A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice

Stephen Bodnar
Tübingen Center for Digital Education, University of Tübingen, Germany
stephen.bodnar@uni-tuebingen.de

Abstract

ICALL systems that offer grammar exercises with authentic texts have the potential to motivate learners, but finding suitable documents can be problematic because of the low number of target grammar forms they typically contain. Meanwhile, research showing the ability of Large Language Models (LLMs) to rewrite texts in controlled ways is emerging, and this begs the question of whether or not they can be used to modify authentic L2 texts to increase their suitability for grammar learning. In this paper we present a tool we have developed to explore this idea.The authoring tool employs a lexical database to create prompts that instruct an LLM to insert specific target forms into the text. We share our plans to evaluate the quality of the automatically modified texts based on human judgments from native speakers.

1 Introduction

Perhaps because learning grammar is sometimes perceived as boring by students (e.g., Jean and Simard, 2011), researchers have explored a variety of techniques for spicing up computerised grammar practice. For example, Colling et al. (2024) developed a student dashboard that highlighted the relevance of practice exercises to communicative tasks. Adding gamification elements to make grammar practice more exciting or enjoyable is another possibility (Strik et al., 2013). Occasionally researchers develop speech-interactive grammar practice to help develop oral proficiency (Drozdova et al., 2013). Still another approach is to contextualise the practice by situating it within an interesting mystery narrative (Cornillie et al., 2013). The work we present here connects with previous work in Intelligent Computer-Assisted Language Learning (ICALL) that contextualises grammar practice through the use of authentic texts (e.g., Meurers, Ziai, Amaral, Boyd, Dimitrov, Metcalf, and Ott 2010).

In the next section we draw on the instructed L2 learning literature to build a case for why and how authentic texts can be used to contextualise grammar practice. Next, we review past work in ICALL that uses authentic texts to deliver grammar practice. We then discuss some of the challenges with using authentic texts for grammar practice, and follow by suggesting that LLMs as a tool for rewriting texts may be effective for increasing the availability of authentic texts suitable for grammar practice. Section 3 outlines a high-level method for using LLMs to increase the number of target linguistic forms in a document. In Section 4, we present an authoring tool we have developed that employs this method to support L2 French instruction targeting grammatical gender and gender-predictive noun suffixes. Section 5 presents our plans to evaluate the method and tool, and in Section 6 we discuss current limitations with our proposal.

2 Background

2.1 Authentic texts in grammar instruction - why and how ?

Pedagogically speaking, the use of authentic texts[1] as contexts for grammar practice is interesting for both compelling motivational and linguistic reasons.

One frequently given reason is related to motivation. In their survey of authentic materials in foreign language learning, Gilmore (2007) lists

[1] As Gilmore points out, the term 'authentic' has been defined differently in the literature, and depending on the researcher can include or exclude text that has been modified for educational purposes. In the present paper, we follow Gillmore (2007) and adopt Morrow's (1977) definition, i.e. authentic material is "a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort" (p.13) and may include texts modified for instructional purposes.

Stephen Bodnar. A prototype authoring tool for editing authentic texts using LLMs to increase support for contextualised L2 grammar practice. *Proceedings of the 14th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2025)*. University of Tartu Library. 63–71.

some of the more common rationales: authentic texts are inherently more interesting because their purpose is "to communicate a message rather than highlight target language" (p. 107); authentic texts are challenging but overcoming the challenges can in itself be motivating; instruction using authentic texts allows more freedom to choose material that matches with specific learner interests; authentic texts can be seen as giving learners an opportunity to leave the sandboxed world of textbooks and work with 'real' material intended for native speakers (see also Berwald, 1987). Although these claims seem plausible, and are compatible with L2 motivation theories (e.g., those related to self-efficacy, self-determination, or Gardner's notion of integrative orientation; see Dörnyei and Ushioda, 2011, pp. 16, 23-25, 41), Gilmore (2007) points out that there is also disagreement in the literature and that few empirical studies exist, leaving the link between authentic materials and motivation as an important area for future work.

