International Journal on Advanced Computer Engineering and Communication Technology

ISSN: 2278-5140

Volume 15 Issue 2 (2026) | Pages: 349-357

Explainable AI for English Grammar Teaching: Making Neural Decisions Transparent to Learners

Parul Berwal[1], Kavita Berwal[2]

[1]Department of Computer Science, Shri Vishwakarma skill university
[2]Department of English, Haryana Govt School TGT English, Leharpur
[1]prlberwal3@gmail.com, [2]kavitaberwal3@gmail.com

Peer Review Information

Type: *Article*
Received: *12 Aug 2026*
Revised: *09 Sept 2026*
Accepted: *16 Oct 2026*
Published: *19 Nov 2026*

Abstract

There are tools for grammar checking, grammatical error correction (GEC) and conversational tutors that are now powered by neural networks and that give verdicts without reasons. An unstated correction is pedagogically thin for learners, as they have to notice, understand and internalise rules. The paper is an integrative review based on secondary data comprised of 40 peer-reviewed and archival papers published from 1988 to 2022 covering explainable AI (XAI), neural-language-model interpretability, GEC and second-language-acquisition (SLA) studies dedicated to feedback research. These are guided by three research questions: What are the different types of explanation techniques available and how do they fit into the context of teaching grammar? What should explanations be doing, as determined by the pedagogy? What challenges exist that make the deployment of such techniques difficult? The review reveals that (a) explanation methods vary significantly in faithfulness and learner intelligibility, and explanation displayed in a learner does not necessarily constitute evidence of noticing and explicit feedback in SLA; (b) in the corpus of studies reviewed, there is very little direct experimental evidence of XAI in the context of grammar learning, and therefore, conclusions have to be largely based on neighboring evidence from AE and IT in the context of writing; and (c) direct experimental evidence of XAI in the context of grammar learning is relatively scarce in the corpus reviewed, and therefore, conclusions have to be mostly based on neighboring evidence from AE and IT in the context of writing, with (d) being the conclusion that attention-weight displays are an unreliable basis for learner facing explanations, and that unfaithful explanations and cognitive overload and over-reliance are the main risks. The paper suggests the four layer Transparent Grammar Explanation Framework (TGEF) and the Research agenda.

Keywords: Explainable AI; Grammar Teaching; Grammatical Error Correction; Corrective Feedback; Transparency; English Language Learning; Neural Language Models

How to Cite This Article

Berwal, P., & Berwal, K. (2026). Explainable AI for English grammar teaching: Making neural decisions transparent to learners. *International Journal on Advanced Computer Engineering and Communication Technology,* 15(2), 349–357.

2278-5140/ © 2026 The Authors. This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0).

Introduction

The neural models are impacting English grammar instruction. State-of-the-art grammatical error correction has been based on transformer architectures (Vaswani et al., 2017) and on pre-trained encoders like BERT (Devlin et al., 2019), and shared tasks evaluating grammatical error correction performance (Bryant et al., 2019) have seen substantial advances. Now learners get immediate feedback and corrections from writing assistants, chatbots and tutoring platforms. It is not, however, a measure of instruction. If a learner sees go silently replaced by goes, they are given an answer but not an explanation, and the teacher is unable to determine if the system is appropriate for the right reason. The significance of this gap is related to the findings of decades of SLA research, which indicate that learning grammar relies on conscious noticing of the difference between what is produced and the target form (Schmidt, 1990) and that explicit, well-targeted grammar instruction has significant and lasting effects (Norris & Ortega, 2000; Ellis, 2006). Feedback is most effective when it provides the students with information about the direction of their learning, their progress and what they need to do next (Hattie & Timperley, 2007). A black-box correction provides just the first part of that message. For the latter, tools for explainable AI (XAI) have been explained as not an optional luxury, but a necessity (Adadi & Berrada, 2018; Arrieta et al., 2020; Guidotti et al., 2018), and explainable AI in education has been argued to be a necessity (Conati et al., 2018; Khosravi et al., 2022). However, explanation techniques were developed primarily for developers and auditors, and not so much for language learners and their potential use has not been fully explored. This paper fills this void. They are: (1) a structured synthesis of 40 secondary sources related to XAI, NLP interpretability, GEC and SLA; (2) a comparative appraisal of the explanation methods with regard to pedagogical criteria; (3) the framework of TGEF and a research agenda for transparent grammar teaching.

