I Can Explain This: Enhancing Grammatical Error Correction with Explanatory Feedback in Icelandic

Atli Jasonarson, Steinþór Steingrímsson

The Árni Magnússon Institute for Icelandic Studies

Abstract

We present an automated written corrective feedback system that connects a grammatical error correction model and the Icelandic Official Spelling Rules. The system is intended to aid users in correcting their spelling as well as to provide them with explanations for the suggestions the model makes, along with references to the relevant rules in the spelling rules. By providing the users with these references, we hope to help them better understand the Icelandic spelling rules. While we do not aim to establish statistically significant comparisons, we provide a qualitative discussion contrasting our system with two large language models, GPT-4o and Claude 3.5 Sonnet, to highlight the challenges of explainability in grammatical error correction. Furthermore, we demonstrate that these large language models exhibit a problematic tendency to offer incomplete or even inaccurate explanations for their edits.

Keywords

grammatical error correction, computer assisted language learning, automated written corrective feedback, Icelandic

1. Introduction

Grammatical error correction (GEC) is the process of automatically correcting errors in text, whether they are purely grammatical, orthographic or semantic. The field has developed significantly in the past decade, with the current state of the art being neural machine translation[1] systems (Bryant et al. 2023). These systems have been shown to outperform human annotators by a considerable margin (Qorib and Ng 2022), but as discussed in Bryant et al. (2023), this can be difficult to measure as GEC is a subjective task which often has low inter-annotator agreement.

Despite these systems' capabilities with regard to GEC, they have their limitations, as discussed in Bryant et al. (2023). These limitations include their tendencies to perform well in one domain, e.g. instruction manuals, but not another, e.g. scientific articles; their underperformance with regard to semantic errors, including multi-word expressions and collocations; and their insensitivity to errors which span longer context, e.g. across sentences.

Another limitation of LLMs, at least when predictability is desired, is their inclination to hallucinate, which Xu, Jain, and Kankanhalli (2024) define as "models generat[ing] plausible but

The 9th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2025), March 5–7, 2025, Tartu, Estonia.
atli.jasonarson@arnastofnun.is (A. Jasonarson); steinthor.steingrimsson@arnastofnun.is (S. Steingrímsson)
https://github.com/atlijas (A. Jasonarson); https://steinst.is/ (S. Steingrímsson)
0009-0002-0965-8514 (A. Jasonarson); 0000-0002-9776-9507 (S. Steingrímsson)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

[1]GEC can be seen as a translation of sorts, with the original text being "translated" to a corrected version of it.

factually incorrect or nonsensical information" (p. 1). Furthermore, they argue that, despite the many works that have aimed to reduce them, it is impossible to eliminate hallucinations in LLMs. A similar case is made in Banerjee, Agarwal, and Singla 2024, where the authors argue that an "understanding of structural hallucinations is vital for the responsible use of these powerful tools".

The research on pedagogical relevance of automated written corrective feedback (AWCF) is still developing, but its effect on students' writing has been shown to be positive, although some of its aspects still remain unclear, e.g. on which features it should focus (e.g. grammar, lexis or structure), how explicit it should be and from where it should come (i.e. from teachers or peers) (El Ebyary and Windeatt 2010). Some of the positive effects of AWCF reported are increased writing accuracy (Barrot 2023) and learner autonomy (Zhang and Hyland 2018), as well as positively influencing error correction, especially when the feedback is explicit, as opposed to generic (Ranalli 2018).

With the limitations of LLMs, as well as specialized GEC models, in mind, as well as the possible pedagogic importance of explainable spellchecking, we build a system which points out errors it knows with confidence while maintaining caution regarding others. Our system is aimed at children from the ages of 12–16 and L2 learners of Icelandic. We believe it is more important to provide them with correct and reliable information, instead of relying on an LLM to generate a sound explanation for a given grammatical error, since LLMs, despite their capabilities, are prone to hallucations, as discussed earlier.

