Using Predictive Text for Grammatical Error Correction in Second Language Learning

Katerina Korre
Athens University of Economics and Business
MSc in Digital Methods for the Humanities

Contents of this presentation

1

Introduction

What is Grammatical Error Correction and how can it contribute to Second Language Learning?

What is Grammatical Error Correction?

- Grammatical Error Correction (GEC) is the task of automatically correcting different types of errors, such as spelling, punctuation, and grammatical errors, in written texts.
- It requires a system to use the erroneous sentence and transform it into the correct version of it.
- Such systems can also assist second language (L2) learners to improve their writing skills in their target language.

How can GEC assist L2 learning and teaching?

- Opportunity to create “tailored” applications suited to the learner and educator’s needs.
- Self-teaching becomes more effective.
- It can allow the creation of curricula that will be adapted to the student’s needs.
- Alleviates educator’s workload.

GEC approaches

Rules

- Assure that the sentences follow specific manually coded grammar rules.

+ Simple to implement and precise.

- Can’t cover all error types, e.g. Strong tea vs Powerful tea.

Machine Learning Classifiers

- Instead of rules they define features. Then each feature is weighed.

+ More flexible than rules.

- Can focus on only one error type at a time.

GEC approaches

Language Models

- Use N-gram models to predict the probability of sequence of words.

\+ Easy to implement and can handle all errors.

\- Probability does not mean grammaticality and vice versa.

Machine Translation

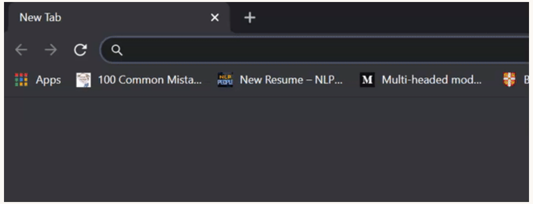
- Statistical Machine Translation (SMT)-uses parallel annotated data.

\+ Can handle all error types.
\- Corpus and context dependent.

- Neural Machine Translation (NMT)-encoder decoder mechanism.

\+ Faster and more fluent than SMT.

Predictive Text



Studies have shown that predictive text:

- Can help in faster and more accurate typing (Waldron, Wood, and Kemp, 2017).
- Improve spelling and grammatical skills (Waldron, Wood, and Kemp, 2017).
- Assist children with learning difficulties (Newell, Booth, & Beattie, 2006).

BUT

- ❖ Success is dependent on age and cohort effects (Kalman, Kave, and Umanski, 2015)

Aims of the study

1

Language Modeling

Evaluate two language models on the prediction of the correct token.

2

Experiment

Use predictive text with real L2 learners.

2

Data Analysis

Datasets

- The Cambridge English Write & Improve corpus
- The LOCNESS corpus
- The National University of Singapore Corpus of Learner English (NUCLE)
- The First Certificate in English corpus (FCE)
- Lang-8 Corpus of Learner English

Format

`S This are gramamtical sentence.` → Original sentence

Start and end token offsets

Error type

Correction

Optional or required correction?

Comment field

```
A 1 2|||R:VERB:SVA|||is|||REQUIRED|||-NONE-|||0
A 2 2|||M:DET|||a|||REQUIRED|||-NONE-|||0
A 2 3|||R:SPELL|||grammatical|||REQUIRED|||-NONE-|||0
A -1 -1|||noop|||-NONE-|||REQUIRED|||-NONE-|||1
```

Annotator's ID

No changes

Data Analysis

R(Replacement)	M(Missing)	U(Unnecessary)	UKN(Unknown)
60-65%	20-25%	10-15%	2-3%

Data Analysis

Error Relative Location

For the relative location calculation and statistical significance :