Turning to linguistic reasons for using authentic texts, from a perspective focused on grammar learning, a well-chosen authentic text can serve as a good basis for providing the essential ingredients for acquiring new grammar knowledge through practice. First, exposure to relevant L2 input, in this case target language grammar forms, is available from the text, and can be combined with input enhancement techniques (e.g., highlighting) to help learners notice specific word forms and linguistic structures[2] (Ziegler et al., 2017). Second, opportunities to produce L2 output can be made available by drawing on the text to create written or spoken grammar activities in the classroom (e.g., Lyster, 2018) or computerised self-study (see Section 2.2). These exercises in turn serve as opportunities for learners to receive corrective feedback on their output to push them to develop their linguistic accuracy.

From a more holistic instructional perspective, it is important to point out that these opportunities for L2 input and output practice take place within a meaningful context, i.e. the authentic text whose main purpose is to communicate a message. Combining content and grammar practice together helps to ensure that both communication and linguistic accuracy goals receive support and neither is left behind (Lyster, 2018).

Unfortunately, practical constraints often stand in the way of implementing the kind of contextualised grammar practice described above. When the focus of instruction is on meaning and attention to grammar is given incidentally, for example in response to questions from learners, instruction can take place without special attention to the linguistic structures present in a text. However, to support instruction with specific grammar learning goals, what is needed are texts featuring many instances of specific linguistic structures, which is possible but can be challenging (see Section 2.2.2). Another practical issue is that, unlike textbook material, authentic texts do not come with the accompanying comprehension and grammar practice exercises. These additional materials can be developed by instructors or material developers, but it is of course extra work. In the next section we review research efforts aimed at developing technology that makes it easier to use authentic texts in grammar instruction.

2.2 ICALL systems supporting grammar practice with authentic texts

2.2.1 Existing systems

ICALL tools targeting grammar practice with authentic texts tend to provide support for one of two tasks, namely 1) helping to find suitable texts and 2) creating accompanying exercise sets.

Tools that help with finding suitable texts combine automated linguistic profilers with search interfaces (e.g., Hagiwara et al., 2021; Dittrich et al., 2019; Chinkina and Meurers, 2016). The linguistic profilers use NLP pipelines to analyze documents and obtain fine-grained information such as how frequently different POS tags, verb tenses, clause types, and other grammatical phenomena appear in a text. The search interfaces allow users to locate documents based on keywords and desired linguistic criteria. For example, in FLAIR (Chinkina and Meurers, 2016) users can specify that they are interested in documents related to the keyword 'weekend' that also feature verbs in the simple past or which contain Wh- questions. Search interfaces often use highlighting to help users quickly locate strings in the document that satisfy their search criteria, and in this way get an indication of how useful a text is for teaching particular linguistic structures. The tools target differ-

[2] Similar to (Ziegler et al., 2017), we use the term 'linguistic structure' to refer to abstract grammatical structures (e.g., a noun phrase consisting of a determiner and noun), and the term 'form' to refer to surface language instances of these structures (e.g., *a bicycle*, *un vélo*).

ent languages, with FLAIR (Chinkina and Meurers, 2016) supporting English, Octanove Learn supporting English and Chinese (Hagiwara et al., 2021), and KANSAS targeting German (Dittrich et al., 2019). Often these tools include information on the CEFR level to give a global characterisation of the L2 proficiency range a text is suitable for.

A larger number of systems have been developed that accept authentic text as input and create accompanying exercises. A helpful observation made by Heck and Meurers (2022) is that these systems use the authentic material in mainly two different ways. On the one hand, there are systems that in a sense mine the authentic material to identify seed sentences, and transform these into individual, stand-alone exercise or test items where the larger meaningful context surrounding the exercise is discarded, and the meaningful context is limited to the item itself (e.g. Baptista et al., 2016; Chalvin et al., 2013; Aldabe et al., 2006). On the other hand, there are systems that aim to leave the authentic material intact and present it to learners as one coherent whole and integrate grammar exercises into the text presentation and thereby make it interactive.

The distinction between limited-context and full-context systems is important because it helps us see that it is the latter full-context systems that best align with instructional methods that push learners to attend to meaning and form (see Section 2.1).

