Grammatical error detection in transcriptions of spoken English

Andrew Caines[1] **Christian Bentz**[2] **Kate Knill**[3] **Marek Rei**[4,1] **Paula Buttery**[1]

[1] ALTA Institute & Computer Laboratory, University of Cambridge, U.K.
{andrew.caines|paula.buttery}@cl.cam.ac.uk
[2] Department of General Linguistics, University of Tübingen, Germany
chris@christianbentz.de
[3] ALTA Institute & Engineering Department, University of Cambridge, U.K.
kate.knill@eng.cam.ac.uk
[4] Department of Computing, Imperial College London, U.K.
marek.rei@imperial.ac.uk

Abstract

We describe the collection of transcription corrections and grammatical error annotations for the CROWDED Corpus of spoken English monologues on business topics. The corpus recordings were crowdsourced from native speakers of English and learners of English with German as their first language. The new transcriptions and annotations are obtained from different crowdworkers: we analyse the 1108 new crowdworker submissions and propose that they can be used for automatic transcription post-editing and grammatical error correction for speech. To further explore the data we train grammatical error detection models with various configurations including pre-trained and contextual word representations as input, additional features and auxiliary objectives, and extra training data from written error-annotated corpora. We find that a model concatenating pre-trained and contextual word representations as input performs best, and that additional information does not lead to further performance gains.

1 Introduction

We introduce a new resource for speech-centric natural language processing (speech NLP) – more than a thousand transcriptions and error annotations for 383 distinct recordings from the CROWDED Corpus (Caines et al., 2016). CROWDED is a crowdsourced English corpus of short monologues on business topics, recorded by both native and non-native speakers. It was created in response to the lack of speech corpora freely available for research use, and the lack of appropriate native speaker reference corpora with which language learners' exam monologues can be compared. In this new project, crowdworkers were asked to first correct existing speech transcriptions and then to edit the resulting transcriptions to make them more fluent. These new annotations enable both post-editing of noisy speech transcriptions – such as might come from automatic speech recognisers – and also grammatical error correction for spoken English.

There has been a marked increase in openly-available NLP resources in recent years, especially for English, but these have on the whole been sourced from written texts. Public resources for speech NLP, on the other hand, are relatively scarce, even for English – though English is again by far the best served in this respect. Obtaining linguistic data from large, distributed, online workers ('crowdsourcing') has become a well-established practice. With the contribution of these new annotations, we note that the CROWDED Corpus now features a thousand recordings which have all been transcribed, and of which almost 40% now have improved transcriptions and error annotations thanks to this work. The entire corpus has been collated through crowdsourcing means.

In this paper we describe the method for collecting the new data, analyse the annotations received, and report on some initial grammatical error detection experiments (GED). In the GED experiments we trial various configurations which have been successful in GED for written texts. We found that we can identify errors in the transcriptions fairly reliably using a publicly-available sequence labeller adapted to take contextual word representations as additional input, similar to previous work on GED in written corpora (Rei and Yannakoudakis, 2016; Bell et al., 2019). All new transcriptions and annotations referred to in this paper are made available on the corpus website[1].

2 Related Work

Compared to the relatively abundant amounts of written text corpora, there are fewer speech corpora in which both transcriptions and audio files have been made available, even for the English language. Some well-known exceptions include the British National Corpus (BNC Consortium, 2001), the Switchboard Corpus (Godfrey et al., 1992), and AMI Corpus (Carletta et al., 2006). However, it is now almost thirty years since the BNC and Switchboard recordings were made, Switchboard is only available for a fee, and all three contain spontaneous dialogues – either on general (BNC and Switchboard) or business topics (AMI) – whereas we have a research interest in assessing spoken monologues in language exams. In terms of non-native speaker English, one of the few freely available corpora is The NICT Japanese Learner English Corpus (JLE; Izumi et al. (2004)). Again, this corpus contains transcriptions of conversations about general topics, which are certainly of interest but not the type of data we need to work with relating to our other work on speech assessment (Wang et al., 2018; Craighead et al., 2020): namely, short monologue responses by learners of English.

One reason why speech corpora are rare is that their preparation is a labour-intensive process: human effort is required to obtain recordings, transcriptions and optionally annotations. Crowdsourcing has been widely used for corpus annotation, whether word senses (Lopez de Lacalle and Agirre, 2015), system evaluation (Rayner et al., 2011), grammatical errors (Madnani et al., 2011), as well as speech transcription (Evanini et al., 2010). In the case of the CROWDED Corpus, crowdsourcing was used for the whole data collection process, from recording to transcription and annotation, and was found to be cost-effective (Caines et al., 2016).

