# ERRANT: Assessing and Improving Grammatical Error Type Classification

**Katerina Korre**
Department of Informatics
Athens University of Economics
and Business, Greece
katkorre95@aueb.gr

**John Pavlopoulos**
Department of Computer
and System Sciences
Stockholm University, Sweden
ioannis@dsv.su.se

## Abstract

Grammatical Error Correction (GEC) is the task of correcting different types of errors in written texts. To manage this task, large amounts of annotated data that contain erroneous sentences are required. This data, however, is usually annotated according to each annotator's standards, making it difficult to manage multiple sets of data at the same time. The recently introduced Error Annotation Toolkit (ERRANT) tackled this problem by presenting a way to automatically annotate data that contain grammatical errors, while also providing a standardisation for annotation. ERRANT extracts the errors and classifies them into error types, in the form of an edit that can be used in the creation of GEC systems, as well as for grammatical error analysis. However, we observe that certain errors are falsely or ambiguously classified. This could obstruct any qualitative or quantitative grammatical error type analysis, as the results would be inaccurate. In this work, we use a sample of the FCE coprus (Yannakoudakis et al., 2011) for secondary error type annotation and we show that up to 39% of the annotations of the most frequent type should be re-classified. Our corrections will be publicly released, so that they can serve as the starting point of a broader, collaborative, ongoing correction process.

## 1 Introduction

Grammatical Error Correction (GEC) is the task of correcting different types of errors in written texts, usually by taking erroneous sentences as input and transforming them into correct ones. This can be achieved with a variety or even combination of techniques, such as language modeling (Bryant and Briscoe, 2018), statistical machine translation (Katsumata and Komachi, 2019), and neural machine translation (Grundkiewicz and Junczys-Dowmunt, 2018). An important step that is usually taken in these techniques is error tagging, namely "when all errors in the corpus have been annotated with the help of a standardized system of error tags" (Granger, 2003). Error tagging (or error classification) is of utmost importance as it contributes to sentence transformations in a GEC system, when the error is mapped to the correction through special tags, such as in (Omelianchuk et al., 2020). The most popular error tagger to date is the grammatical ERRor ANnotation Toolkit (ERRANT), which automatically extracts and categorizes errors from parallel original and corrected texts (Bryant et al., 2017). By employing a rule-based classifier, ERRANT is able to expand to other languages, such as German (Boyd, 2018), Spanish (Davidson et al., 2020) and Czech (Náplava and Straka, 2019). This fact makes it particularly important for second language (L2) learning, where it can provide automatic evaluation of GEC systems in several languages (Boyd, 2018; Náplava and Straka, 2019; Davidson et al., 2020). This work suggests an ERRANT improvement, by observing a major shortcoming that currently applies and suggesting the way for it to be addressed. More specific, the contributions of this work are summarised to the following:

- We demonstrate a number of false or ambiguous classifications, using a sample of the FCE dataset (Yannakoudakis et al., 2011). Although the error classifier has been evaluated to some degree (Bryant et al., 2017), we firmly believe that more investigation is needed.





*Proceedings of LaTeCH-CLfL 2020*, pages 85–89
Barcelona, Spain (Online), December 12, 2020.

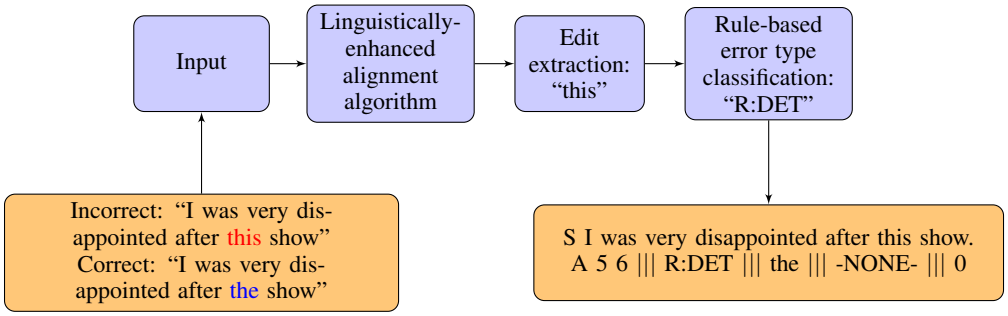


Figure 1: ERRANT system demonstration. After the input, the linguistically enhanced-algorithm aligns the two parallel sentences by making sure that items with similar linguistic properties are aligned. R:DET means that the determiner 'this' needs to be replaced with the determiner 'the'.