Our system focuses on explaining orthographic errors, as the Icelandic Official Spelling Rules are readily available, comprehensive and represented in a systematic way, while we leave grammatical and semantics ones mostly untouched. We only point out inflectional changes the GEC model makes, without providing possible explanations for them, and instead direct the users to external resources. This paper aims to describe the system and discuss the explainability of different GEC systems and not to establish statistically significant comparisons, for that we refer the reader to Ármannsson et al. (2025). The comparison to the LLMs will therefore be limited to an analysis of how well the systems do with regard to spelling.

2. Related Work

The best-known AWCF tool today is perhaps Grammarly, which claims to provide "personalized suggestions to ensure your writing and reputation shine" and offers "full-paragraph rewrites" as well as help to "[r]each Your Goals at School and Beyond".[2] According to Koltovskaia (2020), "Grammarly [...] could serve as a useful resource for writing assessment in L2 classrooms if active engagement is in place", but "students with low language proficiency may not be able to utilize Grammarly effectively as their lack of linguistic competence can prevent them from adequately understanding AWCF".

Another AWCF application is DanProof (Bick 2015) which covers about 35 error types, which are connected to an error definition and an explanation, as well as references to external resources, such as online exercises. DanProof finds ⅔ of errors in Danish texts, with a precision

[2]https://www.grammarly.com/features

of 91.7%. The authors highlight their inclusion of pedagogical comments and discussed the need for testing the software in-classroom, but we are not aware of further work on DanProof.

Other AWCF tools include, but are not limited to, the text revision functionality in text editors such as Google Docs and Microsoft Word, and, following some prompting, LLMs with a web interface, such as ChatGPT or Claude, can be utilized as such.

3. System description

The two main building blocks of our system are a grammatical error correction (GEC) model (Ingólfsdóttir et al. 2023), which is a fine-tuned version of the byte-level sequence-to-sequence model ByT5 (Xue et al. 2022), and a post-hoc explanation system which analyzes the model's changes to a given text and connects them to the Icelandic Official Spelling Rules (IOSR), the latter being our main contribution. The demo's front-end is a web interface[3] whose functionality is simple: It prompts the user to insert a text which is sent to the back-end, where it is corrected by the GEC model, and displays a corrected version of each sentence, alongside any erroneous tokens they may contain and, if applicable, a reference to the IOSR.

3.1. The Icelandic Official Spelling Rules

The latest update to the Icelandic Official Spelling Rules was published in 2016 and on punctuation in 2018 (Sigtryggsson 2021). Icelandic orthography is conservative and grounded in tradition. It has been argued that it is an important part of the continuity of the Icelandic Language and in turn an important part of Icelandic language policy (Kristinsson 2017). The normalized orthography, codified in the 19th century, follows in large part the origin principle which means that in spite of phonological changes, modern orthography is very much in line with normalized orthography for old Icelandic.

3.2. Post-hoc explanations

Seeing as the GEC model we use was only trained to correct text but not to explain the changes it makes, we generate the explanations retroactively. We do this by comparing the original input from the user and the edited output from the model and finding the differences.

The first step is to roughly align the tokens, i.e. words and punctuation marks, in the original text and their corrected counterparts. We do this using Python's SequenceMatcher[4]. The alignment is then finalized by comparing the roughly-aligned tokens and checking whether their difference might stem from a known spelling error covered in IOSR, e.g. if an original token is **hesturin* and it has been roughly aligned with the corrected token *hesturinn* ('the horse')[5]. The tool pairs the two tokens together and determines that the spelling error is *n4nn*, meaning a word was written with one *n* where there should have been two. This is done by applying hand-written rules to a given token pair, which in the current version of our system only cover 6 of the 33 rules found in the IOSR.

[3] https://ritun.arnastofnun.is

[4] https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher

[5] This specific example is covered in rule 12.2 in the IOSR: https://ritreglur.arnastofnun.is/#12.2

If the difference between two tokens can not be connected to a rule in the IOSR, the system cautiously suggests that the original token might be erroneous and points the user towards appropriate and available lexicographical online resources, such as *The Language Usage Database*[6] if the tokens share a lemma, i.e. if the system suspects the model's change is an inflectional one, and *m.is*, a website aimed at people learning Icelandic, where various resources such as *The Dictionary of Contemporary Icelandic Language*[7] and *The Database of Icelandic Inflections*[8] have been published together. This is necessary as the IOSR can not possibly cover everything the model corrects, especially when the endless possibilities of typographical errors are taken into account.