A prominent example of a full-context system is the *Working with English Real Texts interactively* (WeRTi) tool that transforms web texts into an interactive web page where parts of the original document become different kinds of interactive practice items, for example fill-in-the-blank items (Meurers et al., 2010). A number of other full-context systems have been developed, including a browser plugin called VIEW for Russian (Reynolds et al., 2014) and North Saami (Antonsen and Argese, 2018), the Language Muse Activity Palette (Burstein et al., 2017) and the AGREE system (Chan et al., 2022) for English , the COLLIE e-learning platform targeting French (Bodnar, 2022), and an extension of FLAIR that adds exercise generation features (Heck and Meurers, 2022).

Summarising, ICALL researchers have developed a number of innovative search and exercise generation tools that help lower the barrier to creating full-context grammar exercises that offer both L2 input and output practice. Some of these tools are freely available online, which is an important step for more wide-spread adoption that can help the field to make a real-world impact on L2 instruction, as well as inspire the development of new tools that target so-far unsupported languages.

2.2.2 Challenges with using authentic text

Arguments against using authentic texts have been presented in the literature on automatic exercise generation, but the points are not so much critiques of the instructional validity or usefulness of full-context systems but instead more related to practical difficulties. One common point is that authentic materials such as language corpora often do not naturally contain a sufficient number or variety of target linguistic structures (Aldabe et al., 2006). A second point is that the sentences in authentic materials can be very complex and more suitable for intermediate and advanced learners (Perez-Beltrachini et al., 2012).

In a nutshell, these views are arguing that finding suitable material to support contextualised grammar practice is difficult. This can be for at least two reasons, either suitable documents exist but are difficult to find, or suitable documents are very rare. For the former, certainly tools like FLAIR can be helpful for locating relevant documents if they exist. However, based on our own recent experience crawling RSS feeds to build a database of documents suitable for practising French grammatical gender, we would tend to agree with others that documents that are naturally suitable for grammar instruction targeting specific word forms or structures can be rare, and in this case linguistically-aware search tools unfortunately do not have much to offer.

One way to handle this problem is to adopt a more pragmatic perspective and aim for a compromise in which we accept that only stand-alone practice items are feasible, but when creating them try to include as much context as possible. An advantage of this approach is that we are no longer constrained to text from the same document; instead, the systems are free to search through multiple documents and cherry pick, producing more practice items and opportunities for learners.

Another way to handle this problem is to consider editing authentic texts to make them more suitable, by for example, carefully introducing in-

stances of target grammar forms that will be the focus of a lesson. To our knowledge no tool exists that helps authors adapt an existing text to make it more suitable for grammar practice. While costly, employing human authors to edit a text to, for example, include more pedagogically desirable linguistic structures is possible. However, clearly some form of technological support that lowers the barrier to contextualising grammar practice with authentic texts would be welcome.

2.3 LLMs as a tool for rewriting texts to support grammar practice

Interest in how LLMs can be used to perform useful everyday tasks has increased in recent years (Yang et al., 2024), with some researchers exploring their potential for editing or rewriting text (Shu et al., 2024). In one study, researchers working in the area of search advertising have begun to investigate whether or not LLMs can rewrite texts to blend in advertisements into chatbot responses so that they appear seamlessly, a technique known as native advertising (Zelch et al., 2024).

The impressive capabilities of LLMs and in particular the emerging findings that LLMs can be effective tools for rewriting beg the question of whether or not an LLM approach could be used to address some of the challenges with using authentic texts for grammar instruction discussed above (see Section 2.2.2). These abilities suggest that LLMs might be able to modify authentic texts to make them more pedagogically useful. Such an approach would need to find a balance between maintaining the authenticity of the text as much as as possible, while inserting or substituting target linguistic forms into the text, and possibly deleting sections of text, to seamlessly blend in the modifications. Doing so would require developing prompts that instruct an LLM to perform the needed edits, and measures (automatic or human judgements) for determining the degree to which a modified text has been improved. In the next section we propose a method for using LLMs to edit authentic texts to support grammar practice.

