Negative language transfer in learner English: A new dataset

Leticia Farias Wanderley **Nicole Zhao** **Carrie Demmans Epp**
EdTeKLA Research Group
Department of Computing Science
University of Alberta
{fariaswa,zzhao1,cdemmansepp}@ualberta.ca

Abstract

Automatic personalized corrective feedback can help language learners from different backgrounds better acquire a new language. This paper introduces a learner English dataset in which learner errors are accompanied by information about possible error sources. This dataset contains manually annotated error causes for learner writing errors. These causes tie learner mistakes to structures from their first languages, when the rules in English and in the first language diverge. This new dataset will enable second language acquisition researchers to computationally analyze a large quantity of learner errors that are related to language transfer from the learners' first language. The dataset can also be applied in personalizing grammatical error correction systems according to the learners' first language and in providing feedback that is informed by the cause of an error.

1 Introduction

English has become an international language. It is the lingua-franca that unites native speakers of other languages around the world (Lysandrou and Lysandrou, 2003). For that reason, it is not hard to believe that the teaching of English as a Second Language[1] has caught a lot of attention from the research community (Caine, 2008). Over the years, computational linguistics researchers have collected corpora containing text written by language learners. These corpora have made possible several advances in language teaching, such as automatic writing assessment (Rahimi et al., 2017; Yannakoudakis et al., 2011) and automatic error detection and correction (Chollampatt et al., 2016; Nadejde and Tetreault, 2019; Omelianchuk et al., 2020).

Although learner corpora are used to model grammatical error correction systems, they are not as often employed in the enhancement of learner feedback. Language learners benefit from direct corrective feedback (Sheen, 2007). Moreover, feedback that makes them reflect upon their errors and distinguish a cause for their mistakes correlates to increased performance (Demmans Epp and McCalla, 2011; Sheen, 2007; Shintani and Ellis, 2013; Karim and Nassaji, 2020). In this paper, we introduce a learner English dataset enhanced with error cause information and concrete examples of learner errors that relate to the learners' first language. It has the potential to help create computational models that provide personalized feedback to English language learners based on the learners' native languages. This new dataset can be accessed by following the instructions described in our research group's repository[2].

The dataset presented in this paper contains supplementary explanations for errors made by Chinese native speakers when writing in English. Chinese learners represent a growing share of the English as a Second Language market. A nationwide language survey from the Chinese government reports that at the beginning of 2001 at least one third of China's population was learning a new language and out of those, 93% were learning English (Wei and Su, 2012). These numbers have only seemed to increase in recent years. The latest survey of international students in the US that was conducted by the Institute of International Education (2020) shows that 35% of these students come from China. With that in mind, it is reasonable to say that this large portion of English learners can benefit from receiving personalized feedback on their writing errors.

[1] Throughout this manuscript, we use the term second language to refer to any additional language beyond the mother tongue, whether the speaker is in a second or foreign language learning context.

[2] https://github.com/EdTeKLA/LanguageTransfer

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3129–3142
June 6–11, 2021.

1.1 Grammatical error correction

One computational task that can benefit from the contrast between first (L1) and second language (L2) is Grammatical Error Correction (GEC). In this task, the objective is to find and correct grammatical errors in learner text (Ng et al., 2013, 2014; Bryant et al., 2019). Since the GEC task was introduced in 2013, many types of grammatical errors have been added. The BEA-2019 Shared Task upgraded the task's error pool by adding new test sets containing essays written by learners from a more diverse set of nationalities. This update is meaningful as it exposes the GEC models to a more general set of error types. In the previous tasks, the essays analyzed were written by South-East Asian students and due to that, the distribution of grammatical error types in the dataset was skewed towards that group's most common mistakes (Bryant et al., 2019).

Grammatical error correction research shows that GEC systems benefit from L1 specific learner data. Rozovskaya and Roth (2011) used L1 specific learner data to adapt a Naïve Bayes GEC system. They applied priors extracted from L1 specific learner error distributions to improve the correction of preposition replacement errors. Chollampatt et al. (2016) used L1-specific data from Russian, Spanish, and Chinese learners to adapt a general GEC model. The resulting adapted models outperformed their general counterpart. Nadejde and Tetreault (2019) expand on this topic by adapting general GEC models to L1-specific and proficiency-specific learner data. Their experimental setup covered twelve different L1s and five proficiency levels. Both L1 and proficiency adaptations outperformed the baseline, and the models which achieved the best performance were the ones that were adapted to both features at the same time.

1.2 Writing feedback

Direct corrective feedback, such as grammatical error correction, helps language learners improve their writing proficiency (Liaqat et al., 2020; Sheen, 2007). In addition to that, feedback that contrasts erroneous utterances with correct ones facilitates the acquisition of accurate language structures. This facilitation occurs both when the feedback is applied to L1-transfer and non-transfer errors (Tomasello and Herron, 1989). Fine-tuning error feedback by contrasting L1 and L2 has been shown to increase learners' language understanding and awareness (Kupferberg, 1999; Han, 2001).