Literature Review

Explainable AI: Concepts And Methods

There are a number of definitions of interpretability and a warning by Lipton (2018) that the word covers different objectives of causality and transferability and trust. Doshi-Velez and Kim (2017) suggest assessing interpretability on three levels: application-grounded, human-grounded and functionally grounded. Typically, methods are divided into models that are inherently interpretable and post-hoc explanations (Guidotti et al., 2018). According to Rudin (2019), interpretable models ought to be used in place of black-box models, which are then explained after the fact, for high-stakes decisions, especially when a grammar tool has an impact on assessment. Post-hoc families are those that include local surrogate models like LIME (Ribeiro et al., 2016), Shapley-value attribution (Lundberg & Lee, 2017), path-integrated gradients (Sundararajan et al., 2017) and counterfactual explanations (Wachter et al., 2017). Building on social sciences, Miller (2019) demonstrates that the explanations that humans give are contrastive, selective and social, a key takeaway for any design that is intended for learners.

Interpreting Neural Language Models

A wide range of attention analysis, feature attribution, probing and behavioural testing techniques have been developed, as evidenced by surveys of explainable NLP (Danilevsky et al., 2020; Madsen et al., 2022) and analysis methods for neural language processing (Belinkov & Glass, 2019). There is disagreement about the concept of attention: Jain and Wallace (2019) demonstrated that there is no correlation between attention weights and gradient-based importance in many cases and that different attention distributions can lead to the same predictions, while Wiegreffe and Pinter (2019) argued that this would depend on the definition and testing of the explanation. Jacovi and Goldberg (2020), therefore, emphasize the difference between plausibility (convincing to people), and faithfulness (true to the model's reasoning), which is a fundamental one in education, as a plausible but unfaithful explanation can teach the wrong rule. Probing results show that pre-trained encoders do contain information that is relevant to grammar. BERT's layers recover the classical NLP pipeline in a more or less expected fashion from part-of-speech to syntax to semantics (Tenney et al., 2019); syntactic tree structure can also be recovered with a linear transformation of representations (Hewitt & Manning, 2019); and BERT is reported to encode phrase-level information in lower layers and syntactic information in middle layers (Rogers et al., 2020). The results are more varied regarding behaviour, with recurrent networks showing good performance on many examples of subject–verb agreement but failing on a complex structure (Linzen et al., 2016); and BLiMP's 67 sub-datasets of 1,000 minimal pairs, representing 12 phenomena, showing variable performance across different phenomena (Warstadt et al., 2020). This irregularity implies that sometimes explanations will follow incorrect corrections, which learners need to be shielded from.

Grammatical Error Correction and Learner Feedback

Modern GEC covers the sequence-to-sequence rewriting and sequence-tagging techniques. To make each edit an explicit and inspectable operation and to accelerate the inference process, GECToR predicts edit tags at the token level using a Transformer encoder and applies them iteratively (Omelianchuk et al., 2020). ERRANT automatically classified the edits into approximately 25 primary error types (Bryant et al., 2017), thus providing a middle ground between model output and grammar teaching vocabulary. Lyster and Ranta (1997) discovered that although recasts were the most frequent feedback move in class, they were less likely to result in learner uptake than

moves such as elicitation and metalinguistic cues. Research on automated feedback in writing indicates conflicting findings: automated feedback is less specific than that of the instructor (Dikli & Bleyle, 2014); the value of automated feedback hinges on their integration in instruction (Li et al., 2015); and learners cope better with accurate automated feedback than with inaccurate feedback (Ranalli, 2018). The overall evidence from intelligent tutoring systems is encouraging (Kulik & Fletcher, 2016); however, the evidence has not yet separated the contribution of explanation.

Research Gap

Three lines of research have thus emerged fairly independently: XAI methods designed for developers, NLP interpretability focusing on models, and research on SLA feedback focusing on learners. There is little in the way of work relating to them for grammar instruction, and evidence of the effectiveness of the formats of explanation for learning is mostly indirect. A caveat from the human-centred XAI literature: Poursabzi-Sangdeh et al. (2021) discovered that while providing transparency to a model facilitated users to simulate model predictions, it did not help to identify model errors. The present review attempts to fill the lacuna by bringing together the strands.