Even rarer are freely-available speech corpora with grammatical error annotations: indeed it is the first one we know of, because NICT JLE does offer error-annotated transcriptions but does not make the audio files available. Grammatical error annotations can be used to train and evaluate machine learning models in the GED task, in which classifiers are required to identify grammatical errors in word sequences. Most of the previous work on GED has tended to focus on written corpora (Foster and Vogel, 2004; De Felice and Pulman, 2008; Tetreault and Chodorow, 2008; Rei and Yannakoudakis, 2016; Kasewa et al., 2018; Bell et al., 2019). There has been less GED research with spoken corpora but that is more a reflection of the scarcity of appropriate data than a lack of interest. There has been work to assess grammatical accuracy by spoken CALL systems ('computer-assisted language learning') where such judgements might be rule-based (Lee et al., 2014) or compared to a reference if the task is suitably constrained (de Vries et al., 2015). There have been several shared tasks requiring binary acceptability judgements (correct/incorrect) at the utterance level (Baur et al., 2017; Baur et al., 2018; Baur et al., 2019). Treating spoken GED as a word classification task is a more recent approach (Knill et al., 2019; Lu et al., 2019), one which applies the sequence labelling approach for written GED of Rei and Yannakoudakis (2016) and Rei (2017) to speech transcriptions. It involves pre-training models on large written datasets – e.g. the FCE Corpus (Yannakoudakis et al., 2011) – and then fine-tuning to the target spoken corpus.

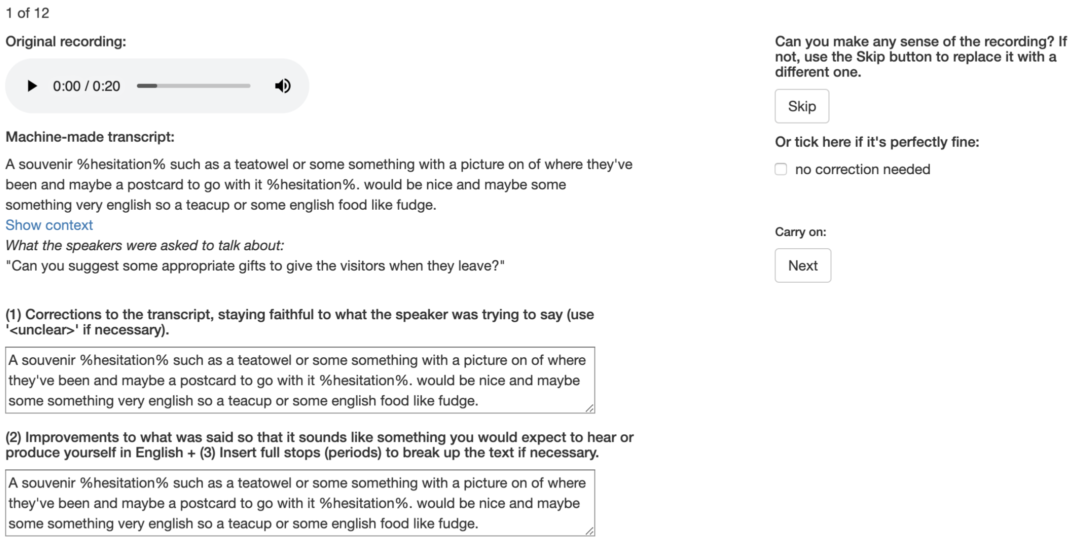
3 The CROWDED Corpus

We use the CROWDED Corpus for our experiments (Caines et al., 2016). The first release of the CROWDED Corpus included one thousand short recordings in German and English from eighty people, transcribed into thirty-four thousand word tokens. All speech recordings and transcriptions were collected through crowdsourcing platforms (Crowdee[2] and CrowdFlower[3] respectively), and it was demon-

[1] https://www.ortolang.fr/market/corpora/ortolang-000913

[2] http://www.crowdee.de/en

[3] http://www.crowdflower.com



1 of 12

Original recording:

Machine-made transcript:

A souvenir %hesitation% such as a teatowel or some something with a picture on of where they've been and maybe a postcard to go with it %hesitation%. would be nice and maybe some something very english so a teacup or some english food like fudge.

Show context

What the speakers were asked to talk about:

"Can you suggest some appropriate gifts to give the visitors when they leave?"

(1) Corrections to the transcript, staying faithful to what the speaker was trying to say (use '<unclear>' if necessary).

A souvenir %hesitation% such as a teatowel or some something with a picture on of where they've been and maybe a postcard to go with it %hesitation%. would be nice and maybe some something very english so a teacup or some english food like fudge.

(2) Improvements to what was said so that it sounds like something you would expect to hear or produce yourself in English + (3) Insert full stops (periods) to break up the text if necessary.

A souvenir %hesitation% such as a teatowel or some something with a picture on of where they've been and maybe a postcard to go with it %hesitation%. would be nice and maybe some something very english so a teacup or some english food like fudge.

Can you make any sense of the recording? If not, use the Skip button to replace it with a different one.