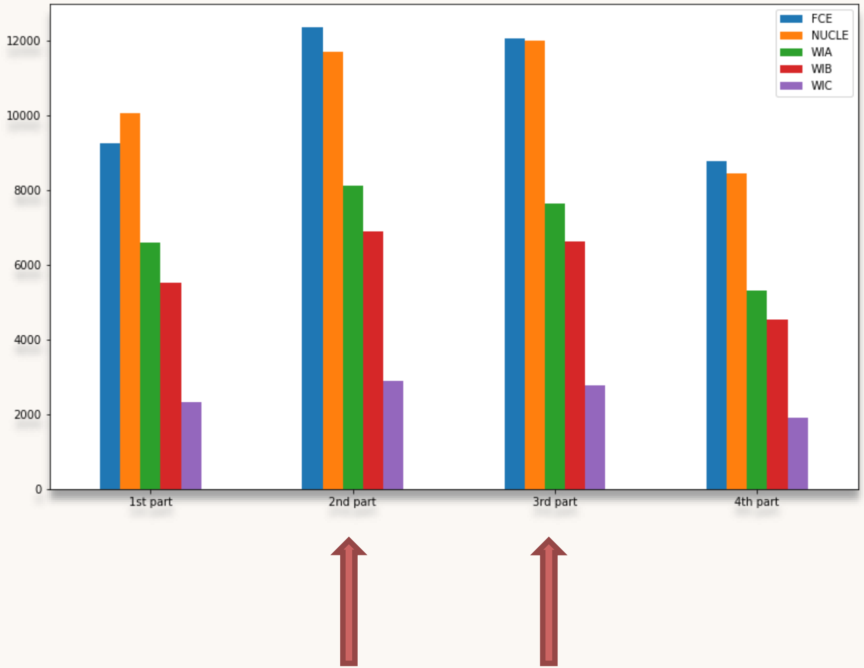
1. I divided the location of the error with the number of the total words.

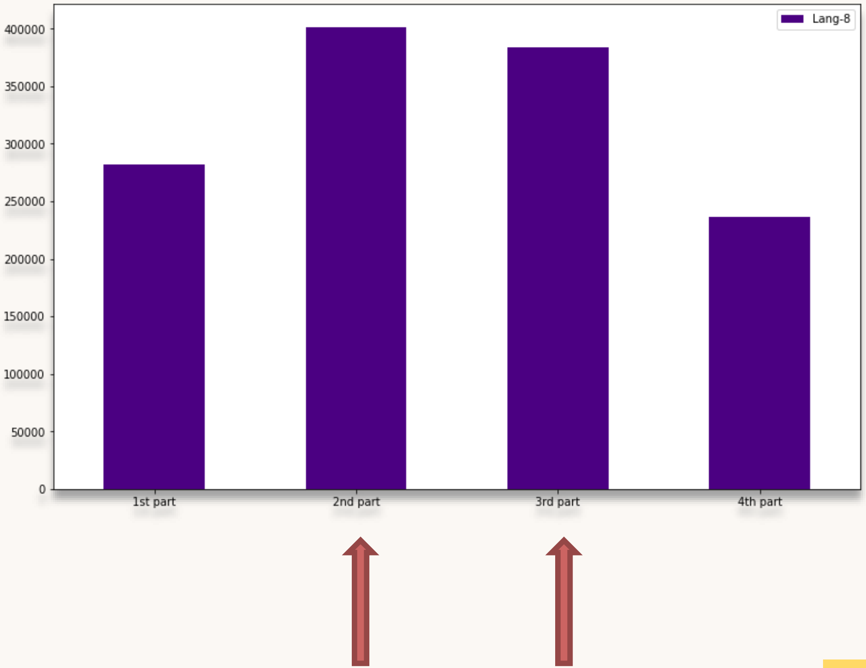
Sentence	Offsets	Total Word Count	Relative Location
I was very disappointed after this show.	5 6	7	0.71