3 Proposed method

We assume that the input is an authentic L2 text t an instructor would like to use (e.g., because it is on an interesting topic) for providing practice on a specific linguistic structure s. We also assume that a reference linguistic profile p is available to specify what an ideal document should look like from a linguistic point of view, that is, the number and variety of target forms needed to support learning. Lastly, we assume an LLM service is accessible via a remote API. Then, the procedure we propose consists of four steps:

- Step 1: Profile the input document t to count the number of occurrences of target grammatical forms; compare these with counts in the reference linguistic profile to obtain the difference n.
- Step 2: Generate a set of n target-form strings needed to reduce the difference to zero, where a target form is a surface-language realization, e.g., if the linguistic structure is a verb phrase requiring verbs in the Simple Past, a target form string could be "I went", or "She saw".
- Step 3: Modify an LLM prompt template by inserting instructions to seamlessly blend in each grammatical target string, send the prompt to the LLM API, and store the result.
- Step 4: Profile the rewritten document and compare the resulting text with the reference linguistic profile, and repeat / manually adjust if necessary until n is negligible.

To help make clear how the method would work and what resources other than an LLM are needed, we outline how we are currently using this method in the context of the COLLIE e-learning platform (Bodnar, 2022), which provides instruction on French grammatical gender and includes an exercise generation pipeline.

A prerequisite of the proposed method is a reference linguistic profile. Its purpose is to indicate whether or not COLLIE would be able to generate an exercise with a suitable number of items covering the target structures and including a variety of forms. Specifying these criteria requires pedagogical consideration and should take into account the amount of time available for a lesson and its learning objectives. In our case, we obtain a linguistic profile by processing texts created in a previous human-led instructional intervention with COLLIE's NLP pipeline (see Bodnar, 2022). However, a profile could also be created without a reference document, for example by providing users with a settings panel similar to those used in

ICALL search tools (see Section 2.2.1) that allows users to specify different linguistic criteria.

The profiling stage (Step 1) requires the ability to automatically detect target linguistic structures. In our review in Section 2.2.1 we saw that this technology is already available (e.g., FLAIR). In our case we implement this ability using an NLP pipeline that detects French singular nouns featuring gender-predictive suffixes along with their determiners (e.g., *une potion*, *un bateau*; see Lyster, 2006). The pipeline, implemented in Java, detects these forms using the output of a dependency parser from the Stanford CoreNLP toolkit (Manning et al., 2014) and the *Lexique* database, the latter to ensure that nouns with target suffixes actually have the expected gender (for details, see Bodnar, 2022).

Step 2 above requires some generation capabilities, however, note that the goal here is to generate short strings containing target forms; we rely on the LLM in Step 3 to blend these into the text. To accomplish this, we propose using computational linguistic resources that offer precise control for generating only the needed target forms. In the case of French, *Lexique* (New et al., 2004) is a comprehensive database containing information on grammatical gender for over 45,000 nouns and is freely available; we use this resource to build a list of strings consisting of singular nouns with gender-predictive suffixes preceded by a determiner [3]. For other instructional targets, NLG tools like Simple NLG (Gatt and Reiter, 2009; SimpleNLG) and GramEx (Perez-Beltrachini et al., 2012), or corpus-mining approaches (see Section 2.2.1) could be used to obtain the short strings featuring the target forms.

Step 3 involves selecting an LLM service and developing a suitable prompt template. The LLM service used in the prototype is the OpenAI API platform[4] with the ChatGPT 3.5 Turbo model. Although this is not the most recent model available, it offers competitive performance on rewrite tasks (Shu et al., 2024). Based on our experience with the prompt shown in Figure 1, the model appears to perform well enough to be used in the prototype, and it has the advantage of being relatively inexpensive, which is important during tool development, when testing new features and fixing bugs require many API calls. Clearly, however, different LLM models and prompt formulations are parameters that should be explored in a future evaluation.

```
1 Please rewrite the text below to include the string
        'la cuisine'.
2 Also please rewrite the text below to include the string
        'une chose'.
3 ...
4 Also please rewrite the text below to include the string
        'un bateau'.
5 Please combine the rewrites into one coherent text.
6 Text: <TEXT>
```

Figure 1: An example of the current LLM prompt template we are using.