Or tick here if it's perfectly fine:

□ no correction needed

Carry on:

Figure 1: R Shiny web-app for CROWDED transcription correction and error annotation.

strated that crowdsourcing is a fast, feasible and cost-effective method of corpus collation.

The speakers responded to 20 questions designed to prompt spontaneous monologues of up to one minute about imagined business scenarios. For instance, in the scenario of being asked to give advice about setting up a new shop, speakers were asked to respond to prompts such as the following: *what do you think is the best location for a retail business? what are the most effective ways to advertise a new shop? why is it important to find good suppliers?* There were 4 different scenarios: setting up a new shop, preparing for business visitors from a foreign country, running a taxi company, and the pros and cons of sponsoring sports events. Each scenario had 5 associated questions. Speakers who identified themselves as speakers of both languages were prompted to answer 10 questions in English and 10 in German; in the monolingual setting, speakers were asked to respond to all 20 questions in English.

There was a quality control filter on the collected recordings, flagging and removing any with poor audio quality or in which the speaker failed to respond appropriately. Approved recordings were passed to CrowdFlower workers for transcription, but only the English transcriptions were returned from CrowdFlower with acceptable quality: the German recordings were not successfully transcribed, either because crowdworkers falsely claimed to know the language, or the job settings requiring German competence did not properly filter the worker pool. Therefore we focus on the English section of the corpus in the remainder of the paper, and intend to revisit the German data in future work. In what follows we describe the subset of recordings which have been error annotated since the initial release of the corpus, and report on our experiments to automatically detect grammatical errors in the transcriptions.

3.1 Crowdsourcing transcription correction & error annotation

Each recording in the CROWDED Corpus has been transcribed twice but it is not clear how to resolve this into a single transcription version. In this new work, we use an ASR-based method for the merger of different transcription versions developed by van Dalen et al. (2015). In that work the method was shown to produce a combined transcription which is more accurate than either version on its own, and the same is true for our data (section 3.2). We then uploaded the merged transcriptions to the Prolific platform (Palan and Schitter, 2018), along with the audio recordings, where crowdworkers reviewed the new transcriptions and edited them where necessary. We required that workers were educated to at least GCSE level (a U.K. exam aimed at 16 year olds) or equivalent, had an approval rate of at least 95% from previous studies, and listed English as a first language. Prolific requires a fair rate of pay to workers, above or equivalent to the U.K.'s minimum hourly wage, but outputs tend to be better quality than from other crowdsourcing services (Peer et al., 2017). We asked workers to correct and annotate 12 transcriptions as a unit of work, a task estimated to take 30 minutes, for which they were paid £3.10

(equivalent to approximately US$4 at the time). Noting that there is a service fee payable to Prolific, in common with other services, our funding allowed us to pay 100 workers in total.

One drawback of crowdsourcing from the researcher's point of view is the lack of training and contact time with workers. On the other hand its main advantages are the scale and speed of data collection, along with evaluations from population groups one might not normally reach in campus-based studies (Paolacci and Chandler, 2014; Difallah et al., 2018). Another issue is the bursty nature of responses: upon publishing the task there tends to be a rush of early responses. The consequence is that the central record of transcription correction does not keep pace with the issuing of work, and so a few transcriptions are issued to workers many times, whereas the intention was to limit the number of new annotations per item and thereby achieve greater coverage of the original corpus. This is a flaw which we will address in future work either with a redesign of the workflow, or a shift to more expensive but more evenly-paced and controllable local annotation.

Besides transcription correction, we asked Prolific workers to apply minimal grammatical error corrections to their updated transcriptions, in order to make them "sound like something you would expect to hear or produce yourself in English". With this statement we intended to convey the error correction task to crowdworkers in a straightforward way without the use of jargon: to alter the text into something linguistically acceptable in that worker's judgement, without referring to notions of grammar or 'correctness' which may have particularly strong connotations for some.

We developed the annotation web-app in R Shiny (Chang et al., 2020), adapted from Caines et al. (2017a) with a simple user-interface and text instructions kept to a minimum so as not to overload the workers with information. The relevant audio recordings were provided for unlimited playback, and workers were presented with the 'machine-made' transcription which was formed from the original two CrowdFlower transcriptions, along with the prompt for context. There were two text boxes, firstly for the worker to edit the existing transcription so that it better matched the recording, and the second being a copy of that updated transcription ready for the second task of error correction. A screenshot of the transcription correction and error annotation web-app is shown in Figure 1.

After quality checks and capping the number of submissions for a single item at 10, we reduced our 1200 submissions from Prolific (12 submissions from 100 workers) to 1108 submissions covering 383 unique recordings (mean: 2.9 per recording) which were selected at random from the original thousand recordings in the CROWDED Corpus. Figure 2 shows a histogram of transcriptions and annotations per recording. This is the set of data we work with in the remainder of this paper, totalling 39.7K word tokens (mean: 35.9 per transcription).