2. Beginning (0.00-0.25), middle-left (0.25-0.50), middle-right (0.50-0.75), or end (0.75-1.00) of the sentence.

Data Analysis

Error Relative Location





Data Analysis

Error Relative Location

3. Sampled 1000 sentences

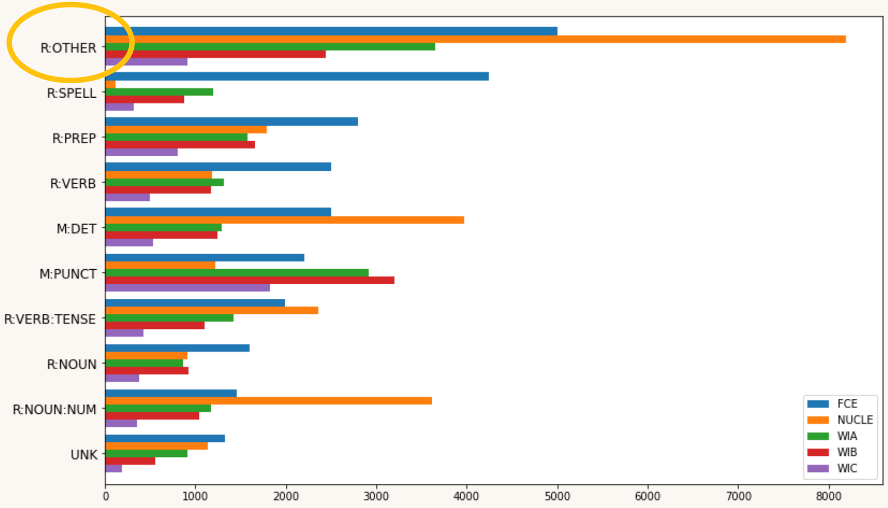
4. Calculated whether the mistake falls into the two middle parts or not.

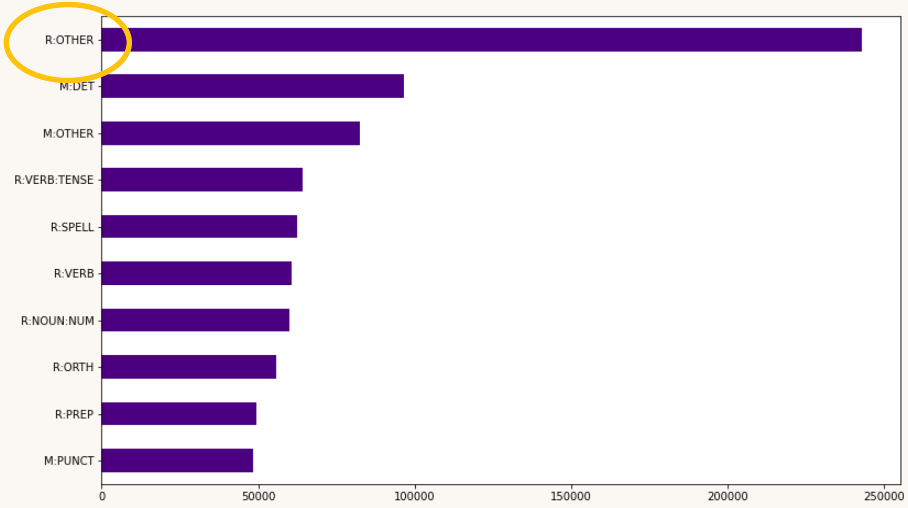
5. Repeated the experiment 1000 times using a for loop.

6. Result: *p*-value>0.05 → our hypothesis that the location of the errors is usually in the middle is not statistically significant.

Data Analysis

Error Type Frequency

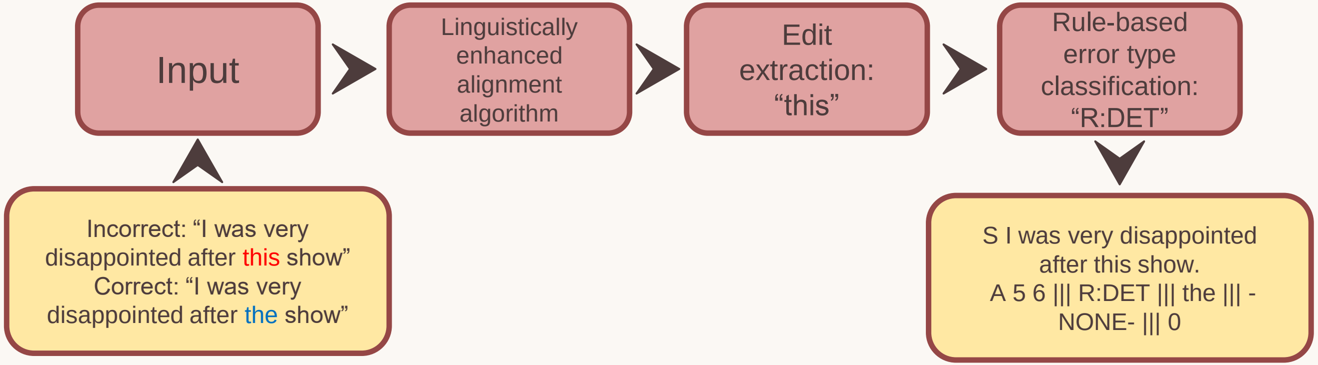




Data Analysis

ERRANT

- What is **ERRANT?**
    - Error Annotation Toolkit (Bryant et al., 2017).
    - Automatically extracts and categorizes errors from parallel original and corrected texts.