Advances in learner data annotation foster language transfer research by providing details that can be used to inform the contrast between learners' L1s and L2s and possibly further explain incorrect learner utterances. Highlighting this contrast is beneficial to learners as it has the potential of increasing their metalinguistic awareness. That is, it can improve the learners' capacity to think about language as an object. It supports their ability to recognise the mismatch between their L1 and L2 as well as their ability to refrain from incorrectly using L1 rules in L2 utterances (Wanderley and Demmans Epp, 2020).

Considering the importance of feedback for learners, Nagata (2019) introduced the task of feedback comment generation. In this task, the objective is to automatically generate feedback for learner essays. Along with this new task, the author introduced a dataset that contains learner essays and their respective annotated feedback. The annotation available in this new dataset contains feedback regarding preposition usage errors and text organization. It also contains annotation samples in which the feedback praises the learners' writing. While our annotation procedure focused on annotating Chinese L1 learner errors, with a special focus on whether those errors were related to negative language transfer, our datasets may complement the one described by Nagata (2019). As, ultimately, both efforts aim to provide more personalized feedback to language learners.

1.3 Native language identification

The differences between L1s and English can provide valuable features that help identify learners' L1s. Information about learner errors and their association with the learners' L1s can be useful in tasks such as native language identification. This task takes advantage of latent signals in non-native written data to identify the authors' L1s (Tetreault et al., 2013). Wong and Dras (2009) apply the contrastive analysis hypothesis (Lado, 1957), which correlates the learner's more common errors to the divergences between L1 and L2 in a native language identification task. They analyzed three types of syntactic errors and found evidence that the contrastive analysis hypothesis can aid in L1 detection. The distribution of learner errors alone can also be employed in native language Iidentification. Flanagan et al. (2015) showed that writing

error patterns performed well as features in the prediction of learners' native languages.

1.4 Negative language transfer

The correlation between L1s and writing error patterns happens because language learners, sometimes unknowingly, use certain strategies when they are learning how to communicate in a new language. One of those strategies is called language transfer. Language transfer, or cross-linguistic effects, is a subject that has been studied since 1957 when Robert Lado defined the phenomenon and its effects on second language acquisition (Lado, 1957). According to Lado, second language learners rely on their first languages when forming utterances in the second language. They tend to transfer morphological, syntactical, and semantic paradigms that they are accustomed to from their L1 when attempting to communicate in the L2. When learners transfer patterns from their L1 and those patterns are not valid in the L2, it results in negative language transfer. Since Lado's book was published, language transfer evidence has consistently been reported by language teachers, linguists, and second language acquisition researchers (Swan and Smith, 2001). This body of evidence supports the theory that learners' L1s influence their L2 learning.

1.5 Learner data

English learner data is amply available online, especially due to endeavours like the aforementioned native language identification and grammatical error correction tasks. However, it is considerably more difficult to find learner data that highlights the differences between the learners' L1s and English, and how these differences influence learners' mistakes. Learner English lacks large and accessible corpora like the MERLIN corpus, a dataset of Italian, German, and Czech learner essays in which errors are annotated with several characteristics of learner language and their potential causes (Boyd et al., 2014). This corpus contains features derived from sources such as language teachers, reference textbooks, and second language acquisition research. Some of these features (e.g., capitalization errors by German native speakers and negation errors in Czech) can be associated with the learner's L1 (Boyd et al., 2014).

There have been efforts to enhance English learner data. Meaningful work that provides syntactic analyses for learner English was introduced by Berzak et al. (2016); they created a manually annotated syntactic treebank for learner English. The Treebank of Learner English they created aims to facilitate second language acquisition research and research on the processing of ungrammatical language. It contains part-of-speech tags and dependency parse trees for erroneous learner English sentences, as well as the same features for their corrected counterparts.

Although computational tasks and previous research on learner English have shared several learner datasets, these datasets do not contain information about linguistic phenomena, such as negative language transfer. It is well-known that learner error patterns and L1 versus English distinctions can aid both computational tasks and language learning, e.g., Nadejde and Tetreault (2019); Flanagan et al. (2015); Karim and Nassaji (2020). In the present paper, we introduce an enhanced learner English dataset, manually annotated with error cause features that highlight the differences between English and the learners' L1, Chinese. The goal of this dataset is to inform learner error feedback with metalinguistic details that can aid learning and to support computational linguistics tasks that take into account native language influence on learner English.

2 Creating a negative language transfer dataset

2.1 The FCE dataset

The negative language transfer annotation proposed in this paper builds on the collection of error annotated learner essays described by Yannakoudakis et al. (2011). These essays were written by English as a Second Language learners while taking the First Certificate in English (FCE) test, an upper-intermediate level English certification. The dataset contains essays written between the years 2000 and 2001 by 1244 distinct learners. Each essay in the dataset contains the answers to two FCE writing tasks. There is one essay script per learner amounting to 1244 essays in the dataset. Each script has on average 409 words (SD = 96). In total, the essays contain more than 500K words.

