AUTOMATIC GRAMMATICAL ERROR DETECTION OF NON-NATIVE SPOKEN LEARNER ENGLISH

K.M. Knill[1]*, M.J.F. Gales*[1]*, P.P. Manakul*[1]*, A.P. Caines*[2]

[1]ALTA Institute / Engineering Department
[2]ALTA Institute / Computer Science and Technology Department
Cambridge University, UK
Email: {kate.knill,mjfg,pm574}@eng.cam.ac.uk, apc38@cam.ac.uk

ABSTRACT

Automatic language assessment and learning systems are required to support the global growth in English language learning. They need to be able to provide reliable and meaningful feedback to help learners develop their skills. This paper considers the question of detecting "grammatical" errors in non-native spoken English as a first step to providing feedback on a learner's use of the language. A state-of-the-art deep learning based grammatical error detection (GED) system designed for written texts is investigated on free speaking tasks across the full range of proficiency grades with a mix of first languages (L1s). This presents a number of challenges. Free speech contains disfluencies that disrupt the spoken language flow but are not grammatical errors. The lower the level of the learner the more these both will occur which makes the underlying task of automatic transcription harder. The baseline written GED system is seen to perform less well on manually transcribed spoken language. When the GED model is fine-tuned to free speech data from the target domain the spoken system is able to match the written performance. Given the current state-of-the-art in ASR, however, and the ability to detect disfluencies grammatical error feedback from automated transcriptions remains a challenge.

Index Terms— Spoken language assessment, CALL, grammatical error detection

1. INTRODUCTION

Automatic systems that enable assessment and feedback of learners of a language are becoming increasingly popular. One important aspect of these systems is to provide reliable, meaningful feedback to learners on errors they are making. This feedback can then be used independently, or under the supervision of a teacher, by the learner to improve their proficiency. A growing number of applications are available to non-native learners to improve their English speaking skills by providing feedback on aspects such as pronunciation and fluency. It would be beneficial for a computer assisted language learning (CALL) system to provide the learner with feedback on their use of English, one aspect of which is their choice of grammar. This is a challenging problem. Even for non-learner speech there is still open debate on defining what is correct spoken grammar: we do not generally speak in sentences; we hesitate; repeat ourselves; we do not enunciate clearly; and use more than our words, such as intonation and gesture, to communicate our message. These effects are accentuated in learner speech.

Currently automatic systems to provide feedback on spoken learner grammar are focused on very constrained scenarios and/or the approaches are not suitable for conversational and spontaneous speech e.g. [1, 2, 3]. For written texts, however, there have been significant developments in recent years in the ability to detect (and correct) the full range of grammatical errors partly due to the development of deep learning based approaches [4]. Applications such as Grammarly and Write & Improve[1] help the learner over a broad range of writing styles. An interesting question is whether these systems can work on spoken utterances. This paper presents initial investigations into applying a state-of-the-art grammatical error detection (GED) system [5, 6] to non-native English learner speech from a range of L1s. The system is applied to free speaking tasks, where a learner is engaged in a conversation or talks for up to a minute in response to a prompt.

In free speaking scenarios the learner's spoken language is unknown in advance so the spoken GED system must rely on automatic speech recognition (ASR) to provide a transcription of the learner's speech. The non-grammatical and disfluent learner speech leads to a mismatch with the language model. Combined with an acoustic mismatch due to e.g. pronunciation errors, this leads to poorer ASR performance than for native speakers [7]. This increases the challenge of producing reliable grammatical feedback since reporting on an ASR error will be confusing to the learner. It is therefore important that false positive errors arising from ASR errors are minimised. To assess the effect of relying on ASR output the GED system is investigated on both manual and ASR transcriptions.

The deep-learning-based, sequence-labeller GED system is presented in Section 2. The corpora used for training and test and the experimental setup are described in Sections 3 and 4, respectively. Experimental results are given in Section 5, followed by conclusions in Section 6.

2. GED SYSTEM

The grammatical error detection (GED) system chosen for this work is a state-of-the-art bidirectional recurrrent neural network based framework[2] proposed for detecting all kinds of grammatical errors in learner writing [6, 5]. It treats GED as a sequence labelling task with each token in the input sequence assigned a label, indicating

This paper reports on research supported by Cambridge Assessment, University of Cambridge. Thanks to Cambridge English Language Assessment for supporting this research and providing access to the BULATS data.

