can adequately evaluate student responses.

To fill part of this newly emerged gap, this paper presents `ShAnEL-2` (**Sh**ort-**An**swer **E**xercises for **L2** learning), a multilingual benchmarking dataset for the educational domain of second language acquisition (SLA). The dataset includes 1,185 authentic student responses to 237 short-answer exercises across nine grammar topics (three per language). Each response comes with a binary correction (i.e. labelled as *correct* or *incorrect*) and an error annotation provided by expert SLA teachers.

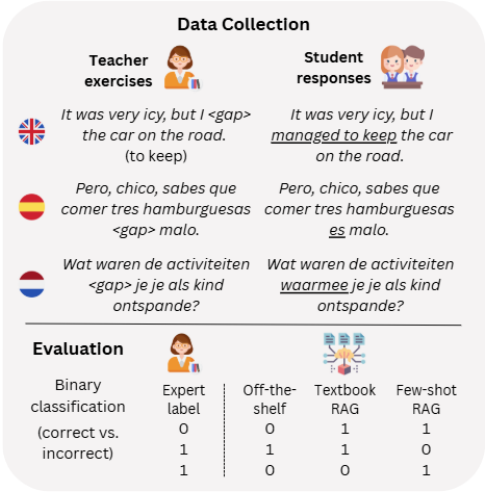

Figure 1: Visualisation of the methodology.

We evaluate model performance using two metrics that carry distinct pedagogical implications: `accuracy` and `recall` (of the *incorrect* class). High accuracy is essential when the model serves directly as a learning tool for students, as both false positives (correct responses flagged as incorrect) and false negatives (incorrect responses labelled as correct) undermine the learning experience. High recall is particularly important when the model is used as a correction assistant for teachers: detecting all incorrect responses is paramount, as teachers can subsequently filter out any false positives before returning the exercises to their students.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6764–6771
11–16 May 2026.

Exercise type (topic)	Instruction	Target sentence	Example solution	Additional possible solutions
Gap-filling (*can* versus *may* + alternatives)	Complete the following sentence using: can, could (have), be able to, manage to, succeed in, may (have), might (have). It may be necessary to use passives or negatives.	I <gap> a horse when I was twenty, but I am too old for it now. (to ride)	could ride	(1) may have been able to ride; (2) might have been able to ride
Rephrasing (conditionals)	Rewrite the following sentence as a conditional. Make sure the original meaning is preserved.	She underwent plastic surgery because she didn't know about the consequences.	If she had known about the consequences, she would not have undergone plastic surgery.	(1) Had she known [...] she would not have undergone [...]; (2) She wouldn't have undergone [...] if she had known [...]

Table 1: Example of gap-filling and rephrasing exercise taken from ShAnEL-2.

As a secondary contribution, we use the ShAnEL-2 dataset to answer the following research questions:

- RQ1 – How well do pretrained LLMs perform at automatically correcting short-answer language learning exercises? This setup, referred to as **off-the-shelf**, serves as our baseline.
- RQ2 – Do more dedicated retrieval-augmented generation (RAG) techniques allow us to beat this baseline? We analyse two setups: (1) retrieving information from coursebook materials containing theory on the grammar topic (**textbook RAG**) and (2) consulting a database containing previously corrected exercises on the same topic (**few-shot RAG**).

Figure 1 presents a simplified visualisation of the complete process, from dataset creation to testing an LLM on the dataset across three different setups. To the best of our knowledge, this kind of methodology has not yet been explored for the specific purpose of automatically evaluating short-answer exercises in an SLA context.

2. Related Research

2.1. Second Language Acquisition

In addition to highlighting the use of *implicit* learning activities (e.g., reading a book), various SLA studies have also argued in favour of including *explicit* activities (e.g., gap-filling exercises) in L2 study programmes (DeKeyser, 2003; Ellis, 2002; Nation, 2019). Although learning gains in explicit learning can vary depending on the activity type (Webb et al., 2020), there is ample empirical evidence indicating that "focused L2 instruction results in large target-oriented gains, that explicit types of instruction are more effective than implicit types, and that the effectiveness of L2 instruction is durable" (Ellis, 2002, p. 145).