4. Comparison with LLMs

Large language models have shown remarkable capabilities when it comes to GEC (Rothe et al. 2021; Omelianchuk et al. 2020), even surpassing human annotators (Qorib and Ng 2022). We are not aware of many studies into their ability to explain grammatical corrections, although as Cortez, Norris, and Duman (2024) and Kaneko and Okazaki (2024) demonstrate, LLMs are capable of doing so for English.

We extracted 10 random sentences from a combined set of four Icelandic error corpora (Ingason et al. 2021a, 2021b). The sentences in the corpora have all been manually corrected and annotated with the errors they contain. The 10 sentences we extracted contain two errors each and the error types are five in total: lowercase letter instead of an uppercase one, an uppercase letter instead of a lowercase one, two words have been merged, a letter is missing, or an extra letter has been added. In total, there are 44 error types in the corpora.

We used a zero-shot approach by prompting GPT-4o[9] and Claude 3.5 Sonnet[10] to correct each of the sentences, and explain their changes, using the same prompt. We used our system's web interface to gather corrections and explanations for comparison.[11]

4.1. Evaluation

To evaluate the systems' performances, we labelled their corrections and explanations with one of four labels, based on the annotations of the 10 manually corrected sentences:

- **Missing**
  - The system did not return a correction/explanation found in the corpus annotation.
- **Correct**
  - The system's correction or explanation matched the annotation in the error corpus.
- **Incorrect**
  - The system introduced an error in its correction or provided an erroneous explanation.

[6] https://malfar.arnastofnun.is/
[7] https://islenskordabok.arnastofnun.is/
[8] https://bin.arnastofnun.is
[9] Specifically: gpt-4o-2024-08-06
[10] Specifically: claude-3-5-sonnet-20240620
[11] Prompt available in appendix.

- **Neutral**
  - The system made a change that was unnecessary according to the Icelandic language standard.

Table 1
The systems' correction scores

	GPT-4o	Claude 3.5	Ours
Missing	3	0	2
Incorrect	7	5	1
Correct	9	14	17
Neutral	2	5	0

Table 2
The systems' explanation scores

	GPT-4o	Claude 3.5	Ours
Missing	3	0	2
Incorrect	8	8	1
Correct	9	13	8
Neutral	1	1	10

Although the examples we used are too few to draw any meaningful conclusions about the models' GEC capabilities, it seems as if the one we used is the most performant one when correcting Icelandic texts, perhaps unsurprisingly, as it is the only one of the three that was trained for that very purpose.

The reason the numbers do not necessarily match between the two tables, i.e. why Claude 3.5 has 13 correct explanations and 14 correct corrections, is that the models sometimes generate a correct correction while providing a wrong explanation for it, and vice versa. Additionally, the models occasionally generate corrections and explanations even when they are unnecessary.

The analysis of the explanations generated by our system shows that about half are labelled as neutral. This label applies to explanations linked to correct corrections that do not reference the relevant IOSR rule, but instead point to the online resources discussed earlier. This high proportion of neutral explanations stems from two factors: the system currently addresses only 6 of the 33 IOSR rules, and when no matching rule is found for a token pair, the system conservatively flags a potential issue and directs the user to external resources.

Although our system has fewer correct explanations for the underlying model's changes, it only has one incorrect explanation, the reason being that it does not know the fictional name *Snorkur*[12].

4.2. Incorrect explanations

The two LLMs we tested, GPT-4o and Claude 3.5 Sonnet, returned eight incorrect explanations each. These vary in severity, but must be considered highly undesirable, especially when considering the target group of our system: children and L2 learners of Icelandic. Some examples of these errors are as follows (Icelandic → English translations are ours, as well as rephrasing and emphasis for the sake of clarity):

GPT-4o

- "Because *Lilja Rafney* is a full name, *Rafney* should be in the genitive case, *Rafneyju*"
  - Names in Icelandic are declined together and the given genitive form is incorrect, as it should be *Rafneyjar*.

