Neural Predictive Text for Grammatical Error Prevention

Anonymous ACL submission

Abstract

In this paper we study the potential of two neural language models, an LSTM and an autoregressive language model (GPT-2), to predict possible correction tokens in erroneous sentences and to predict the next token in randomly sliced correct sentences, in the aim of establishing a new Grammatical Error Correction (GEC) subarea, for which we coin the term Grammatical Error Prevention (GEP). Systems that could assist in GEP, such as language models, are expected to predict elements and therefore prevent grammatical errors in advance. Our findings show that GPT-2 can predict 29% of the correct tokens with one prediction. Accuracy rises up to 44% when the top 3 predictions are considered. To test the pedagogical capacity of such a model, we also experimented with real English as a second language (ESL) learners. By equipping GPT-2 to generate text that functions as potential continuation of the learners' sentences, we created a small corpus of the learners' writings and analyzed their errors along with their frequencies.

1 Introduction

Grammatical Error Correction (GEC) is the task of correcting different types of errors, such as spelling, punctuation, and grammatical errors, in written texts. The procedure of correcting a sentence that contains an error usually requires a system to use the erroneous sentence and transform it into the correct version of it. We suggest, however, that GEC could possibly function "proactively", by providing the correct continuation of the sentence and thus preventing the error from happening. In this case, we suggest the term **Grammatical Error Prevention (GEP)**, as it is descriptive of our aim and because the task is differentiated from GEC. GEP can be achieved with predictive text systems.

This study is concerned with the pedagogical aspect of GEP in second language (L2) learning, and in particular in learning and teaching English as a second language (ESL). Given that English is spoken by around 20% of the world's population as a foreign language,[1] there is an urgent need for new pedagogical methods that comply with a new technological framework, and which can offer adequate assistance both to learner and to educator. Due to the autonomy that GEC systems are rapidly acquiring, L2 learning applications will be able to both aid the learner's self-study and self-evaluation, and at the same time alleviate the educator's workload, such as correcting essays. Given that not all learners prefer a supervised pedagogical method, self-teaching through the assistance of GEC and GEP systems becomes even more effective as learners will have the ability to prevent errors.

The increasing popularity of automatically handling errors as a Natural Language Processing (NLP) topic is proven by the two most recent shared tasks, CoNLL-2014 (Ng et al., 2014) and BEA-2019 (Bryant et al., 2019). These two shared tasks involved the development of GEC systems that would correct the sentences of a multi-set of data of different groups of learners, and which consequently contained a great variety of grammatical errors. Not only did the two shared tasks present state-of-the-art systems in the field of GEC, but also brought into the spotlight several weaknesses that still afflict modern systems, such as handling sentences that contain multiple errors.

Most of the systems that participated in the shared tasks used transformation methods to correct the erroneous sentences. Given that GEC usually operates after the user has made an error, the question that arises here is how effectively these errors could be predicted in advance, and whether those predictions would facilitate the task of GEC systems. Those questions drove our decision to focus on *predicting the potential corrections of er-*

[1] https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf

roneous sentences.

The chosen approach in this study is predictive text. First of all, the idea is to check whether language modeling can be successful at predicting the correction of an erroneous sentence. By making this prediction accurately and before the error occurs, the error is avoided. Usually, predictive keyboards are evaluated disregarding the difficulty of the (correct) token to be predicted. For example, predicting what preposition follows the verb *come (on, to, for, out,...)* is more difficult than predicting the preposition of the verb *focus (on)*. Thus, a question remains unanswered: are language models more or less effective in predicting tokens that people are having difficulty with? Such tokens might as well be called 'commonly confused words' or 'confusion sets', according to Rozovskaya and Roth (2010) who made an attempt on the issue of prepositions using discriminative classifiers. In this study we are tackling the same issue in the neural era, by considering all kinds of errors as well as zooming in on prepositions. Specifically, we attempt to evaluate two neural language models on this task.

The language models used are a Recurrent Neural Network (LSTM) and a state-of-the-art unsupervised neural language model (GPT-2). The second aim of this study is to verify the hypothesis that predictive text can help in ESL learning. In this perspective, we experiment with the most successful language model by using it with real ESL learners. We show that GPT-2 is the most efficient language model when it comes to predicting potential correct tokens in a sentence.