3.2 Error-annotated CROWDED dataset

Of the 1108 new transcriptions we received from Prolific workers, 80% had been updated in some way compared to the merged versions of the original CrowdFlower transcriptions. We aligned the merged and corrected transcriptions using the ERRANT toolkit (Bryant et al., 2017), which lists the edit operations to transform one text into another. These lists indicate that across the whole dataset there were 12 edits for every 100 tokens in the original transcriptions – whether a replacement, deletion or insertion. For transcriptions from the native speakers of English in the dataset, corrections were applied to 7% of word tokens; whereas for learners of English the correction rate was 17%, indicating that transcription is a harder and consequently more error-prone task with learner English speech.

Taking these new Prolific transcriptions as the ground truth, we can calculate the word error rate (WER) of the original CROWDED transcriptions obtained from CrowdFlower. Recall that each recording was transcribed twice in the original study: let us randomly assign each version to a transcription set for comparison with the new transcriptions. Thus the mean WER is 19.4% for version 1 averaged across all new versions from Prolific, and for version 2 it is 18.5%. WER for the merged transcriptions drops to 12%, confirming the benefit of that method.

On average there are just under 3 Prolific submissions for each original CROWDED transcription, with a median of 2, minimum of 1 and maximum of 10. Within each recording's set of updated transcriptions, similarity scores are good, indicating general agreement. We calculate one-versus-rest string distances

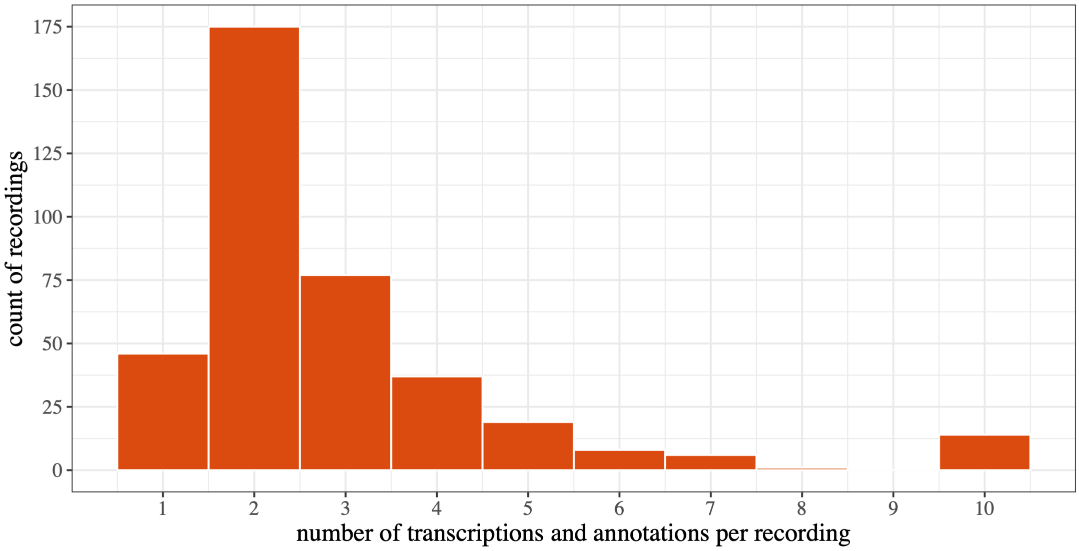


Figure 2: Histogram of transcriptions and annotations per recording.

using optimal string alignment – the restricted Damerau-Levenshtein distance (van der Loo, 2014) for each transcription within the set of other transcriptions for that recording. On average the string distance within transcription sets is 18 characters, set against an average transcription length of 200 characters. This is good, but also underlines the fact that transcription is a subjective process: not all transcribers perceive the same words, especially when the speaker is unclear or the audio is degraded[4].

Meanwhile, 80% of the corrected transcriptions were edited for grammatical errors in some fashion, at an average of 21 edits per 100 word tokens. Again we align and type the identified errors using ERRANT. Table 2 in the appendix imitates Table 4 in Bryant et al. (2019), listing the frequency of error types as proportional distributions. We present these statistics for the whole dataset, and then the native speakers of English and learners of English separately, with statistics from the FCE Corpus for comparison.

Note that the edit rate in grammatical error correction of the transcriptions is not greatly different between the native speaker and learner groups at 18.9% and 23.9% respectively. The native speakers do produce more word tokens per recording (51.3 versus 27.0) as might be expected in comparing fluent native speakers of any language to learners with varying levels of proficiency. The main differences in terms of edit types are that unnecessary word tokens occur in native speaker transcriptions more than they do in learner transcriptions, where the replacement error type is the most common.