- We suggest re-classifications of the detected faulty items. In specific, we estimate that 39% of what has been classified as error type OTHER (the most frequent type), should have been classified to other, known error types (e.g., R:VERB).

- We publicly release our detected false classifications and our suggested re-classifications, in order to initiate a collaborative, ongoing correction process of improving the FCE dataset, which we will use for a future robust training of machine learning classifiers. In this way, we believe that any ERRANT evaluation scorers can be improved (e.g., ERRANT was employed by the most recent Grammatical Error Correction shared task: BEA-2019 (Bryant et al., 2019)).

We will first present our approach to analysing the mis-classification problem. Then we will discuss our observations on mis-classification frequencies and patterns, along with possible implications in GEC.

## 2 Methodology

For the purposes of this study, we are only concerned with the FCE corpus (Yannakoudakis et al., 2011). We used the FCE data file from the BEA-2019 shared task which was in M2 format and included all the extracted edits, error types and corrections. A thorough exploratory data analysis showed that the most frequent error type was R:OTHER (see Figure 2), meaning that something in the sentence needs to be replaced with something else that does not fit into a certain category. Also, there were errors of type M:OTHER and U:OTHER, i.e. something is missing and something is unnecessary, respectively. We focused our analysis only on sentences containing the most frequent error type, namely OTHER. We

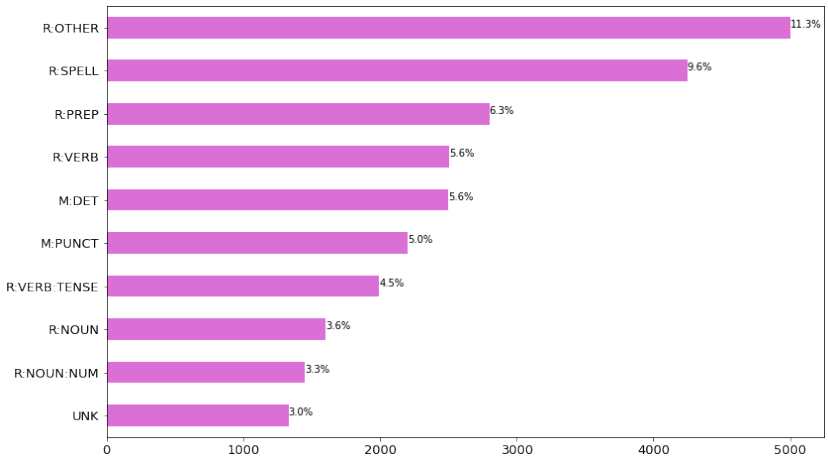


Figure 2: 21 most frequent error types in the FCE dataset, where R:OTHER type errors comprise the most frequent error type.

| Code | Meaning | Description | Example |
|---|---|---|---|
| **ADJ:FORM** | Adjective Form | Comparative/Superlative adjective errors | more easy (easier) |
| **ORTH** | Orthography | Case and/or whitespace errors | Bestfriend (instead of best friend) |
| **VERB:INFL** | Verb inflection | Missaplication of tense morphology | getted (got), fliped (flipped) |

Table 1: Three main error categories selected out of the 25 presented in (Bryant et al., 2017), serving as examples of the classification.

sampled the first 100 sentences from the FCE corpus that contain OTHER type errors (incl. M:OTHER and U:OTHER) and we manually re-labeled each of them. All of our re-classifications are publicly released as an XLSX file,[1] along with the original uncorrected sentences, the starting and ending offsets, the suggested correction, and any comments.