In Step 4, content authors examine the output from the LLM and decide whether or not the result is satisfactory. We assume that the success of the LLM-performed edits will vary, and that an iterative workflow will be necessary (see Figure 2).

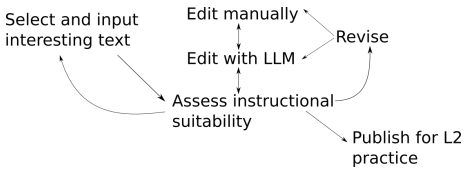

Figure 2: Authoring workflow with an LLM-enhanced authoring tool.

Comparing longer documents manually would be tedious and a productivity bottleneck. In the next section we present an authoring tool we have developed to support users during potentially multiple rounds of document modification.

4 Authoring tool prototype for French grammar practice

To explore the method described in Section 3 we have implemented a prototype authoring tool designed to assist authors with editing authentic texts to better support instruction targeting specific grammatical phenomena. The tool combines the linguistic profiling of ICALL search tools (see Section 2.2) with a prompt generation feature that makes it easy to generate specific instructions that can be used by an LLM to insert target forms into an authentic text. The tool incorporates dashboard-inspired visual elements so that authors can quickly assess the status of the documents in their collection. Figure 3 shows the current user interface.

[3] The current implementation selects nouns based on their suffixes without consideration of semantic fit; implementing a semantic fit criterion and investigating its impact on the quality of edited texts would be an interesting future direction.

[4] https://platform.openai.com/

The left pane shows the documents in the author's collection. Each document is displayed with bar graphs that indicate the readiness of a document for supporting a specific grammar instruction target. Colors communicate a document's status, with green bars indicating that a criterion has been met, and blue that more work is needed. The first two bars show the number of words in the text and the number of target forms, both relative to a desired value specified in the reference linguistic profile. The third bar provides a measure of a document's support for practice with a variety of target structures. We define this score, which we refer to as the "**c**overage of **t**arget **s**tructures" or *cts* score, for a document d and a set of target linguistic structures s, as

$$cts(d) = \frac{\sum_{i=1}^{len(s)} min(\frac{num\_target\_forms_i}{des\_num\_target\_forms_i}, 1)}{len(s)}$$

where *len(s)* is the number of distinct linguistic structures to practice in the lesson, $num\_target\_forms_i$ is the number of forms found in the document for the ith target linguistic structure, and $des\_num\_target\_forms_i$ is the desired number of forms for the ith target linguistic structure specified in the reference linguistic profile.

To give a concrete example using grammatical gender with predictive suffixes, a lesson may ask a student to practice forming noun phrases with single nouns featuring the three suffixes *-tion* (typically feminine), and *-eau* and *-age* (typically masculine). In this case, a document should score well when each of the suffixes are present in the document with the needed counts (defined in the reference linguistic profile), and no one suffix that happens to frequently occur should be allowed to compensate for other suffixes that are lacking.

The right pane is where an author can work on a text to make it more suitable for grammar instruction. Fine-grained information for each of the structures a learner should practice is available using the same bar graph format, again with target thresholds for criteria values (target word count, number of items, and number of instances of noun phrases for each gender-predictive suffix) set from the reference linguistic profile. Using these, an author can quickly understand the strengths and weakness of the document.

The tool provides authors with support in addressing weaknesses in the document by making available the method proposed in Section 3. Authors can generate an LLM prompt with a button click; the prompt can be modified before being sent to the remote LLM service. Once received, the generated text is saved to a database and tagged with a new version, in case a rollback is needed. The document text can also be edited manually.