Each essay in the FCE dataset was manually annotated with the learners' errors. These errors are categorized with error types that follow the error coding described by Nicholls (2003). Most errors in the dataset are also accompanied by corrections suggested by the annotators. The few er-

Incorrect utterance	Correct utterance	Negative language transfer?	Likely reason for the mistake
This are only my immature views.	These are only my immature views.	Yes	Used singular form instead of plural form
In a result of this...	As a result of this...	No	Chinese doesn't use the word in this context but learner included it

Table 1: Negative language transfer and error cause annotation examples for a speaker of Chinese

Likely reason for the mistake	Description
Used singular form instead of plural form	Chinese does not mark plurals in most pronouns other than the anaphoric ones. Hence, determiner type pronouns are not inflected properly. Usually learners will use the singular form of the determiner (this would be negative transfer) but other times they use the plural (not negative transfer) form.
Chinese doesn't use the word in this context but learner included it	The structure of the translation of this phrase in Chinese does not include prepositions. Sometimes the learner will use a random preposition that they feel would fit the context, instead of just omitting it.
Chinese uses commas to mark the end of a complete thought	Unlike English, commas in Chinese are added only to aid in comprehension and are not actually required. Chinese commas mark a change in thought but continuation in topic, similar to a period marking the end of a sentence. Sentences containing subordinate clauses are seen as "one thought" and hence do not need any punctuation like a comma to separate them.
Overcorrection (along with unnecessary pronoun errors)	Chinese uses pronouns less than English so learners will overcompensate by using pronouns in places where they feel like there should be one.

Table 2: Error cause description examples

rors which are not accompanied by corrections are the ones that caused the FCE annotators to be uncertain about their appropriate correction. Along with the error annotation, the dataset includes metadata such as the learners' L1, age range, essay score, and overall exam score. Sixteen different L1s are represented in the FCE dataset.

2.2 Negative language transfer dataset

There are 66 essays written by 66 distinct Chinese native speakers in the FCE dataset. These essays amount to a total of 30K words. Each essay contains on average 468 words (SD = 101).

We enhanced the essays written by Chinese native speakers in the FCE dataset by adding information that associates the learners' L1 rules to the annotated writing errors. Each error in this subset of FCE essays is classified as being related to language transfer or not. For an error to be categorized as negative language transfer, there has to be concrete evidence that English and Chinese rules diverge for that specific sentence structure. The categorization of an error as negative language transfer is an indicator that the error was the learner's attempt, conscious or not, to apply one or more L1 rules while writing in English. Along with the binary negative language transfer classification, each error in this dataset is annotated with a possible reason for its occurrence. Whether that reason is related to language transfer or not, all errors are accompanied by a short sentence describing one of their possible causes. Table 1 presents examples of learner errors, their negative language transfer label, and possible error causes.

The FCE dataset augmented with error cause annotations as described above is complemented by a new learner English dataset. This dataset catalogues the error cause categories and provides more substantial descriptions for each error cause, as well as exemplar sentences in English and in the learner's L1 that highlight the different language rules possibly related to the mistake. The error cause categories used in this dataset are the same as the ones used in the FCE dataset error cause annotations. Maintaining this link means that if a

	Accompanied by corrections	Not accompanied by corrections	Total
Negative language transfer errors	1797	94	1891
Not negative language transfer errors	1276	113	1389
Spelling errors	292		292
Omitted errors	12		12
Combined	3377	207	3584

Table 3: Descriptive statistics for the negative language transfer dataset

Error type	Error description	Total	Negative language transfer	Not negative language transfer
RP	replace punctuation	336	228 (67.86%)	108 (32.14%)
TV	incorrect tense of verb	267	185 (69.29%)	82 (30.71%)
RV	replace verb	230	81 (35.22%)	149 (64.78%)
MD	missing determiner	209	206 (98.56%)	3 (1.44%)
RT	replace preposition	209	118 (56.46%)	91 (43.54%)

Table 4: Distribution of negative language transfer errors across the most frequent error categories. An extended version of this table containing all error types is available in Appendix A

user needs more information about a specific error cause, they can consult the error cause descriptions in this dataset and find more details and analyses regarding the error in question. Table 2 provides error cause exemplars and their respective descriptions in the new dataset.

In total, 269 possible error causes have been identified for the errors made by Chinese native speakers. Each possible error cause in the dataset occurs on average 11 times (SD = 26); 110 of the error causes were only found once. The most common negative language transfer error cause was "Chinese uses commas to mark the end of a complete thought". This error cause occurs 270 times and refers to the disparity in punctuation usage patterns between English and Chinese — an example of negative language transfer. The non-negative language transfer error cause that is most frequent in the dataset is "Overcorrection", found 186 times. This possible error cause indicates that learners may have used known English patterns where they were not necessary, in a failed attempt to conform to English grammatical rules.

2.3 Dataset statistics

Table 3 presents the statistics of the negative language transfer dataset. There are 3584 errors in the Chinese L1 dataset. Of those errors, 52.76% are tagged as negative language transfer and 38.76% are tagged as non-transfer errors. The remaining 8.48% were left unlabelled in the dataset for one of two reasons: they were spelling errors or they were omitted due to, for example, the correction proposed not being enough to amend the error or the error being tagged as incorrect because of an English variety divergence, e.g., the learner sentence was correct according to American English rules but not according to British English rules.