[1]https://www.grammarly.com, https://writeandimprove.com

[2]https://github.com/marekrei/sequence-labeler

if it is correct or incorrect in this context. Given a sequence of tokens $w_{1:N} = \{w_1, \ldots, w_N\}$ the system predicts a probability distribution over the possible labels for each token. For example,

	Internet	was	something	amazing	for	me	.
	i	c	c	c	c	c	c
$P(i)$	0.98	0.04	0.03	0.03	0.05	0.02	0.01
$P(c)$	0.02	0.96	0.97	0.97	0.95	0.98	0.99

Every token in the input sequence is mapped to a vector representation, yielding $\mathbf{x}_{1:N} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. This word embedding is optimised during training. To help the model handle uncommon words, for each token, w_i, every character is mapped to a character embedding and combined into a character-based token representation, $\mathbf{m}_i$, using a bidirectional LSTM [6]. The token and character vector representations are concatenated together to form a single vector representation, $\tilde{\mathbf{x}}_{1:N} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N\}$, where $\tilde{\mathbf{x}}_i^\mathsf{T} = [\mathbf{x}_i^\mathsf{T}\, \mathbf{m}_i^\mathsf{T}]$.

A bidirectional LSTM [8] composes $\tilde{\mathbf{x}}$ into context-specific representations of each token,

$$\overrightarrow{\mathbf{h}}_i = \texttt{LSTM}(\tilde{\mathbf{x}}_i, \overrightarrow{\mathbf{h}}_{(i-1)}); \quad \overleftarrow{\mathbf{h}}_i = \texttt{LSTM}(\tilde{\mathbf{x}}_i, \overleftarrow{\mathbf{h}}_{(i+1)}) \tag{1}$$

$$\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_i^\mathsf{T}\ \overleftarrow{\mathbf{h}}_i^\mathsf{T}]^\mathsf{T} \tag{2}$$

where $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ are the ith hidden state of the forward and backward LSTM, respectively, and $\mathbf{h}_i$ is the concatenation of both hidden states. The concatenated representation is passed to a feed-forward hidden layer with a $\tanh$ the non-linear activation function. Finally a softmax layer is used to predict the probability of each token being grammatically correct or incorrect (the 2-dimensional vector $\mathbf{y}_i$):

$$P(\mathbf{y}_i|w_1, \ldots, w_N) = \text{softmax}(\mathbf{W}_\text{o} \tanh(\mathbf{W}_\text{d}\mathbf{h}_i)) \tag{3}$$

The model is optimised by minimising cross-entropy between the predicted label distributions and the annotated labels. This is equivalent to minimising the negative log probability of the correct labels.

3. CORPORA

The main focus of this work is to build a GED system for free speaking non-native learner English, across the full range of CEFR proficiency levels [9] from beginner (A1) to proficient (C2). A test set derived from BULATS [10] Business English assessment speaking tests has been manually annotated for this purpose. To provide a comparison of spoken GED with this proprietary BULATS test set, the publicly available NICT Japanese Learner English (JLE) Corpus [11] is also used, which is derived from spoken English oral proficiency test interviews. There is no audio available for the NICT-JLE corpus so only experiments on manual transcriptions of the spoken learner data are possible. As can be seen from Table 1, neither of the spoken corpora are very large and preliminary sequence labeller models performed poorly when only trained on these corpora. To train the baseline system the written Cambridge Learner Corpus (CLC) [12] is adopted following the written GED work in [5, 6]. The grammatical error annotation scheme for BULATS [13] was derived from the CLC scheme [12]. NICT-JLE uses its own scheme [11].

BULATS: The spoken BULATS Business English assessment test [10] consists of read speech and free speaking components, with the candidate responding to prompts. Only responses to the main free speaking parts of the test from 226 candidates were used for this work. In BULATS sections C and D each utterance is up to 60 seconds of spontaneous speech, on topics such as describing the sales shown in a graph. Section E is made up of 5x 20 second responses to sub-questions related to an overall topic e.g. questions about organising a conference. The test set consists of a total of 1438 responses from these 3 sections. Speakers are approximately evenly distributed across CEFR grades A1-C (C1 and C2 are merged) and first language (L1s) of which there are 6: Arabic, Dutch, French, Polish, Thai, Vietnamese. Two sets of manual transcription, grammatical error markup and meta-data annotations are being produced starting from the same ASR transcription [13]. On a subset of 230 recordings across the 2 annotations, Cohen's kappa [14] for transcription is 0.898, grammatical error detection 0.752, and error type agreement 0.784 i.e. 'moderate' and 'strong' agreement [15]. For this paper all annotations were taken from a single transcriber as the second set were incomplete. 51 grammatical error types were marked of which 40 occurred more than 10×. ASR transcriptions were produced using the graphemic stacked hybrid DNN+LSTM-HMM joint decoding system, System 2 in [7] which was trained on 330 hours of non-native learner speech data from 8.5K BULATS candidates.