Research Questions and Objectives

- RQ1. Which explanation techniques can expose the decisions of neural grammar models, and how suitable is each for grammar teaching?
- RQ2. What do SLA, feedback and learning-science sources imply about what a learner-facing explanation should do?
- RQ3. What technical, pedagogical and ethical challenges limit transparent AI grammar tutoring?

The corresponding objectives are to map the evidence base, to appraise methods against pedagogical criteria, and to propose and justify a design framework.

Methodology

Design. The study is a narrative review of an integrated nature, as it is conducted with secondary data only, no learners were recruited and a new model was not trained. Integrative designs are appropriate to topics that require conceptual, technical and empirical work from multiple disciplines to be integrated.

Source selection. Publicly indexed sources such as the ACL Anthology, ACM Digital Library, IEEE Xplore, Elsevier, Cambridge, Wiley, Sage and Taylor & Francis journals and the arXiv repository were targeted for purposive sampling. The items included were those that (i) discussed at least one of four strands: XAI foundations, neural NLP interpretability, GEC and SLA/feedback/learning theory with an educational-technology connection, (ii) were either peer-reviewed or are widely cited archival reports and (iii) are written in English. The outcome is a collection of 40 sources (Figure 1). Since the selection was done on a purposive basis and not by an exhaustive database search, the review cannot claim to be complete and is not a systematic review.

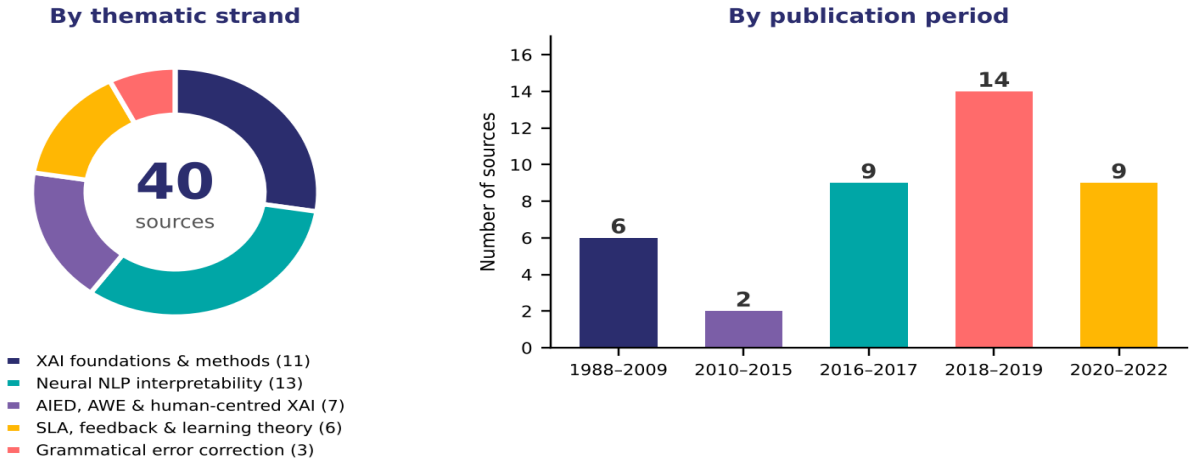


Fig. 1. *Profile of the 40-source corpus by thematic strand (left) and publication period (right). Counts derived from the reference list of this paper.*

Analysis. The claims were coded by source type, main claim and relation to the three research questions, and the claims were synthesised thematically (Table 1). Five pedagogical criteria were defined through literature to compare the explanation techniques: faithfulness to the model (Jacovi & Goldberg, 2020), learner intelligibility (Miller, 2019), actionability (Hattie & Timperley, 2007) and link to grammar

rules (Ellis, 2006) as well as computational cost. Each criterion was rated as Low, Medium or High for technique. These are the author's judgments; they are not empirical measures based on the literature cited, but are rather a qualitative ranking.