Data Analysis

ERRANT

- Why is ERRANT important?

  - Contribution in Grammatical Error Correction (GEC).

  - Standardization of the annotation of learner corpora.

  - Facilitation in second language learning.

Research Question

How accurate is **ERRANT** regarding **error type classification**?

“What does this mean?”

If ERRANT cannot classify errors adequately enough, then:

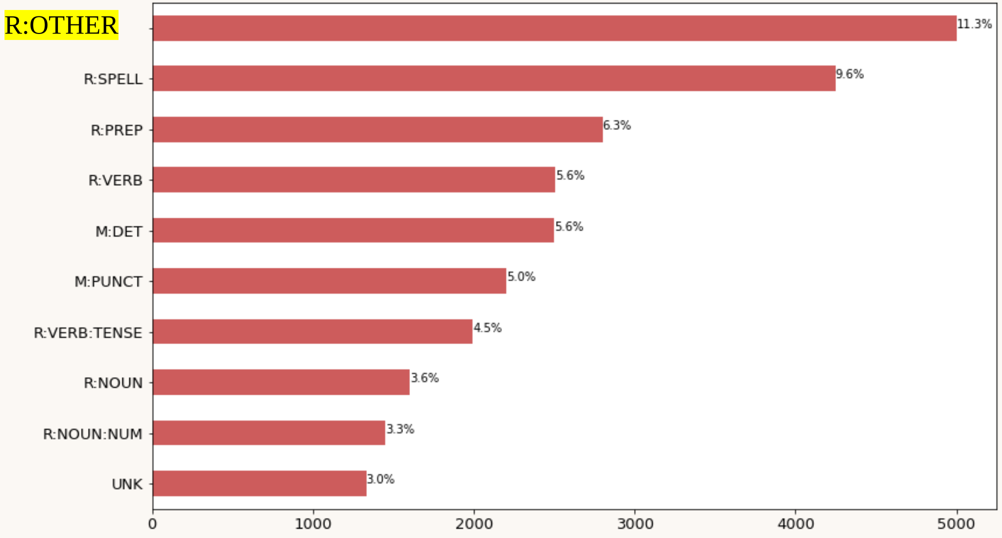
- Valuable information about the errors is lost and, therefore, the toolkit cannot be used for education purposes.

- GEC systems that have used ERRANT in their evaluation might be mis-judging their performance.

Data & Method

FCE corpus (Yiannakoudakis et al., 2011)

The most frequent error type is R:OTHER (something needs to be replaced with something else)



Data & Method

1. Sample the first 100 sentences from the FCE corpus that contain OTHER type errors (incl. M:OTHER and U:OTHER).

2. Manually re-label each of them.

3. Compare data before and after re-classification.

4. Publicly release an XLSX file, with the original uncorrected sentences, the starting and ending offsets, the suggested correction, and any comments.