Corpus	Spoken/ Written	# Wds	# Uniq Wds	Audio
BULATS	Spoken	71.8K	4.1K	Yes
NICT-JLE	Spoken	135.3K	5.6K	No
CLC	Written	14.1M	79.1K	No

Table 1. Corpora used in experiments. CLC train and development set from FCE, IELTS, BULATS, CPE and CAE.

NICT-JLE: The NICT Japanese Learner English (JLE) Corpus (v4.1) [11] consists of manual transcriptions from a spoken English oral proficiency test, ACTFL-ALC SST (Standard Speaking Test). The test consists of a Japanese L1 candidate being interviewed by a native or proficient speaker of English. 167 of the interviews have been annotated with grammatical errors and disfluencies. For this paper only the candidate side is considered. The candidates' SST scores correspond to grades A1-B2 on the CEFR scale [9].

CLC: The Cambridge Learner Corpus (CLC) [12] is a corpus of written text responses by non-native English learners to examinations from Cambridge Assessment. It has been manually annotated with grammatical errors and consists of various exams corresponding to different proficiency levels of candidates. The publicly available CLC FCE-public Dataset [16] consists of 1,244 exam scripts written by candidates sitting the Cambridge ESOL First Certificate in English (FCE) examination in 2000 and 2001. The same FCE-public test and dev sets defined in [16] were used for evaluation and tuning, respectively. For training, the FCE-public training set was augmented with further FCE test data and written tests from IELTS, BULATS, CPE and CAE [5], giving around 27.5K candidates.

4. EXPERIMENTAL SETUP

Spoken utterances have a number of different attributes to written text: disfluencies affect the flow of the spoken words; there is no punctuation so there can be no grammatical punctuation errors but also there is no segmentation to mark sentence and phrase boundaries; and there can be no spelling mistakes. In addition, ASR output is typically in a single case so capitalisation cues which might help identify grammatical errors are unavailable. Each of these requires some modifications to the standard written text GED setup.

By definition, a disfluency cannot carry a grammatical error so the system should ignore or mark it as correct. Hesitations and partial words are simple to identify from the ASR output so are straightforward to remove from the speech transcription. Discourse markers,

false starts and restarts are harder to automatically detect and spoken meta-data extraction remains an open research topic e.g. [17, 18, 19]. Initial experiments showed that removing words marked with these 3 types of disfluency from the GED input manual speech sequence improved the GED performance with an increase of 13 in $F_{0.5}$ from 36.7 to 49.7 for the NICT-JLE test and of 3 for BULATS. The lower gain for BULATS may be due to some possible disfluencies not being removed as for this data a rich set of tags were used which included tags such as "unnecessary". Identification and removal of appropriate disfluencies from these additional tags should help in the future. Disfluency removal was carried out for all the experiments in the next section.

Since there are no spelling mistakes or punctuation in speech these were removed from the CLC data. The CLC texts were then converted to lower case and all punctuation removed to mimic ASR output. In the public FCE database words following a missing item are labelled with the corresponding GE tag [16]. Any such tags falling on final sentence punctuation were moved to the preceding word. BULATS responses are up to 1 minute in length with a maximum of 180 words observed. To handle overly long segments, since there is no punctuation the BULATS data was manually segmented into annotator defined "speech units" [13]. The same segmentations were applied to both the manual and ASR transcriptions. Unlike BULATS, the NICT-JLE utterances are conversational turns so tend not to be too long, with a maximum length of 62 words, and no further segmentation was applied. The vector representation of each word in the sequence labeler was initialised using Google's 3 million word 300-dimensional word2vec word embedding [3].

The GED model trained on the CLC was fine-tuned to either the NICT-JLE or BULATS data set by running further training epochs on a subset of the respective data. Each evaluation set was split into 10 distinct speaker subsets. A 10-fold cross-validation experiment was run where 8 sets were used for fine-tuning, one for GED model development, and one for evaluation. Instead of averaging the 10 results, the results were scored as a single output by combining all 10 evaluation blocks together, each of which had been held-out.

