Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5994–6000
Marseille, 20-25 June 2022

ProQE: Proficiency-wise Quality Estimation dataset for Grammatical Error Correction

Yujin Takahashi[1], **Masahiro Kaneko**[2], **Masato Mita**[3,1], **Mamoru Komachi**[1]
[1]Tokyo Metropolitan University, [2]Tokyo Institute of Technology, [3]RIKEN
takahashi-yujin@ed.tmu.ac.jp, masahiro.kaneko@nlp.c.titech.ac.jp, masato.mita@riken.jp, komachi@tmu.ac.jp

Abstract

This study investigates how supervised quality estimation (QE) models of grammatical error correction (GEC) are affected by the learners' proficiency with the data. QE models for GEC evaluations in prior work have obtained a high correlation with manual evaluations. However, when functioning in a real-world context, the data used for the reported results have limitations because prior works were biased toward data by learners with relatively high proficiency levels. To address this issue, we created a QE dataset that includes multiple proficiency levels and explored the necessity of performing proficiency-wise evaluation for QE of GEC. Our experiments demonstrated that differences in evaluation dataset proficiency affect the performance of QE models, and proficiency-wise evaluation helps create more robust models.

Keywords: Grammatical Error Correction, Evaluation, Quality Estimation, Proficiency, Corpus

1. Introduction

Grammatical error correction (GEC) refers to the task of correcting a variety of grammatical errors in text written by non-native speakers learning a language. Thus, the main application of a GEC system is to assist language learners with their writing. The performance of GEC systems has been improving, but there is still room for improvements. The precision of the state-of-the-art system is approximately 75% and the recall is below 50% on the CoNLL-2014 test set (Stahlberg and Kumar, 2021). Therefore, it is difficult for learners to judge whether they can trust the GEC system's output, as it may be potentially misleading in some cases.

The quality estimation (QE) of GEC, which is the task of estimating the quality of GEC system outputs, has attracted attention as a key method to address the aforementioned difficulties in real-world scenarios (Heilman et al., 2014; Chollampatt and Ng, 2018b; Asano et al., 2017; Yoshimura et al., 2020). Quality estimates of the system's output can help learners and instructors decide whether to adopt or ignore the system's correction. Recently, it has been reported that QE using Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) was highly correlated with human evaluations (Yoshimura et al., 2020).

However, the current evaluation in QE of GEC is unfortunately insufficient because it is restricted to a corpus limited to certain proficiency levels. For example, the data used in prior works (Asano et al., 2017; Yoshimura et al., 2020) were biased toward data generated for learners with relatively high proficiency levels, such as CoNLL-2014 (Ng et al., 2014). By contrast, it is assumed that the proficiency level of language learners varies from beginners to advanced learners in real-world use cases. Given the nature of the QE task, QE systems need to estimate quality with high accuracy for sentences written by learners with low proficiency who may have difficulty trusting the system's output. However, the current evaluation cannot address this issue. If research is conducted based on such a limited evaluation, the QE systems will be overfitted to a specific type or genre of written English. We consider proficiency levels as one of the critical factors affecting quality estimation, but other factors such as the topic of the essays can also be considered.

To overcome these limitations and improve the real-world applicability of GEC with QE, we present a new dataset, **Pro**ficiency-wise **Q**uality **E**stimation (**ProQE**) dataset, which includes multiple proficiency levels designed for proficiency-wise evaluation. Using our dataset, we first explore the necessity of proficiency-wise evaluation based on the hypothesis that it is insufficient to evaluate only a specific proficiency level as a methodology for evaluating QE models. Then, we reveal how QE models are affected by the learner's proficiency.

Thus, the contributions of this work are three-fold: (1) We propose a novel dataset designed for proficiency-wise evaluation in QE of GEC to accommodate the detailed analysis of the impact of learner's proficiency on models' performance. (2) Using this dataset, we demonstrated that results varied when the evaluation data contained different proficiency levels and that we could create more robust QE models based on our findings. (3) Our dataset[1] will be made public to allow researchers in the community to conduct proficiency-wise evaluations efficiently.