The rest of the paper is structured as follows. First, we discuss related work on predictive text. Then, we present some information about the data used in this study. Section 4 demonstrates the data preparation, the methods, the experimental set up for assessing the language models, and the findings. Section 5 is concerned with the use and evaluation of predictive text in ESL. Finally, we summarize our findings and discuss future work.

2 Related Work

Predictive keyboard is omnipresent in all of our digital devices, from computers to tablets and mobile phones. The speed and convenience that it provides during typing has now made it an integral part of any writing tasks. Despite the common misconception that such conveniences might impair one's language abilities, a misconception mainly based on the fact that writing tools like predictive keyboard might reduce the activity of the brain, new studies suggest otherwise.

Predictive text can not only help in faster typing but it can also, in conjunction with auto-correct, improve the user's spelling and grammatical skills. Waldron et al. (2017) observed that predictive text used in text messaging can influence the quality of errors primary school students made, as well as that university students made significantly fewer grammatical mistakes when using predictive text. Cohort effects and age, however, can influence the capacity of such tools. Kalman et al. (2015) conducted an experiment focusing on the use of predictive keyboard in younger and older age groups, in terms of speed and accuracy. As expected, there were differences in the scores of the two groups with the younger group typing faster and with a greater variation of keys, while the older groups typed more slowly and with less variation of keys. These findings suggest that a better understanding of the variables can contribute to personalized Human-Computer Interaction (HCI) designs (Gajos et al., 2012). Moreover, given that language and cognitive ability are interrelated, such studies can provide information on markers for cognitive decline or even injury (Kalman et al., 2012).

The benefits of predictive text in relation to cognitive skills can be traced from the 1990s, when PAL, a predictive computer program, was used in a classroom environment consisting of children with learning difficulties (Newell et al., 2006). PAL works differently from a usual predictive keyboard. "It exploits the redundancy in natural language to reduce the number of character entries necessary to produce a piece of text" (Newell et al., 2006, p.23). In this way PAL manages to offer some predictions that function as the continuation of the user's sentence, reducing thus the typing time. In Newell et al. (2006), 8 out of 9 study cases showed very positive results. PAL helped produce higher quality writings with reduced spelling errors, while it also enhanced the children's confidence and motivation.

In terms of the quality of writing with a predictive text, Arnold et al. (2020) underline that aside from speed and accuracy, it is mandatory that we evaluate the effect intelligent text has on the content written. More specifically, their findings show that when the users were presented with the predicted options, they tended to write predictable sentences with fewer words. The two studies bring the two

sides of the coin to the limelight, and address the potential benefits and shortcomings of predictive text. It is obvious then that there are certain effects of the predictive text on the native language users. The next question that needs to be addressed is what the effects of the predictive text in L2 learners are. A very interesting hypothesis is that L2 learners will be influenced differently from native speakers, if we take into account that the former group does not anticipate information during processing to the same degree the latter group does (Kaan, 2014). How ESL learners react is one of the objectives of this study.

3 Data

Our data set comprises the corpora used in the BEA-2019[2] shared task (Bryant et al., 2019) and which were in M2 format (see Table 2). This study presents a detailed description of the data by combining the analysis conducted for the BEA-2019 shared task paper with our analysis.

Dataset	Total sentences	Erroneous sentences
FCE	28,350	18,045
Lang-8	1,037,561	497,703
NUCLE	21,835	21,835
W&I(A)	10,493	8,330
W&I(B)	13,032	9,243
W&I(C)	10,783	5,472
Total	1,122,054	560,628

Table 1: Almost 50% of the sentences contained errors in total. Note that the NUCLE dataset comprised only erroneous sentences.

We considered error types a vital aspect of the data, especially when it comes to ESL. Error types can provide great insights into the learners' learning pace and error patterns, and, therefore the study of error types can equip educators with information for a curriculum that fits each learner individually. To obtain a better idea on the error types of this data set, we worked out some frequency ratios. The top ten frequencies of each error type per dataset, upon pre-processing, are shown in Figure 1.