Findings

RQ1: Explanation Techniques And Their Fit For Grammar Teaching

This appraisal is summarised in Figure 2. Attention displays: These are cheap and easy to use, but the attention weight is not necessarily causal importance of the correction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), and should not be presented to learners as evidence of the causal importance of the correction. Gradient and path-integrated attribution (Sundararajan et al., 2017) provides more faithfulness guarantees in the form of axiomatic properties and yet still provides the token-level scores that need to be translated to grammar-level. Model agnostic methods like LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017) have the benefit of being model agnostic, but incur the cost of repeated sampling and their results can be unpredictable in the context of text. Two families are outstanding for teaching. As in human explanation (Miller, 2019) and the noticing-the-gap principle (Schmidt, 1990), counterfactual/ minimal-pair explanations (Wachter et al., 2017; Warstadt et al., 2020) demonstrate how a little change changes the verdict (go to goes). The edit-tag and rule traces from tagging-based GEC (Omelianchuk et al., 2020) are intrinsically inspectable and correspond to grammar categories, in line with Rudin's (2019) desire for interpretable designs. Probing classifiers (Tenney et al., 2019; Hewitt & Manning, 2019) show a teacher or model developer what the model has learned, but they do not provide insights into a specific decision, so are not suited to learners.

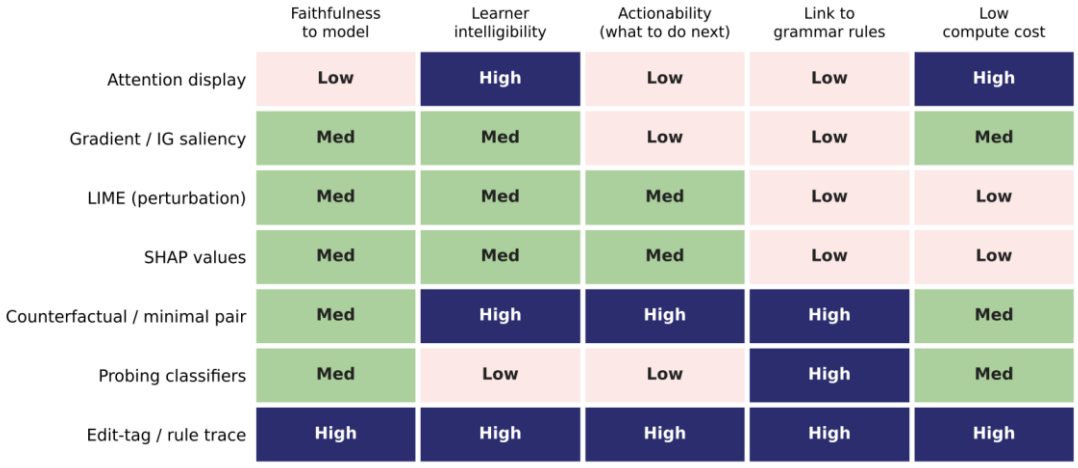


Fig. 2. *Appraisal of seven explanation techniques against five pedagogical criteria (author's qualitative synthesis of the cited literature; ordinal, not empirical).*

RQ2: What Pedagogy Implies Explanations Should Do

Five implications emerge. To promote noticing, explanations must call attention to the specific difference between the output of learners and the target form (Schmidt, 1990). Second, they should be explicit and grammar-related; explicit instruction has been found to have a large and long-lasting impact (Norris & Ortega, 2000) and grammar instruction benefits from both explicit and meaning-focused instruction (Ellis, 2006). Third, they need to be invitations to action; those prompts that elicit self-correction result in more uptake than recasts alone (Lyster & Ranta, 1997), and feedback is best when it suggests next steps (Hattie & Timperley, 2007). Fourth, they must reduce the cognitive load – an explanation that is more complex than what working memory can cope with during learning is not beneficial (Sweller, 1988), and the more transparent the explanation, the more efficient is not necessarily a better human judgement (Poursabzi-Sangdeh et al., 2021). Fifth, they should be adaptable and reliable, as claimed in the literature of open-learner-modelling and educational XAI (Conati et al., 2018; Khosravi et al., 2022); and intelligent tutoring literature reveals that well-designed systems can boost achievement (Kulik & Fletcher, 2016).