The distribution of error types (the middle section of Table 2) is broadly similar between native speaker and learner groups, with more ‘other’ errors in native speaker transcriptions than in learner transcriptions, in which there are more determiner, preposition and verb errors. In other words, there are more of the formal errors in learner speech which are typically found in written learner corpora: for comparison the most common error types in the FCE Corpus (Yannakoudakis et al., 2011) are ‘other’, prepositions and determiners (Bryant et al., 2019), though the ‘other’ type is much less frequent at just 13.3% of the errors in that corpus. Punctuation and spelling are the next most frequent error types in the FCE, each forming more than 9% of the total edit count.

In our CROWDED annotations the most frequent error types are ‘other’, punctuation and nouns, followed by determiners, orthography and verb. The ‘other’ group is the largest by far, one reason being that the ERRANT typology was designed for written language, and so disfluencies such as filled pauses, partial words, and false starts fall into this category (Caines et al., 2017b). It may seem odd that punctuation and orthography errors feature in the correction of speech transcriptions, but the former type of edit was invited in the Shiny web-app with the request to “insert full-stops (periods) to break up the text

[4] For any readers curious about this, we have prepared a web-app which invites the user to transcribe an audio recording and view string distances between their transcription and the crowdsourced transcriptions in the corpus: https://cainesap.shinyapps.io/crowded

if necessary” (Figure 1). Again, we did not aim to be overly prescriptive in defining this task, adhering to previous work indicating that ‘speech-unit delimitation’ in transcriptions is an intuitive task which depends on a feel for appropriate delimitation based on some combination of syntax, semantics and prosody (Moore et al., 2016). One could more definitely require that the units be syntactically or semantically coherent but this was more instruction than we wished to give in a crowdsourcing interaction.

Orthography edits relate to the speech-unit delimitation task: they are changes in character casing due to the insertion of full-stops by the crowdworkers, and as such are not errors made by the speaker but rather a step towards making the transcriptions more human-readable. Included in the ‘other’ error category are filled pauses (‘er’, ‘um’, etc) which are filtered before GED because we can remove these with a rule. Note that filled pauses are a common occurrence in naturalistic speech, and hence are produced as much by native speakers – 273 instances, or 7% of the transcription edits made for this group – as by the learners (256 instances, or 5.6% of edits).

4 Grammatical error detection

Automatic grammatical error detection in natural language is a well-developed area of research which tends to involve one of several established datasets: for example, the FCE Corpus (Yannakoudakis et al., 2011), CoNLL-2014 (Ng et al., 2014) , JFLEG (Napoles et al., 2017), and the Write & Improve Corpus (W&I; Bryant et al. (2019)) among others. These corpora all contain written essays, whether written for language exams (FCE, JFLEG) or for practice and learning (CoNLL-2014, W&I). Thus techniques for GED are advanced and tuned to the written domain. We evaluate how well these methods transfer to spoken language.

GED is usually treated as a separate task to grammatical error *correction* (GEC) – which involves proposing edits to the original text. GEC could be an area for future work with the CROWDED Corpus, but at first we wish to explore GED for this speech dataset, anticipating that performance will be quite different to GED on written texts in which error types more often relate to word forms, punctuation and spelling (Table 2).

The state-of-the-art approach to GED involves sequence labelling word tokens as correct or incorrect with a bi-directional LSTM: the original model (Rei and Yannakoudakis, 2016) has evolved to include forwards and backwards language modelling objectives (Rei, 2017) and contextual word representations concatenated to pre-trained word-level representations (Bell et al., 2019). In addition, multi-task learning has proven effective for GED, with auxiliary predictions of error types, part-of-speech tags and grammatical relations aiding model performance (Rei and Yannakoudakis, 2017). We take these insights forward to GED in the CROWDED Corpus, running a series of experiments with modifications to the publicly available sequence labeller released with Rei (2017)[5]. Note that $F_{0.5}$ has been the standard evaluation metric for GED since Ng et al. (2014), and it weights precision twice as much as recall.

4.1 Data pre-processing

As explained in section 3.2 the Prolific transcriptions and error-corrected versions were tokenized and aligned with ERRANT: we then converted the resulting M^2 format files into CoNLL-style tables in readiness for the sequence labeller. In the first column of the tables are the word tokens, one on each line, while in the final column is a ‘correct’ (c) or ‘incorrect’ (i) label for that token. Note that ‘missing’ error types are carried by the word token following the missing item, and that error labels are shared across tokens if they have been split from a single white-space delimited token (e.g. an erroneous “it’s” would carry an ‘i’ label on both “it” and “’s”).

We created ten train-development-test data splits in order to carry out ten-fold cross-validation in our GED experiments. Transcriptions were assigned to data splits in batches associated with distinct recordings. That is, where we have multiple transcriptions and annotations for a single recording, these are placed together in the same split. For each fold, recording sets were randomly selected from the 383 in the corpus until we had filled the development and test splits. As there are 1108 transcriptions in the corpus, we sought out a minimum of 110 transcriptions for the development and test splits in each fold.

