Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4984–4991
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Enriching Grammatical Error Correction Resources for Modern Greek

Katerina Korre[†], **John Pavlopoulos**[‡]
[†]Università di Bologna, Forlì, Italy
[‡]Athens University of Economics and Business, Athens, Greece
[†]aikaterini.korre2@unibo.it, [‡]annis@aueb.gr

Abstract

Grammatical Error Correction (GEC), a task of Natural Language Processing (NLP), is challenging for under-represented languages. This issue is most prominent in languages other than English. This paper addresses the issue of data and system sparsity for GEC purposes in the modern Greek Language. Following the most popular current approaches in GEC, we develop and test an MT5 multilingual text-to-text transformer for Greek. To our knowledge this the first attempt to create a fully-fledged GEC model for Greek. Our evaluation shows that our system reaches up to 52.63% F0.5 score on part of the Greek Native Corpus (GNC), which is 16% below the winning system of the BEA-19 shared task on English GEC. In addition, we provide an extended version of the Greek Learner Corpus (GLC), on which our model reaches up to 22.76% F0.5. Previous versions did not include corrections with the annotations which hindered the potential development of efficient GEC systems. For that reason we provide a new set of corrections. This new dataset facilitates an exploration of the generalisation abilities and robustness of our system, given that the assessment is conducted on learner data while the training on native data.

Keywords: GEC, MT5, Encoder-Decoder

1. Introduction

This article addresses the issue of low-resourced languages in Grammatical Error Correction (GEC) by providing the prerequisites for the expansion of GEC in the Modern Greek language. In the past decade, especially since the Helping Our Own (HOO) shared task (Dale and Kilgarriff, 2011), the performance of GEC systems has increased greatly, while a variety of new datasets has emerged. The advances in GEC, however, are mostly focused on the English language with progress in other languages being quite limited.

There have been some recent attempts to enrich the resources for GEC purposes in the following languages: Spanish (Davidson et al., 2020); German (Boyd, 2018); Russian (Rozovskaya and Roth, 2019); Czech (Náplava and Straka, 2019a); Chinese (Rao et al., 2018); Arabic (Solyman et al., 2019). Our work focuses on the Modern Greek language, a highly inflectional language that consists of many declinable parts of speech that produce a vast set of morphological word forms. As Gakis et al. (2016) explain, from a single verb or adjective lemma more than 100 word forms can be produced (including both active and passive word forms, and the comparative and superlative forms). In the same study, it becomes clear that the development of Natural Language Processing systems for Modern Greek is a challenging task as Greek is also a "free-word-order" language, which gives its speakers a freedom of use that can often lead to multiple errors.

We address the issue of the scarcity of resources for GEC purposes in Greek by calibrating an MT5 multilingual text-to-text transformer (Xue et al., 2020). The pre-trained MT5 model was fine-tuned on a recently published Modern Greek dataset, the Greek Native Corpus (GNC) (Korre et al., 2021). The evaluation of the model was conducted with the Greek version of the Error Annotation Toolkit scorer (ERRANT) (Bryant et al., 2017). There were two test sets for the evaluation: Part of GNC, and part of the Greek Learner Corpus (GLC) (Tantos and Papadopoulou, 2018), on which we built by adding corrections, as the original dataset provided only XML error typing annotation, along with demographic metadata about the learners who produced the text. The performance of our system can achieve up to 52% F0.5 score for GNC, which is 16% less than the winning system in the most recent GEC shared task in English, the BEA-2019 (Bryant et al., 2019). The F0.5 score for the GLC is much lower, approximately 23%, possibly owning to the fact that the errors in this dataset were much more frequent compared to GNC.

2. Related Work

Automatic grammatical error correction can be defined as the task of automatically generating corrections and feedback on a person's writing. There are three main approaches for the creation of a GEC system so far, rule-based, classification-based and approaches based on machine translation.

