Enriching Grammatical Error Correction Resources for Modern Greek

LREC 2022

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

Introduction

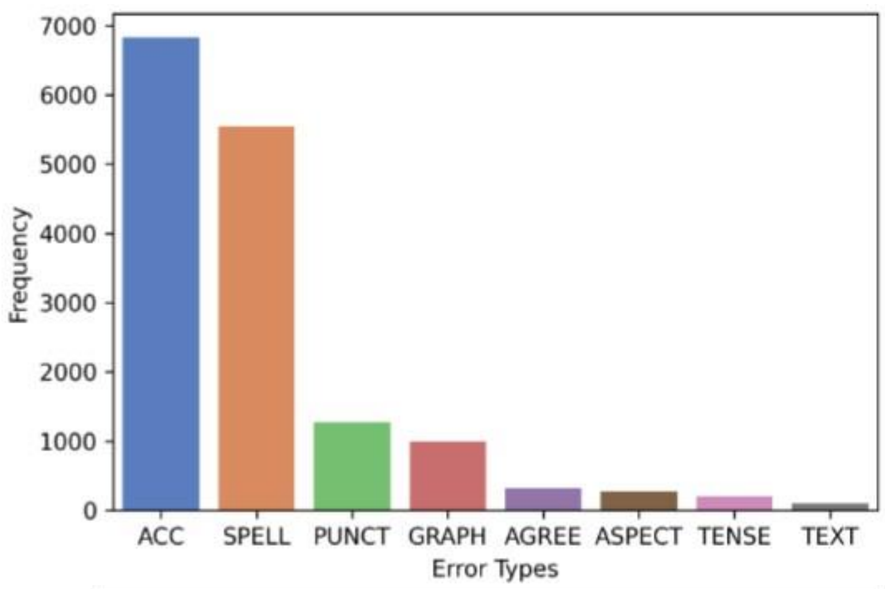
- Issue: **scarcity of resources** for GEC purposes in Greek.
- Approach: **MT5 multilingual text-to-text transforme**r (Xue et al., 2020).
- Evaluation with the **Greek version of the Error Annotation Toolkit scorer (ERRANT)** (Bryant et al., 2017).

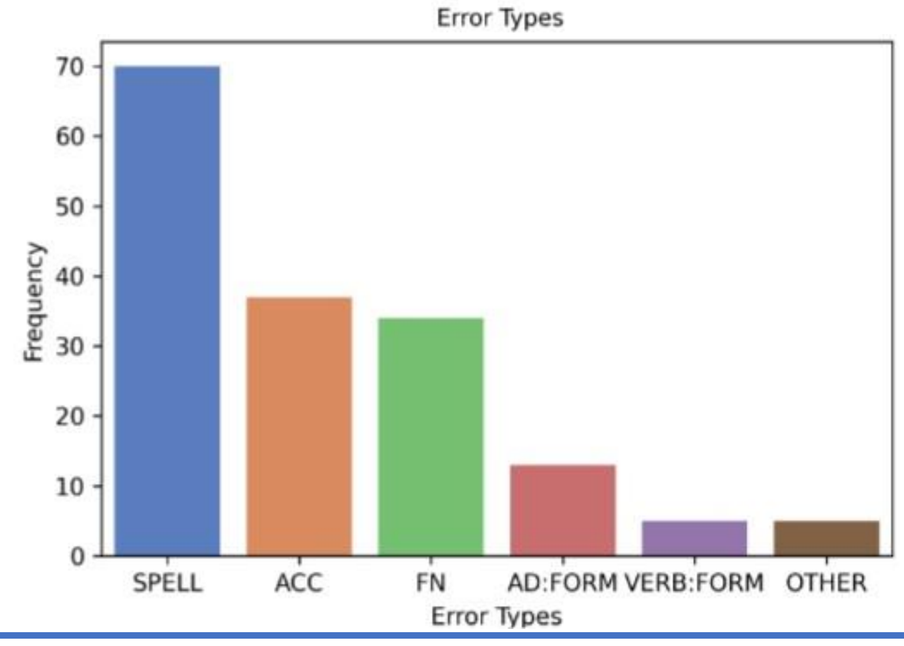
Data

- **Greek Native Corpus** (Korre et al., 2021)
- **Greek Learner Corpus** (Tantos and Papadopoulou, 2018) → **GLC2**

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.
Translation	**Transliteration**
Once upon a time, there were three birds and while the mum was leaving, a cat approached and was looking at them weirdly.	Mia fora ki enan kairo itan tria poulia kai etsi opos efevge i mama plisiase mia gata meta tous koitouse perierga.

Error type statistics for GLC and GNC





Inter-annotator agreement

Sample annotation of 30 sentences to calculate inter-annotator agreement → 29.29%

Potential reasons:

- GLC challenging: Great quantity of errors & fluency issues.
- Multiple errors → multiple corrections

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

Method

MT5 employs 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.

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

Results

	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

Error Analysis

- MT5 performs well: **accent and spelling** errors. under-performs when the sentences are more complex
- **Reduces the length** or **modifies** other **parts** of the sentence. E.g., Το φαινόμενο αυτό αποτελεί θέμα 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].

Conclusions

- **Second highest** compared to published results in F0.5 and only 16 percent units below the state of the art in English GEC.
- The performance **drops** when it is evaluated on a learners' dataset, possibly due to high error frequency.

Limitations and future work

- Inter-annotator agreement was very low for the GLC2 annotation.
- Small dataset sizes → Synthetic data.
- BASE version of MT5, due to constrained resources → better results with larger available models.

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.