In all datasets presented in Fig 1, we can see that R:OTHER error type occupies first or second

```
S This are gramamtical sentence.
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
```

Table 2: ERRANT M2 format example. The line starting with S is the original sentence, while the ones starting with A are the edit annotations. The edits contain the start and end token offsets, the error type, the correction, a flag indicating whether the edit is required or optional, a comment field, and a unique annotator ID. A 'noop' edit indicates that no changes were made to the sentence.

positions in terms of frequency. Bryant et al. (2017, p. 795) define this category as "[e]rrors that do not fall into any other category (e.g. paraphrasing)". After a qualitative analysis, it became apparent that this particular error category contained errors that could have been included in other, more concrete categories, in which the error type would be more adequately described. One justification for this, according to Bryant et al. (2019, p. 55), is that "certain edits are longer and noisier...and do not fit into a more discriminate ERRANT category". Korre and Pavlopoulos (2020) describe the issue in greater detail, suggesting that it might even affect the evaluation of the performance of the systems that used ERRANT.

Keeping in mind that each data set comes from a different demographic of learners, it is expected that the frequencies of error types vary among data sets. For example, a great portion of error types is assigned as M:PUNCT, namely missing punctuation. More specifically, and as it is also mentioned in Bryant et al. (2019), in NUCLE punctuation errors occur at a percentage of 5% while in W&I it rises to 20%, when we add the percentages of each individual subset.[3] This is also visualized in Fig. 1, where in two out of three W&I datasets, the most frequent error type is M:PUNCT. This difference might be due to the fact that W&I has a wider range of learners. Another observation made by Bryant et al. (2019) was that noun number (NOUN:NUM) errors occur twice the times in NUCLE compared to the rest of the datasets. Similarly to subject-verb agreement (SVA) errors, NOUN:NUM was among the five targeted error categories in ConLL-2013 shared task, hence the higher proportion.

[2]The BEA-2019 train data set consists of the FCE (Yannakoudakis et al., 2011), Lang-8 (Mizumoto et al., 2012; Tajiri et al., 2012), Write and Improve (Yannakoudakis et al., 2018), and Nucle corpus (Dahlmeier et al., 2013). Part of FCE was also used as an evaluation and development set. Part of the Write and Improve + LOCNESS corpus was used for development only.

[3]W&I is divided into three subsets: A, B and C according to CEFR levels.

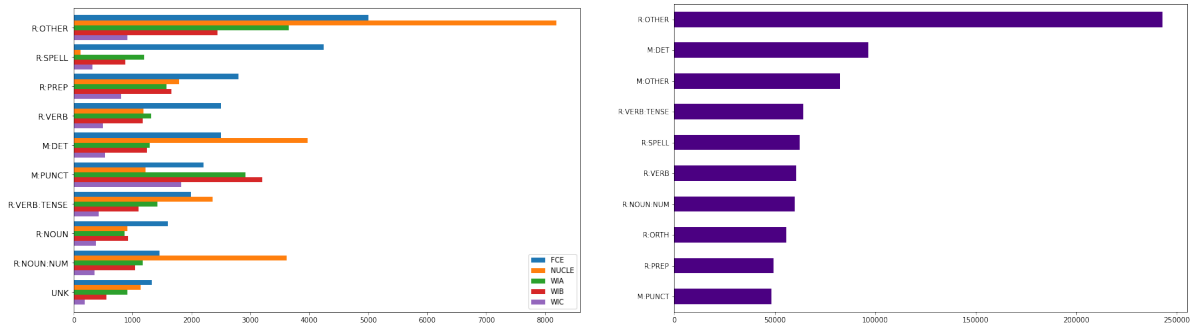

Figure 1: The ten most frequent error types among Lang-8 (right) and other data sets (left). Lang-8 frequencies are presented separately, as they are too disproportionate to present along with the others. R:OTHER is the most frequent error type in four out of all six datasets. In WIB and WIC M:PUNCT is the most frequent type.

4 Assessing Language Models on Predicting ESL learner Errors and Random Tokens

In this study, we focused on whether neural language models can efficiently predict the correct token. For this purpose, we equipped two different neural models: an LSTM and GPT-2.