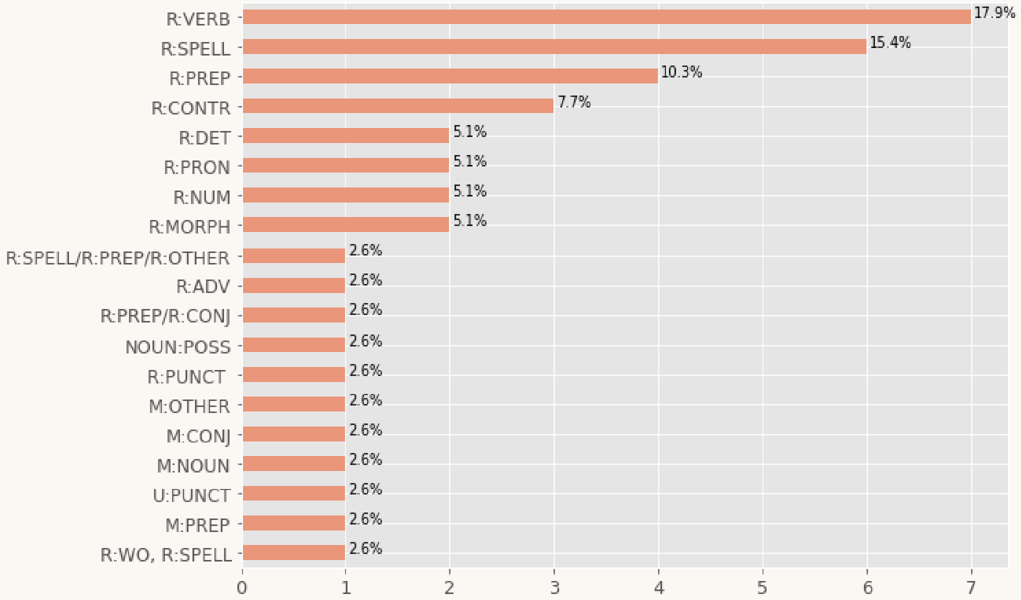
Findings

	FCE Sentence	Offsets	Correction	Old type	New type
1	There was only a person who used to call her by this name.	3 4	one	R:OTHER	R:DET
2	Your sincerely	0 1	Yours	R:OTHER	R:PRON/ R:SPELL

Findings

- 39% of the errors could have been placed in 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.



Improving ERRANT

Our suggestions:

- Introducing more grammar rules that will allow a more thorough classification (e.g. numerals can be determiners)
- Qualitative evaluation by linguists could ensure the quality of the classification and provide professional feedback.

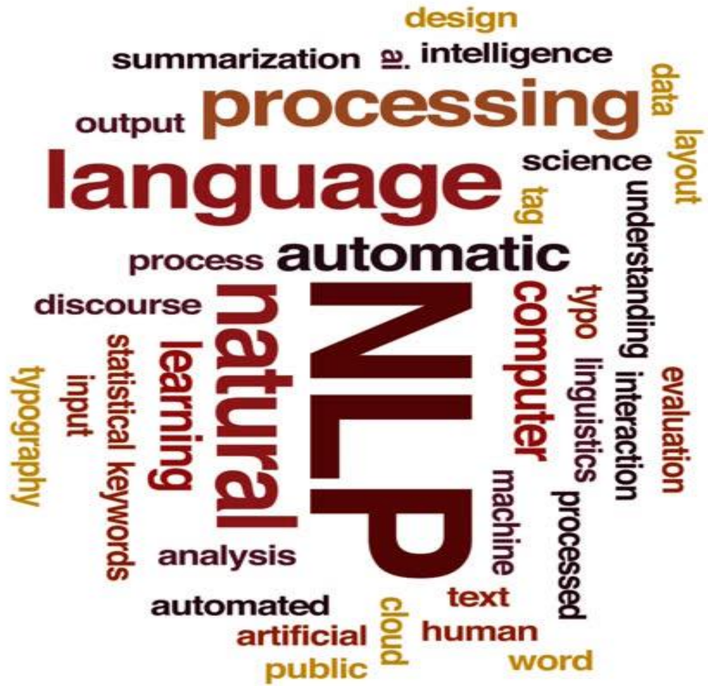
Improving ERRANT

Our next steps:

- Examine and evaluate more types of errors extracted with ERRANT (i.e. re-classify more OTHER type errors, as well as check other categories).
- Development of a larger reference dataset, that 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.
- Design a more systematic and thorough error classification system, by employing transfer learning and deep learning approaches (Korre and Pavlopoulos, 2020).

3

Language Model Prediction



Training and test sets

- BEA-2019 datasets
- Used the offsets and corrections to recreate a corrected version of the sentences that contain preposition replacement errors (R:PREP).
- Used the FCE test set as the test set.
- Sliced the sentence just before the error occurred.

Language Model Prediction

Language Models

Statistical Language Model

Pavlopoulos & Papapetrou, 2020.

Autoregressive Language Model

GPT-2, Radford et al., 2019.