[12]The Icelandic version of Snorken, the character from the Moomin books.

- "*eignartjón* should be *eignatjón*, as the latter doesn't exist in Icelandic"
  - This is wrong, as both terms are used, as noted in The Dictionary of Contemporary Icelandic[13].
- "*landlæknir* should be *Landlæknir*, as it is the name of an official position"
  - While the correction was correct (the word was at the beginning of a reference and should therefore be titleized), the explanation is incorrect, and the fact that *landlæknir* should be lowercased is even specifically mentioned in rule 1.3.2 in the IOSR.

Claude 3.5 Sonnett

- "*var* should be plural, *voru*, when its subject is a number larger than one, in this case 471"
  - While the correction is a neutral one, the explanation is incorrect or, at best, incomplete, as when a number in Icelandic is a subject and ends with *1*, e.g. *471*,[14] the following verb is usually singular, the exception being when a conjoined noun phrase acts as a subject.
- "There should be a comma between two sentences connected with *og*"
  - This is incorrect, as mentioned in rule 21 of the IOSR. A comma should only be used to join two sentences if they are not connected with a conjunction, which *og* ('and') is.
- "*unga fólki* was changed to *unga fólkið*, as the context demands the definite article and not the indefinite one"
  - While the correction was correct (a definite article is appropriate in the context), the explanation is incorrect, as there is no indefinite article in Icelandic.

These examples demonstrate that, while LLMs have the capability to provide automatic written corrective feedback, they are unreliable, at least in the context of Icelandic. This unreliability extends to both the accuracy of the corrections and the explanations they provide, limiting their effectiveness.

5. Conclusion and Further Work

We are not aware of many studies giving experimental results for the use of explained GEC vs. unexplained GEC. However, Rimbar (2017) found that students do not internalize corrections made by spellcheckers and that their effect is limited on the cognitive level, and the results published by Darvishi et al. (2024) indicate that students have a tendency to rely on, rather than learn from, AI assistance.

As previously noted, our tool currently addresses 6 of the 33 IOSR rules, with plans to incorporate the rest. Alongside this, we will gather feedback from teachers and students on the tool's usability and effectiveness, focusing on the interface and its usefulness for language learning. We will ask teachers if the error correction explanations help students internalize spelling rules, and if not, consider alternative approaches. Additionally, we will seek their views on whether precision or recall should be prioritized in GEC tasks to better align the tool with pedagogical needs.

[13] https://islenskordabok.arnastofnun.is/ord/65919
[14] Although not if the number ends with *11*.

Finally, when the work on our tools is completed, we intend to run a study to investigate whether students are more likely to internalize spelling rules when using explained GEC as opposed to when they use GEC that does not explain the errors made by the user.

References

Ármannsson, Bjarki, Hinrik Hafsteinsson, Jóhannes B. Sigtryggsson, Atli Jasonarson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2025. "Playing by the Rules: A Benchmark Set for Standardized Icelandic Orthography." In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa).* Tallinn, Estonia.

Banerjee, Sourav, Ayushi Agarwal, and Saloni Singla. 2024. "LLMs Will Always Hallucinate, and We Need to Live With This." *arXiv preprint arXiv:2409.05746.*

Barrot, Jessie S. 2023. "Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy." *Computer Assisted Language Learning* 36 (4): 584–607.

Bick, Eckhard. 2015. "DanProof: Pedagogical Spell and Grammar Checking for Danish." In *Proceedings of the International Conference Recent Advances in Natural Language Processing,* 55–62. Hissar, Bulgaria. https://aclanthology.org/R15-1008.

Bryant, Christopher, 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* 49, no. 3 (September): 643–701. ISSN: 0891-2017. https://doi.org/10.1162/coli%5C_a%5C_00478.