4.1 Data Preparation

GPT-2 is already pre-trained on curated-by-humans text, which was extracted from 8 million web pages (Radford et al., 2019). LSTM was trained on a data set, which was formed by using the corrections of each train data set from the BEA-2019 (see Table 1) to re-create a corrected text. We opted for correcting, and later using for training, only the sentences that contained replacement errors. Sentences with more than one errors were also eliminated. We then concatenated all the data sets to form an ESL training set of approx. 122k sentences. For the evaluation, we used two different test sets. For the first set, we concatenated the FCE and the Write & Improve + LOCNESS (W&I+L) test sets of BEA-2019 and sliced the sentences just before the error occurred. The second set involved using the corrected version of the same sentences but slicing them at random points and not before the error. Each evaluation data set consists of 1000 sentences.

4.2 Methods

The first language model or LM we used in this study was an LSTM (Hochreiter and Schmidhuber, 1997). At each time step s the LSTM learns a hidden state h_s as a non-linear combination (weight matrix W) of the input word x_s and the previous hidden state h_{s-1}. More formally:

$$
\begin{aligned}
i_s &= \sigma(W_i \cdot [x_s, h_{s-1}] + b_i) \\
f_s &= \sigma(W_f \cdot [x_s, h_{s-1}] + b_f) \\
o_s &= \sigma(W_o \cdot [x_s, h_{s-1}] + b_o) \\
q_s &= \tanh(W_q \cdot [x_s, h_{s-1}] + b_q) \\
c_s &= f_s \cdot c_{s-1} + i_s \cdot q_s \\
h_s &= o_s \cdot \tanh(c_s),
\end{aligned}
\tag{1}
$$

where i_s is the input gate and f_s is the forget gate that regulates the information from the current (q_s) and previous (c_{s-1}) cells; o_s is the output gate that regulates the information of the new hidden state. The generation of the next word x_{s+1} is a classification task, with $softmax$ yielding a probability distribution over the whole vocabulary and the next word to be generated being the most probable one. We note that a confusion set (e.g., with prepositions) can be used along with $softmax$ to restrict the model to predict only words of the set.

The second language model used was the recently popular GPT-2 (Radford et al., 2019),[4] which is a large transformer-based (Vaswani et al., 2017) language model and a successor of GPT (Radford, 2018). Besides its state of the art performance in language modeling (Radford et al., 2019), the other reason we chose to use GPT-2 is that it is able to perform multiple tasks unsupervised and without requiring a manual training set creation and annotation. Our baseline involved generating random tokens from the vocabulary of the training set and comparing them against the gold references (Table 3).

[4] https://pypi.org/project/next-word-prediction/

4.3 Experimental Results

Each language model was used to predict a random token and the correction token at the sentence location where the error occurred. For the latter, we measured the accuracy against the gold references, which were extracted from the initial M2 file (ACC@1). In principle, other metrics, such as Precision and Recall, could have been used. However, we cannot know all the possible continuations of the sentences to determine whether the given prediction is right or wrong (see Table 4). As mentioned in (Rozovskaya and Roth, 2021), common metrics are not sufficient since there are many ways to correct a sentence and even multiple gold references cannot account for all of them. Therefore, system performance can be significantly higher than what the metrics show. For this reason, and especially because prediction is a more open-ended task than correction, we opted for evaluating our systems with accuracy scores which do not evaluate the systems in the strict sense but give us an idea of the potential of the systems. A second stage involved predicting a greedy selection of 3 tokens and checking if the gold reference is among them, and whether this can elevate the accuracy of the model (ACC@3). We also focused on the performance of the models in preposition prediction, because prepositions are one of the most common and idiosyncratic errors for ESL learners. The sentences with the preposition errors were extracted from the dataset of the erroneous sentences. The results are presented in Table 3.

	ERRONEOUS		RANDOM		PREPOSITION	
	Acc@1	Acc@3	Acc@1	Acc@3	Acc@1	Acc@3
RAND	1	1	1	2	2	1.5
LSTM	5	12	18	29	11	17
GPT-2	17	29	29	44	30	52

Table 3: Accuracy (%) of next token prediction, using one (Acc@1) and three (Acc@3) predictions of LSTM, GPT-2 and a baseline that predicts a random word. All the models are evaluated on mistaken prepositions, mistaken words and words at random (top row).