5. EXPERIMENTAL RESULTS

When an automatic system provides feedback to a learner it is very important that the system has, as far as possible, correctly identified an error. Giving feedback on an error that the system has made will confuse the learner. This means that high precision is required in the GED. $F_{0.5}$ gives double the weighting to precision over recall but does not fully capture the performance at the precision point of interest so precision-recall curves are also presented here.

5.1. Manual annotations

Table 2 gives the $F_{0.5}$ scores for the baseline manually annotated tasks using the GED system trained on the CLC corpora. The FCE written test has higher precision and recall scores than the spoken systems. Note, the FCE $F_{0.5}$ score is lower than in other papers due to the removal of the spelling and punctuation errors which make up 23% of the standard FCE test and are somewhat easier to detect than other grammatical errors. Similarly the NICT-JLE scores are a lot lower than in [2] as [2] only considered unnecessary determiners and on a subset of the data. Figure 1 shows that even at the highest precision/lowest recall region, the precision of the GED on BULATS is ∼0.2 worse than the written test.

[3] https://code.google.com/archive/p/word2vec/

Test		P	R	$F_{0.5}$
Written	FCE	69.9	33.9	57.6
Spoken	NICT-JLE	60.6	28.9	49.7
	+ fine-tune	66.5	35.9	56.8
	BULATS	52.4	27.0	44.1
	+ fine-tune	66.7	33.8	55.8

Table 2. Precision (P), Recall (R) and $F_{0.5}$ scores with a CLC trained GED system and fine-tuned to the test set.

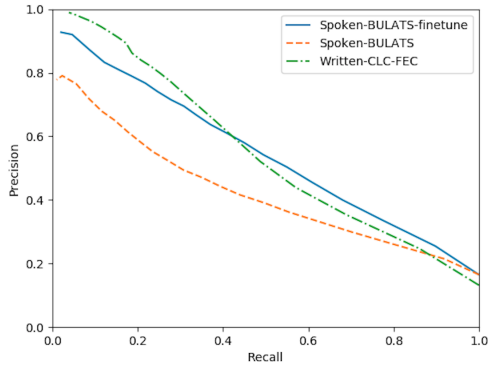

Fig. 1. Precision-recall curves for CLC and BULATS with a CLC trained GED system, and fine-tuned to the BULATS data.

To provide feedback the recall of high precision items needs to be increased. The lack of annotated spoken learner corpora poses a challenge for training the GED system so the CLC-trained GED model was fine-tuned to each target domain. As shown in Table 2 and Figure 1, fine-tuning gives a large performance boost, especially in the high precision region of interest, with BULATS and JLE performing similarly. The $F_{0.5}$ scores increase by 6 to 56.8 for JLE, and by 11 to 55.8 for BULATS. It is interesting to note the larger gain for fine-tuning the BULATS data compared to the NICT-JLE data. Some of these additional gains may be due to the system learning attributes of the single annotator used for the BULATS data. Additional analysis is planned, comparing with other annotators on this data. Applying the fine-tuned JLE models to the BULATS data was found to degrade $F_{0.5}$ by 4. This may be due to the different nature of the conversational speech in NICT-JLE compared to the prompt responses in BULATS and/or the different L1 and grade coverage. For both models over 80% agreement and verb derivation errors were correctly detected on BULATS. Most missing errors had high error rates, except for conjunctions (70%). Replacement errors scored lowest (<40% correct).

A small scale analysis of the false negatives predicted for the BULATS data at the highest precision point indicates that the true precision is higher than the scores suggest. Annotation is difficult due to the nature of the data and disagreement between annotators as to what constitutes a grammatical error. About 40% of the errors could be potentially classed as true positives where a GE was made but not annotated e.g. no missing determiner marked for "... I think you need *taxi*". From the perspective of providing helpful feedback to the learner, of most interest were about 27% of errors which were words that were in the region of an error, that is, the grammatical error was marked on a word next to, and part of a linguistic chunk with, a word with an error tag. For example

... to continue to inform *with* **customer** when we have ...

tag	c	c	c	c	i	c	c	c	c
pred	c	c	c	c	c	i	c	c	c

For feedback purposes it may well be sufficient to highlight the region where the error tag was predicted, in which case these small mis-alignments will be beneficial in providing information.

