ELERRANT: Automatic Grammatical Error Type Classification for Greek

Katerina Korre[†], **Marita Chatzipanagiotou**[†], **John Pavlopoulos**[†♣]
[†] Athens University of Economics and Business, Greece
[♣] Stockholm University, Sweden
katkorre95,marita.xatzh,annis@aueb.gr

Abstract

In this paper, we introduce the Greek version of the automatic annotation tool ERRANT (Bryant et al., 2017), which we named ELERRANT. ERRANT functions as a rule-based error type classifier and was used as the main evaluation tool of the systems participating in the BEA-2019 (Bryant et al., 2019) shared task. Here, we discuss grammatical and morphological differences between English and Greek and how these differences affected the development of ELERRANT. We also introduce the first Greek Native Corpus (GNC) and the Greek WikiEdits Corpus (GWE), two new evaluation datasets with errors from native Greek learners and Wikipedia Talk Pages edits respectively. These two datasets are used for the evaluation of ELERRANT. This paper is a sole fragment of a bigger picture which illustrates the attempt to solve the problem of low-resource languages in NLP, in our case Greek.

1 Introduction

Grammatical Error Correction (GEC) is the task of automatically correcting language mistakes in written texts. These mistakes can vary from grammatical mistakes to punctuation, spelling and morphology of a word. The development of a GEC system usually involves the transformation of an erroneous sentence into its correct version, while also keeping the initial meaning intact. Developing those systems requires error annotated data, which can be either learner data or artificial. High-resource languages, such as English, present a variety of learner data that cover a relatively wide spectrum of language proficiency levels, native language and topics. Some notable examples are the Cambridge English Write & Improve corpus (Yannakoudakis et al., 2018), the LOCNESS corpus (Granger, 1998), and the NUCLE corpus (Dahlmeier et al., 2013). Low-resource languages, on the other hand, are characterized by a scarcity of such corpora, as well as of other GEC resources and Natural Language Processing (NLP) tools. That is also the case of Greek.

Although, Greek is only spoken by approx. 13.5 million people (native),[1] the fact that Greece has a high immigrant population underlines the need for learning Greek as a Second Language (GSL).[2] Therefore, and as technology is being integrated in education (Meurers, 2012; Forcier, 2016), the need for more GEC and NLP tools for Greek becomes evident.

In this paper, we present ELERRANT, an automatic annotation tool, which is based on ERRANT (Bryant et al., 2017). ERRANT produces an annotation mainly consisting of the error location, the error type and the correction of the error, by using an original erroneous sentence along with its correction as input. ERRANT is the first toolkit that not only annotates texts but also provides automatic error typing, offering detailed feedback to Second Language (L2) learners and useful information for language analysis (Bryant et al., 2017). Most importantly, the annotator's workload is relieved and all learner corpora, regardless of size, level and other factors can be annotated in a standardized manner. We believe that its easy application and versatility can encourage the generation of more error annotated datasets in the Greek language, thus tackling the scarcity of resources.

For the evaluation of ELERRANT we developed two datasets: the Greek Native Corpus (GNC) and the Greek Wiki Edits (GWE). GNC comprises native Greek student essays, while GWE comprises sentences extracted from WikiConv (Hua et al.,

[1] https://en.wikipedia.org/wiki/Greek_language
[2] https://eacea.ec.europa.eu/national-policies/eurydice/content/population-demographic-situation-languages-and-religions-33_en

Proceedings of Recent Advances in Natural Language Processing, pages 708–717
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_081

2018). By evaluating ELERRANT on the latter corpus, we also show that the tool has the potential to detect edits and alterations on Wikipedia Talk Pages that are due to grammatical error correction, which can then be automatically white-flagged from being moderated for misinformation (e.g., for words introducing bias). It also paves the way for further analysis of such edits. Both datasets are shared for public use.[3]

The rest of the paper is structured as follows: First, we discuss related work on Error Annotated Data and the original ERRANT. Then, we describe the development of ELERRANT. Section 4 introduces the two new datasets, demonstrates our method of evaluation and presents the findings. Section 5 is concerned with the use and implications of ELERRANT. Finally, we conclude by discussing limitations and future work.