In our dataset, we focus on two different types of explicit learning activities: **gap-filling** and **rephrasing** (see Table 1). Unlike multiple-choice exercises, these two kinds of short-answer exercises can have multiple correct answers: although they usually include one obvious answer (the "example solution"), gap-filling and rephrasing exercises can also have other correct answers that are equally valid but much harder to predefine. This phenomenon – denominated "multiple admissibility" (Katinskaia and Ivanova, 2019) – is illustrated in the "Additional possible solutions" column of Table 1. In other words, automating the correction of gap-filling and rephrasing exercises using a rule-based method (as can be done for multiple-choice exercises) does not guarantee a 0% error rate, which is often required in a pedagogical setting.

2.2. Automated Exercise Correction

Previous research on automated exercise correction in SLA and datasets built for this purpose mostly originated in the domains of grammatical error detection/correction (GED/GEC) and automated corrective feedback generation. Datasets are typically based on error-annotated learner corpora containing essays (e.g., Volodina et al., 2021) or exam responses (e.g., Yannakoudakis et al., 2011). ShAnEL-2 constitutes a valuable addition to this body of existing resources, providing high-quality data for gap-filling and rephrasing exercises.

Regarding models, the development of Transformer-based architectures (Vaswani et al., 2017) caused a shift from rule-based and supervised models to (generative) LLMs, for GED/GEC (Katinskaia and Yangarber, 2021), short-answer grading (Willms and Pado, 2022), as well as feedback generation (Mazzullo and Bulut, 2024).

Especially the use of synthetic data as additional training data, the combination of different pre-trained LLMs, and the use of performance-boosting techniques have shown to push the state of the art (Bryant et al., 2023). The main contribution of the present study in this regard is the exploration of "pedagogically-oriented" RAG techniques with open-source LLMs.

2.3. Retrieval-Augmented Generation

RAG improves LLM performance by integrating relevant external information (Lewis et al., 2020). This approach prioritises effective information retrieval over architectural modifications, as providing LLMs with appropriate contextual information often proves more effective than developing task-specific architectures. RAG has been applied to question-answering systems that retrieve passages from knowledge repositories (Karpukhin et al., 2020), code generation using relevant code snippets (Zhou et al., 2022), and dialogue systems that incorporate conversation history or external knowledge (Shuster et al., 2021).

In educational contexts, RAG addresses domain-specific challenges in intelligent tutoring systems, personalised learning, and automated assessment. Systematic surveys of over 51 studies show that RAG architectures improve factual accuracy and enable dynamic knowledge updates in educational systems without requiring model retraining (Li et al., 2025b). For second language learning, Li et al. (2025a) introduced explanation-based in-context demonstration retrieval for multilingual GEC, showing that retrieving examples based on grammatical error patterns outperforms semantic and BM25-based approaches.

A straightforward yet successful RAG approach is few-shot learning, in which task-specific examples are provided to the LLM by directly incorporating them into the prompt (Brown et al., 2020). In few-shot learning, the model's parameters remain fixed during inference, and the model relies on pattern recognition from the provided examples to generalise to new inputs. This has been proven effective, for example, in machine translation, where providing just five translation examples at inference time can enable models to match or exceed state-of-the-art supervised systems (Garcia et al., 2023).

3. Data Compilation

The ShAnEL-2 dataset focuses on three target languages: L2 Dutch, L2 English, and L2 Spanish. To render the dataset as relevant as possible for our SLA target setting, we asked three experienced L2 teachers (one per language) to (1) choose three different grammar topics and (2) provide, per topic, the coursebook materials and exercises (including the example solution) they use to teach the topic to their students. An overview of the topics and number of exercises per topic is provided in Table 2. As described in Section 2.1, we allowed teachers to provide either **gap-filling** or **rephrasing** activities, as these short-answer exercises constitute an excellent starting point for building a benchmarking dataset for SLA exercise evaluation.