Language Model Prediction

Results

Model	Top prediction	Top 3 predictions
SLM	13%	21%
GPT-2	17%	26%

Language Model Prediction

Secondary experiment

1. Sliced the sentences before words that were not mistaken by the learners.

2. Used the two language models to generate the next token.

Accuracy < 4%

Language Model Prediction

Results

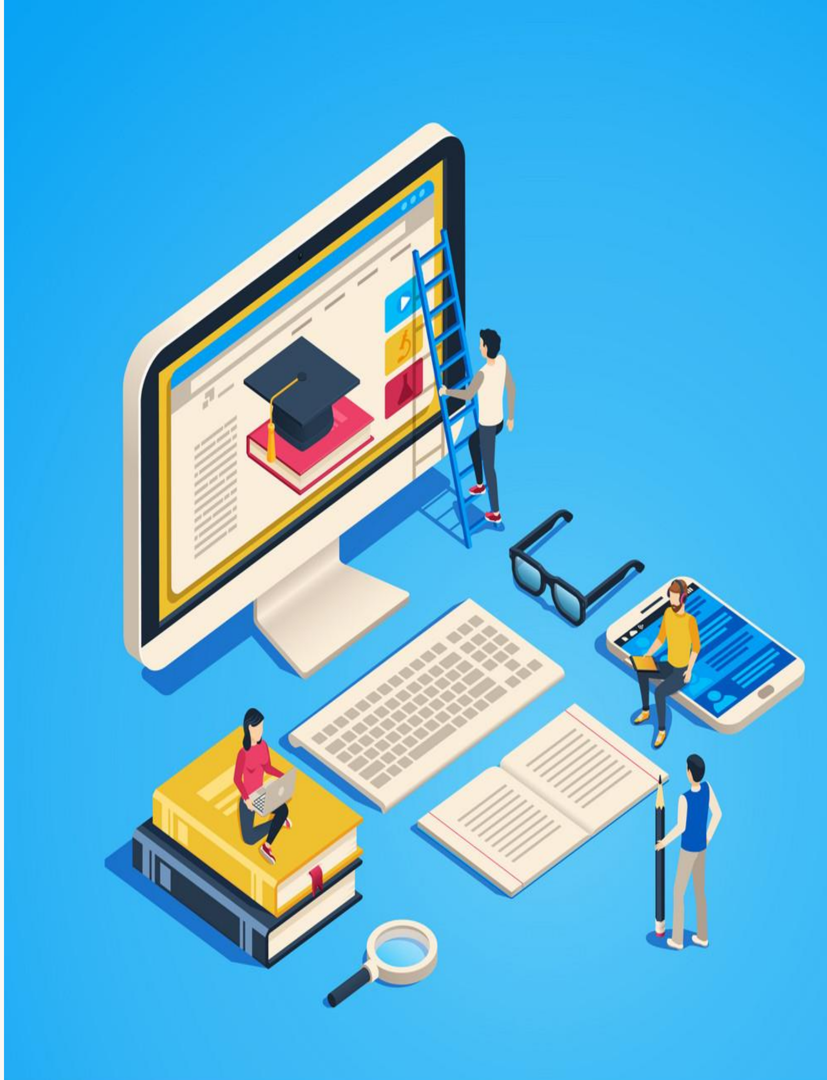
Possible explanations:

- Prepositions can generally be encountered more frequently than other POS.
- Prepositions are more possible to appear in a bigram than other words.
- The suggestions of the system are not always mistaken, but they are simply not the ones of the corrections.

Sentence	Correction	Prediction 1	Prediction 2	Prediction 3
So you are going to come	at	home	across	into

4

Using predictive text in ESL



Using Predictive Text in ESL

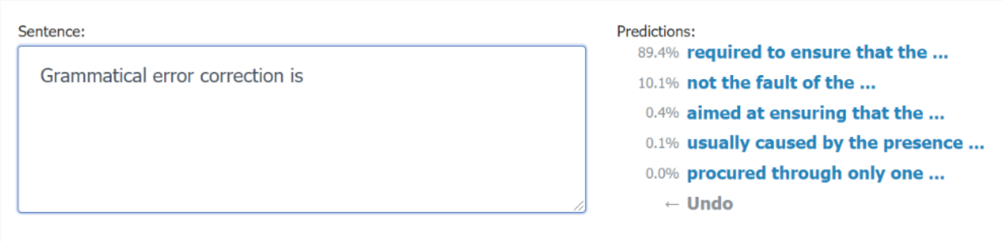
Aims:

- Determine whether predictive text is beneficial for L2.
- More specifically, answer the question whether the predictive text tool can aid ESL learners in their writing

Method

AllenNLP Demo Platform

The platform uses the public 345M parameter of GPT-2 to generate sentences. In this platform, while the user is writing any sort of text, the top predictions appear next to the writing prompt in the form of sentences.