2 Related Work

Error Annotated Data - What for? Error annotated data can be useful in multiple domains ranging from real-life teaching and educational research to NLP tasks, especially in GEC. More specifically, recent advances in GEC often require large amounts of annotated data both for development and evaluation of any given systems. Mita et al. (2019) underline the need for cross-corpora evaluation when it comes to GEC systems, given that the task difficulty depends on factors such as proficiency level and essay topic. Consequently, there is a demand for standardized error-annotated corpora, while also reducing the annotator's workload (Bryant et al., 2017). In addition, error annotated corpora can play a major role in error analysis, which has slowly started to step into CALL (Computer Assisted Language Learning). Until very recently, the staple technique to NLG (Natural Language Generation) for language learning purposes, was to train models on large bodies of correct English (Lee and Seneff, 2008). Although this technique has proven to be effective, a more recent one seems to take into account more parameters when it comes to non-native speakers. This new technique involves relying on two kinds of corpora: a source corpus from non-native texts, and a target corpus, which, in reality, is a corrected version of the source corpus. Meurers (2012) summarizes the benefits on the analysis of learner corpora claiming that the annotation of learner corpora can point out learner language properties thus supporting the aim of improving our understanding of Second Language Acquisition (SLA) and developing instructional methods and materials for SLA purposes.

ERRANT Bryant et al. (2017) attempt to solve the issue of corpora standardization by presenting ERRANT, an automatic annotation tool which serves as both annotator and system-output scorer. ERRANT only needs an erroneous sentence along with its correction to produce an annotation essentially consisting of the location of the errors, the error type, and the correction. ERRANT has paved the way for a new annotation framework, and has therefore been used in the most recent shared task, the BEA-2019 (Bryant et al., 2019), both for annotating the datasets used for the task and for evaluating the system output of the participants per error. The convenience and versatility of ERRANT has led to the adaptation of the tool in more languages, such as German (Boyd, 2018), Spanish (Davidson et al., 2020), Czech (Náplava and Straka, 2019), and Romanian (Cotet et al., 2020). In this paper, we present our Greek version.

Task difficulty Inter-annotator agreement, although a staple in computational linguistics procedures when it comes to evaluation, has been quite controversial regarding GEC. Traditionally, corpora for GEC purposes would be annotated by solely one native annotator providing one gold standard annotation, a practice which automatically renders the research highly biased and even uninformative (Bryant and Ng, 2015; Tetreault and Chodorow, 2008). The obvious solution to the problem would be to recruit multiple annotators and estimate the degree to which they agree on the correction of an error. Yet, this method also proves insufficient, since annotators usually agree up to 70%, a percentage inadequate for system evaluation. The same holds for intra-annotator agreement, where the same annotator does not always agree with themselves ($\varkappa$ scores of about 60%) (Bryant and Ng, 2015).

3 ELERRANT

To adapt ERRANT in the Greek language, we used the original ERRANT classifier as blueprint. Our version uses the Greek Hunspell spellchecker dictionary[4] to detect spelling errors and the Greek

[3]The datasets are shared with CCO licence on: https://github.com/katkorre/elerrant

[4]https://sourceforge.net/projects/grspell/files/hunspell-gr

Error Type	Meaning	Description	Example
AD:FORM	Adverb Form	Errors concerning the form an adverb.	καλός → καλώς
ADJ:FORM*	Adjective Form	Errors concerning the form of an adjective	καλός → καλύτερος
NOUN:FORM	Noun Form	Errors concerning the number,the case or the suffix of a noun.	του νους → του νου
PRON:FORM	Pronoun Form	Errors concerning the number, the case or the suffix of a pronoun.	κάποια → κάποιας
VERB:FORM	Verb Form	Errors concerning the disposition, the voice, the inflection, the tense,the number or the person of a verb.	(εσείς) πηγαίνεται → (εσείς) πηγαίνετε
CONJ	Conjunction	Errors concerning conjunctions.	και → αλλά
PREP	Preposition	Errors concerning prepositions.	από → σε
DET*	Determiner	Errors concerning articles or determiners.	το → του τον → έναν
SPELL	Spelling	Spelling errors.	ευχέρια → ευχέρεια
FN	Final -ν/nu	Final -ν/nu addition or removal.	την → τη / μη → μην
PUNCT	Punctuation	Errors concerning the punctuation.	. → ;
OTHER	Other Errors	An error that does not fit into any other category but can still be corrected.	καμία → για κανένα
ACC	Accentuation	Accentuation addition or removal.	καθηκοντα → καθήκοντα
UNK	Unknown error type	An error that can be detected but not corrected.	usually long error spans
WO	Words Order	Error in words order.	όταν φεύγω έρθεις → όταν έρθεις φεύγω
ORTH*	Orthography	Spacing Errors	γιασένα → για σένα
PART:FORM	Participle Form	Errors concerning the number,the case or the person of a participle.	(πήγε) τρεχόμενος → (πήγε) τρέχοντας
VERB:SVA	Subject Verb Agreement	The subject and the verb to be in person agreement.	(εγώ θα) φύγει → (εγώ θα) φύγω