Table 1. *Extract of key secondary sources and their contribution to the synthesis*

Source	**Type**	**Key contribution**	**Relevance to grammar teaching**
Miller (2019)	Review (social science)	Explanations are contrastive, selective and social	Supports minimal-pair, "why this, not that" design
Jain & Wallace	Empirical	Attention weights are an unreliable stand-	Avoid raw attention as

(2019); Wiegreffe & Pinter (2019)	debate	alone explanation	learner evidence
Jacovi & Goldberg (2020)	Conceptual	Separates faithfulness from plausibility	A plausible but false rule misleads learners
Tenney et al. (2019); Hewitt & Manning (2019)	Probing studies	Encoders represent syntax in structured ways	Models can, in principle, support rule-linked explanation
Warstadt et al. (2020)	Benchmark	67 minimal-pair sets across 12 phenomena; uneven mastery	Flags where explanations may accompany wrong corrections
Omelianchuk et al. (2020)	System paper	Edit-tag GEC with iterative correction	Edits are inspectable units for explanation
Lyster & Ranta (1997)	Classroom study	Prompts yield more uptake than recasts	Explanations should invite learner response
Poursabzi-Sangdeh et al. (2021)	Experiments	Transparency did not improve detection of model errors	Plan for overload and over-reliance

RQ3: Challenges

Faithfulness versus plausibility. The central technical risk are explanations that appear convincing, but which misrepresent the model (Jacovi & Goldberg, 2020). Developers should not rely on plausibility alone since analyzing interpretability requires human-grounded and application-grounded tests (Doshi-Velez & Kim, 2017). Model fallibility. Benchmarks show systematic shortcomings in syntactic competence (Linzen et al., 2016; Warstadt et al., 2020) while automated feedback is only valuable if it is correct (Ranalli, 2018). An attached explanation for a wrong correction can solidify an error. The cognitive and metacognitive load. Giving too many detailed attributions might overload novices (Sweller, 1988; Poursabzi-Sangdeh et al., 2021). Instructional integration. Automated feedback should complement rather than substitute for the teacher's feedback (Li et al., 2015) and may be less specific than teacher feedback (Dikli & Bleyle, 2014). Evidence gap. The sources examined showed no evidence of a direct impact of grammar explanations produced by XAI on learning outcomes, and this is the primary focus of future experiments.

Cross-Cutting Synthesis

f you read through the three research questions, there is a uniform pattern. Techniques that are most effective for pedagogical criteria (counterfactual contrast, edit-tag traces) are also the ones that are most similar to how people explain and how teachers provide feedback, while techniques that are most prevalent in the developer-oriented literature (attention maps, probing, perturbation scores) need to be translated before they can serve learners. The evidence also suggests there is a point of division of labour: learner-facing messages need to be concise, and based on rules; elaborations of messages such as attribution maps, probing results should be placed in a teachers or developers view and therefore be used to support auditing. Last but not least, the benchmark evidence on uneven syntactic competence (Warstadt et al., 2020) suggests that, as long as the syntax quality of the model is low, either the explanation should be disabled or a caution should be added.

Table 3. *Proposed evaluation plan for TGEF, following the three-level evaluation logic of Doshi-Velez and Kim (2017)*

Level	**Question**	**Example measures**
Functionally grounded	Is the explanation faithful to the model?	Deletion and perturbation tests; agreement with edit-tag trace (Jacovi & Goldberg, 2020)
Human-grounded	Can learners and teachers understand it?	Rule recall, rating of clarity, time to comprehend (Poursabzi-Sangdeh et al., 2021)
Application-grounded	Does it improve learning?	Pre/post grammar tests, delayed retention, uptake rate (Lyster & Ranta, 1997; Norris & Ortega, 2000)

Proposed Framework: TGEF

The results of the synthesis are presented in the Transparent Grammar Explanation Framework (TGEF) which classifies transparent grammar tutoring into four layers (Figure 3). Layer 1: Detect & Correct is edit based GEC model which provides an explicit operation type and confidence value for each correction. Layer 2, Explain, generates faithful evidence in the form of a counterfactual contrast, an attribution check and, where possible, the edit-tag trace, while maintaining probing and attention displays in a teacher view. In Layer 3, Translate, evidence is translated to language of the level: rule, example, minimal pair and short practice item, while taking into consideration the limits of cognitive load. Layer 4, Learn & Audit, keeps track of learner uptake, uncertainty and provides a dashboard for teachers to identify any gaps in explanations to be captured and incorporated into the model development..