LSTM was clearly better than a baseline that simply predicts a random word. When evaluating using Acc@1 (evaluating the top prediction), the model performs worse when it is asked to predict the correct token in a position where the learner made a mistake (1st column) and best when it predicts the next word at a random position (3rd column). When the LSTM predicts the next preposition (5th column), Acc@1 is between the other two. The accuracy of LSTM elevates (12-29%) when we evaluate the model using Acc@3; i.e., whether the correct token is within the top 3 predictions of the model (a user, for example, would see 3 options to choose the next word from). GPT-2 performed better, achieving 17-29% Acc@1 and 29-44% Acc@3. Its superiority is probably due to the great amount of native data it is pre-trained on. Interestingly, both neural models seem to perform worse in the sentences where the learner made errors.[5]

In terms of preposition prediction, the accuracy of GPT-2 is even higher. When evaluating the top-3 predictions of the model, the accuracy was 52% while the respective Acc@1 was 30%. What the latter means is that in three out of ten preposition mistakes, the model would have returned the correct preposition as its top prediction. The fraction of prepositions predicted correctly (i.e., the top prediction) by GPT-2 is even higher, reaching 43%. The same applies for LSTM, whose fraction of correctly predicted prepositions was 26%.

What the very high accuracy of GPT-2 in preposition prediction tells us is that for one out of two preposition mistakes made by ESL learners, the correct one would have been included in the top three suggestions of the model. It remains unknown though whether the learner would have chosen the correct preposition or whether the system could lead to more errors overall, an issue which is more thoroughly studied in Section 5.

4.4 Discussion

Next ESL word prediction Elimination experiments that involved variations of the parameters in the language models, or using other corpora for training, did not seem to have a great effect on the performance of the language models. Comparing all error types and prepositions, one possible explanation for these scores lies in the nature of the prepositions, which can be detected more easily than other parts of speech (POS), since prepositions often occur in collocations (e.g., 'look for', 'look out') (Hartrumpf et al., 2006) and can be learnt by models more efficiently. Manual evaluation would give higher scores, because the suggestions of the system are not always mistaken, but they are simply not the ones of the corrections. That is, there may be more than one correct answers per error. This issue is illustrated in Table 4 below.

[5]Preliminary experiments showed that fine-tuning GPT-2 on ESL or native data drops performance.

Sentence	*So you are going to come ...*
CORRECT TOKEN:	at
GPT-2 PREDICTION 1:	home
GPT-2 PREDICTION 2:	across
GPT-2 PREDICTION 3:	into

Table 4: GPT-2 fails to predict the correct token.

Future steps When it comes to preposition prediction, the performance of the language models could be further improved. Elghafari et al. (2010) achieved a 76.5% accuracy with their surface-based n-gram strategy using masked language modeling, emphasising, nonetheless, that there are still issues to tackle, such as the nature of the preposition, i.e., whether it is functional or lexical. To correctly predict a functional preposition, one only needs the context (e.g. Mary is dependant **on** her phone), while for a lexical preposition prediction context is not enough (e.g. He put the box **on/under/behind** the table). To sum up, predictive text definitely has potential in grammatical error prevention. Particularly for prepositions, an optimization of language models can be achieved by taking into account all possible parameters that concern prepositions, (e.g. nature, possible pairings, frequency).

5 Employing predictive text in ESL

The use of predictive text, not only in education but in daily life (i.e. through messaging) as well, is a controversial issue. On the one hand, one could argue that the automatic completion of sentences might lead to restricting and dulling brain activity, consequently affecting the user's language skills, including their grammar (Waldron et al., 2015). On the other hand, there are those who advocate that predictive text systems might in fact enhance the user's ability to generate more creative texts (Waldron et al., 2017). Such a result could therefore mean that predictive text can help acquiring a better command of the language. To explore the potential of predictive text in regards to second language learning, and more specifically to ESL, we conducted an experiment with real English language learners, to determine whether predictive text is beneficial for ESL learners. The experiment and its results are discussed below.