Table 1: ELERRANT and human error type annotation guide. The error types with the asterisk (*) do not exist in the human annotation scheme while the two last error types in the table do not exist in the ELERRANT annotation scheme

SpaCy[5] as the main POS tagger. Due to morphological differences between the two languages (English and Greek), we removed some error categories that exist in the original ERRANT, while adding some new ones. Due to the fact that Greek is a highly inflectional language and most POS have some sort of inflection, we "merged" some error types in order to include as much information about the error as possible. This decision can be regarded as a compromise, since many errors might have more than one overlapping error types (e.g., των γάτα → των γατιών, wrong case and number), therefore by merging the sub-types into the FORM type we preserve the ambiguity and multifacedness of the error.

The main alterations are the following: We added the error type AD:FORM (Adverb Form), to convey errors that mainly concern the comparative and superlative degree of the adverb. The CONTR (Contraction) category has been removed temporarily due to the fact that contractions in Greek can happen in any word starting or ending with a vowel under certain conditions. We are currently developing a dictionary that assembles the most frequent cases of contractions in Greek and we plan to integrate it in future versions. NOUN:INFL (Noun Inflection), NOUN:POSS (Noun Possessive) and NOUN:NUM (Noun Number) are all captured in NOUN:FORM (Noun Form). PART (particles) were also dropped as in the Greek language they have a different function, mainly in tense construction. PRON:FORM (Pronoun Form) was also added, since pronouns are also inflectional. From the verb categories, only VERB, VERB:FORM, and VERB:TENSE have been preserved.

Two new categories We added two more categories from scratch: ACC (Accent) and FN (Final -ν/nu). The accent in Greek is signified with a stress mark (΄) rather than just the intonation of the word when speaking. The maintenance or omission of the final nu in some Greek words (articles, pronouns or particles), due to its frequency as an error even by native speakers, is considered a different error type and not just a spelling error. For the two aforementioned error types, we made two

[5]https://spacy.io/models/el

new corresponding functions and added them in the ELERRANT classifier (see Algorithms 1 and 2). Table 1 demonstrates the final error categories in ELERRANT.

Algorithm 1: Accent error detection

```
Data: chars^orig, chars^corr
Result: label ∈ {R:ACC, M:ACC, U:ACC}
accents = [ά,έ,ή,ί,ό,ύ,ώ];
accents^orig = chars^orig ∩ accents;
accents^corr = chars^corr ∩ accents;
if accents^orig ≠ {} then
    // Only the original word has an accent
    if accents^corrr = {} then
        return U:ACC;
else
    // Only the correction has accent
    if accents^corr ≠ {} then
        return M:ACC;
    else
        // Both words have accents, so compare the
        // number of accents between them
        if len(accents^orig) > 1 then
            if len(accents^corr) = 1 then
                // Redundant accent in the original
                return U:ACC;
        else if len(accents^orig) = 1 then
            if len(accents^corr) > 1 then
                // Missing accent in the original
                return M:ACC;
        else if accents^orig ≠ accents^corr then
            // Same number of accents, yet different
            return R:ACC;
```

Algorithm 2: Final -ν error detection

```
Data: chars^orig, chars^corr
Result: label ∈ {M:FN, U:FN}
// Original token: the corrected + ν
if chars^orig = chars^corr[: −1] then
    if chars^corr[−1] = "ν" then
        // The original is missing the final ν
        return M:FN;
else if chars^corr = chars^orig[: −1] then
    if chars^orig[−1] = "ν" then
        // The other way, unnecessary final ν
        return U:FN;
```

4 Empirical Evaluation

We consider this work an opportunity to introduce two novel datasets: **Greek WikiEdits (GWE)**, and the **Greek Native Corpus (GNC)**, which we use as gold standards in our ELERRANT evaluation. This section first describes our two datasets, their annotation and development process, and presents some statistics and inter-annotator agreement results. Then, the evaluation is discussed and the experimental results are reported, evaluating ELERRANT on both datasets.