Among the learner errors that received a negative language transfer annotation, it is important to make a distinction between errors that are accompanied by corrections and errors that are not. The FCE dataset annotation scheme allowed annotators to highlight errors by enclosing them with "" and "" tags. It also instructed that the suggested corrections for those errors should be enclosed in "<c>" and "</c>" tags. In some situations, the FCE annotators were unsure about the appropriate correction for an error and, hence, did not suggest edits (Bryant, 2019). In these situations, the annotators simply highlighted the errors using "<NS>" and "</NS>" tags. Although these errors are annotated with negative language transfer and error cause information in our dataset, they are kept separate from the other errors due to them not containing any information about error correction. There are 207 errors made by Chinese native speakers that are not accompanied by edits in the FCE dataset. Out of those, 94 are related to negative language transfer and 113 are not.

Each error from the FCE dataset is annotated with an error type. Table 4 presents the negative

language transfer statistics across the most common error types in the dataset. By investigating these types, it is possible to detect recognizable patterns from language that Chinese learners of English use. One of the most problematic grammatical structures for Chinese native speakers, when writing in English, is the placement of determiners before noun phrases. As the Chinese language does not have determiners, Chinese learners have trouble deciding when to use them and when to refrain from using determiners in their writing (Han et al., 2006). This fact is reflected in the proportion of missing determiner (MD) errors that are labelled as negative language transfer in the dataset. Out of the 209 MD errors, 206 (98.56%) are labelled as transfer related errors. An example of a non-negative transfer error for MD is where the learner omits a determiner which specifies the subject, for instance

"I want to ask for [my] money back"

where the word "my" is the missing possessive determiner. Generally speaking, in Chinese, the word "my" (我的 Pinyin: wǒ de) is also used in formal settings. However, it can be omitted to shorten sentences in informal settings. Therefore, this error is not classified as a negative transfer error.

On the other hand, there are error types in the dataset that are rarely associated with negative language transfer. Errors involving the unnecessary usage of determiners, for example, are not related to negative language transfer. They are a result of learners overusing an L2 grammatical structure by placing it where it is not needed (Smith, 1982). Replacement errors, i.e., errors in which the erroneous word needs to be replaced by another word from the same category, tend to be distributed more evenly between negative language transfer and not negative language transfer. These errors are labelled as not negative language transfer when the erroneous structure used by the learner has no parallel in Chinese. That is, it is not possible that the learner is reusing an L1 structure because the structure used only occurs in English.

3 Annotation procedure

3.1 Annotators

The FCE dataset errors were grouped by learner L1 and each error was annotated by one annotator. The Chinese errors' annotator is a native speaker of Mandarin Chinese and English who teaches Chinese as a foreign language. She is also able to read and write in both languages, with higher proficiency writing in English. She speaks multiple dialects of Mandarin originating from South-East China. Furthermore, she has taken linguistics courses on English syntax.

3.2 Annotation

The annotator had access to a dataset containing all the errors made by Chinese native speakers. To facilitate the annotation process, the errors in the datasets were further grouped by error type. The annotator then worked on one error category at a time. For example, she analyzed and annotated all the "wrong verb tense" errors in the dataset before moving on to another error category. This procedure helped keep the annotator focused on a small number of grammatical structures at a time which aided the recognition of common error patterns. In fact, this structured use of error types is one of the reasons presented by Nicholls (2003) for the addition of error type features to learner English datasets.

Beyond the grammatical errors and their types, the annotator had access to more information about the errors, such as the context surrounding the erroneous utterance and the Extensible Markup Language (XML) data extracted from the FCE dataset. These two features proved useful to elucidate semantic errors. A semantic error initially looks like an annotation error, since the utterance's grammatical structure is not problematic. However, when the annotator checked the context around the error, they would often find its cause to be context-related. In the sentence "I have never been to in my life.", the word "never" does not seem incorrect, although it is tagged as such. By looking at the context surrounding this error, "It was the worst show and theatre I have never been to in my life.", it is possible to see that indeed the word "never" should be replaced with the word "ever".

3.3 Ambiguous cases

During the annotation procedure, ambiguous cases were discussed and reviewed among the annotator and the research group in weekly meetings. The annotator highlighted entries that she found hard to label and those were discussed within the group. Such cases ranged from entries that were deemed erroneous by the FCE annotators due to language variety (e.g., British or American idioms), entries that did not have an equivalent structure in the learners' L1 (e.g., hyphenated words, which do not exist in Chinese), and semantic errors (e.g., errors in

	Incorrect utterance	Correct utterance	Ambiguity	Number of cases
British vs American English varieties	We **all would** like to go there.	We **would all** like to go there.	The incorrect version of the sentence is more commonly used in the American variety of English. It is not incorrect in that variety.	18
Chinese does not have an equivalent structure	I'm standing on your **left hand** side.	I'm standing on your **left-hand** side.	Hyphens do not have a parallel structure in Chinese.	17
Semantic errors tagged as structural errors	You **could** find a restaurant.	You **can** find a restaurant.	Although the verb "could" is in the past tense, some learners may choose to use it to indicate respect.	10

Table 5: Ambiguous errors from the FCE dataset

which the grammatical structure is not incorrect, but the utterance does not fit the overall essay context). Table 5 presents examples of errors that were discussed during the annotation process. These errors are considered ambiguous with regards to whether they should be labelled as transfer related.