5.1 Empirical evaluation

Platform The AllenNLP website provides a demo user interface that allows typing and receiving GPT-2-based text completion.[6] With each next word a user types five predictions are shown on the right of the writing prompt, each being a suggested sentence continuation. [7]

Participants The main participants of this study were two ESL learners. Both of them have B2 certification in the English language.[8] The first participant is a 19-year-old female university student from Greece, currently studying English to get higher certification. She has not stopped taking English classes since the age of 9. The second participant is a 50-year-old female, also from Greece. She had not had English classes for several years until she decided to start again for her own reasons. In addition, we experimented with higher level participants (C2) but the impact of the tool was negligible. We considered that such a tool would be more appropriate for learners with lower language proficiency and who tend to make errors more frequently. Therefore we chose to focus on the B2 level learners. Age distribution and native language (Greek) were random since these two were the only available B2 level learners at the time of the experiment.

Instructions The experiment was conducted in three phases. During the first phase, the participants were instructed to write three compositions. They were presented with an array of 8 topics taken from the First Certificate in English past papers, and they could choose whichever topic they wanted. They were instructed to write without any additional tools (e.g. translation tools, dictionaries, or asking for help). The word limit was 140-190 words, as in the examination. However, there was a leeway of 50 words. To write the composition they both used a simple text editor. For the second phase, participants were instructed to choose three from the remaining topics. However, this time they wrote the essay on the AllenNLP demo platform where they were allowed to use the text completion suggested on the right, whenever they saw fit. The third phase involved a small discussion on their

[6] https://demo.allennlp.org/next-token-lm As of April 2021, the platform only provides set examples of sentences and cannot be used experimentally.

[7] The number of predictions is configured by the platform developers and users cannot change it.

[8] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

experience with the platform.[9]

Data Preparation As mentioned above, the L2 essays were in text form. In the text document, each new sentence occupies a separate line. This is a mandatory step because such format is required for the ERRANT[10] tool to work. ERRANT feeds on the original and corrected sentences and produces an M2 output (Table 2). The next step was to put the data into ERRANT. From the output of ERRANT, the error types were saved into 4 lists. Two lists for each learner, one with the error types made when using the tool, and one without it. This will enable the calculation of the total of the error types and the comparison between the essays written with the predictive text tool and those written without it.

5.2 Results and Discussion

The results of the experiment confirmed the initial hypothesis that predictive text can help the learner to some degree, but the success is also quite dependent on the learner's individual characteristics. Noteworthy is the fact that both participants chose almost exactly in the same way which topic to write with the tool and which without it. Some initial statistics show that both participants wrote around 1,000 words in total, yet Learner A made almost twice as many errors as Learner B. In addition, the word count of the essays of Learner A seem to fluctuate much more, with the longest essay being 215 words and the shortest 117 words.

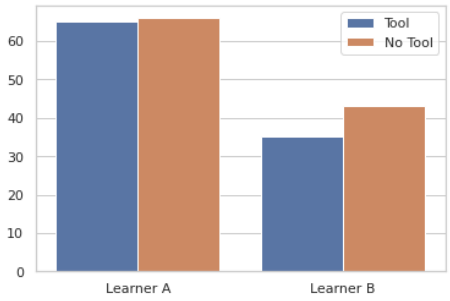

Figure 2: Number of errors with and without using the predictive text tool.

Looking at Figure 2, it is apparent that predictive text yielded different results for each learner. Learner A did not show any significant improvement with the tool, whereas Learner B has made fewer mistakes. The question that follows is what does the outcome depend on? In an attempt to answer this question, it is worth to look at some of the participants' characteristics. Learner A is a 50 year old female without much contact with technology, while Learner B is 19-year-old student, probably having spent a lot of time on her computer or any other electronic device since a young age. Taking this into account, we must think that older people do not have the same familiarity with technology as young people. The fact that they have to type the essay instead of writing it and then also clicking on the most appropriate text continuation might be a challenge. On the other hand, younger people are expected to be facilitated from such applications and platforms, given that they can quickly pick up how to use them. This also confirms the findings of Kalman et al. (2015), in that age and cohort effects can influence once ability to use predictive text.