4.1 Datasets

GWE The first corpus we evaluated ELERRANT on is based on WikiConv (Hua et al., 2018). WikiConv is a multilingual corpus that encompasses the history of conversations on Wikipedia Talk Pages, including comment deletion, modification and restoration. The authors of the respective article kindly provided us with the Greek part of the corpus, which comprises 194,499 Talk Pages. We processed the provided pages so that only sentences with edits remained. Despite the fact that edits do not necessarily regard grammatical errors (e.g., they could be about a corrected date), their employment by Grammatical Error Correction models leads to improvements (Lichtarge et al., 2019). To the best of the authors' knowledge, this is the first dataset with edits of Wikipedia Talk Pages comprising human annotations for grammatical errors. Henceforth, we will refer to this Greek WikiEdits dataset as GWE.

GNC The existing publicly available corpora compilations, that attempt to contribute to the scarcity of Greek resources for NLP, are the Greek Learner Corpus (GLC) (Tantos and Papadopoulou, 2018), and the Electronic Learner Corpus of L2 Greek (Tzimokas, 2010). Both datasets consist of data generated by learners of Greek as a Second Language (GSL). Despite their usefulness, we observe that none of these datasets includes corrections in their annotations. This lack of corrections was intentional, in order to reflect the "error ambiguity", rather than choosing between several corrections, given that an error can be corrected in multiple ways (Tantos and Papadopoulou, 2018). This lack of corrections, however, also means that ELERRANT is effectively inapplicable on them. This gap motivated us to develop another corpus to evaluate ELERRANT on, the first Native Greek Corpus (GNC), designed in the aim of being compatible to ELERRANT and of use to automatic grammatical error correction systems.

The compilation of GNC is currently in progress, but at the time of writing this paper, 227 sentences have been collected and annotated. The GNC comprises essays written by High School students,

whose native language is Greek. The hand-written essays were split into sentences and manually digitalized (no OCR was used). Each sentence may contain none, one or more grammatical errors.

4.2 Annotation Schema

The GWE corpus comprised edits which we considered the "corrections", in order to be able to apply ELERRANT. GNC, however, does not comprise any suggested corrections and thus, ELERRANT is not applicable. Hence, the suggested corrections have to be provided by the annotators. Due to this significant difference between the two datasets, the annotation process differs in the two cases, and in particular concerning the GNC it becomes slightly more complicated. Annotation for both datasets was based on the rule-based error type framework of the original ERRANT. Two main code categories were created. The first category ('Error Description'), consists of the three prefix operation codes [**U**(nnecessary) / **R**(eplacement) / **M**(issing)] which indicate what needs to happen to each erroneous item in order to be corrected, i.e., whether it should be removed from the sentence, whether it should be modified or replaced, or whether an item is missing. The second code category ('Error Type') comprises 16 codes(created based on the 25 codes presented in (Bryant et al., 2017) (see Table 1), which form a simplified classification system of the errors that may occur in the written Greek language. These codes indicate what kind of error we encounter in each sentence and (in most cases) what part of speech the erroneous element that needs to be removed, modified or added is.

GWE Schema The annotation on the GWE was carried out in two steps. First, the two parallel sentences (original and changed) were input in ELERRANT, which helped us extract the tokens of the edits. Then, the annotators classified the edits according to the 'Error Description' and 'Error Type' fields from the annotation schema described above.

GNC Schema The annotation process for the GNC was based on a schema containing four annotation fields. The first field is intended for two mutually - exclusive values [**c**(orrect)/**e**(rroneous)] in order to mark the absence or presence of an error in the sentence. The second is for the annotator to correct the existing error and rewrite the whole sentence providing the corrected string. The two remaining fields are respectively for the 'Error Description' and 'Error Type' codes that were used also at the annotation of GWE.