GEC approaches

Rule-based approaches assure that the sentences follow specific manually coded grammar rules and that they match certain patterns (Bustamante and León, 1996). Classification-based approaches are again error-type specific. Machine learning classifiers work with error-coded corpora and are built to correct the errors (Rozovskaya et al., 2013; Han et al., 2004). Approaches based on Machine Translation are divided into Statistical Machine Translation (SMT) and Neural Machine

Translation (NMT). SMT uses parallel error-annotated data sets and can be used to solve any types of errors (Dahlmeier and Ng, 2012a). An issue with SMT is that it can be dependent of the corpus size, while its efficiency might depend on contextual information (Brockett et al., 2006). NMT, on the other hand, uses an "Encoder-Decoder" mechanism (Yuan and Briscoe, 2016). Specifically, the encoder reads the sentence and encodes it into a vector, while the decoder produces a translation. This is possible because the encoded vector can help predict the next word (Yuan and Briscoe, 2016).

Evaluation of GEC systems

As far as the evaluation of the systems is concerned, there are several evaluation metrics used to assess the performance of the systems. The most used ones are BLEU (Papineni et al., 2002), GLEU (Mutton et al., 2007), and MaxMatch (M2) scorer (Dahlmeier and Ng, 2012b). A recent addition to this list is the ERRANT scorer, which is a modification of the M2 scorer (Bryant et al., 2017). A great leap in GEC was made with the last two shared tasks: CoNLL-14 (Ng et al., 2014) and BEA-19 (Bryant et al., 2019) which offered the option of different tracks regarding the data resources, and which attracted many competing systems. Noteworthy is the fact that, despite the small time gap between the two tasks, the corresponding most popular approaches were different; more specifically, in the CoNLL-14 shared task, there was a greater range of approaches, from rule-based ones and SMT to language models (Ng et al., 2014), while in BEA-19 two-thirds of the total of the participating teams opted for NMT approaches, with the remaining teams using convolutional neural networks or a combination of the two (Bryant et al., 2019).

GEC in under-represented languages

In spite of the great progress in GEC, very little work has focused on low-resource languages. We must note here that we consider the term "low-resource" or "under-represented" task specific. Greek falls into this category since tools and resources for GEC purposes (i.e., error-annotated data) are almost non-existent. A tactic that has been proven efficient generally in low-resource settings is generating synthetic data. Grundkiewicz and Junczys-Dowmunt (2019) used a rule-based approach to insert synthetic errors for English, Russian and German data. A similar approach was used by Náplava and Straka (2019b) who experimented with synthetic data, as well as introducing a new dataset in Czech. A second popular approach addressing low-resource scenarios is deriving data from online sources like Wikipedia (Lichtarge et al., 2019). This approach has the benefit that such data are usually available in multiple languages. Boyd (2018) adopted this approach for the German language with promising results. Other examples of reinforcing under-represented languages in GEC include Rozovskaya and Roth (2019), who introduced an error-tagged corpus of Russian learner writing, as well as tried various state-of-the-art models. For Spanish, Davidson et al. (2020) created a new error-annotated dataset along with a neural-network-based GEC system for Spanish learner writing. Solyman et al. (2019) developed an Arabic GEC model based on multi convolutional layers with an attention mechanism. Finally, a major step was taken for Chinese GEC with the NLPTEA-2018 Shared Task for Chinese Grammatical Error Diagnosis where 13 participating teams developed GEC systems (Rao et al., 2018).

Regarding Modern Greek, any academic research or work on GEC is almost non-existent, at least to the knowledge of the authors of this paper. An exception is the work of Gakis et al. (2016), who did not build a correction system but they created an electronic Greek grammatical checker. According to their findings, their grammatical checker reached almost human accuracy when it came to "pure" grammatical cases. However, accuracy dropped significantly when issues of cohesion, coherence and meaning were involved.