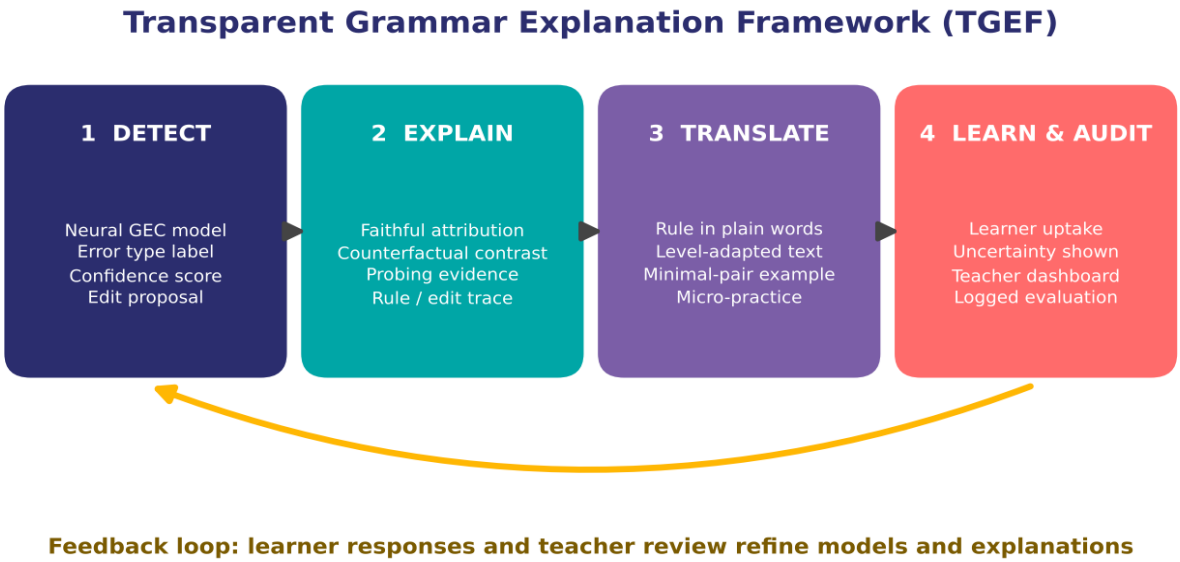


Fig. 3. *The Transparent Grammar Explanation Framework (TGEF): four layers with a continuous feedback loop (author's synthesis).*

Figure 4 shows how Layers 2 and 3 might appear to a learner. The sentence and scores are a constructed illustration of the explanation format, **not** output from a real model. Table 2 maps typical learner questions to explanation forms and to the literature that supports each.

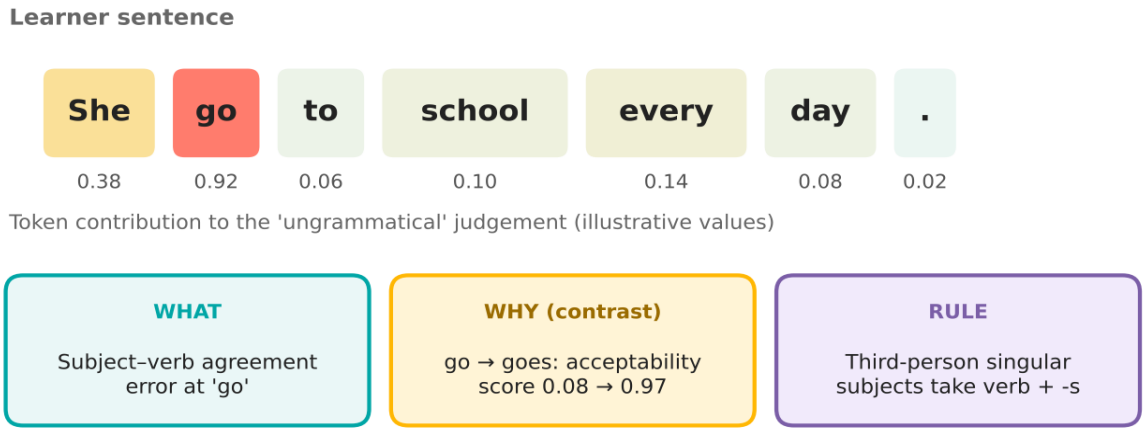


Fig. 4. *Illustrative learner-facing explanation: what (error), why (contrast) and rule. Values are constructed for demonstration only.*