As can also be seen in Table 2, when a sentence contains multiple errors, such as in (a) below, then copies of that sentence are inserted in the corpus (i.e., the second and third), leading to as many records with the same sentence as the errors it comprises. In each entry, the annotator should maintain and correct only one error, while the other errors must be recorded in advance (both in the field of the original and the corrected text) as correct. Respectively, the same should happen if two or more errors are detected in a single word, such as the example of sentence (b) below:

a **Erroneous sentence containing more than one errors**

Αρχικά, από την μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυνχάνοντας έτσι την προστασία τους και προκαλόντας μάλλον αντίθετα αποτελέσματα.

"Initially, it would be good for adults to understand that overprotectiveness inhibits the responsibility of young people and usually provokes strong reactions, thus failing to protect them and causing the opposite effect."

b **Erroneous word containing more than one errors**

(Εμείς) επιχηρούμαι.

"We attempt."

Annotation process We recruited two Greek philology graduates and provided them with 327 sentences, 227 from GNC and the remaining from GWE. In GWE, where each page comprised a single edit, we performed sentence segmentation (based on full stop) and only considered the sentence comprising the edit. We asked the annotators to follow the annotation schemas (see Sec. 4.2), in order to classify (and detect and correct in GNC) all possible grammatical errors.

4.3 Corpus Statistics

Table 3 presents the statistics of the two annotated corpora, by considering the detected errors of the two annotators on (micro) average. Sentences were longer in GWE, compared to GNC. More errors were detected on (micro) average in GNC, which is explained by the fact that each GWE sentence contains almost always exactly one edit. By contrast,

Label	Original Text	Corrected Text	Error Description	Error Type
e	Αρχικά, από την μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυγχάνοντας έτσι την προστασία τους και προκαλώντας μάλλον αντίθετα αποτελέσματα.	Αρχικά, από τη μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυγχάνοντας έτσι την προστασία τους και προκαλώντας μάλλον αντίθετα αποτελέσματα.	R	FN
e	Αρχικά, από τη μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυνχάνοντας έτσι την προστασία τους και προκαλώντας μάλλον αντίθετα αποτελέσματα.	Αρχικά, από τη μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυγχάνοντας έτσι την προστασία τους και προκαλώντας μάλλον αντίθετα αποτελέσματα.	R	SPELL
e	Αρχικά, από τη μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυγχάνοντας έτσι την προστασία τους και προκαλόντας μάλλον αντίθετα αποτελέσματα.	Αρχικά, από τη μεριά των μεγάλων καλό θα ήταν να κατανοήσουν πως ο υπερπροστατευτισμός αναστέλλει την υπευθυνότητα των νέων και προκαλεί συνήθως έντονες αντιδράσεις, αποτυγχάνοντας έτσι την προστασία τους και προκαλώντας μάλλον αντίθετα αποτελέσματα.	R	PART:FORM

Table 2: Annotation sample: three entries of the same sentence annotating each time a different error

	GWE	GNC
ANNOTATED SENTENCES (#)	100	227
ERROR ANNOTATIONS (#)	100	180
TOKENS PER ANNOT. SENTENCE (#)	46.43	21.2
COHEN'S KAPPA (%)	70.02	84.65

Table 3: Overview of GWE (Talk Pages) and the GNC (essays). The respective count is shown per row.

GNC may contain none, one or more errors per sentence. Concerning GNC, 102 (44,9%) and 96 (42,3%) (Annotator I and Annotator II respectively) of all sentences have zero errors, 87 (38,3%) and 93 (40,9%) have exactly one error, 27 (11,9%) and 28 (12,3%) have exactly two errors, and 11 (4,8%) and 10 (4,4%) of all sentences have three or more errors.

Inter-Annotator Agreement In GWE, Cohen's Kappa for the human-annotated edits was 70.02%. Out of the sixteen available codes, only 10 were used by Annotator I and 13 by Annotator II., The Pearson's correlation between the error type frequencies of the two annotators was 99.62%, which indicates that the two frequency distributions coincide to a large extent. Given that the annotators had to annotate the edits with annotation guidelines fitting for error types, they opted for more abstract categories such as OTHER, which was the most frequently annotated type. In GNC, Cohen's Kappa between the two annotators regarding the error type was even higher than GWE, reaching 84.65%. Out of the sixteen available codes, in the annotated texts we encountered the fourteen (UNK and WO tags were not used at all). The distribution of the frequency of occurrence of these fourteen (14) error types also coincides to a large extent between the two annotators, which is reflected in the high error type frequency correlation between the two annotators (99.20%). Any disagreement is mainly due to additional errors and not to the incorrect or different rendering of the codes.