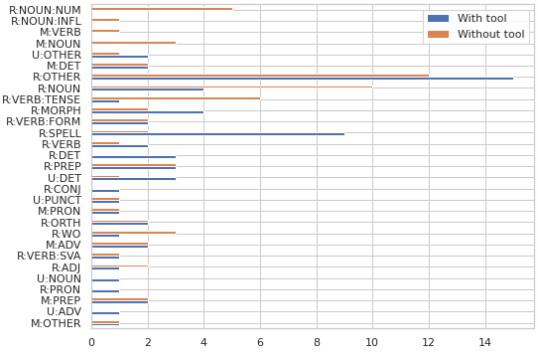

Figure 3: Learner A error types before and after the use of the predictive text tool. Learner A made 24 different errors types with the tool and 24 without the tool.

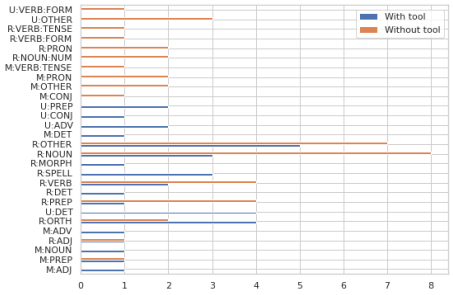

Figure 4: Learner B error types before and after the use of the predictive text tool. Learner B made 17 error types without the tool and 18 error types with the tool.

As far as the error types the learners made are concerned, looking at Figure 3, the first thing that we see is that the most common error type made when using the predictive text tool is R:OTHER, with almost 15 out of the 66 errors. This is not however the true portion of the classification of the

[9] Before the beginning of the experiment both participants filled an ethics form.

[10] https://github.com/chrisjbryant/errant

	Learner A		Learner B	
Topic	Total Words	Total Errors	Total Words	Total Errors
1	150	10.6%	152	6.5%
2	215	16.7%	150	12%
3	117	11.9%	143	5.5%
4	215	14.4%	160	8.8%
5	168	9.5%	202	9.9%
6	160	11.9%	186	4.8%
Total	1025	12.9%	993	8%

Table 5: Basic statistics. The rows in gray indicate that those topics were written with the predictive text tool. Apart from the first topic, which was different, both learners chose in the same way.

errors, because, as already mentioned (Section 3), the automatic annotation tool has a major shortcoming, which prevents a more accurate error type classification. The second, and third most frequent error types are R:SPELL and R:NOUN along with R:MORPH, respectively. The rest of the error types presented less than 4 occurrences in total. Without using the predictive text tool, the most frequent error type is again R:OTHER, however, with twice fewer occurrences than with the tool. R:NOUN, R:VERB:TENSE and R:NOUN:NUM errors follow with ten, six and five occurrences respectively.

The most frequent error type Learner B made without the tool was R:NOUN, with 8 occurrences. R:OTHER was the second most frequent with 7 occurrences. R:OTHER was the most frequent error types with the tool, as well. Spelling (R:ORTH), along with determiner U:DET mistakes were quite frequent, too, when the tool was used.

An observation that could be made when comparing the two figures concerns the consistency of error types, when switching from no tool to tool. Particularly, Learner A made mostly the same types of errors with and without the tool, with only 9 not overlapping. On the other hand, the error type pattern of learner B is completely different with only 7 error types overlapping, meaning that, although fewer mistakes have been made, there is a greater variation of error types and less consistency. A deduction that could be made from this observation is that Learner A wrote the essays as she would without using the tool, and relying on the suggestions as little as possible.

After the completion of the task, the two learners shared their thoughts about their experience of the predictive text tool. 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 also 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 also commented very positively on the time-saving benefit of the tool. Learner B, on the other hand said that although she did not find the tool confusing to use, she found the process of using it time consuming. This is a very interesting comment given that Learner B used to know how to write in blind system, but still considered typing the essay time-consuming.