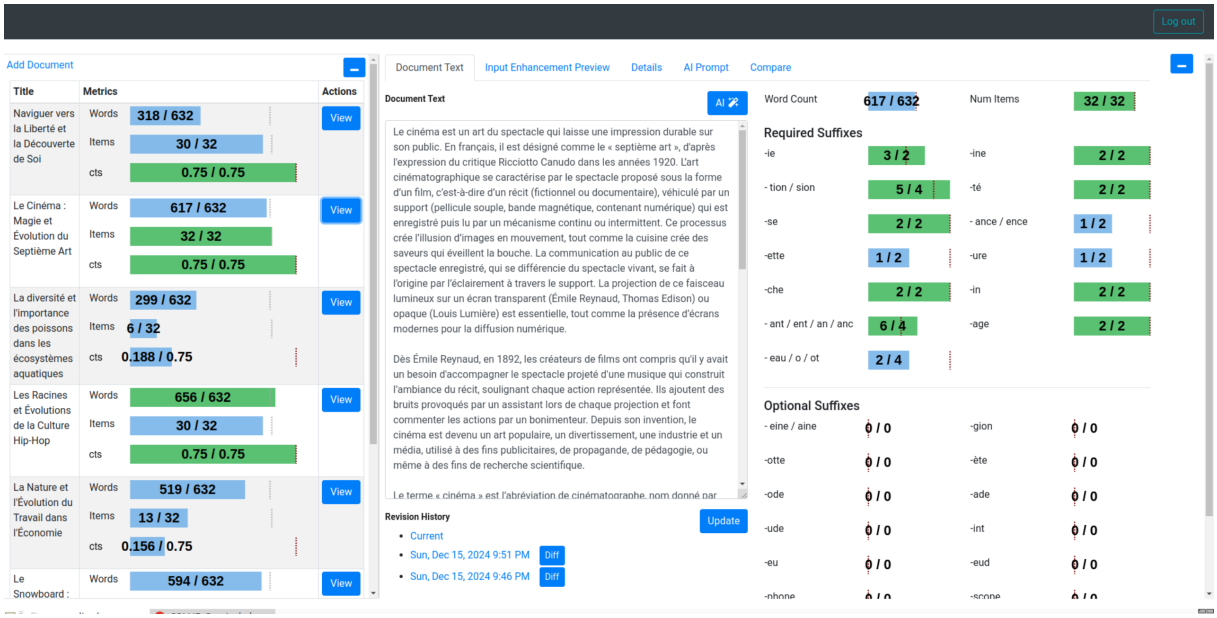

Figure 3: The tool interface allows authors to quickly scan the "readiness" of each document in the author's collection (left, see Section 4). Detailed metrics about specific linguistic shortcomings for a document are also available (right). To address the shortcomings, authors can click on the "AI" button to automatically generate a prompt that instructs an LLM to make specific edits to the document, which is then sent to a remote LLM service. The modified text is saved as a new version, and "Input Enhancement" and "Compare" views are available to help the author quickly see how the text has been edited.

Two other views that support the author worth briefly mentioning are 1) a Diff Viewer component (Ravi, 2024) that allows comparison of the old and new versions for authors to quickly review changes after an iteration of LLM edits, and 2) an Input Enhancement view that highlights target forms to allow the author to quickly locate them in the text.

5 Plans for evaluation

Evaluating the proposed LLM-based method and the authoring tool is an important step that we are currently planning. For the first evaluation we plan on using human judgements by French native speakers to evaluate a set of modified documents. The five measures used by Shu et al. (2024) in their evaluation of LLM performance on rewriting tasks seem to capture all the dimensions of the text quality we would be concerned with:

1. **Instruction success**: whether the rewrite accurately follows the instruction provided.
2. **Content preservation**: whether the rewritten text preserves the essential content and meaning of the source text, regardless of its writing style or quality.
3. **Factuality**: Checks the accuracy and truthfulness of the answer's content.
4. **Coherence**: whether the rewritten text is easy to understand, non-ambiguous, and logically coherent when read by itself (without checking against the source text).
5. **Fluency**: Examines the clarity, grammar, and style of the written answer.

(Shu et al., 2024, p. 18974)

Since the tool's main purpose is to introduce new target linguistic structures into an existing text, it will be important to check whether or not the LLM model actually inserts strings featuring the needed target forms without modifying them (an LLM may try to modify the strings so that they fit better in the text but no longer count as a valid instance of a target linguistic structure).

Of course, ensuring that the strings are present in the text is not, by itself, a good reflection of how well the edits were performed. The whole point of the method is to try to carry out the edit instructions while preserving the original meaning of the text, so that it can continue to serve as a meaningful context for instruction (see Section 2.1). The dimensions of *content preservation* ("Have the messages conveyed by the text changed?") and *coherence* ("Is the text as easy to understand and as logically coherent as it was before the edits?") are therefore important performance criteria for ensuring that the text remains a valid meaningful context.