In other words, inter-annotator agreement in GWE falls within the threshold of 70% (see Section 2), while in the GNC this threshold is exceeded. In such cases, where agreement is not optimum but not extremely disheartening, tools such as ERRANT can be used as moderating tools by pinpointing any great discrepancies, as a third annotator would.

4.4 Experimental Results

Since GNC was annotated by two annotators (and therefore there might be differences in the corrections), we ran ELERRANT on both annotators' corrections separately and compared the output error types against the error types provided by each annotator, which served as our gold standards. For GWE, first, the texts were input into ELERRANT and then the edits were assigned an error type by the annotators. Then, we calculated the accuracy, precision, recall and F1-score separately for each comparison. The three latter metrics were calculated with macro, micro and weighted averages to also be able to see whether ELERRANT performs better at more or less frequent error types. We also calculated and compared the frequencies of the human-annotated error types and those generated by ELERRANT. The process was the same for GNC and GWE.

It must also be noted that to evaluate the performance of ELERRANT against the gold standards of

both datasets, we had to mitigate the differences between the ELERRANT annotation schema and the human annotator schema (see Table 1). The necessary modifications included changing all VERB:* error types to VERB:FORM, ARTORDET to DET, PART:FORM to ADJ:FORM, and temporarily change ADJ:FORM of ELERRANT to AD:FORM.

Evaluation on WikiEdits (GWE) For the evaluation of ELERRANT on GWE, as already discussed, we considered the annotations of each annotator as our ground truth. The results are demonstrated on the right of Table 4. All metrics are low with the accuracy reaching 31% and 27%, precision never exceeding 44.47%, recall 40.03% and F1-score 35.75%. This low performance is explained in part by the inability of ELERRANT to detect specific types (see Fig. 1). Additionally, as mentioned in Section 4.1, the edits in Greek Wikipedia are not necessarily grammatical errors. For instance, edits can be 'vandalisms', attempting to alter the content and context of the respective Wikipedia article.

Evaluation on Learner Data (GNC) GNC results are presented on the left of Table 4. Scores are considerably high for micro and weighted averaging. Error type classification is a multi-class problem and each instance of error type might be encountered in different frequencies. This is also apparent in Figure 2, where we can see that spelling, accent, and final nu errors have the highest frequencies both in terms of ELERRANT annotation and according to human annotators. In both annotator cases, micro scores are higher than macro scores indicating that ELERRANT performs better when it comes to classifying an error type that occurs frequently, while it tends to misclassify less frequent error types. The accuracy scores, 83.91% and 77.30%, by Annotator I and II, respectively, show that ELERRANT can correctly classify the error type approx. eight out of ten times, assigning the most appropriate error type possible, if we consider the annotation as the gold standard, and thus the best possible error type classification. However, and as we can see from the results, there are discrepancies between the two annotators, hence the different accuracy scores (see Section 4.3).

Taking into consideration that ground truth in GEC annotation cannot be as established as in other NLP tasks, due to the fact that one erroneous sentence can have multiple corrections (Napoles et al., 2015; Bryant and Ng, 2015), we also looked at the ELERRANT output manually to pin down where it is lacking exactly. We noticed two major issues: First, Greek SpaCy (which is the core of ELERRANT) does not perform as well as in the English version, assigning wrong POS tags or dependencies. Secondly, there are cases in Greek where a single word can contain two or more errors at the same time, and cannot be solved with the 'FORM' category. These cases are usually a combination of accent and spelling mistakes, spelling mistakes and morphology, accent and morphology, etc. In this case, ELERRANT assigns only one error type disregarding the rest. We hope to solve these issues in future versions.

5 Discussion

Our experimental results showed that ELERRANT performs better when it comes to actual error type classification (GNC) than edit classification (GWE). In terms of scores, the ELERRANT classifier works adequately (83% accuracy), especially when compared to the evaluation of the original ERRANT, for which a manual evaluation rated 86% of the output error types as "GOOD". There is still, however, room for improvement. As mentioned in Section 4, the most important issue is Greek SpaCy which also hinders the development of a more detailed ELERRANT, i.e., with more error types such as Noun-Gender Agreement, Case, etc.