4 Annotation scheme

The annotation scheme was designed to highlight the relationship between the learner error and the learner's L1. Other than the boolean label representing whether an error is related to negative language transfer, each entry carries information about the possible reason behind that learner mistake. Even when the relationship between the error and the learner's L1 is not apparent, the annotation scheme will provide a possible cause for the error. This cause is not related to language transfer.

The error cause feature was heavily influenced by language teacher guides, books that aim to make teachers aware of the learner errors they can encounter in the classroom, e.g., "Learner English: A Teacher's Guide to Interference and other Problems" by Swan and Smith (2001). Guides like these have been written based on years of in-classroom experience and contain information about error causes along with potential learner feedback. These guides were used as a baseline for negative language transfer detection during the annotation process. Other important sources of guidance for the error cause feature annotation were Chinese and English grammar books and guides[3] (Faigley, 2015; Li and Thompson, 1989). These sources allowed the direct contrast of erroneous utterances with language rules and this contrast enabled the derivation of possible causes for learner mistakes.

[3] https://www.grammarly.com/blog/category/handbook/

5 Dataset application

Second language acquisition researchers and language teachers are well acquainted with learner errors that are related to the learners' L1s. These communities have produced comprehensive guides to learner errors and their causes. Learner language guides allow learners and teachers to identify the reasons behind certain error types and, with that, better understand and prevent those mistakes. In some of these guides the reader will find information that connects learners' L1s with common error types committed by native speakers of that language. Our new dataset, enables the use of other indicators, such as linguistic features and proficiency levels, to identify errors related to negative language transfer.

To understand the effect of linguistic features in negative language transfer prediction, we built classification models to predict when a learner error is related to negative language transfer. We wanted to explore the relationship between negative language transfer and the linguistic features of errors, such as part-of-speech (POS) tags and dependency labels, since these features are made available by this new dataset.

5.1 Negative language transfer classification

In this experiment, we used the new negative language transfer dataset to compare the predictive

Incorrect utterance	Error length	Error type	POS tags
This are only my immature views.	1	AGA (anaphoric pronoun agreement error)	DT VBP RB

Table 6: Best performing feature set for the random forest classifier

power of classification models for negative transfer[4]. These models are trained on error features from the new negative language transfer dataset. The models output whether the errors are related to negative transfer. This is a binary classification problem in which most of the features available are categorical. For this reason, we converted the categorical features, such as error types, into one-hot-encoding columns and binary vectors. The one-hot-encoding conversion creates one new column on the dataset for each unique categorical value. The binary vector conversion creates one new column on the dataset containing a binary number in which the position of the digit one corresponds to the category of the entry.

The conversion of categorical features into one-hot-encoding columns increased the number of dimensions in our data. Hence, we decided to experiment with a random forest classifier, a classification model that is known to perform well with high-dimensional data (Xu et al., 2012). For our baseline model we decided to use a logistic regression model trained only on the error type data features. This choice was based on the parallel that can be drawn between the dataset's error type information and the teacher guide descriptions of connections between L1s and specific error patterns. A strong baseline for the experiment relies solely on error types to predict negative language transfer. Both classifiers were trained using the models available in the Python library scikit-learn[5].

Since the new dataset contains actual learner writing, it is possible to extract a wide range of linguistic features from the sentences in the dataset. We used the Python library spaCy[6] to extract dependency labels, Universal Dependencies POS tags (Nivre et al., 2016), and Penn Treebank POS tags (Marcus et al., 1993) from the erroneous utterances and their surrounding tokens. These features were then converted to one-hot-encoding columns and binary vectors, as described above.

Given the wide range of features in the new dataset, we performed an initial step of feature selection to determine the most relevant features to predict the negative language transfer label. The feature selection process consisted in performing 10-fold cross validation with 90% of the dataset as training data. The remaining 10% was held out for testing. We performed the cross validation on all feature set combinations training a random forest classifier on nine folds and testing it on the remaining one. The mean score for each feature set combination was used to select the best performing set. The best performing model in the feature selection process was trained with three features: the error length (the number of words in the error), the error type (described in Nicholls (2003)), and the Penn Treebank POS tags of the erroneous utterance plus the POS tags of the error's two subsequent words. Table 6 presents an example of the features selected. The columns "Error type" and "POS tags trigram" were converted into one-hot-encoding columns during the feature selection, training, and testing processes but are presented here as categorical data for intelligibility.