Since the texts will serve language learners as models of well-formed L2 writing, another important dimension of performance is how linguistically correct the edits are. Shu et al. (2024) use the label 'fluency', which in SLA literature is used to refer to how well language flows (e.g. Housen and Kuiken, 2009), but in our case it seems more important to measure the linguistic correctness, or accuracy, of the modified texts, to investigate if LLMs introduce grammatically incorrect language into their output.

Regarding *factuality*, it is well-known that LLMs can hallucinate, i.e., generate text that includes untrue or misleading information (Huang et al., 2024). Of course, an authentic text could already contain factually untrue information, but the point of including this measure would be to understand whether or not new factually incorrect information is introduced during the editing task.

6 Conclusion

In this paper we presented a new authoring tool aimed at solving a practical issue with using authentic texts to contextualise grammar practice, namely that authentic texts usually do not contain a sufficient number or variety of linguistic structures needed to support L2 input and output practice exercises with a specific grammar target. The tool relies on a method that proposes combining traditional natural language generation, using lexical databases and rule-based tools, with current LLM services to dynamically generate prompts that instruct an LLM to insert strings with specific linguistic structures into the text.

Our experience with the tool so far is encouraging, but to really determine the viability of the approach a formal evaluation is needed. The next step will be to carry out a first evaluation with human judgements using the criteria presented in Section 5, possibly while also exploring the impact of different prompt formulations, and different LLM service providers and models.

Assuming that the method is successful, two other issues may arise. A first issue has to do with the current high computational cost of using LLMs: the best performing LLMs cannot be self-hosted due to their high computational cost which means that our tool currently depends on paid

LLM service providers. This places a limit on how many documents an organisation can rewrite before hitting budget limits. A second issue has to do with copyrighted materials. While authentic texts that are in public domain or released with permissive licenses allowing derivatives shouldn't be an issue, it seems likely that many useful texts will be copyrighted; even if use for educational purposes is permitted, rewriting the material seems to go one step further and could be problematic. These are issues that need further consideration.

References

Itziar Aldabe, Maddalen Lopez de Lacalle, Montse Maritxalar, Edurne Martinez, and Larraitz Uria. 2006. ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In *Intelligent Tutoring Systems*, volume 4053 of *Lecture Notes in Computer Science*, pages 584–594. Springer.

L. Antonsen and C. Argese. 2018. Using authentic texts for grammar exercises for a minority language. In *Linköping Electronic Conference Proceedings*, page 152.

J. Baptista, S. Lourenco, and N. J. Mamede. 2016. Automatic generation of exercises on passive transformation in Portuguese. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 4965–4972.

Jean-Pierre Berwald. 1987. *Teaching Foreign Languages with Realia and Other Authentic Materials. ERIC Q&A*. Distributed by ERIC Clearinghouse, S.l. Sponsoring Agency: Office of Educational Research and Improvement (ED), Washington, DC.

Stephen Bodnar. 2022. The instructional effectiveness of automatically generated exercises for learning French grammatical gender: preliminary results. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 10–22, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating language activities in real-time for english learners using language muse. In *Proceedings of the Fourth ACM Conference on Learning @ Scale, L@S 2017, Cambridge, MA, USA, April 20-21, 2017*, pages 213–215. ACM.

Antoine Chalvin, Egle Eensoo, and François Stuck. 2013. Mining a parallel corpus for automatic generation of Estonian grammar exercises. In *Third biennial conference on electronic lexicography (eLex 2013) "Electronic lexicography in the 21st century: thinking outside the paper"*, pages 280–295, Tallinn, Estonia.

Sophia Chan, Swapna Somasundaran, Debanjan Ghosh, and Mengxuan Zhao. 2022. AGReE: A system for generating automated grammar reading exercises. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–177, Abu Dhabi, UAE. Association for Computational Linguistics.

Maria Chinkina and Detmar Meurers. 2016. Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 188–198, San Diego, CA. Association for Computational Linguistics.