2. Related Work

We were motivated by the issue of robustness in the GEC community. Prior works focused on improving the performance of GEC systems on CoNLL-2014 (Ng et al., 2014). However, it has been recently reported that a single corpus evaluation is insufficient because

[1] https://github.com/tmu-nlp/ProQE

the input of GEC (i.e., the learner text) is diverse with respect to the learner's proficiency and native language (L1), and the system performance is sensitive to these factors (Mita et al., 2019). Additionally, and with similar motivation, the BEA-2019 Shared task (Bryant et al., 2019) was held and provided the research community with proficiency-wise data for more robust evaluations. We conjecture that a similar situation occurs in QE of GEC datasets.

The evaluation methods for GEC include reference-based and reference-less evaluations. Reference-based metrics, such as Max Match (Dahlmeier and Ng, 2012), ERRANT (Bryant et al., 2017), and GLEU (Napoles et al., 2015), face a fundamental problem in that it is difficult to include all possible references, even if they are grammatically correct (Choshen and Abend, 2018). As a result, prior works have proposed reference-less evaluations to address these limitations. Napoles et al. (2015) first presented a reference-less evaluation using grammatical error detection tools and linguistic features. Asano et al. (2017) alternatively combined three sub-metrics (grammaticality, fluency, and meaning preservation) and achieved a higher correlation with manual evaluations than the previous reference-based metrics. Moreover, Yoshimura et al. (2020) optimized each sub-metric for manual evaluation and obtained much significant improvements using BERT-based QE models. In GEC, reference-less evaluations are a part of QE because it estimates the quality of the system's output without requiring gold-standard references.

3. The ProQE dataset

Dataset Design To analyze the impact of the learners' proficiency level on the QE model, we require a dataset that is differentiated by proficiency level. To serve this purpose, we considered that the data with proficiency information should ideally be sourced from the Common European Framework of Reference for Languages (CEFR)[2], an international index for assessing language proficiency. Hence, we selected the Write & Improve (W&I) and The Louvain Corpus of Native English Essays (LOCNESS) (Bryant et al., 2019; Granger, 1998) datasets because they meet our requirements, and their use is prevalent in the GEC community. W&I includes non-native English students' writing across three different proficiency levels: beginner (A), intermediate (B), and advanced (C). Furthermore, LOCNESS contains essays written by native (N) English students. As requirements of system outputs, we considered that the systems must be diverse and commonly used since we need to ensure the applicability in a practical use case. Hence, we adopted five diverse and commonly used GEC systems, (SMT (Grundkiewicz and Junczys-Dowmunt, 2018), RNN (Luong et al., 2015), CNN (Chollampatt and Ng, 2018a), Transformer (Vaswani et al., 2017), and Transformer with a copy mechanism (Zhao et al., 2019)), to obtain system outputs. We selected unique pairs from these outputs for annotating the source and correction sentence pairs, respectively.

Please rate each of the following 50 sentences. You should evaluate using two-sentences; an original sentence (S1) and a corrected sentence (S2). Please rate the S2 (a corrected sentence) on a scale of 5 overall. The evaluation criteria for each label are as follows.

label	criteria
4	S2 corrected all errors completely. (S2 has perfect grammar.)
3	S2 corrected serious errors and fixed many minor errors. (S2 has one or more minor grammatical errors.)
2	S2 corrected serious errors but contains minor errors and/or incorrect minor corrections.
1	S2 did not correct serious and minor errors and/or contains serious incorrect corrections.
0	S1 is an incomplete sentence. (S1 cannot be corrected.)

Note: S1 has a possibility of containing no grammatical errors.

Figure 1: Task description.

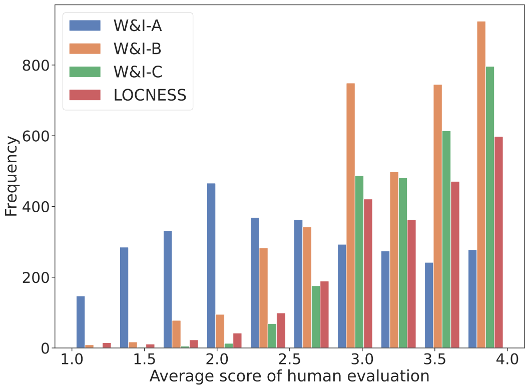


Figure 2: Histogram of each average score.