Cortez, S Magalí López, Mark Josef Norris, and Steve Duman. 2024. "GMEG-EXP: A Dataset of Human-and LLM-Generated Explanations of Grammatical and Fluency Edits." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 7785–7800.

Darvishi, Ali, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. "Impact of AI assistance on student agency." *Computers & Education* 210:104967.

El Ebyary, Khaled, and Scott Windeatt. 2010. "The impact of computer-based feedback on students' written work." *International Journal of English Studies* 10 (2): 121–142.

Ingason, Anton Karl, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, and Xindan Xu. 2021a. *The Icelandic Error Corpus (IceEC). Version 1.1.*

Ingason, Anton Karl, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, and Xindan Xu. 2021b. *The Icelandic Specialized Error Corpora. Version 1.0.*

Ingólfsdóttir, Svanhvít Lilja, Pétur Orri Ragnarsson, Haukur Páll Jónsson, Haukur Barri Símonarson, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2023. "Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora." In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 7299–7316. Toronto, Canada. https://aclanthology.org/2023.acl-long.402.

Kaneko, Masahiro, and Naoaki Okazaki. 2024. "Controlled Generation with Prompt Insertion for Natural Language Explanations in Grammatical Error Correction." In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 3955–3961. Torino, Italia. https://aclanthology.org/2024.lrec-main.350/.

Koltovskaia, Svetlana. 2020. "Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study." *Assessing Writing* 44:100450.

Kristinsson, Ari Páll. 2017. *Málheimar. Sitthvað um málstefnu og málnotkun.* Reykjavík: Háskólaútgáfan.

Omelianchuk, Kostiantyn, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. "GECToR – Grammatical Error Correction: Tag, Not Rewrite." In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications,* 163–170. Seattle, WA, USA → Online. https://aclanthology.org/2020.bea-1.16.

Qorib, Muhammad Reza, and Hwee Tou Ng. 2022. "Grammatical error correction: Are we there yet?" In *Proceedings of the 29th International Conference on Computational Linguistics,* 2794–2800.

Ranalli, Jim. 2018. "Automated written corrective feedback: How well can students make use of it?" *Computer Assisted Language Learning* 31 (7): 653–674.

Rimbar, Hazelynn. 2017. "The influence of spell-checkers on students' ability to generate repairs of spelling errors." *Journal of Nusantara Studies (JONUS)* 2 (1): 1–12.

Rothe, Sascha, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. "A Simple Recipe for Multilingual Grammatical Error Correction." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers),* 702–707. Online. https://doi.org/10.18653/v1/2021.acl-short.89.

Sigtryggsson, Jóhannes B. 2021. "Retskrivning og normering i nutidens Island – nogle tanker." *Sprog i Norden* 49, no. 1 (August): 67–78. https://tidsskrift.dk/sin/article/view/114946.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli. 2024. "Hallucination is inevitable: An innate limitation of large language models." *arXiv preprint arXiv:2401.11817.*

Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models." *Transactions of the Association for Computational Linguistics* (Cambridge, MA) 10:291–306. https://aclanthology.org/2022.tacl-1.17.

Zhang, Zhe Victor, and Ken Hyland. 2018. "Student engagement with teacher and automated feedback on L2 writing." *Assessing Writing* 36:90–102.

A. Prompts

The following is the prompt used for GPT-4o and Claude 3.5. The English translation was not part of the prompt and is provided solely to aid in understanding the original content.

"Þú ert sérfræðingur í íslensku máli. Leiðréttu villur í eftirfarandi texta, einungis með tilliti til málfars, stafsetningar og greinarmerkjasetningar, og útskýrðu breytingarnar sem þú gerir. Athugaðu að þú átt ekki að breyta merkingu textans eða stíl hans. Hér er textinn sem þú átt að leiðrétta: [sentence]."

English translation: "You are an expert in the Icelandic language. Correct any errors in the following text, only with regard to grammar, spelling, and punctuation, and explain the changes you make. Bear in mind that you should not change the meaning of the text or its style: [sentence]."