	Acc	P	R
Logistic regression baseline	0.72	0.79	0.73
Random forest model	**0.78**	**0.82**	**0.79**

Table 7: Accuracy, precision, and recall results on the held-out test set for the baseline logistic regression model trained with the error type information and the random forest classifier trained with the best performing feature set

After feature selection, a random forest classifier was trained on the three best performing features using 90% of the dataset, i.e., 2952 error instances. This model accurately classified 78.04% of the test samples as negative language transfer or not. The baseline model, a logistic regression model trained on the error type features, achieved 72.56% accuracy on the test set. Table 7 presents the accuracy, precision, and recall scores yielded by both baseline and random forest models on the test set, which contained 328 error instances.

Analyzing the models' outputs, it becomes clear

[4]The experimental code is available in our research group's repository (see footnote 2).

[5]https://scikit-learn.org/

[6]https://spacy.io/

that more information about learners' incorrect utterances captures more of the language transfer phenomenon. One of the most common baseline model misclassifications was the prediction of "replace punctuation" errors as negative language transfer when they were not negative transfer related. The model misclassified 30% of the "replace punctuation" errors. Although Chinese learners are known to replace periods with commas incorrectly (Liu, 2011), the error category by itself is not enough to make an accurate classification. In contrast, the random forest classifier mislabelled 16% of the "replace punctuation" errors. The random forest approach misclassified entries as negative transfer and not negative transfer related, demonstrating that this approach does not simply associate one error type with one output label. Another error category in which the random forest classifier outperformed the baseline model was when classifying the "wrong verb tense" errors. The error type features do not provide information about the verb tense that was used incorrectly, but the POS tags extracted from the incorrect utterance do. This extra information helps the random forest classifier make more accurate predictions about negative language transfer related errors.

These results suggest that negative language transfer classification can benefit from features other than the error type. Furthermore, it shows that linguistic features are important in the identification of negative language transfer errors.

6 Conclusion

Our dataset is the first we are aware of that annotates a large amount of learner English data with negative language transfer features and error causes. It has the potential to improve the performance of computational linguistics tasks, such as native language transfer identification and grammatical error correction. More importantly, its content can benefit English teachers and learners by making more personalized error feedback available. Another potential application of the dataset is in the automatic detection of negative language transfer. This application could help provide real-time L1-informed feedback to English learners.

Our research group is currently working on annotating errors from other L1 learner groups. We are also expanding our annotation process to learner data from sources other than the FCE dataset, such as the Lang-8 English corpus (Mizumoto et al., 2011). With that, we hope to broaden the scope of English learners supported by L1 informed error feedback.

Ethical considerations

The new datasets presented in this paper are built on top of the FCE dataset described in Yannakoudakis et al. (2011). The FCE dataset contains anonymised essays from the First Certificate in English test-takers between the years of 2000 and 2001. The essays' meta-data contains information about the age range and native language of the learners. Although the original dataset description does not address how the learners' consent was obtained, Cambridge Assessment should be governed by the same consent procedures as other UK researchers. The candidate privacy policy[7] from the Cambridge Assessment website states that the learners' data could be used in "developing and delivering publications and other resources that support learning".

The annotation procedure described in this paper was performed by undergraduate and graduate students as part of individual project courses and research assistantships, respectively. All three authors[8] are fluent in at least one variety of English. Two of them also have deep familiarity with other English varieties. This knowledge was used to ensure that the negative language transfer annotations do not reinforce existing power structures around language varieties and standard forms of English. That said, cases of linguistic imperialism are likely to remain in the dataset.

Another facet that may limit these datasets' applicability is the fact that the FCE annotated essays were collected 20 years ago. As both the Chinese and English languages have evolved and, possibly intersected over time, the occurrence of some negative language transfer errors may have decreased. For example, the prevalence of determiner omis-

[7] https://www.cambridgeenglish.org/ch/fr/footer/data-protection/candidates/

[8] **Demmans Epp** has specialized in computer-assisted language learning and she has taught English as a second or foreign language in a variety of educational contexts. She has training in several areas of linguistics that include language acquisition and sociolinguistics. She is a first language speaker of Canadian English and has experience working in American and British English contexts.
Wanderley speaks Portuguese as her first language and is proficient in English. She has experience working in British and Canadian English contexts.
Zhao grew up speaking Mandarin Chinese and is familiar with various dialects of Chinese. She received her education in American and Canadian English and is familiar with British English to a certain extent.

sion errors in Chinese L1 English communication has been targeted from an instructional perspective in recent years which may have lowered the occurrence rate of this negative language transfer error. On the English language front, the usage of the pronoun "they" has changed and may have rendered some of the entries in our dataset obsolete.

The new dataset's main purpose is to aid English language learning by providing personalized error causes according to the learner's L1. It aims to help English as a Second Language learners acquire a better understanding of the English language by contrasting it to the learner's L1. Although this type of information tends to be helpful to language learners, there might be learners who do not benefit from it. The data available in the dataset was reviewed by our research group to ensure clarity and correctness. We do not foresee additional risks stemming from the usage of the new dataset.

Acknowledgements

We acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2018-03834], and the Social Sciences and Humanities Research Council (SSHRC). We would also like to acknowledge the anonymous reviewers for their insightful and valuable feedback.

References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–1288, Reykjavik, Iceland. European Language Resources Association (ELRA).