L2	Grammar topic	#exercises
EN	*Can* versus *may* (+ alternatives)	25
	Conditionals	25
	Expressions of the future	19
ES	Past tenses	29
	Ser versus *estar*	14
	The subjunctive	22
NL	Indirect speech	25
	Relative clauses	40
	Present perfect	38

Table 2: Number of exercises per grammar topic.

Teachers were given full freedom in selecting grammar topics, resulting in nine distinct topics. We are aware that, as a result, it is impossible to directly compare performance results of models tested on ShAnEL-2 across the three target languages. However, as each language has its own grammatical challenges and L2 curriculum design, we believe that prioritising cross-linguistic comparability would have jeopardised the dataset's real-world nature. Moreover, even if the same grammar topics are chosen, this does not necessarily mean that the results can be compared directly. In the case of "past tenses", for example, there are considerable differences in how these tenses are formed across the three languages, which can make the topic easier or harder depending on the L2 (e.g., Spanish makes an important distinction between the *imperfecto* ['imperfect'] and *pretérito simple* ['simple past'], while Dutch and English do not).

Next, we recruited 15 L2 learners (five per target language) and asked them to complete all exercises for their particular L2. Finally, the three L2 teachers were asked to perform a binary evaluation of the student responses, labelling them as either *correct* or *incorrect*. One teacher per language was deemed sufficient, as the short-answer exercises under consideration have objectively determinable correct or incorrect labels, leaving little room for subjective interpretation. Teachers were instructed to disregard the severity of the errors: as soon as the response contained a mistake, they were required to label the response as *incorrect*. For

rephrasing exercises, teachers were also asked to provide error annotations by indicating the sentence parts containing the error and suggesting a possible correction[1]. As a result, we obtained a final dataset containing 1,185 authentic student responses, labelled as either *correct* (847 instances) or *incorrect* (338 instances) and provided with an error annotation.[2]

4. Initial Experiment

In this section, we report on the results of our initial experiment using `ShAnEL-2`. Based on the results from a pilot study (Degraeuwe and Moerman, 2025), we opted for using open-source, multilingual generative models from the `Gemma 3` family released by Google (`gemma-3-4b-it` and `gemma-3-27b-it`). The choice for open-source LLMs was deliberate, as these models can be run locally, thereby ensuring full control over the data submitted to them. As a result, they *a priori* avoid the privacy issues (student responses can contain personal and/or sensitive data) and copyright issues (adding copyrighted coursebook materials to the prompt) that often arise when using proprietary generative LLMs in an educational setting.

As mentioned in the introduction, we tested the models in three setups, described below. The task was identical across the setups: models were instructed to output a binary judgement for each exercise, labelling the exercises either as *correct* or *incorrect*. These predictions were then compared against the true labels included in the dataset to assess the performance of the models. An overview of the three different setups, their prompts[3], and augmentation can be found in Table 4.

- **Off-the-shelf** – In this baseline setup, the prompt ("base prompt" hereafter) submitted to the LLM contains the exercise instruction, exercise item, student response, example solution, and a brief description of the task.
- **Textbook RAG** – For this setup, the base prompt was tailored to the SLA setting by adding the coursebook materials[4] provided by the L2 teachers (see Section 3) to the prompt as knowledge for the model to use upon correcting the exercises.
- **Few-shot RAG** – For this setup, the base prompt was tailored to our SLA setting by adding four teacher-corrected responses to the prompt as examples for the model to learn from. The added examples correspond to the responses from the four other learners on the same exercise item as the target response.