3. Data

The datasets used for the purposes of this paper were the Greek Native Corpus (Korre et al., 2021) and the Greek Learner Corpus (Tantos and Papadopoulou, 2018), dubbed GNC and GLC respectively. The former is a collection of essays (358 sentences) written by Greek high school students, which were digitalized and annotated, providing both corrections and error types following a Greek adaptation of the ERRANT annotation schema (Bryant et al., 2017). Despite the fact that the GNC dataset can be used for GEC purposes, no work in literature has reported respective results. GLC is a compilation of essays written by learners of Greek as a second language (GSL learners). Given that it includes only the error type annotation and not any corrections, it cannot be used for GEC purposes. Therefore, and to be able to use GLC as an evaluation set, and for it to be potentially more versatile for GEC purposes, we decided to proceed with the correction of the sentences by a Greek philologist, extending it to the Greek Learner Corpus Corrections, or GLC2. The two datasets are the first Greek datasets to comprise expert-generated corrections, while GLC is also the first complete dataset with texts authored by GSL learners. This is very important since most current NMT approaches such as Encoder-Decoders need great amounts of data (error-tagged data included) in order to be able to generalize properly (Kiyono et al., 2020). Information about the datasets, as well as an exploratory analysis is presented next.

3.1. GLC2

The Greek Learner Corpus (GLC) was originally compiled by Tantos and Papadopoulou (2018). The motive behind the compilation of GLC had three aspects. First,

Original	Corrected
Μια φορά κι έναν καιρό ήταν τρεια πουλιά και έτσι όπος έφεβγε η μαμά πλησίασε μια γάτα μετά τους κοιτούσε περίεργα	Μια φορά κι έναν καιρό ήταν τρία πουλιά και έτσι όπως έφευγε η μαμά πλησίασε μια γάτα μετά τους κοιτούσε περίεργα.

Table 1: GLC sentence example with corrections. All three mistakes are spelling mistakes. The fluency of the writing is also unsatisfactory.

to present the difficulties when it comes to error annotating second language learner datasets. Second, to provide an overview of the use of an error annotation scheme by using the UAM corpus tool, a tool that provides the environment for annotation of text corpora. Finally, to highlight the importance of learner corpora by adopting stand-off annotation strategies that adhere to the Graph Annotation Framework (GrAF) format of Linguistic Annotation Framework (LAF). The reasons behind using a stand-off approach were the following:

- It allows multiple annotation efforts on the same data but for various error annotation schemes (Tantos and Papadopoulou, 2018).
- It allows inter-annotator agreement to be qualitatively and quantitatively checked in an easier way (Tantos and Papadopoulou, 2018).

Despite these benefits, one great drawback is the lack of corrections. It is understandable that from a linguistic perspective, grammatical error correction cannot be objective (Rozovskaya and Roth, 2021), and therefore, the existence of one correction for an erroneous sentence would be utopian. Yet, not providing corrections renders the development of automatic grammatical correction models almost impossible. In other words, a stand-off annotation can be effective when it comes to automatic error typing but not in aiding grammatical error correction.

The original Greek Learner Corpus (GLC)

According to the authors of the original GLC paper (Tantos and Papadopoulou, 2018), "[t]he GLC is the first learner corpus of Greek assembled from written productions of learners in the first and secondary education levels, which aims at both providing a more user-oriented error annotation scheme and employing standardized means for the development and exploitation of language resources". With regard to the demographic information of the corpus, the texts were produced by 7-to-12-year-old learners of Greek. Approximately 500 texts (around 33,000 words) were annotated by six annotators and demographic metadata are also included for sociolinguistic purposes.

Figure 1 shows that spelling and accent errors are the most frequent error types occurring approx. 6-7 times more frequently than grapheme and punctuation errors, which are the third and fourth most frequent error types, respectively. Tense, agreement and aspect errors occur less than 500 times in GLC.

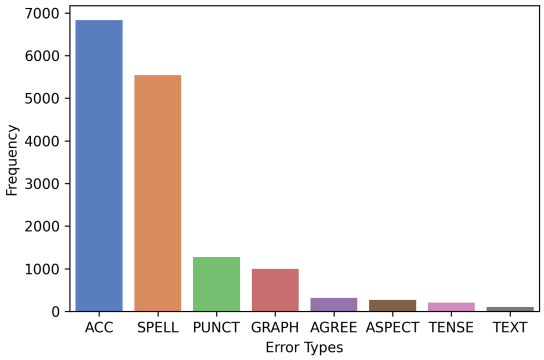

Figure 1: Frequency of the most common error types in GLC. Specifically, error types with frequency over 100 occurrences are demonstrated.