Participants

Learner A

- 50-year-old female
- English language learner.
- B2 certification.

Learner B

- 19-year-old female university student
- English language learner.
- B2 certification.

Experiment Procedure

Instructions

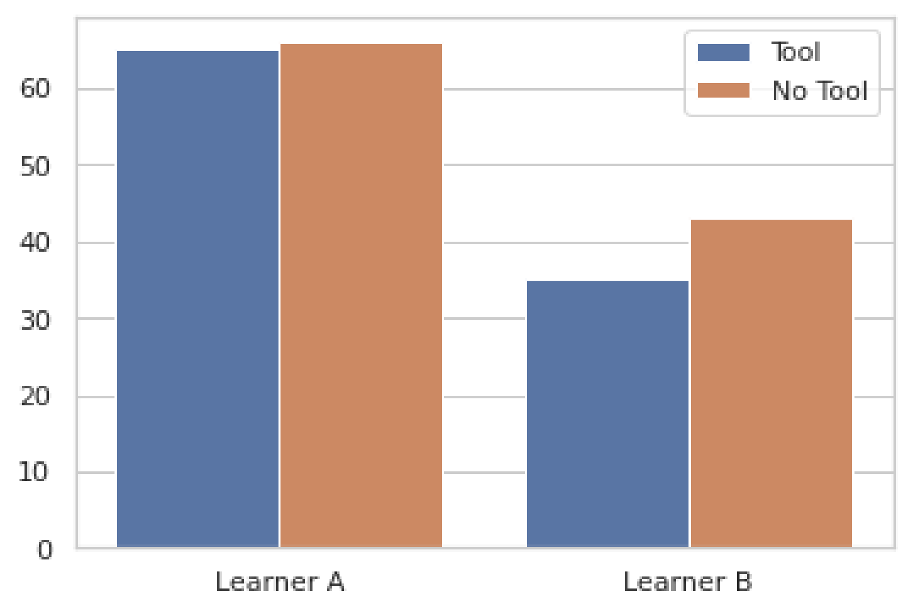
1. Participants were presented with an array of 8 essay topics from the FCE exam.
2. Participants were instructed to write 3 topics of their choosing without any additional tools (e.g. translation tools, dictionaries, or asking for help).
3. Participants were instructed to write 3 topics of their choosing on the AllenNLP demo platform where they could use the sentence suggestions on the right, whenever they saw fit.

Data Preparation

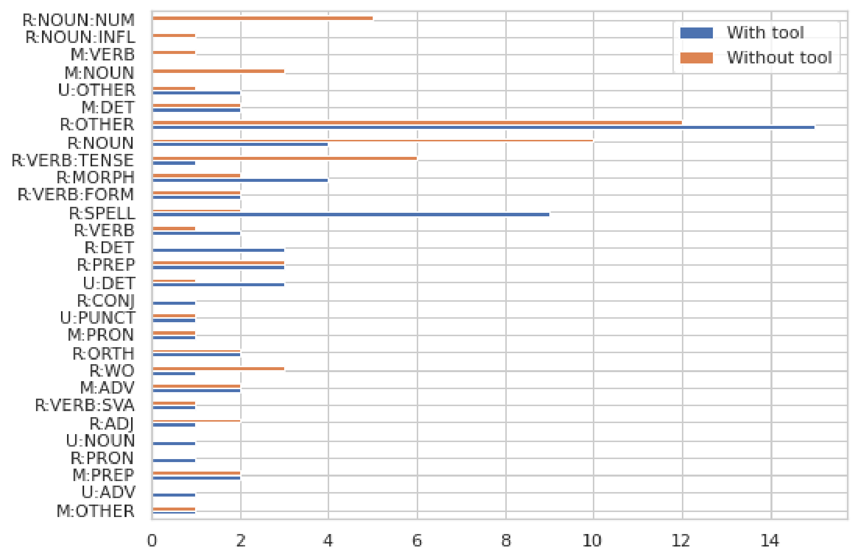
4. Input essays into ERRANT
5. Calculation frequencies of error types.
6. Comparison between the essays written with the predictive text tool and those written without it.