Across all setups, the model receives the same core information available to a teacher: the exercise instruction, the exercise item, the example solution, and, in the few-shot RAG setup, corrected responses from other students on the same exercise item. Our goal is to assess how well LLMs can leverage this pedagogical knowledge to arrive at the correct judgment. Importantly, we deliberately chose not to exclude cases where the student response exactly matches the example solution. While such exact matches could be trivially handled by a rule-based filter in a real-life deployment pipeline, we include them in our experiment because our objective is to evaluate the model's overall reasoning ability across the full range of responses a teacher encounters.

To assess performance, we computed two metrics: `accuracy` and `recall`. Obtaining high accuracy is crucial when the model is used directly as a learning tool by L2 learners (i.e. it is equally important that correct exercises are not flagged as incorrect and that incorrect exercises are not labelled as correct). Obtaining high recall is essential when the model is used by L2 teachers: here, recognising all incorrect exercises is most important, since the teacher can still manually filter out correct exercises that were previously identified as incorrect and return them to the students.

Table 3 presents the results of the initial experiment. Regarding RQ1, the results indicate that pretrained models achieve good performance but are insufficient for direct use in a pedagogical setting (overall accuracy of 0.83 and recall of 0.9). Regarding RQ2, the results suggest that the "SLA-based" RAG approaches are generally able to outperform the off-the-shelf baseline (in particular, few-shot RAG). With an average accuracy of 0.9 and recall (for the *incorrect* class) of 0.95, few-shot RAG with `27b-it` outperforms the other setups by 0.07 (accuracy) and up to 0.05 (recall) across the three languages. This finding implies that adding relevant examples to the prompt is more useful than adding extensive theoretical descriptions of the grammar topic tackled in the exercise.

The difference in performance between textbook RAG and few-shot RAG is especially noticeable with the smaller `4b-it` model. In the textbook RAG setup, the `4b-it` model fails to recognise incorrect responses and instead predicts *correct* almost exclusively, as evidenced by the near-zero recall scores (0.12, 0.04, and 0.03 for EN, ES, and

[1] For gap-filling exercises, this step was not necessary, as it is known that the error occurs in the gap.

[2] The dataset (including participant metadata) is made publicly available in a GitHub repository `https://github.com/JasperD-UGent/ShAnEL-2` under an ODC-By licence.

[3] The full prompts for all setups are available in the dataset repository.

[4] For illustrative purposes, we share a small excerpt of these copyrighted materials in the dataset repository.

Setup	EN				ES				NL				Overall			
	4b-it		27b-it		4b-it		27b-it		4b-it		27b-it		4b-it		27b-it	
	A	R	A	R	A	R	A	R	A	R	A	R	A	R	A	R
Off-the-shelf	.85	.89	.86	.82	.74	.75	.87	**.93**	.75	.39	.75	.96	.78	.67	.83	.90
Textbook RAG	.72	.12	.9	.92	.72	.04	.82	**.93**	.74	.03	.76	.92	.73	.07	.83	.92
Few-shot RAG	**.91**	**.91**	**.94**	**.97**	**.89**	**.88**	**.93**	**.93**	**.82**	**.86**	**.83**	**.97**	**.87**	**.89**	**.90**	**.95**

Table 3: Performance of Gemma models on ShAnEL-2. "A" stands for accuracy and "R" for recall of the *incorrect* class.

System setup	Prompt components	Example augmentation (beyond baseline)
Off-the-shelf (= baseline)	**Base prompt:** • Task description • Exercise instruction • Exercise item • Student response • Example solution • Output format	*None* (relies on LLM's pretrained knowledge)
Textbook RAG	Base prompt + • **Coursebook materials**	Grammar theory: • COULD = past ability as a state • WAS ABLE TO = past ability as an event Example sentences: • "He could swim very well at the age of six." • "After the accident, she was able to swim ashore."
Few-shot RAG	Base prompt + • **Four teacher corrections** (from peers on same item)	Example 1 - Response "can" → {"correct": false, "corrected_response": "could"} Example 2 - Response "could" → {"correct": true, "corrected_response": null} [two more examples]

Table 4: Overview of the three system setups for automated exercise correction.