## 3 Results & Discussion

According to our re-classification, 39% of the errors could have been placed in other categories (i.e., 39 errors out of the sample of 100 sentences with one error each). Given that OTHER is the most frequent error type, a large number of sentences of the FCE corpus could potentially be re-classified to other categories. If this percentage applied to the whole FCE dataset, this would mean that 2724 out of the 6984 OTHER errors, are currently mistakenly tagged as OTHER.

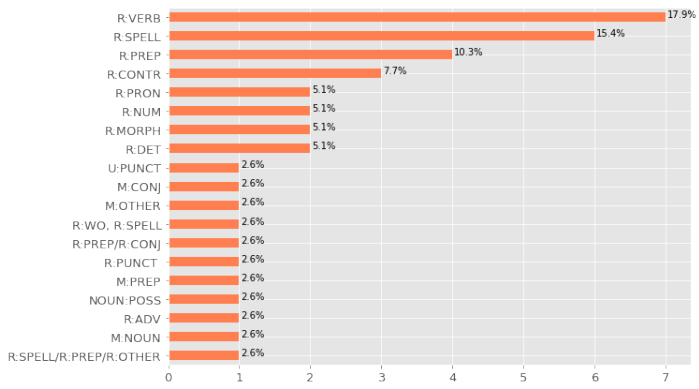


Figure 3: Error type frequencies (prev. tagged as OTHER).

The most frequent error type that was classified as OTHER was R:VERB, namely a word in the sentence has to be replaced with a verb. Spelling mistakes (R:SPELL) were also very common, accounting for about 15% of the sample. Preposition replacements (R:PREP) comprised about 10% of the sample. There were errors that were placed in other categories, as well, but as the figure shows, they account for smaller percentages.

A more qualitative demonstration is presented in Table 2. Examples 1, 3 and 4 in Table 2 contain preposition replacement errors. Example 1 and 4 are cases that possibly reflect a greater issue of ERRANT. In particular, ERRANT seems to find it easier to properly classify errors that belong to the same part of speech, or POS in short, as their correction, possibly as a result of its linguistically-enhanced alignment figure, which aligns items that are similar linguistically (Bryant et al., 2017).

In the examples, the words ‘because’ and ‘and’ are conjunctions and need to be replaced with the prepositions ‘for’ and ‘at’ respectively. Therefore, we are dealing with different POS. ERRANT ignores the option to classify the errors as R:PREP (our suggestion), and classifies them as R:OTHER instead. The specific mis-classification could be explained if we take into consideration the linguistically-enhanced alignment algorithm, which aligns linguistically similar items (see Figure 1). Because conjunctions and prepositions are different POS, ERRANT fails to assign the correct error type.

This is not the case for example 3 where the wrong preposition is replaced with a correct preposition, yet ERRANT does not provide the correct classification again. ERRANT seems to be also neglecting grammatical rules which have possibly not been implemented during the creation of ERRANT (see Figure 1 for the annotation process). For example, in sentence 2, the original sentence contains a wrong determiner ‘a’ in ‘a person’ and needs to be substituted with ‘one’. In this case, the cardinal number ‘one’ becomes the determiner, hence the suggested error classification R:DET. Example 5 clearly con-

[1] https://github.com/katkorre/ERRANT-reclassification

tains a spelling mistake, but has been overlooked by ERRANT and has been put in the R:OTHER category. The error in example 6 was re-classified from R:OTHER to R:VERB. A hypothesis for the initial mis-classification could be that 'put in' is a phrasal verb, and again the linguistically-enhanced alignment algorithm prevented the correct classification. The last example could be re-classified either as R:PRON or as R:SPELL. The inability of the tool to choose between the annotation could be the reason behind the mis-classificaion.