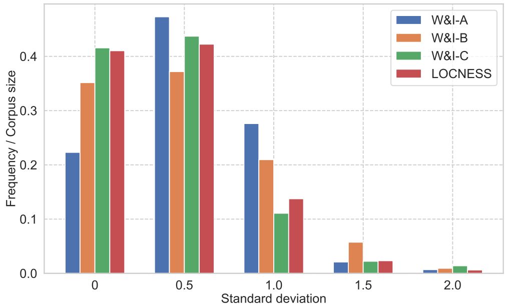


Figure 3: Histogram of each average score.

Annotation We used Amazon Mechanical Turk[3] for annotation and recruited three native English speak-

[2] https://www.cambridgeenglish.org/exams-and-tests/cefr/

[3] https://www.mturk.com/

	example sentence
source output scores	I live at Patras a big city in Greece. I live **in** Patras big cities in Greece. 2, 1, 1 (avg. score: 1.33)
source output scores	Let me allow to enter in to your cabin. Let me allow **you** to enter your cabin. 4, 4, 4 (avg. score: 4.00)

Table 1: Example annotations

level	A	B	C	N	avg.
sent.	3,049	3,740	2,644	2,233	2,917

Table 2: Number of sentences in each created dataset.

ers as annotators. First, we assigned unique sentence pairs except for duplicate corrected sentences to the annotators. Then, the annotators used both the source and the corrected sentence to evaluate the quality of the correction. Figures 1 show an actual task description[4]. Note that we did not annotate each of the submetrics unlike Yoshimura et al. (2020) because their report showed substantial weighting of one metric and its high correlation with the holistic score. Therefore, we created the five-scale evaluation criteria (4: Perfect, 3: Comprehensible, 2: Somewhat comprehensible, 1: Incomprehensible, 0: Other) to evaluate the overall correction quality. Figure 2 shows the distribution of the evaluation scores. The beginner level data were not biased for every score, but the higher scores were prevalent for the intermediate and higher levels. For reference, we provide examples of annotations in Table 1. Using the above procedure, we obtained 3,049 sentences for A, 3,740 sentences for B, 2,644 sentences for C, and 2,233 sentences for N. Table 2 lists the number of sentences for each created dataset.

Ethical Consideration The annotator was given two sentences in a question, one before and one after the correction, and evaluated the quality of the correction. Each task included 50 questions; we paid $1.3 per task, which resulted in an average hourly wage of $7.8. The entire corpus collection took approximately five weeks.

4. Proficiency-wise Evaluation

The main objective of our experiment is to investigate the necessity of proficiency-wise evaluation based on the hypothesis that it is insufficient to evaluate only a specific proficiency level as a methodology for evaluating QE models. Thus, following the methodology introduced by Mita et al. (2019), we prepare a wide variety of QE models and examine whether the performance ranking of the QE models changes between each proficiency level.

4.1. Configurations

Evaluation We report Pearson's correlation coefficients and Spearman's rank correlation coefficients. All scores were evaluated using a five-fold cross-validation because the evaluation results were highly dependent on the seed value of the data division. We used Yoshimura et al. (2020)'s implementation [5] implementation for QE models.

Baseline Models We adopted BERT used in Yoshimura et al. (2020), which achieved the highest correlation with the manual evaluations, as the base architecture for our baseline models. We used Hugging Face's implementation [6] for classification models. In these experiments, we arranged six models in terms of two aspects: (1) writer's proficiency and (2) data size. Specifically, for the first, we employed a total of four different models by fine-tuning BERT for each proficiency level: BEGINNER, INTERMEDIATE, ADVANCED, and NATIVE. To avoid differences in the number of sentences, we unified the number of sentences to 2,233 based on the dataset with the fewest number of sentences[7]. For the second, we employed two models: one fine-tuned with the combined data of all the proficiency levels (MIXED) and the other fine-tuned with the data randomly sampled from the combined data so that the data size is the same as other models except for MIXED (RANDSAMP). We performed a grid search for hyperparameters to maximize Pearson's correlation coefficient.