Limitations Due to the small number of B2-level participants and their limited native language diversity, the results cannot be generalised with regard to B2-level learners. Instead, the study confirms previous research (Waldron et al., 2017; Kalman et al., 2015; Newell et al., 2006) and illustrates the potential of GEP as a new GEC subarea.

6 Conclusion

This paper has examined the potential of a new GEC area, namely GEP and which employs predictive text. The potential of GEP is also studied in conjunction with its use by ESL learners. The study first involved evaluating the accuracy of neural language models regarding the prediction of mistaken tokens by ESL learners, which we showed that it could reach up to 44%. The study then involved an experiment with real ESL learners who were called to write essays with and without using predictive text. Our findings showed that predictive text can help in the reduction of grammatical errors, but this also depends on the learners personal characteristics and cohort effects, such as age and familiarity with technology. Future work comprises an endeavour to build predictive text specific to ESL learning and testing it longitudinally in a classroom setting.

7 Ethical Considerations

ESL experiment Before the beginning of the experiment, all participants filled in and signed an ethics form. The ethics form required noting down demographic details of the participants and ensuring their consent. In the ethics form, participants are asked for confirmation of their understanding of the research purposes and are assured of their right to withdraw any time they would like. They are also reassured that their details will remain confidential.

Predictive Text Tool The predictive text tool can prove to be very useful especially in settings where language learners are self-taught or in cases of distance learning where indirect feedback is not possible. Especially, in case the tool provides a specific word type prediction (e.g., preposition, conjunction, etc.) learners will be able to observe the behavior of each type in the sentence individually. Taking into account our experiment results, we also believe that this tool would be the most beneficial for intermediate level of ESL learners, because higher-level learners tend to rarely make mistakes, while lower-level learners might not have the judgement to properly use the tool. More specifically, a learner might have a word in mind that is in fact correct, and wants to cross-check it with the predictive tool. The predictions of the tool however might not include the word the learner was thinking, falsely leading them to think that the word they thought of was wrong.

References

Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 128–138, New York, NY, USA. Association for Computing Machinery.

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.

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.

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.

Anas Elghafari, Detmar Meurers, and Holger Wunsch. 2010. Exploring the data-driven prediction of prepositions in English. In *Coling 2010: Posters*, pages 267–275, Beijing, China. Coling 2010 Organizing Committee.

Krzysztof Z. Gajos, Amy Hurst, and Leah Findlater. 2012. Personalized dynamic accessibility. *Interactions*, 19(2):69–73.

Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. 2006. Semantic interpretation of prepositions for nlp applications. *Proceedings of the third ACL-SIGSEM workshop on prepositions*, pages 29–36.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Edith Kaan. 2014. Predictive sentence processing in l2 and l1: What is different? *Linguistic Approaches to Bilingualism*, 4:257–282.

Yoram Kalman, Gitit Kavé, and Daniil Umanski. 2015. Writing in a digital world: Self-correction while typing in younger and older adults. *International journal of environmental research and public health*, 12:12723–12734.

Yoram M. Kalman, Kathleen Geraghty, Cynthia K. Thompson, and Darren Gergle. 2012. Detecting linguistic hci markers in an online aphasia support group. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 65–70.

Katerina Korre and John Pavlopoulos. 2020. ERRANT: Assessing and improving grammatical error type classification. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 85–89, Online.

Tomoya Mizumoto, Yuta Hayashibe, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2012. The effect of learner corpus size in grammatical error correction of ESL writings. In *Proceedings of COLING 2012: Posters*, pages 863–872, Mumbai, India.

Alan Newell, Lynda Booth, and William Beattie. 2006. Predictive text entry with pal and children with learning difficulties. *British Journal of Educational Technology*, 22:23 – 40.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 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. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 961–970.

Alla Rozovskaya and Dan Roth. 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. Association for Computational Linguistics.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202, Jeju Island, Korea.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sam Waldron, Nenagh Kemp, Beverly Plester, and Clare Wood. 2015. *Texting Behavior and Language Skills in Children and Adults*.

Sam Waldron, Clare Wood, and Nenagh Kemp. 2017. Use of predictive text in text messaging over the course of a year and its relationship with spelling, orthographic processing and grammar. *Journal of Research in Reading*, 40(4):384–402.

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

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.