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

Most of the folds have 110 or 111 transcriptions in development and test: the largest such split contains 115 transcriptions. We make our pseudo-random splits available for the sake of reproducibility.

We opted for cross-validation rather than a single train-development-test split because, with several recordings each being associated with many transcriptions, there is a danger of over-concentrating the smaller splits (development and test) with many similar texts. In the scenario of only having one dataset split, conclusions about the generalisability of our GED models would have therefore been limited. We add extra information to the text files from several sources: morpho-syntactic labels obtained from a parser, n-gram frequencies from several corpora, the identification of complex words, ERRANT, and prosodic features from the CROWDED audio recordings.

Specifically, the morpho-syntactic labels come from the pre-trained English Web Treebank (UD v2.4) parsing model for UDPipe (Bies et al., 2012; Nivre et al., 2019; Straka and Straková, 2017; Wijffels, 2019). For each word token we obtain a lemma, Universal part-of-speech tag (UPOS), Penn Treebank part-of-speech tag (XPOS), head token number and dependency relation directly from UDPipe's output. The motivation for preparing such information was that certain error types may occur with tell-tale morpho-syntactic signals, and furthermore predicting such labels has been of benefit in multi-task learning approaches to GED (Rei and Yannakoudakis, 2017).

The n-gram frequencies were obtained for values of $n = \{1, 2, 3\}$ from the following corpora: the CROWDED Corpus itself, the British National Corpus (BNC Consortium, 2001), and the One Billion Word Benchmark (Chelba et al., 2014). In this case there were 6 values per corpus: a unigram frequency, two bigram frequencies with the target word in both first and second position of the gram, and three trigram frequencies with the target word in all three positions. The intuition here is that frequency information may be useful in identifying ungrammatical word sequences: if the n-gram is low frequency, that might indicate an error.

For each word token in the CROWDED Corpus we added a binary complexity label obtained from a model pre-trained on separate data (Gooding and Kochmar, 2019). Words in the complexity training data were labelled as complex or not by twenty crowdworkers (Yimam et al., 2017), and the model is currently state-of-the-art for the complex word identification task. We expect that complex words are more likely to be involved in or around grammatical errors. In total 4485 word tokens were identified as complex – for instance, 'guarantee', 'presentation', and 'suppliers'.

Error types were obtained from ERRANT (Bryant et al., 2017) based on the alignment of the Prolific transcription and its error-corrected version. Predicting error types, such as `R:NOUN` (replace noun), `M:DET` (missing determiner), and so on, as an auxiliary task improved GED performance in previous work (Rei and Yannakoudakis, 2017).

Finally, we extracted a number of prosodic values associated with each word token from the audio recordings. First we force aligned the transcriptions with the audio using the SPPAS toolkit (Bigi, 2015). Based on the resulting token start and end timestamps we could then calculate token durations, durations of any pauses preceding or following word tokens, and a number of values relating to the pitch and amplitude of the speaker's voice measured in 10 millisecond increments.

Thus for each token we collect the speaker's initial and final fundamental frequency (F0), the minimum and maximum, mean and standard deviation. We do the same for voice amplitude, or energy (E). Values of F0 and E are first smoothed with a 5-point median filter (Fried et al., 2019) in common with prosodic feature extraction described in previous work (Lee and Glass, 2012; Moore et al., 2016). The motivation for collecting such values is that speakers may display certain prosodic patterns – such as pausing before or afterwards – where they find speech production difficult and therefore may produce a grammatical error, or where they realise that they have recently or are in the process of making an error.

4.2 Experiment configuration

Our initial experiments did not involve additional features or auxiliary tasks: fundamentally we initialise input word vectors with pre-trained word representations and train the sequence labeller to predict whether each word token is a grammatical error. At first we ran several hyperparameter tuning experiments with manual search, recognising that grid search or random search might be more thorough, but

Model	**P**	**R**	**F$_{0.5}$**	**Hours**
GloVe$_{300}$	.386 (.115)	.111 (.058)	.227 (.063)	**1.18**
GloVe$_{300}$+BERT$_{\text{BASE}}$	**.488 (.070)**	**.258 (.059)**	**.405 (.040)**	2.52

Table 1: CROWDED Corpus GED with a sequence labeller and pre-trained (GloVe$_{300}$) or contextual (BERT$_{\text{BASE}}$) word representations as input. Precision, recall and F$_{0.5}$ are averaged over 10-fold cross-validation executed 10 times (standard deviations in brackets), with average time per run in hours.

also wishing to keep computational cost to a minimum (Strubell et al., 2019). Furthermore we tuned hyperparameters based on intuition and experience so as to offer good coverage of likely optimal values.