4.2. Result

Table 3 shows the result of the proficiency-wise evaluation. The evaluation results reveal that the models' rankings vary considerably depending on the proficiency, indicating that it is insufficient to evaluate only a specific proficiency level as a methodology for evaluating QE models.

Difference in Proficiency The top group in Table 3 shows the effect of differences in proficiency levels. Experimental results show that each QE model performed better correlations when the learners' proficiency levels in the data were consistent at each stage of fine-tuning and evaluation. Except for the evaluation data at the beginner level, there were no significant differences in the correlations. By contrast, we confirmed that there was a large difference between BEGINNER and NATIVE in the beginner-level evaluation data (e.g., 0.70±.05 vs. 0.59±.05 in Pearson's correlation). We provide a detailed analysis of this observation in Section 5.1.

Importance of Data Size The second group in Table 3 shows the effect of differences in data size. Al-

[4] Screenshot of actual working in Appendix A

[5] https://github.com/kokeman/SOME

[6] https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification

[7] For datasets with more than that many sentences, we randomly selected 2,233 sentences.

	Pearson's correlation				Spearman's rank correlation			
Model	A	B	C	N	A	B	C	N
Beginner	**0.70±.05**	0.54±.03	0.51±.05	0.60±.05	**0.69±.05**	0.55±.03	0.52±.05	0.60±.05
Intermediate	0.63±.03	0.58±.06	0.52±.06	0.57±.05	0.65±.04	0.58±.05	0.52±.06	0.57±.04
Advanced	0.63±.02	0.52±.04	0.52±.05	0.58±.05	0.66±.05	0.53±.02	0.52±.05	0.58±.02
Native	0.59±.05	0.54±.05	0.49±.05	**0.61±.04**	0.63±.07	0.56±.03	0.50±.07	0.60±.03
RandSamp	0.59±.02	0.53±.06	0.51±.04	0.57±.05	0.63±.04	0.54±.05	0.51±.04	0.57±.05
Mixed	0.68±.06	**0.58±.05**	**0.57±.03**	0.60±.02	0.69±.06	**0.60±.05**	**0.57±.05**	**0.61±.05**
Mixed+Tag	**0.72±.06**	**0.61±.04**	**0.60±.05**	**0.62±.02**	**0.71±.06**	**0.63±.02**	**0.60±.04**	**0.63±.03**

Table 3: Result of the proficiency-wise evaluation.

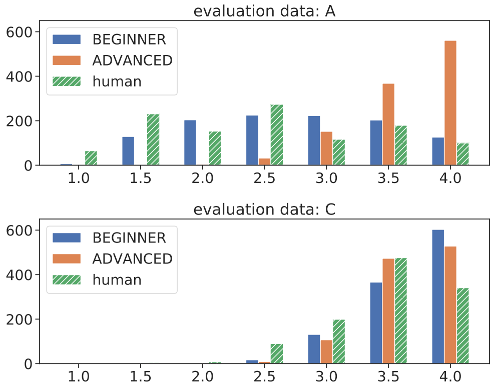


Figure 4: Distribution of prediction scores by QE.

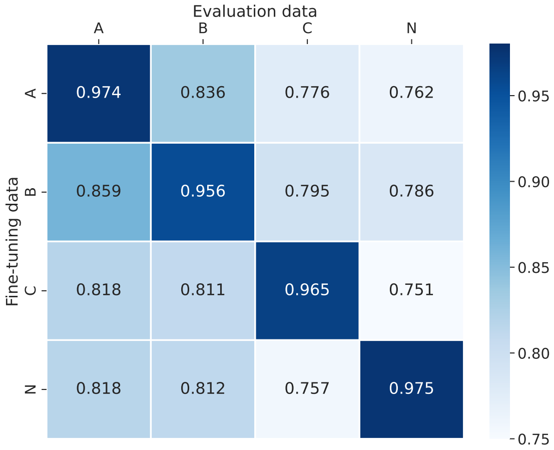


Figure 5: Overlap of vocabulary between the fine-tuning and evaluation data.

though the performance of RandSamp was the lowest, or nearly the lowest score for all of the evaluation data, the performance of Mixed was competitive for almost all of the evaluation data. Mixed obtained improvement by increasing the data size, but it did not outperform the best model for each proficiency level in some cases.