Table 2. *Learner questions mapped to explanation forms and supporting evidence*

Learner question	Explanation form	Evidence base
What did I get wrong?	Highlighted token plus error-type label	Noticing (Schmidt, 1990); ERRANT types (Bryant et al., 2017)
Why is it wrong?	Counterfactual contrast: "go → goes"	Contrastive explanation (Miller, 2019); counterfactuals (Wachter et al., 2017)

What is the rule?	Plain-language rule linked to the edit	Explicit instruction (Norris & Ortega, 2000; Ellis, 2006)
How sure is the system?	Confidence indicator and caution flag	Faithfulness and trust (Jacovi & Goldberg, 2020; Rudin, 2019)
What should I do next?	Micro-practice or self-correction prompt	Prompts and uptake (Lyster & Ranta, 1997); feedback (Hattie & Timperley, 2007)

Worked scenario (illustrative)

For an intermediate student, let's look at this: “She go to school every day, and she is liking mathematics.” A tool that uses the conventional method could silently give a corrected sentence. According to TGEF, layer 1 suggests two changes, and classifies them as error types (subject–verb agreement, stative verb in the progressive). Layer 2 provides for every edit a counterfactual comparison and a confidence value, and if the confidence is low (which can occur with less common structures, for instance, as in Warstadt et al., 2020), the system outputs a warning rather than a rule. At the learner's level, Layer 3 translates the evidence into a short sentence, such as a one-line rule, a minimal pair and a cue to get the learner to fix a second sentence that has the same pattern. Layer 4 indicates if the learner has corrected the new sentence and the teacher's dashboard adds up such indications, pinpointing the rules that are most challenging for the class. There are three design decisions to make note of. First, the learner is given a chance to try to fix the problem before checking the answer, and thereby turn the feedback from a recast into a prompt (Lyster & Ranta, 1997). Secondly, the explanation is given in the language of the course materials, thus facilitating explicit learning (Norris & Ortega, 2000). Thirdly, attention maps are not shown as evidence by the system, as the relationship between attention maps and the model's reasoning is questionable (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). The scenario is not an empirical finding but it is offered as a frame of reference.

Discussion

The synthesis implies that it is not so promising to show everything that a model “knows” but rather a few but faithful, contrasting and rule-related messages. This corresponds to the view of Miller (2019) that explanation is a social and selective process, and to the proposal of Rudin (2019) for an inspectable design in situations of high stakes. The upshot for teachers is to consider AI explanations as a draft, e.g., benchmark evidence indicates that teaching competence is weaker in less frequently used constructions (Warstadt et al., 2020). For the developers, edit-tag architectures (Omelianchuk et al., 2020) and a standardised error typing approach (Bryant et al., 2017) offer a good starting point, while any post-hoc layer ought to be validated for faithfulness (Jacovi & Goldberg, 2020). From a researcher point of view, a controlled study explaining the impact of different explanation types on retention and transfer and calibration of trust would be the most valuable. Transparency alone has not been proven to lead to better human decision quality (Poursabzi-Sangdeh et al., 2021).

Implications for Practice

Classroom teachers can make responsible use of existing tools by posing the following questions when using any AI correction: Does the system identify the type of error? Does it illustrate a contrast or provide an example which the learner can test? Is there a possibility for the learner to try the fix before seeing it? When tools are not enough, teachers can provide explanations, thus maintaining the centrality of the human factor in the teaching and learning process (Li et al., 2015). Curriculum designers can make an explanation vocabulary consistent with the grammar syllabus so that the explanations in the tool are the same as those in the textbook, which will lessen the burden of the translation of the explanation for learners. Platform developers should include log documentation of explanation views and learner responses, which is more than just an occasional or exceptional event, but should be routine at the human-grounded and application-grounded levels of evaluation (Doshi-Velez & Kim, 2017).