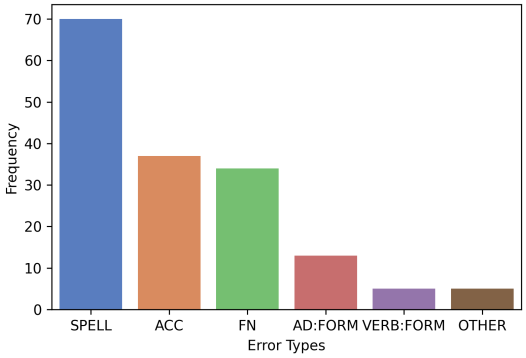

Figure 2: Frequency of the most common error types in GNC. Specifically, error types with frequency over 5 occurrences are demonstrated.

Comparing the two datasets, there are structural differences as two different annotation schemas have been adopted. As mentioned in the previous section, GLC follows a stand-off XML annotation, while GNC is annotated according to the error typing schema of ELERRANT (Korre et al., 2021). In spite of those differences, we can see in Figures 1 and 2 that the two most common error types are spelling and accent errors, indicating that both native speakers and learners find the use of stresses as well as using the correct spelling challenging, regardless of the language competence. What seems to reveal the language proficiency level is the average number of errors per sentence which in GNC is a

barely one error per sentence, while for GLC the number goes as high as up to seven errors per sentence.

Extending GLC with new error corrections
With regard to the correction procedure, we recruited a Greek language philologist to provide the corrections for the 500 GLC texts according to her judgment. Before proceeding to the corrections, the texts were split into 1,524 sentences. Due to the challenging nature of the texts, the annotator was instructed to focus mainly on grammatical errors and not as much on meaning, coherence and cohesion. Punctuation was also not considered a major error, while accent was. Out of the 1,524 sentences, only 159 were correct (approx. 10%). Thirty randomly selected erroneous sentences were annotated by one of the authors, who is also a graduate of English and Greek language and literature, in order to measure inter-annotator agreement. The annotation procedure was the same as the one followed by the first annotator. The percentage agreement was very low, only 29.29%. One possible reason for this low score is the fact that the GLC texts are very challenging for annotators since they are derived from very young GSL learners. Not only must the annotators deal with a great quantity of errors but also with meaning and fluency issues despite the low levels of language proficiency. Therefore, multiple error can mean multiple corrections which lead to low inter-annotator agreement. For example, For example,

- (Original) Με μια φιλι μου ελεγε πολλις πλακες.
- (A) Μία φίλη μου έλεγε πολλές πλάκες.
- (B) Με μία φίλη μου λέγαμε πολλές πλάκες.

This problem becomes even more evident when this score is compared to the GNC inter-annotator agreement kappa score that reached up to approximately 85% (Korre et al., 2021).

4. Empirical Analysis

Despite the fact that two Greek language resources exist in literature (Korre et al., 2021; Tantos and Papadopoulou, 2018), no published work has attempted to use them for benchmarking GEC systems. This is mainly due to two reasons. First, GLC (Tantos and Papadopoulou, 2018) lacks corrections, which is an integral part to train GEC systems. Secondly, GNC (Korre et al., 2021) was introduced only very recently while the authors did not experiment with any GEC system. Our work addresses this gap by extending GLC so that it comprises corrections and by using it, along with GNC, to benchmark GEC.

4.1. Methods

For the purposes of the study we employed a multilingual variant of the Text-to-Text Transfer Transformer (T5), which has achieved state-of-the-art results on various NLP tasks in English (Raffel et al., 2019). Its multilingual MT5 variant (Xue et al., 2020) was pretrained on a dataset covering 101 languages, following the paradigm of the more general solution of pretraining on multiple languages (Liu et al., 2020; Conneau et al., 2019)
Both T5 and MT5 employ an Encoder Decoder Transformer (Vaswani et al., 2017) that is pre-trained with masked language modeling by masking consecutive spans of input tokens and then trying to reconstruct them. T5 was pre-trained on 750GB of English-language text that was sourced from the public Common Crawl web.[1] MT5 was pre-trained on data from all of the 71 monthly web scrapes released so far by Common Crawl, which is more than the source data used by T5. Pages with few characters and ones including bad words were excluded. The pages were then grouped into the 101 automatically detected languages.[2] The most frequent language was English, followed by Russian and Spanish. Greek was on the 20th position, with 43 billion tokens extracted from 42 million pages. Given that Greek is a well supported language in the data that was used to pre-train MT5, this model is suitable candidate to be used for downstream NLP tasks in the Greek language.