NL, respectively). This behaviour disappears when scaling to the 27b-it model, which achieves recall scores of 0.92, 0.93, and 0.92 in the same setup. We hypothesise that this discrepancy is due to the larger model's superior ability to process the extensive coursebook materials included in the context window, although a thorough investigation of this phenomenon falls outside the scope of the current study and warrants further research. The overall accuracy of textbook RAG with 4b-it (0.73) is also lower than that of the off-the-shelf baseline (0.78), further confirming that the smaller model struggles to leverage the added textbook context.

Regarding model sizes, the results indicate that, though margins are occasionally small, the 4b-it model is consistently outperformed by its large-parameter counterpart. Particularly for recognising incorrect responses, the 27b-it model performs considerably better, as shown by the significantly lower recall values (only for English, in off-the-shelf, a higher recall is obtained with the 4b-it model). As for languages, the highest accuracy is achieved for English and Spanish (both high-resource languages), although it is worth noting that the highest recall score is obtained for lower-resource Dutch (0.97 for few-shot RAG with 27b-it).

5. Discussion and Conclusion

In this paper, we aimed to contribute to the effective implementation of GenAI in the EdTech domain. We focused on the task of automated short-answer

grading, to which we made two main contributions: (1) the release of the multilingual ShAnEL-2 dataset containing 1,185 authentic L2 learner responses to gap-filling and rephrasing exercises and (2) a study in which we tested Gemma 3 on ShAnEL-2 in three different setups. With a 90% accuracy and 95% recall, few-shot RAG outperforms both the off-the-shelf and textbook RAG setup, highlighting that (1) providing pretrained LLMs with pedagogically relevant data is beneficial and (2) LLMs learn better from examples than from "reasoning" over context (which confirms findings from previous research; Aycock et al., 2025).

Interpreting the results from an educational perspective, we can conclude that open-source generative LLMs show great potential to be used in the process of automatically correcting short-answer language learning exercises. Especially as a tool for teachers, they show real-life viability: if the near-perfect recall score of 0.95 can be pushed to 1, this would mean that teachers can assume that all responses labelled as correct are indeed correct and only have to filter out the false positives from the list of responses labelled as incorrect, a much less time-consuming effort.

6. Limitations

We acknowledge that, by benchmarking standards, the ShAnEL-2 dataset is relatively small-sized. Moreover, the dataset exhibits class imbalance: 847 responses (71.5%) are labelled *correct*, and only 338 (28.5%) are labelled *incorrect*, which should be taken into account when interpreting the reported performance metrics. Additionally, the languages included in the dataset have a Eurocentric focus. Yet, given its high authenticity (real-life exercises, responses from actual L2 learners, and corrections from expert teachers), we believe the dataset constitutes a valuable addition to the domain of automated short-answer grading. Moreover, we plan to release a v2.0 of the dataset in the near future, with more exercises and student responses.

Secondly, it should be noted that the task considered for the initial experiment (i.e. binary *correct* versus *incorrect* labelling) is relatively straightforward. To gain a deeper understanding of the pedagogical capabilities of GenAI models in an SLA setting, future research should analyse (1) if the models are able to detect all errors in the response (i.e. GED, see Section 2.2), (2) if their corrections for exercises they labelled as *incorrect* are accurate (i.e. GEC), and (3) if the feedback they generate is pedagogically suitable. The first type of analysis can be performed semi-automatically thanks to the error annotations included in ShAnEL-2; for the second and third type of analysis, we will develop an evaluation procedure with L2 teachers as the participants.

Finally, only one LLM (family) was included in the study. While this decision was based on previous research (Degraeuwe and Moerman, 2025; in which Gemma clearly outperformed the Llama and Mistral families), testing more LLMs in future research could be useful to put our findings in a broader perspective.