As far as the GWE is concerned, a detection and edit classification by ELERRANT is possible, provided that both the edit and the proper form of the text exist. Moreover, the addition of further edit categories is necessary because the current version of ELERRANT is based on error type classification and not edit classification. Categories such as SYNONYM and NAME might give a better insight into how the edits affect the text.

Table 5 illustrates the potential, as well as the problem of applying ELERRANT to GWE. The first edit is a name replacement, which does not entail a grammatical error, yet it does affect the meaning of the sentence. ELERRANT incorrectly classified the edit as a spelling error, while the two annotators placed it in the more general category OTHER. If both ELERRANT and the human-annotator scheme provided a category such as R:NAME, the performance of the system would have been better and the human annotation would have been more accurate leading to a more accurate and descriptive ground truth. When the edit is a grammatical error ELER-

	GNC						GWE					
	ANNOTATOR I			ANNOTATOR II			ANNOTATOR I			ANNOTATOR II		
	Macro	Micro	Weighted	Macro	Micro	Weighted	Macro	Micro	Weighted	Macro	Micro	Weighted
Precision	58.15	86.90	81.10	50.57	80.79	73.13	38.79	40.26	44.47	19.76	36.00	39.07
Recall	60.92	83.91	83.91	50.89	77.30	77.30	40.03	31.00	31.00	24.69	27.00	27.00
F1	58.54	85.38	82.17	49.59	79.01	74.74	35.75	35.03	33.65	19.54	30.86	28.96
Accuracy	-	83.91	-	-	77.30	-	-	31.00	-	-	27.00	-

Table 4: ELERRANT evaluation using Precision, Recall, F1, Accuracy in classifying GWE and GNC error types.

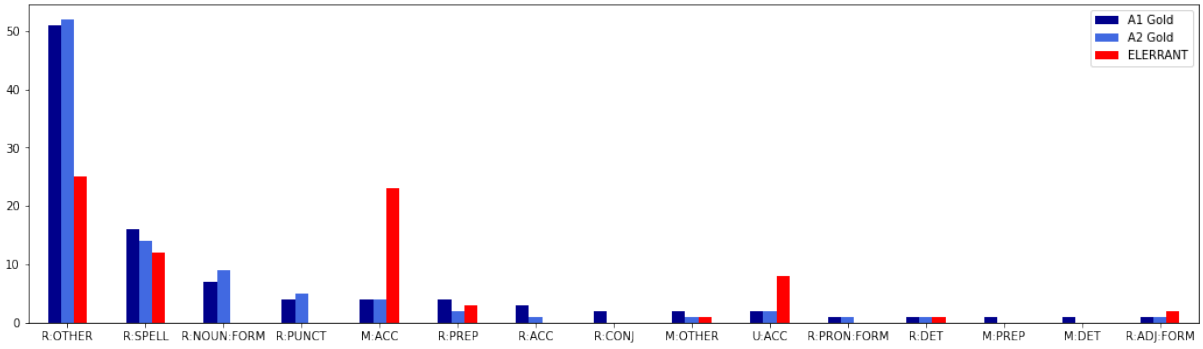

Figure 1: Frequencies of error types on GWE inferred by ELERRANT (A1 ELERRANT, A2 ELERRANT) against those of the two annotators (A1 Gold, A2 Gold).

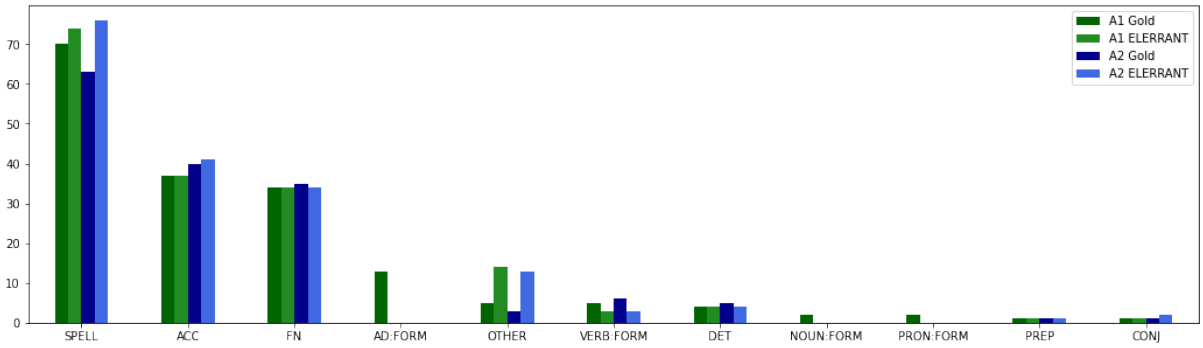

Figure 2: Frequencies of error types on GNC inferred by ELERRANT (A1 ELERRANT, A2 ELERRANT) against those of the two annotators (A1 Gold, A2 Gold).