The state-of-the-art for GED involves contextual word representations concatenated to pre-trained representations. We tried several different pre-trained word representations as input to the model: English fastText vectors trained on 600B tokens from Common Crawl (Mikolov et al., 2018), Wikipedia2Vec trained on an April 2018 English Wikipedia dump (Yamada et al., 2020), and English GloVe trained on 840B tokens from Common Crawl (Pennington et al., 2014).

We also tried several different types of contextual word representation, obtained from the HuggingFace Transformers library using Flair NLP (Wolf et al., 2019; Akbik et al., 2019). These were namely: BERT$_{\text{BASE}}$ and BERT$_{\text{LARGE}}$ (Devlin et al., 2019), ELMo (Peters et al., 2018), FLAIR news (Akbik et al., 2018), GPT2 large (Radford et al., 2019), RoBERTa large (Liu et al., 2019), Transformer-XL (Dai et al., 2019), and XLNet (Yang et al., 2019). We compared the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and AdaDelta with a learning rate of 1.0 (Zeiler, 2012), and also tried different batch sizes of 16, 32 and 64.

The extra information described in section 4.1 was used for various experiments with additional features concatenated to the input embeddings, or as auxiliary objectives in multi-task learning settings. The additional features were treated as discrete or real values: real values were optionally normalised to values with a mean of 0 and standard deviation of 1, and discrete features were encoded as one-hot vectors (dimensionality reduction was attempted with no positive impact on performance). For auxiliary objectives we experimented with weights of 1.0, 0.1 and 0.01. As in both Knill et al. (2019) and Lu et al. (2019), we introduce error-annotated learner essays as additional training data – namely the FCE, W&I and JFLEG corpora. We try three settings for each corpus: training on the written data only and evaluating on CROWDED, or fine-tuning a pre-trained written GED model on CROWDED training texts, or combining the written and CROWDED corpora for training from the outset.

Recall that we set up our dataset for 10-fold cross-validation. As well as 10-folds, all experiments involve 10 different random seeds to provide us with a fair picture of variance in model training and performance. F$_{0.5}$ is the primary metric and results are reported as the mean of all seeds and folds (i.e. 10 folds x 10 seeds = 100 values). Average time is reported per experimental run (i.e. per set of cross-validation experiments, or per random seed).

4.3 Results

In terms of hyperparameter search, the best performing model involves contextual word representations from BERT$_{\text{BASE}}$ concatenated to pre-trained GloVe representations. The AdaDelta optimizer with a learning rate of 1.0 out-performed Adam with a learning rate of 0.001, and a batch size of 32 was better than 64 or 16. We label this best model 'GloVe$_{300}$+BERT$_{\text{BASE}}$'. This result is compared to a baseline GloVe$_{300}$ approach without BERT$_{\text{BASE}}$ representations in summary Table 1. The difference between the two models is statistically significant (Wilcoxon signed-rank test: $p < 0.001$).

Full hyperparameter experiments are reported in Table 3 (appendix). Note that GloVe$_{300}$+BERT$_{\text{BASE}}$ is not the best across the board but rather has the best combination of precision and recall: experiments with both the Adam optimizer and a larger batch size of 64 achieve higher precision; a smaller batch size of 16 achieves higher recall. Nevertheless GloVe$_{300}$+BERT$_{\text{BASE}}$ is the best performing model overall.

We report results with additional features and auxiliary objectives in the appendix in Table 4 (second table section downwards). In summary, no additional features or auxiliary objectives out-performed the