7. Ethical Considerations

The 15 L2 learners who participated in the study were offered a small financial compensation of 25 euros for completing the exercises. Apart from their native language, the number of years they had been studying the L2, and their proficiency level, no personal data were gathered. Participants were informed that their responses would be used for scientific research in a pseudonymised fashion, with the abovementioned personal data linked to a unique ID (e.g., "EN_1" for the first L2 English participant).

The research presented in this paper involves the use of LLMs. As a result, any potential EdTech[5] applications developed based on this research should take into account the inherent limitations of LLMs, such as hallucinations and different types of bias stemming from data they were trained on (e.g., gender bias and Western perspective).

8. Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government - department EWI. Additionally, we wish to express our sincere gratitude to the anonymous reviewers for their valuable feedback and to the L2 teachers and learners for their participation in the study.

9. Bibliographical References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can LLMs Really Learn to Translate a Low-Resource Language from One Grammar Book? _eprint: 2409.19151.

Aras Bozkurt. 2023. Unleashing the Potential of Generative AI, Conversational Agents and Chatbots in Educational Praxis: A Systematic Review and Bibliometric Analysis of GenAI in Education. *Open Praxis*, 15(4):261–270.

[5] EdTech stands for educational technology.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, pages 1–59.

Fiona Concannon, Eamon Costello, Orna Farrell, Tom Farrelly, and Leigh Graves Wolf. 2023. Editorial: There's an AI for that: Rhetoric, reality, and reflections on EdTech in the dawn of GenAI. *Irish Journal of Technology Enhanced Learning*, 7(1).

Jasper Degraeuwe and Thomas Moerman. 2025. Retrieval-augmented Generation for Automated Written Corrective Feedback: From Dataset to Human Evaluation. In *35th Meeting of Computational Linguistics in The Netherlands (CLIN 35), Abstracts*, Leuven, Belgium.

Robert DeKeyser. 2003. Implicit and explicit learning. *The handbook of second language acquisition*, pages 312–348.

Nick C. Ellis. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. ArXiv:2302.01398 [cs].

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *EMNLP (1)*, pages 6769–6781.

Anisia Katinskaia and Sardana Ivanova. 2019. Multiple Admissibility: Judging Grammaticality using Unlabeled Data in Language Learning. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 12–22, Florence, Italy. Association for Computational Linguistics.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing Grammatical Correctness in Language Learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–146, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, 33:9459–9474.

Wei Li, Wen Luo, Guangyue Peng, and Houfeng Wang. 2025a. Explanation based In-Context Demonstrations Retrieval for Multilingual Grammatical Error Correction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4881–4897, Albuquerque, New Mexico. Association for Computational Linguistics.

Zongxi Li, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang. 2025b. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8:100417.

Elisabetta Mazzullo and Okan Bulut. 2024. Automated Feedback Generation for Open-Ended Questions: Insights from Fine-Tuned LLMs. In *Proceedings of Machine Learning Research*, pages 1–18.

I.S.P. Nation. 2019. The Different Aspects of Vocabulary Knowledge. In Stuart Webb, editor, *The Routledge Handbook of Vocabulary Studies*, pages 15–29. Routledge, London.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. ArXiv:2104.07567 [cs].

Lazar Stošić. 2015. The importance of educational technology in teaching. *International Journal of Cognitive Research in Science, Engineering and Education*, 3(1):111–114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elena Volodina, Yousuf Ali Mohammed, and Julia Klezl. 2021. DaLAJ – a dataset for linguistic acceptability judgments for Swedish. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 28–37, Online. LiU Electronic Press.

Stuart Webb, Akifumi Yanagisawa, and Takumi Uchihara. 2020. How Effective Are Intentional Vocabulary-Learning Activities? A Meta-Analysis. *The Modern Language Journal*, 104(4):715–738.

Nico Willms and Ulrike Pado. 2022. A Transformer for SAG: What Does it Grade? In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 114–122, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Shuyan Zhou, Uri Alon, Frank F. Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2022. Docprompting: Generating code by retrieving the docs. *arXiv preprint arXiv: 2207.05987*.