5.2. ASR annotations

The results presented above are based on manual transcription of the test data. In practice the transcriptions will be produced by an automatic speech recognition (ASR) system. As seen in Table 3, learners at lower proficiency levels say a lot less than higher grade learners. In spite of this lower word count, the grammatical error rate (GER) increases with decreasing grade. Grammatical errors cause mismatches with the ASR language model so more GEs would be expected to increase the recognition word error rate (WER) and this is reflected in the WERs presented in Table 3. The less proficient the speaker, the higher the WER[4]. These recognition errors introduce additional challenges into the spoken GED task. From a feedback perspective it is important that feedback is given on actual grammatical errors rather than a false error detected on an ASR error.

	A1	A2	B1	B2	C	Tot
# Words	5326	10991	16226	19206	20002	71751
% GER	21.2	18.9	17.8	16.9	12.6	16.5
% WER	39.0	30.7	24.3	23.6	21.0	25.2

Table 3. ASR word error rate (WER) and grammatical error rate (GER) scored against manual transcriptions.

To assess the effect of grammatical errors and disfluencies on ASR output the manual and automatic BULATS transcriptions were aligned using a modified Damerau-Levenshtein algorithm (restricted edit distance version) from the ERRANT toolkit[5] [20, 21]. To improve the alignment, token transportation was disabled and the costs for a word error changed to be 3 for insertion or deletion and $4 + \Delta$ for a substitution. A character-level Levenshtein distance was used to calculate Δ with range $[0 : 1]$. Following the transcription alignment, the associated GE and disfluency tags were ported to the aligned ASR word. If the ASR word was an error, i.e. it did not match the manually transcribed word, then the GE was changed to correct. This approach was chosen to reflect the need to not give feedback on ASR errors. When ASR is used there may be words missed (deleted) in the recognised output. If a GE occurs on this word then it will not be possible to detect it, for example the incorrect determiner form (FD) here:

Manual	the job we can offer is {FD a} engineering job
ASR	the job we can offer is [[del]] engineering job

	Num.	%WER
Overall	71751	25.2%
"Fluent"	52698	19.3%
Grammatical Error	10348	29.1%
Disfluency	2524	36.4%

Table 4. WER breakdown by error type excluding ASR insertions, filled pauses, inaudible and partial words.

Table 4 shows that where speech contains a grammatical error or disfluency the WER increases. This is due to the language model being poorly matched in these cases even though it is partially trained on transcriptions from the BULATS training corpus [22]. This will increase the mismatch with the GED model which is reflected in the GED performance shown in Figure 2 and Table 5. The baseline CLC trained GED system performance is much lower than for the manual transcriptions with 26.6 $F_{0.5}$, compared to 44.1 for the latter. As for the manual transcriptions, fine-tuning the GED model yields better precision, particularly at the high precision area of interest. Using the ASR transcriptions for fine-tuning (Finetune-ASR) gave slightly better GED performance than taking the model fine-tuned on the manual transcriptions (Finetune-Man) with $F_{0.5}$ of 37.3 and 35.1, respectively. By matching the fine-tuning training and test data transcriptions some of the errors in the latter are mitigated.

GED Model	P	R	$F_{0.5}$
CLC	28.1	22.0	26.6
+ fine-tune Man	37.0	29.2	35.1
+ fine-tune ASR	46.7	20.7	37.3

Table 5. Precision (P), Recall (R) and $F_{0.5}$ scores for ASR transcriptions with a CLC trained GED system and fine-tuned to BULATS manual (Man) and ASR transcriptions.

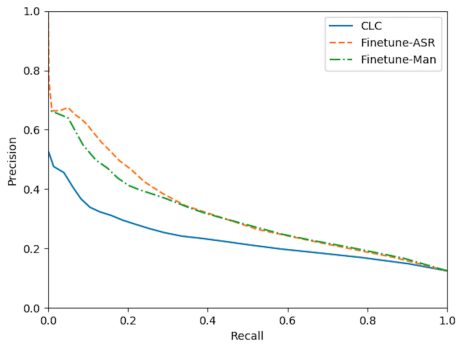

Fig. 2. Precision-recall curves on BULATS ASR transcriptions.