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 Jack Bryant. 2019. *Automatic annotation of error types for grammatical error correction (Doctoral thesis)*. Ph.D. thesis, University of Cambridge.

Tonje M Caine. 2008. Do you speak global?: The spread of English and the implications for English language teaching. *Canadian Journal for New Scholars in Education/Revue canadienne des jeunes chercheures et chercheurs en éducation*, 1(1).

Shamil Chollampatt, Duc Tam Hoang, and Hwee Tou Ng. 2016. Adapting grammatical error correction based on the native language of writers with neural network joint models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1901–1911, Austin, Texas. Association for Computational Linguistics.

Carrie Demmans Epp and Gordon McCalla. 2011. Protutor: Historic open learner models for pronunciation tutoring. In *Artificial Intelligence in Education*, pages 441–443, Berlin, Heidelberg. Springer Berlin Heidelberg.

L. Faigley. 2015. *Little Penguin Handbook*. Pearson Australia.

Brendan Flanagan, Chengjiu Yin, Takahiko Suzuki, and Sachio Hirokawa. 2015. Prediction of learner native language by writing error pattern. In *Learning and Collaboration Technologies*, pages 87–96, Cham. Springer International Publishing.

Nan Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, 12.

Zhao Hong Han. 2001. Fine-tuning corrective feedback. *Foreign Language Annals*, 34.

Institute of International Education. 2020. International student totals by place of origin, 2000/01-2019/20.

Khaled Karim and Hossein Nassaji. 2020. The revision and transfer effects of direct and indirect comprehensive corrective feedback on esl students' writing. *Language Teaching Research*, 24(4):519–539.

Irit Kupferberg. 1999. The cognitive turn of contrastive analysis: Empirical evidence. *Language Awareness*, 8.

Robert Lado. 1957. *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. University of Michigan Press.

C.N. Li and S.A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. Number v. 3 in Linguistics: Asian studies. University of California Press.

Amna Liaqat, Cosmin Munteanu, and Carrie Demmans Epp. 2020. Collaborating with Mature English Language Learners to Combine Peer and Automated Feedback: a User-Centered Approach to Designing Writing Support. *International Journal of Artificial Intelligence in Education*, pages 1–42.

Xing Liu. 2011. The not–so–humble "Chinese comma": Improving English CFL students' understanding of multi–clause sentences. In *Proceedings of the 9th New York International Conference on Teaching Chinese*.

Photis Lysandrou and Yvonne Lysandrou. 2003. Global English and proregression: understanding English language spread in the contemporary era. *Economy and Society*, 32(2):207–233.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Maria Nadejde and Joel Tetreault. 2019. Personalizing grammatical error correction: Adaptation to proficiency level and L1. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 27–33, Hong Kong, China. Association for Computational Linguistics.

Ryo Nagata. 2019. Toward a task of feedback comment generation for writing learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China. Association for Computational Linguistics.

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.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA â†' Online. Association for Computational Linguistics.

Zahra Rahimi, Diane Litman, Richard Correnti, Elaine Wang, and Lindsay Clare Matsumura. 2017. Assessing students' use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA. Association for Computational Linguistics.

Younghee Sheen. 2007. The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles. *TESOL Quarterly*, 41(2):255–283.

Natsuko Shintani and Rod Ellis. 2013. The comparative effect of direct written corrective feedback and metalinguistic explanation on learners' explicit and implicit knowledge of the English indefinite article. *Journal of Second Language Writing*, 22(3):286–306.

Karen L. Smith. 1982. Avoidance, overuse, and misuse: Three trial and error learning strategies of second language learners. *Hispania*, 65.

Michael Swan and Bernard Smith. 2001. *Learner English: A Teacher's Guide to Interference and Other Problems*, 2 edition. Cambridge Handbooks for Language Teachers. Cambridge University Press.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the Eighth*

Workshop on Innovative Use of NLP for Building Educational Applications, pages 48–57, Atlanta, Georgia. Association for Computational Linguistics.

Michael Tomasello and Carol Herron. 1989. Feedback for Language Transfer Errors The Garden Path Technique. *Studies in Second Language Acquisition*, 11.

Leticia Farias Wanderley and Carrie Demmans Epp. 2020. Identifying negative language transfer in writing to increase English as a Second Language learners' metalinguistic awareness. In *Companion Proceedings 10th International Conference on Learning Analytics & Knowledge (LAK20)*.

Rining Wei and Jinzhi Su. 2012. The Statistics of English in China. *English Today*, 28:10–14.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia.

Baoxun Xu, Joshua Zhexue Huang, Graham Williams, Mark Junjie Li, and Yunming Ye. 2012. Hybrid random forests: Advantages of mixed trees in classifying text data. In *Advances in Knowledge Discovery and Data Mining*, pages 147–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

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. Association for Computational Linguistics.