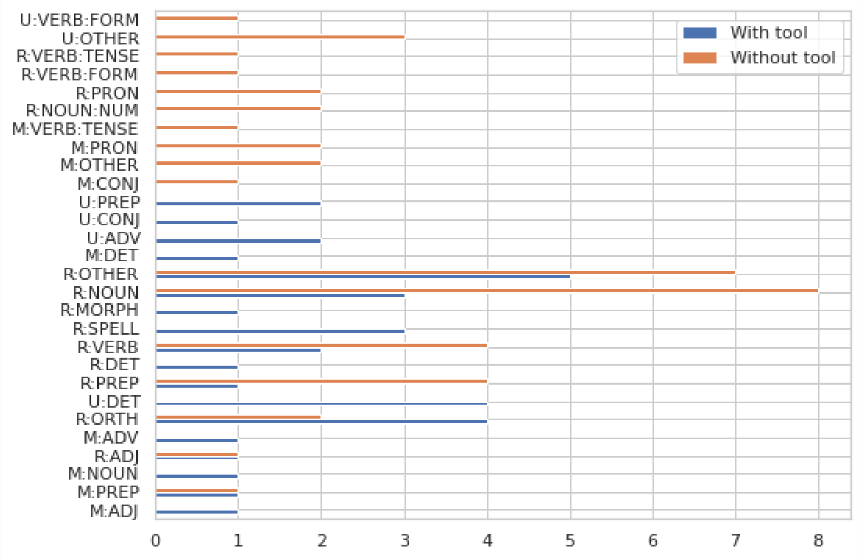
Experiment Results

- Learner A did not show significant improvement, compared to Learner B.
- Age and cohort effects can influence once ability to use predictive text (Kalman et al.,2015)



Experiment Results





Experiment Results

Participants' opinion:

- Learner B was very supportive of the use of such tools in class. She claimed that the tool helped her write much faster and that she wished that she could use it during examinations

.
- She underlined that even though the tool presented some "ready-to-use" sentences, she could learn from it because it suggested syntactical combinations and vocabulary that she had not encountered before.
- She commented very positively on the time-saving benefit of the tool.
- Learner A said that although she did not find the tool confusing to use, she found the process of using it time consuming.
- She also complained that it sometimes "lagged".

5

Conclusion

Use predictive text to complete

The following sentence:

"2021 is going to be..."

Conclusion

This thesis has examined the potential of language modeling and predictive text generation, in Grammatical Error Correction for Second Language Learning.

- Language models reach 15% to 25% accuracy.
- Predictive text tools can help in the reduction of grammatical mistakes in the learners writing.
- However, this is also dependent on the learners' personal characteristics and cohort effects.

Conclusion

Limitations

- ERRANT re-classification was conducted only by one professional.
- LMs might have performed better if trained with other corpora.
- No time profiling during the experiment with the ESL learners.

Conclusion

Future Endeavors

- Improve the errant classification process.
- Create a predicting text tool, specific to L2 learning, and testing it in a classroom setting and by also using a control group and time-profiling, will provide a more detailed image on whether predictive text can assist ESL.

References

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2020). Predictive text encourages predictable writing. In Proceedings of the 25th international conference on intelligent user interfaces (p. 128–138). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3377325.3377523 doi: 10.1145/3377325.3377523

Bryant, C., & Briscoe, T. (2018, June). 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 (pp. 247–253). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-0529 doi: 10.18653/v1/W18-0529

Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019, August). The BEA2019 shared task on grammatical error correction. In Proceedings of the fourteenth workshop on innovative use of nlp for building educational applications (pp. 52–75). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-4406 doi: 10.18653/v1/W19-4406

Bryant, C., Felice, M., & Briscoe, T. (2017, July). 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) (pp. 793–805). Vancouver, Canada: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P17-1074 doi: 10.18653/v1/P17-1074

Dahlmeier, D., & Ng, H. T. (2012, June). 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 (pp. 568–572). Montreal, Canada: Association for Computational Linguistics. Retrieved from ´ https://www.aclweb.org/anthology/N12-1067

Jurafsky, D., & Martin, J. H. (2009). Speech and Language Processing : an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall/Pearson education international. Retrieved from http://books.google.de/books?id=crxYPgAACAAJ

Thanks!

Do you have any questions?

katkorre95@gmail.com
katkorre95@aueb.com

Links: https://github.com/katkorre/ERRANT-reclassification
https://www.cl.cam.ac.uk/research/nl/bea2019st/
https://demo.allennlp.org/next-token-lm?text=AllenNLP%20is