4.2. Evaluation

For the evaluation of the model we used the ELERRANT scorer. ELERRANT is the Greek version of ERRANT (Error Annotation Toolkit) developed by Bryant et al. (2017) and apart from providing automatic error type annotation, it offers a scorer which was also used as the main scorer in BEA-19 system evaluation. The scorer works by comparing the edits (changes) in the hypothesis against the edits in each respective reference and measuring the overlap. A true positive includes any edit with the same span and correction in both hypothesis and reference while unmatched edits constitute false positives (FP) and false negatives (FN) respectively (Bryant et al., 2017). For the evaluation of our system, the gold references were created by using the source text and the correction by human annotators as input to ELERRANT to create the gold M2 file, while the hypothesis M2 file was created by inputting the source text and the system output.

4.3. Experimental Results and Analysis

MT5 was fine-tuned for 300 epochs, with a patience of 2 epochs and a max length of 36 tokens, on 277 randomly selected GNC sentences.[3] The trained model was evaluated on 20 GNC sentences and on 200 randomly selected GLC2 sentences. Table 4.3 presents the results of MT5, fine-tuned on GNC and evaluated on GNC and GLC2, along with GEC scores reported in

[1] https://commoncrawl.org/

[2] https://github.com/google/cld3

[3] We employed a train/val/test split of 90/5/5, and discarded from the 322 training sentences ones with length more than 36 tokens (approximately 13% of the training data).

literature. MT5 fine-tuned on GNC achieved an F0.5 of 52.63%, [4] which is the second best performance and only 5 percent units lower than the best performing system for Arabic (Solyman et al., 2019). Although not shown here, this score is only 16% lower than the best performing system for the well represented English language (Bryant et al., 2019).
Preliminary experiments involved training with the GLC2 corrections, using artificial data to augment the current datasets and experimenting with the maximum length of the sentences. However, all of these experiments seemed to worsen the performance.

	P	R	F05
MT5@GNC[MCCV]	45.11	62.47	47.66
MT5@GNC	50.00	66.67	52.63
MT5@GLC2	28.45	12.64	22.76
Davidson et al. (2020)	25.40	15.30	22.40
Boyd (2018)	51.99	29.73	45.22
Náplava and Straka (2019b)	63.26	27.50	50.20
Rozovskaya and Roth (2019)	38.00	7.50	21.00
Solyman et al. (2019)	70.23	72.10	71.14

Table 2: Precision (P), Recall (R) and F0.5 percent scores of MT5 fine-tuned on GNC (Monte Carlo Cross validation denoted with MCCV
and evaluated on GNC and GLC2, along with GEC systems for languages other than Greek. From top to bottom: Spanish (Davidson et al., 2020), German (Boyd, 2018), Czech (Náplava and Straka, 2019b), Russian (Rozovskaya and Roth, 2019), Arabic (Solyman et al., 2019)

The results obtained from the MT5 look very promising yet there is great deviation between the two datasets we evaluated our model on, vis. GNC and GLC2. This is due to the nature of the two datasets. As we mentioned in Section 3, GNC is a corpus with compiled essays by native Greek speakers. Automatically, this suggests that the fluency of the text is very high, as well as there being fewer and more distinct errors per sentence. GLC on the other hand, is a corpus containing texts written by learners of Greek as a second language, who are also of very young age (7-12 years old), suggesting that their level of fluency should be relatively low and that they should be more prone to errors than the native speakers of GNC. However, we observe that the scores on GLC are not the lowest ones, with the models for Russian and Spanish presenting the lowest scores.