5. Discussion and Analysis

5.1. On the Impact at the Beginner Level

We found that the impact of proficiency was significant, especially at the beginner level. One reason for this effect might be related to the distribution of the scores of the manual evaluation because the distribution differed significantly between the beginner and other proficiency levels (Figure 2). Therefore, we assumed that the QE models fine-tuned on data from the high proficiency level tended to produce high prediction values, even for low-quality inputs.

To verify this hypothesis, we confirmed the distribution of the manual evaluation values and the predicted values of the QE model for the evaluation data of the two proficiency levels: beginner and advanced (Figure 4). In the case of the beginner level, Beginner had a distribution similar to that of the manual evaluation. Still, Advanced only produced predictions of 2.5 or higher. By contrast, the QE model fine-tuned at any proficiency level showed almost no significant difference for the evaluation data from the advanced level. Thus, the bias in the distribution of the manual evaluation caused the QE model to produce overestimations for low-quality input.

5.2. Proficiency and Vocabulary Overlap

Although the distributions of manual evaluation from intermediate to native were similar, the correlation tended to be slightly higher when the proficiency level matched the fine-tuning and evaluation stages. We presumed that this is because of the influence of vocabulary differences caused by increased proficiency levels. Hence, we investigated the overlap of words in each data.

Figure 5 shows the overlap of the vocabulary between fine-tuning and evaluation data. Because each model increased the vocabulary size of evaluation data during the fine-tuning phase, we assumed that this contributed to slight differences. Thus, one reason for this slight difference in the correlation results may be differences in vocabulary.

5.3. Seeking a More Robust QE Model

Our analysis found that proficiency information and data size contributed to improving the performance. Based on our findings, we examined whether adding proficiency information to the MIXED model could improve its performance. Specifically, we created a setting, MIXED+TAG, in which each sentence was prefixed with a proficiency tag (e.g., `[A]`) for fine-tuning. The bottom of Table 3 shows the results of these settings. We confirmed improved correlations with the evaluation data for all proficiency. In particular, we verified an increase in beginner level evaluation data, which slightly improved for the others. This result demonstrated that proficiency-wise evaluation could help create robust QE models.

6. Conclusions

This study performed a proficiency-wise evaluation using the ProQE dataset and presented the necessity of proficiency-wise evaluation for QE of GEC. Furthermore, we showed it to help create robust QE models based on the results. Since we considered only the impact of proficiency in this study, the topic's effect is unknown, but we will be handling it in future studies. To facilitate more research on GEC with QE, we will make the dataset freely available.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 22H03651.

7. Bibliographical References

Asano, H., Mizumoto, T., and Inui, K. (2017). Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Bryant, C., Felice, M., and Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July. Association for Computational Linguistics.

Chollampatt, S. and Ng, H. T. (2018a). A multilayer convolutional encoder-decoder neural network for grammatical error correction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, February.

Chollampatt, S. and Ng, H. T. (2018b). Neural quality estimation of grammatical error correction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2528–2539, Brussels, Belgium, October-November. Association for Computational Linguistics.

Choshen, L. and Abend, O. (2018). Inherent biases in reference-based evaluation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Melbourne, Australia, July. Association for Computational Linguistics.