A Distribution of negative language transfer errors across all error types

Error type	Error description	Total	Negative language transfer	Not negative language transfer
AGA	anaphor agreement error	19	2 (10.53%)	17 (89.47%)
AGD	determiner agreement error	4	0 (0.00%)	4 (100.00%)
AGN	noun agreement error	62	41 (66.13%)	21 (33.87%)
AGQ	quantifier agreement error	7	7 (100.00%)	0 (0.00%)
AGV	verb agreement error	88	75 (85.23%)	13 (14.77%)
AS	argument structure error	18	6 (33.33%)	12 (66.67%)
CD	wrong determiner because of noun countability	1	1 (100.00%)	0 (0.00%)
CE	complex error	4	2 (50.00%)	2 (50.00%)
CL	collocation or tautology error	4	0 (0.00%)	4 (100.00%)
CN	countability of noun error	12	10 (83.33%)	2 (16.67%)
CQ	wrong quantifier because of noun countability	4	3 (75.00%)	1 (25.00%)
DA	derivation of anaphor error	20	1 (5.00%)	19 (95.00%)
DD	derivation of determiner error	12	0 (0.00%)	12 (100.00%)
DJ	derivation of adjective error	50	43 (86.00%)	7 (14.00%)
DN	derivation of noun error	35	23 (65.71%)	12 (34.29%)
DV	derivation of verb error	7	5 (71.43%)	2 (28.57%)
DY	derivation of adverb error	19	8 (42.11%)	11 (57.89%)
FA	wrong anaphor form	2	1 (50.00%)	1 (50.00%)
FD	incorrect determiner form	13	0 (0.00%)	13 (100.00%)
FJ	wrong adjective form	4	3 (75.00%)	1 (25.00%)
FN	wrong noun form	93	73 (78.49%)	20 (21.51%)
FV	wrong verb form	132	58 (43.94%)	74 (56.06%)
FY	wrong adverb form	1	1 (100.00%)	0 (0.00%)
ID	idiom wrong	17	1 (5.88%)	16 (94.12%)
IJ	incorrect adjective inflection	2	0 (0.00%)	2 (100.00%)
IN	incorrect noun inflection	9	5 (55.56%)	4 (44.44%)
IQ	incorrect quantifier inflection	1	1 (100.00%)	0 (0.00%)
IV	incorrect verb inflection	13	0 (0.00%)	13 (100.00%)
L	inappropriate register	12	5 (41.67%)	7 (58.33%)
M	missing error	61	42 (68.85%)	19 (31.15%)
MA	missing anaphor	62	41 (66.13%)	21 (33.87%)
MC	missing link word	20	17 (85.00%)	3 (15.00%)
MD	missing determiner	209	206 (98.56%)	3 (1.44%)
MJ	missing adjective	4	2 (50.00%)	2 (50.00%)
MN	missing noun	18	8 (44.44%)	10 (55.56%)
MP	missing punctuation	151	138 (91.39%)	13 (8.61%)
MQ	missing quantifier	3	2 (66.67%)	1 (33.33%)
MT	missing preposition	73	67 (91.78%)	6 (8.22%)
MV	missing verb	54	47 (87.04%)	7 (12.96%)
MY	missing adverb	13	12 (92.31%)	1 (7.69%)
R	replace error	187	101 (54.01%)	86 (45.99%)
RA	replace anaphor	33	10 (30.30%)	23 (69.70%)
RC	replace link word	10	5 (50.00%)	5 (50.00%)
RD	replace determiner	42	18 (42.86%)	24 (57.14%)
RJ	replace adjective	41	18 (43.90%)	23 (56.10%)

Error type	Error description	Total	Negative language transfer	Not negative language transfer
RN	replace noun	123	54 (43.90%)	69 (56.10%)
RP	replace punctuation	336	228 (67.86%)	108 (32.14%)
RQ	replace quantifier	14	7 (50.00%)	7 (50.00%)
RT	replace preposition	209	118 (56.46%)	91 (43.54%)
RV	replace verb	230	81 (35.22%)	149 (64.78%)
RY	replace adverb	66	24 (36.36%)	42 (63.64%)
S	spelling error	1	0 (0.00%)	1 (100.00%)
TV	incorrect tense of verb	267	185 (69.29%)	82 (30.71%)
U	unnecessary error	25	8 (32.00%)	17 (68.00%)
UA	unnecessary anaphor	20	0 (0.00%)	20 (100.00%)
UC	unnecessary link word	14	1 (7.14%)	13 (92.86%)
UD	unnecessary determiner	75	4 (5.33%)	71 (94.67%)
UJ	unnecessary adjective	6	2 (33.33%)	4 (66.67%)
UN	unnecessary noun	9	8 (88.89%)	1 (11.11%)
UP	unnecessary punctuation	65	6 (9.23%)	59 (90.77%)
UQ	unnecessary quantifier	3	1 (33.33%)	2 (66.67%)
UT	unnecessary preposition	59	7 (11.86%)	52 (88.14%)
UV	unnecessary verb	44	16 (36.36%)	28 (63.64%)
UY	unnecessary adverb	16	4 (25.00%)	12 (75.00%)
W	word order error	44	24 (54.55%)	20 (45.45%)
X	incorrect negative formation	8	4 (50.00%)	4 (50.00%)

Table 8: Negative language transfer counts across error types