4.4. Error Analysis

A manual inspection of the system predictions reveals that MT5 performs well when it comes to accent and spelling errors, but under-performs when the sentences are more complex. When this happens, the model resorts to either reducing the length of the sentence or modifying other parts of the sentence apart from the initial error. In the latter case, the sentence is not necessarily erroneous as demonstrated in the following example, however the meaning of the sentence does change and might create coherence issues:

- Το φαινόμενο αυτό αποτελεί θεμα μεγάλης ανησυχίας στην εποχή μας [This phenomenon is a matter of great concern nowadays].

- Τέλος αυτό αποτελεί θέμα μεγάλης ανησυχίας στην εποχή μας [Finally, this is a matter of great concern nowadays].

4.5. Discussion

Our experimental findings show that MT5, fine-tuned and evaluated on native Greek data, achieves a considerably high performance in GEC, the second highest compared to published results in F0.5 and only 16 percent units below the state of the art in English GEC. This is a promising result, especially under the light of the small size of the training set used in this study (358 sentences; see Section 3).
The performance of our system in GEC in Greek drops when it is evaluated on a learners' dataset. One possible explanation for this result lies in the nature of the learners' dataset and from the fact that we did not train our model on learner data. Compared to the native dataset, the learner dataset contains a higher number of errors per sentence, as shown in Section 3. Consequently, this indicates that there are a lot of factors which must be taken into consideration in GEC experimental setups. For example, the model could be adapted depending on the level of the language of the dataset. Apart from the proficiency level, other factors that could influence the performance of the model, such as demographic information, should be explored further.
The case of GEC in Modern Greek is an idiosyncratic one, given the complexity of the language, which renders the development of GEC systems difficult and which might discourage the NLP community from providing resources and tools. This complexity is also manifested through the low inter-annotator agreement in GLC2, which then comes to show that the evaluation of the model can be problematic considering that there could be more than one ground-truths (see Section 3). This, however, is a phenomenon that we encounter even in high-resourced languages (Bryant and Ng, 2015). The authors' hope is that the promising results presented in this study will encourage the development of more language resources and tools to assist Greek GEC.
Possible limitations of our study are outlined below:

- By contrast to GNC, inter-annotator agreement was very low for the GLC2 annotation. Enriching GLC2 with corrections provided by more annotators could reveal that the system performed higher than reported for this dataset.

[4] The performance drops approx. five units when we use Monte Carlo cross validation with three repetitions.

- Both Greek datasets studied in this work are small in size. This fact constraints the system performance. Synthetic data could be explored to assist with model training.
- We used the BASE version of MT5, due to constrained resources, but better results are expected with larger available models.

5. Conclusion

All in all, in this paper we presented the issue of low-resourced languages in Grammatical Error Correction and we offered two contributions: an enhanced dataset in Greek, GLC2, and an Encoder-Decoder GEC model for the Greek language. The model was trained on the corrected version of GNC (Greek Native Corpus), while it was evaluated both on GNC and on part of GLC2. The results were promising, with 52.63% F0.5 score on GNC, only 16% lower than the best performing model of the BEA-19 shared task. Performance dropped when the model was evaluated on GLC2, reaching only 22.76%. This is due to the nature of the dataset as it contained a high number of errors due to the fact that it is derived from texts written by very young learners of Greek as a second language.
In future work, we would like to further expand GLC2 by adding texts by learners of Greek as a second language of higher levels and test our model on that level, as well as recruiting more annotators for the same task to avoid bias. In addition, expanding and adding synthetic errors to GNC will allow us to render it more versatile for GEC purposes. Finally, since the results of the two datasets are quite different, the development of different tools according to language proficiency would be another pathway to explore.

6. Acknowledgements

This work was supported by the framework of the PON programme FSE REACT-EU, Ref. DOT1303118.We also acknowledge the work of Maria Fasoi from the Athens University of Economics and Business, Greece, who provided us with the GLC corrections.

7. References

Boyd, A. (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.

Brockett, C., Dolan, W. B., and Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 249–256, Sydney, Australia, July. Association for Computational Linguistics.