Dahlmeier, D. and Ng, H. T. (2012). Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572, Montréal, Canada, June. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Grundkiewicz, R. and Junczys-Dowmunt, M. (2018). Near human-level performance in grammatical error correction with hybrid machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 284–290, New Orleans, Louisiana, June. Association for Computational Linguistics.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180, Baltimore, Maryland, June. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Mita, M., Mizumoto, T., Kaneko, M., Nagata, R., and Inui, K. (2019). Cross-corpora evaluation and analysis of grammatical error correction models — is single-corpus evaluation enough? In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1309–1314, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Napoles, C., Sakaguchi, K., Post, M., and Tetreault, J. (2015). Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference* *on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China, July. Association for Computational Linguistics.

Stahlberg, F. and Kumar, S. (2021). Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online, April. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yoshimura, R., Kaneko, M., Kajiwara, T., and Komachi, M. (2020). SOME: Reference-less submetrics optimized for manual evaluations of grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6516–6522, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Zhao, W., Wang, L., Shen, K., Jia, R., and Liu, J. (2019). Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June. Association for Computational Linguistics.

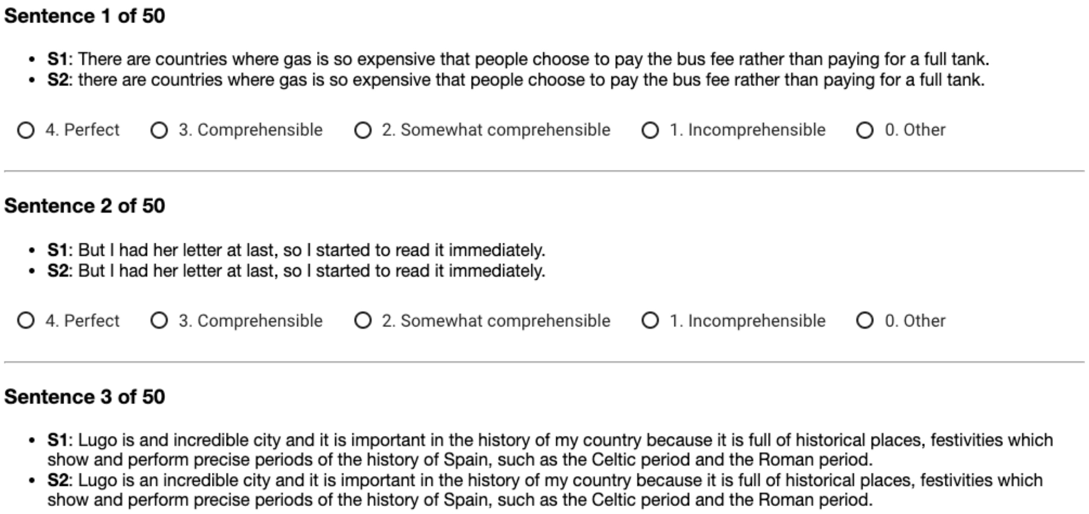
8. Language Resource References

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *BEA*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on Computer*, pages 3–18. Addison Wesley Longman, London and New York.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June. Association for Computational Linguistics.

A. Screenshot of Actual Task



Sentence 1 of 50

- **S1**: There are countries where gas is so expensive that people choose to pay the bus fee rather than paying for a full tank.
- **S2**: there are countries where gas is so expensive that people choose to pay the bus fee rather than paying for a full tank.

○ 4. Perfect ○ 3. Comprehensible ○ 2. Somewhat comprehensible ○ 1. Incomprehensible ○ 0. Other

Sentence 2 of 50

- **S1**: But I had her letter at last, so I started to read it immediately.
- **S2**: But I had her letter at last, so I started to read it immediately.

○ 4. Perfect ○ 3. Comprehensible ○ 2. Somewhat comprehensible ○ 1. Incomprehensible ○ 0. Other

Sentence 3 of 50

- **S1**: Lugo is and incredible city and it is important in the history of my country because it is full of historical places, festivities which show and perform precise periods of the history of Spain, such as the Celtic period and the Roman period.
- **S2**: Lugo is an incredible city and it is important in the history of my country because it is full of historical places, festivities which show and perform precise periods of the history of Spain, such as the Celtic period and the Roman period.

Figure 6: Screenshot of actual task on Amazon Mechanical Turk.