Implications for Research

The review reveals that there are three bodies of knowledge which are ready for integration and have not been tested together. Researchers in NLP can then provide faithfulness tests for the grammatical phenomena, SLA researchers can identify which component of which explanation is likely to be important for which structure and level of proficiency (Ellis, 2006), and educational technologists can design classroom trials to make estimates about the effects. Cumulative evidence is thus most likely to be achieved through cross-disciplinary teams.

Ethical Considerations

Morally, it is not transparency that should be confused with being correct. The fluent explanation may produce an unnecessary feeling of confidence and the learners may rely on the system rather than developing their own grammatical judgement. Design responses, such as

uncertainty displays and teacher prompts asking learners to try to correct first are based on the feedback and tutoring literature (Hattie & Timperley, 2007; Kulik & Fletcher, 2016).

Limitations

Secondary data have been used and a purposive sample of 40 sources has been taken; other relevant work may not have been included, and the sample is skewed towards outlets that published English language and/or computationally oriented work. The appraisal in figure 2 is based on the author's qualitative assessment and not measured impacts. Findings on learners outcomes are infered from neighboring domains (automated writing evaluation, intelligent tutoring) and the TGEF framework and Figure 4 are conceptual proposals that have yet to be empirically tested. The field is also rapidly evolving and many advancements have been made after 2022 in the field of large language models, which is not reflected in the corpus.

Conclusion and Future Research

- Neural Grammar Tools Will Not Understand By Themselves If They Are Not Explained As A Pedagogical Act. Based On The Reviewed Evidence, The Most Promising Explanations Are Contrastive, Rule-Linked And Faithful Explanations, While Attention-Only Displays And Unvalidated Post-Hoc Explanations And Information Overload Are The Greatest Dangers. The TGEF Gives A Systematic Model Decision - Learner Understanding Pathway. Future Work Should:
- Run Randomised Classroom Studies Comparing Explanation Formats On Retention, Transfer And Uptake;
- Develop Faithfulness And Learner-Comprehension Benchmarks Specific To Grammar Explanation;
- Study Learner Calibration Of Trust When The System Is Occasionally Wrong; And
- Extend The Framework to multilingual and L1-sensitive explanation, including large language model tutors.

References

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.
3. Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics, 7*, 49–72.
4. Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 52–75). Association for Computational Linguistics.
5. Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the ACL* (pp. 793–805). Association for Computational Linguistics.
6. Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). *AI in education needs interpretable machine learning: Lessons from open learner modelling* (arXiv:1807.00154). arXiv.
7. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. In *Proceedings of AACL-IJCNLP 2020* (pp. 447–459). Association for Computational Linguistics.
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics.
9. Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17.
10. Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (arXiv:1702.08608). arXiv.
11. Ellis, R. (2006). Current issues in the teaching of grammar: An SLA perspective. *TESOL Quarterly, 40*(1), 83–107.
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), Article 93.
13. Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
14. Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of NAACL-HLT 2019* (pp. 4129–4138). Association for Computational Linguistics.
15. Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the ACL* (pp. 4198–4205). Association for Computational Linguistics.
16. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of NAACL-HLT 2019* (pp. 3543–3556). Association for Computational Linguistics.
17. Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., … Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074.

18. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*(1), 42–78.
19. Li, Z., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18.
20. Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics, 4*, 521–535.
21. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57.
22. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).
23. Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition, 19*(1), 37–66.
24. Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural NLP: A survey. *ACM Computing Surveys, 55*(8), Article 155.
25. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.
26. Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(3), 417–528.
27. Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzhanskyi, O. (2020). GECToR – Grammatical error correction: Tag, not rewrite. In *Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–170). Association for Computational Linguistics.
28. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 237). ACM.
29. Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674.
30. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.
31. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866.
32. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
33. Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158.
34. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR 70, pp. 3319–3328).
35. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
36. Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the ACL* (pp. 4593–4601). Association for Computational Linguistics.
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30* (pp. 5998–6008).
38. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.
39. Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics, 8*, 377–392.
40. Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 11–20). Association for Computational Linguistics.