| No | FCE Sentence | Offsets | Correction | Old type | New type |
|---|---|---|---|---|---|
| 1 | On the other hand , the theatre restaurant was closed because unknown reasons. | 10 11 | for | R:OTHER | R:PREP |
| 2 | There was only a person who used to call her by this name . | 3 4 | one | R:OTHER | R:DET |
| 3 | I want to thak you for preparing such a good programme for us and especially for taking us to the river trip to Greenwich.. | 18 19 | on | R:OTHER | R:PREP |
| 4 | It is in the Central Exhibition Hall and will start and ten o'clock and finish at five o'clock in the evening . | 10 11 | at | R:OTHER | R:PREP |
| 5 | What are you dong here, why are n't you at school ? ' | 3 4 | doing | R:OTHER | R:SPELL |
| 6 | Also supermarket owners have put in a vast amount of money to find out the best way to place goods in order to get the most profit . | 4 6 | invested | R:OTHER | R:VERB |
| 7 | Your sincerely | 0 1 | Yours | R:OTHER | R:PRON/ R:SPELL |

Table 2: Example FCE sentences that are tagged as OTHER (5th column), along with their token-based offsets (3rd column, also highlighted in red in the text) and corrections (4th column). The last column presents our suggested re-classification.

Issues like the aforementioned must not be ignored. A more robust categorization might possibly lead to a more accurate grammatical error detection and, consequently, more efficient grammatical error correction systems.

ERRANT was used in the most recent Grammatical Error Correction shared task (BEA-2019), where all system output was automatically annotated with the scorer of the toolkit (Bryant et al., 2019). Then, the automatically inferred error type was used by the participants to evaluate their performance per type. What this means, however, is that the participants are now misjudging their systems. If we assume the existence of an oracle system that always detects correctly the error type in a (FCE) sentence, then approx. 20% of the correctly detected R:VERB errors (see Fig. 3) would be considered as OTHER errors that were miss-classified, hindering the true performance of the system for the R:VERB category.

## 4 Conclusion

ERRANT has definitely provided an alternative, and to some degree, efficient way of annotating datasets for GEC. This is particularly important for GEC systems to be able to assess their own performance and be improved. However, we show that there is still much room for improvement regarding error type classification. Although standardizing corpora can alleviate the annotators from some of the time-consuming labour, incorrect automatic classification might deprive a GEC system from useful information. Especially, in the case of teaching, where automatic feedback is gradually gaining ground, a precise error type classification is mandatory.

In the foreground, more grammar rules should be introduced during the configuration of ERRANT. This will allow a more thorough classification, and therefore more efficient error detection and correction systems. In addition, a qualitative evaluation by linguists could ensure the quality of the classification and provide professional feedback. We release our sample of second order FCE annotations, to pose the ground for the development of a larger reference dataset. Potentially, this could be used either as

a ground truth evaluation set (e.g., by rule-based systems) or as a training set by more robust machine learning classifiers.

Our next research step would be to delve into the issue of false or ambiguous error type classification further by examining and evaluating more types of errors extracted with ERRANT. We would also like to design a more systematic and thorough error classification system, by employing transfer learning and deep learning approaches.

## References


Adriane Boyd. 2018. Using Wikipedia edits in low resource grammatical error correction. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 79–84, Brussels, Belgium, November. Association for Computational Linguistics.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana, June. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 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.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy, August. Association for Computational Linguistics.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing NLP tools with a new corpus of learner Spanish. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7238–7243, Marseille, France, May. European Language Resources Association.

Sylviane Granger. 2003. Error-tagged learner corpora and call: A promising synergy. *CALICO Journal*, 20:465–480, 01.

Roman Grundkiewicz and Marcin Junczys-Dowmunt. 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.

Satoru Katsumata and Mamoru Komachi. 2019. Towards unsupervised grammatical error correction using statistical machine translation with synthetic comparable corpus.

Jakub Náplava and Milan Straka. 2019. Grammatical error correction in low-resource scenarios.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector – grammatical error correction: Tag, not rewrite.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.