Leona Colling, Ines Pieronczyk, Cora Parrisius, Heiko Holz, Stephen Bodnar, Florian Nuxoll, and Detmar Meurers. 2024. Towards task-oriented icall: A criterion-referenced learner dashboard organising digital practice. In *Proceedings of the 16th International Conference on Computer Supported Education - Volume 1: EKM*, pages 668–679. INSTICC, SciTePress.

Frederik Cornillie, Ruben Lagatie, Mieke Vandewaetere, Geraldine Clarebout, and Piet Desmet. 2013. Tools that detectives use: In search of learner-related determinants for usage of optional feedback in a written murder mystery. *CALICO Journal*, 30:22–45.

Sabrina Dittrich, Zarah Weiss, Hannes Schröter, and Detmar Meurers. 2019. Integrating large-scale web data and curated corpus data in a search engine supporting German literacy education. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 41–56, Turku, Finland. LiU Electronic Press.

Zoltán Dörnyei and Ema Ushioda. 2011. *Teaching and researching motivation (2nd ed.)*. Harlow: Longman.

Polina Drozdova, Catia Cucchiarini, and Helmer Strik. 2013. L2 syntax acquisition: the effect of oral and written computer assisted practice. In *14th Annual Conference of the International Speech Communication Association, INTERSPEECH 2013, Lyon, France, August 25-29, 2013*, pages 982–986. ISCA.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: a realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, ENLG '09, page 90–93, USA. Association for Computational Linguistics.

Alex Gilmore. 2007. Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2):97–118.

Masato Hagiwara, Joshua Tanner, and Keisuke Sakaguchi. 2021. Grammartagger: A multilingual, minimally-supervised grammar profiler for language education.

Tanja Heck and Detmar Meurers. 2022. Parametrizable exercise generation from authentic texts: Effectively targeting the language means on the curriculum. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 154–166, Seattle, Washington. Association for Computational Linguistics.

A. Housen and F. Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4):461–473.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Gladys Jean and Daphnée Simard. 2011. Grammar teaching and learning in l2: Necessary, but boring? *Foreign Language Annals*, 44(3):467–494.

R. Lyster. 2018. *Content-based Language Teaching*. New York: Routledge.

Roy Lyster. 2006. Predictability in french gender attribution: A corpus analysis. *Journal of French Language Studies*, 16(1):69–92.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.

Boris New, Christophe Pallier, and Ludovic Brysbaert, Marc andFerrand. 2004. Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, &amp Computers*, 36(3):516–524.

Laura Perez-Beltrachini, Claire Gardent, and German Kruszewski. 2012. Generating Grammar Exercises. In *NAACL-HLT 7th Workshop on Innovative Use of NLP for Building Educational Applications*, Montreal, Canada.

Pranesh Ravi. 2024. https://praneshravi.in/react-diff-viewer/. Accessed on December 16th, 2024. [link].

Robert Reynolds, Eduard Schaf, and Detmar Meurers. 2014. A VIEW of Russian: Visual input enhancement and adaptive feedback. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 98–112, Uppsala, Sweden. LiU Electronic Press.

Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Yinxiao Liu, Simon Tong, Jindong Chen, and Lei Meng. 2024. Rewritelm: An instruction-tuned large language model for text rewriting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18970–18980.

SimpleNLG. https://github.com/simplenlg/simplenlg. Accessed on December 15th, 2024. [link].

Helmer Strik, Polina Drozdova, and Catia Cucchiarini. 2013. Gobl: Games online for basic language learning. In *Speech and Language Technology in Education (SLaTE 2013)*, pages 137–142.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Ines Zelch, Matthias Hagen, and Martin Potthast. 2024. A user study on the acceptance of native advertising in generative ir. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*, CHIIR '24, page 142–152, New York, NY, USA. Association for Computing Machinery.

Nicole Ziegler, Detmar Meurers, Patrick Rebuschat, Simón Ruiz, José L. Moreno-Vega, Maria Chinkina, Wenjing Li, and Sarah Grey. 2017. Interdisciplinary research at the intersection of call, nlp, and sla: Methodological implications from an input enhancement project. *Language Learning*, 67(S1):209–231.