Bryant, C. and Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? In *ACL*.

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.

Bryant, C., Felice, M., Andersen, Ø. E., and Briscoe, T. (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.

Bustamante, F. R. and León, F. S. (1996). GramCheck: A grammar and style checker. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Dahlmeier, D. and Ng, H. (2012a). A beam-search decoder for grammatical error correction. pages 568–578, 07.

Dahlmeier, D. and Ng, H. T. (2012b). 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.

Dale, R. and Kilgarriff, A. (2011). Helping our own: The HOO 2011 pilot shared task. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 242–249, Nancy, France, September. Association for Computational Linguistics.

Davidson, S., Yamada, A., Fernandez Mira, P., Carando, A., Sanchez Gutierrez, C. H., and Sagae, K. (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.

Gakis, P., Panagiotakopoulos, C., Sgarbas, K., Tsalidis, C., and Verykios, V. (2016). Design and construction of the greek grammar checker. *Digital Scholarship in the Humanities*, 32:fqw025, 07.

Grundkiewicz, R. and Junczys-Dowmunt, M. (2019). Minimally-augmented grammatical error correction. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 357–363, Hong Kong, China, November. Association for Computational Linguistics.

Han, N.-R., Chodorow, M., and Leacock, C. (2004). Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Kiyono, S., Suzuki, J., Mizumoto, T., and Inui, K. (2020). Massive exploration of pseudo data for grammatical error correction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2134–2145.

Korre, K., Chatzipanagiotou, M., and Pavlopoulos, J. (2021). ELERRANT: Automatic grammatical error type classification for Greek. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 708–717, Held Online, September. INCOMA Ltd.

Lichtarge, J., Alberti, C., Kumar, S., Shazeer, N., Parmar, N., and Tong, S. (2019). Corpora generation for grammatical error correction. 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 3291–3301, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Mutton, A., Dras, M., Wan, S., and Dale, R. (2007). GLEU: Automatic evaluation of sentence-level fluency. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 344–351, Prague, Czech Republic, June. Association for Computational Linguistics.

Náplava, J. and Straka, M. (2019a). CUNI system for the building educational applications 2019 shared task: Grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 183–190, Florence, Italy, August. Association for Computational Linguistics.

Náplava, J. and Straka, M. (2019b). Grammatical error correction in low-resource scenarios. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 346–356, Hong Kong, China, November. Association for Computational Linguistics.

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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rao, G., Gong, Q., Zhang, B., and Xun, E. (2018). Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Rozovskaya, A. and Roth, D. (2019). Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17, March.

Rozovskaya, A. and Roth, D. (2021). How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online, April. Association for Computational Linguistics.

Rozovskaya, A., Chang, K.-W., Sammons, M., and Roth, D. (2013). The University of Illinois system in the CoNLL-2013 shared task. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 13–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Solyman, A., Wang, Z., and Tao, Q. (2019). Proposed model for arabic grammar error correction based on convolutional neural network. In *2019 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE)*, pages 1–6.

Tantos, A. and Papadopoulou, D., (2018). *Stand-off annotation in learner corpora: compiling the Greek Learner Corpus (GLC)*, pages 15–40. 10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Yuan, Z. and Briscoe, T. (2016). Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

380–386, San Diego, California, June. Association for Computational Linguistics.

Appendix

Abbreviations for error types

Abbreviation	Meaning
ACC	Accent
SPELL	Spelling
PUNCT	Punctuation
GRAPH	Grapheme
AGREE	Agreement
ASPECT	Aspect
TENSE	Tense
TEXT	Text
FN	Final n (ν)
AD:FORM	Adverb or Adjective Form
VERB:FORM	Verb form
OTHER	Other

Table 3: Abbreviations for error types found in Figures 1 and 2. For a more detailed explanation of the error types see Korre et al. (2021) and Tantos and Papadopoulou (2018)

		Train	Dev	Test
Sentences	GNC	322	18	18
	GLC	-	-	200
Tokens	GNC	8440	451	427
	GLC	-	-	3976

Table 4: Number of tokens and sentences for train/dev/test