Original/Edit	ELERRANT	Annotator I	Annotator II
Σομαλίας/Γροιλανδίας Somalia / Greenland	R:SPELL	R:OTHER	R:OTHER
η/ή the (definite feminine article)/ or	U:ACC	U:ACC	U:ACC
ημερήσιο/ημερήσιπ daily	R:SPELL	R:SPELL	R:NOUN:FORM

Table 5: Example annotated sentences from GWE.

RANT performs better, as we can see in the second example. Finally, there is the case where ELERRANT is right and the human-annotator is wrong, which also indicates that an automatic annotation tool such as ERRANT and ELERRANT can provide more accurate, less biased annotation.

6 Conclusion

This paper presented ELERRANT, the Greek version of the automatic grammatical error type annotation tool ERRANT. With this work we also introduced two new datasets: the GNC and GWE, which can be used for GEC and edit classification purposes. Both our datasets are released for public use. In GNC, our findings showed that ELERRANT achieves an accuracy of 77.30 %- 83.91%, confirming that it can be an effective tool, reducing the scarcity problem of low-resource languages, such as Greek. In GWE, the overall performance was much lower, mainly because Wikipedia edits are not necessarily due to GEC. However, despite this low performance, we observe that ELERRANT can still be helpful to Wikipedia moderators who can use it to shortlist edits that are likely due to GEC and thus white-list them. Furthermore, ELERRANT could be updated to capture more error types that are common in GWE, which we will consider in future work, along with the expansion of our datasets.

7 Ethical Considerations

All texts for the compilation of the GNC dataset were obtained with the consent of the original authors. In case of underage authors, adult parents or guardians gave their consent. Authors were thoroughly informed about the purpose of the study, and became completely aware that the produced texts would be anonymously published.

Acknowledgments

We would like to thank Jeffrey Sorensen (Jigsaw) for contributing to this research by sharing with us the Greek part of the WikiConv corpus, which led to the creation of GWE. We would also like to thank Eleutheria Stroumbouli and Maria Fasoi (Athens University of Economics and Business) for correcting and annotating the GNC and GWE datasets.

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. 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. 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. Association for Computational Linguistics.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

Teodor-Mihai Cotet, Stefan Ruseti, and Mihai Dascalu. 2020. Neural grammatical error correction for romanian. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 625–631.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. 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. European Language Resources Association.

Laurie Forcier. 2016. Intelligence unleashed: An argument for AI in education.

Sylviane. Granger. 1998. The computerized learner corpus: a versatile new source of data for SLA research.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823.

John Lee and Stephanie Seneff. 2008. An analysis of grammatical errors in non-native speech in english. In *2008 IEEE Spoken Language Technology Workshop*, pages 89–92.

Jared Lichtarge, Chris Alberti, Shankar Kumar, Noam Shazeer, Niki Parmar, and Simon Tong. 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.

Detmar Meurers. 2012. *Natural Language Processing and Language Learning*. American Cancer Society.

Masato Mita, Tomoya Mizumoto, Masahiro Kaneko, Ryo Nagata, and Kentaro Inui. 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. Association for Computational Linguistics.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 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. Association for Computational Linguistics.

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

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

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 865–872, Manchester, UK. Coling 2008 Organizing Committee.

Dimitrios Tzimokas. 2010. Ηλετρονικό σώμα κειμένων (ΗΣΚ) εκμάθησης της νέας ελληνικής ως δεύτερης ξένης γλώσσας προς ένα ερευνητικό και διδακτικό εργαλείο [electronic learner corpus of l2 greek: Towards a research and teaching tool]. In *Proceedings of 30th Annual Meeting of the Department of Linguistics.*, pages 602–616, Thessaloniki.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31(3):251–267.