6. CONCLUSIONS

The performance of grammar error detection (GED) systems for non-native English learner writing has reached a good level across GE types on a broad range of writing tasks and proficiency levels. In contrast most spoken GED systems have focused on limited tasks and/or GEs. This paper, therefore, considered if a state-of-the-art bidirectional LSTM based written GED system could be applied to non-native English open free speaking tasks (interviews (NICT-JLE) and prompted talking on topics (BULATS)) across all learner proficiency levels. The baseline written GED system was applied to manual transcriptions of the spontaneous spoken responses for both tasks. The spoken GED performance level was lower than for writing. By fine-tuning to the task transcriptions, however, this domain mismatch was overcome and equivalent performance observed. In contrast, applying the system to automatic transcriptions of the BULATS data proved challenging. ASR errors were shown to increase when there are grammatical errors and disfluencies, leading to false positives which degraded GED performance, even after fine-tuning. Future work will look at boosting the quantity of speech training data by incorporating learner errors into native speech corpora and improving meta-data detection to assist with removing disfluencies.

7. ACKNOWLEDGEMENTS

Thanks to Marek Rei for assistance with his Sequence Labeler toolkit and to him and Helen Yannakoudakis for discussions about GED. Yiting Edie Lu aligned the manual/ASR transcriptions.

[4]Two crowd-sourcer inter-annotator error rate is ∼25%.

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

8. REFERENCES

[1] John Lee and Stephanie Seneff, "An analysis of grammatical errors in non-native speech in English," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2008.

[2] Kyusong Lee et al., "Grammatical error correction based on learner comprehension model in oral conversation," in *Proc. of the IEEE Workshop on Spoken Language Technology (SLT)*, 2014, pp. 283–287.

[3] Bart Penning de Vries et al., "Spoken grammar practice and feedback in an ASR-based CALL system," *Computer Assisted Language Learning*, vol. 28, no. 6, pp. 550–576, 2015.

[4] Hwee Tou Ng et al., "The CoNLL-2014 Shared Task on Grammatical Error Correction," in *Proc. of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (CoNLL-2014 Shared Task)*, 2014.

[5] Marek Rei and Helen Yannakoudakis, "Compositional Sequence Labeling Models for Error Detection in Learner Writing," in *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016)*, 2016.

[6] Marek Rei, Gamal K.O. Crichton, and Sampo Pyysalo, "Attending to characters in neural sequence labeling models," in *Proc. of the 26th International Conference on Computational Linguistics (COLING-2016)*, 2016.

[7] K.M. Knill et al., "Impact of ASR Performance on Free Speaking Language Assessment," in *Proc. of INTERSPEECH*, 2018, pp. 1641–1645.

[8] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[9] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, Cambridge University Press, 2001.

[10] Lucy Chambers and Kate Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011.

[11] Emi Izumi, K. Uchimoto, and H. Isahara, "The NICT JLE Corpus Exploiting the language learners' speech database for research and education," *International Journal of The Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, May 2004.

[12] Diane Nicholls, "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT," in *Proc. of the Corpus Linguistics 2003 conference; UCREL technical paper number 16.*, 2003.

[13] Andrew Caines, Diane Nicholls, and Paula Buttery, "Annotating errors and disfluencies in transcriptions of speech," Tech. Rep. UCAM-CL-TR-915, University of Cambridge Computer Laboratory, Dec 2017.

[14] Jacob Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[15] Mary L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, pp. 276–282, 2012.

[16] H. Yannakoudakis, T. Briscoe, and B. Medlock, "A new dataset and method for automatically grading ESOL texts," in *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 180–189.

[17] Matthew Honnibal and Mark Johnson, "Joint incremental disfluency detection and dependency parsing," *Transactions of the Association for Computational Linguistics*, vol. 2, no. 1, pp. 131–142, 2014.

[18] Russell Moore et al., "Incremental Dependency Parsing and Disfluency Detection in Spoken Learner English," in *Text, Speech, and Dialogue - 18th International Conference, (TSD)*, 2015, pp. 470–479.

[19] Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi, "Disfluency Detection Using a Bidirectional LSTM," in *Proc. of INTERSPEECH*, 2016, pp. 2523–2527.

[20] Mariano Felice, Christopher Bryant, and Ted Briscoe, "Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments," in *Proc. of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016.

[21] Christopher Bryant, Mariano Felice, and Ted Briscoe, "Automatic annotation and evaluation of Error Types for Grammatical Error Correction," in *Proc. of 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017.

[22] K.M. Knill et al., "Use of Graphemic Lexicons for Spoken Language Assessment," in *Proc. of INTERSPEECH*, 2017, pp. 2774–2778.