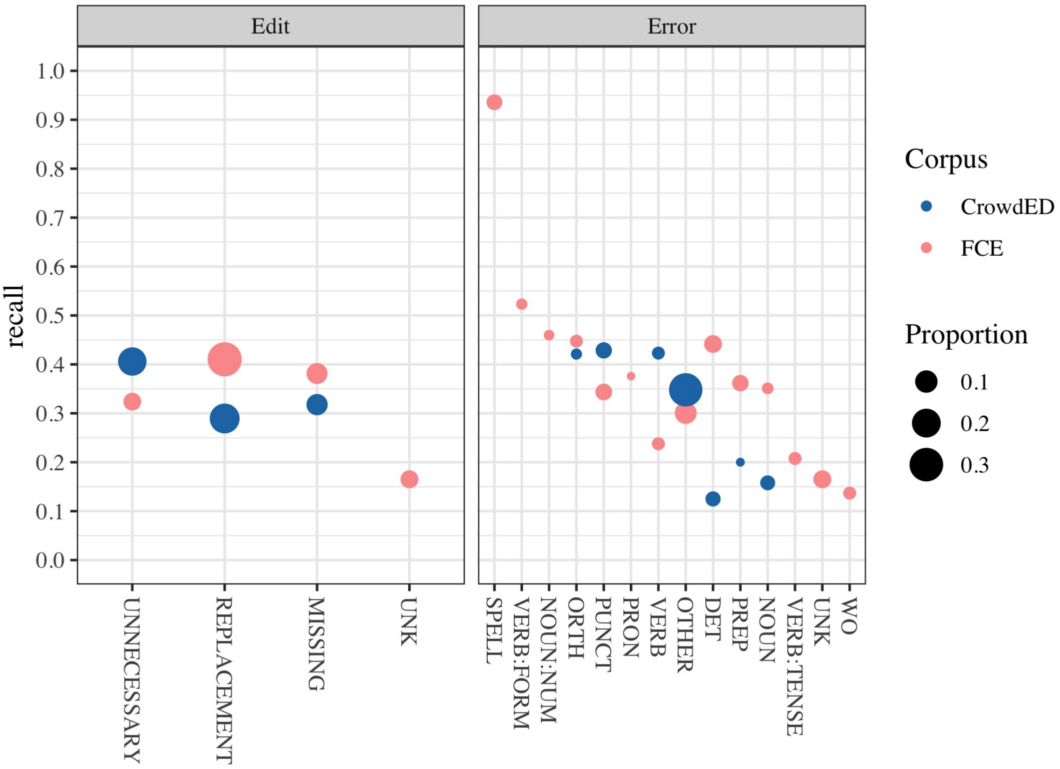


Figure 3: Recall for each edit and error type in the CrowdED and FCE corpora using the GloVe$_{300}$+BERT$_{\text{BASE}}$ GED model. Selected types with at least 1% of edits or errors in each corpus.

GloVe$_{300}$+BERT$_{\text{BASE}}$ model, which we take to mean that BERT representations are very strong, and any additional information for a dataset of this size only adds noise. Or it may be that different feature types or auxiliary objectives are required for spoken error detection. We infer this conclusion from experiments with a GloVe$_{300}$ model (no BERT) in which additional features *do* improve performance (top section, Table 4) so are evidently not unhelpful in themselves.

Similarly, experiments with extra training data from written corpora do show improvement over the GloVe$_{300}$ model (no BERT) but not over the best GloVe$_{300}$+BERT$_{\text{BASE}}$ model (appendix Table 5), whether combining written and spoken corpora from the outset or pre-training on written texts and fine-tuning on spoken data. It is possible that larger such corpora are needed to show any gain over a +BERT model: Knill et al. (2019) and Lu et al. (2019) use the 14 million word Cambridge Learner Corpus (Nicholls, 2003) and do see improvements in GED on other spoken corpora (NICT JLE and a Cambridge Assessment dataset) after fine-tuning.

4.4 Analysis

State-of-the-art performance with a similar model to GloVe$_{300}$+BERT$_{\text{BASE}}$ yields an F$_{0.5}$ of .573 for the FCE Corpus (Bell et al., 2019). Performance on CROWDED data is quite a bit lower: about .4 at best. We analyse performance on each of the edit and error types listed in Table 2, calculating accuracy of error detection for each type in the FCE test set and one of the ten CROWDED test sets (i.e. calculating recall). Figure 3 shows the edits and errors separately, including only those types which represent at least 1% of the edits in the test set. The proportion each type represents is indicated by datapoint size, CROWDED points are dark and FCE points are light, and the y-axis shows recall.

The first indicator of worse performance is the difference between CROWDED and FCE recall for the 'replacement' edit type (the majority edit type). The 'missing' edit type is also worse for CROWDED while 'unnecessary' edits are detected with better recall in CROWDED than FCE. Recall on the majority error type, 'other', is a little better in CROWDED than FCE, but for determiners and nouns it is notably worse. Also we see that spelling recall is very high for the FCE, whereas this error type is absent from CROWDED (except for some transcription anomalies). Of course this plot only tells part of the story: precision of GED is much higher on the FCE (currently .650 state-of-the-art) and therefore we are predicting more false positives in CROWDED which merits further investigation in future work.

5 Conclusion

In this paper we have presented a new resource for speech NLP: 1108 separate corrected transcriptions and grammatical error annotations for 383 distinct English recordings from the CROWDED Corpus. These are available for research use[6] and complement the existing CROWDED audio files and original transcriptions available in the same place. These data enable further research into automatic post-editing of speech transcriptions for readability (inserting full-stops and orthographic correction), which we have not explored here but could do in future work. In addition the error annotations allow experiments in grammatical error detection and correction (GED and GEC).

We undertook GED experiments in this work, using methods shown to be highly effective on written corpora. We find that a combination of contextual and pre-trained word representations as inputs to a bi-directional LSTM lead to good performance in sequence labelling word tokens in speech transcriptions as correct or incorrect. Performance is still some way off that for written GED, which perhaps may be accounted for by the small size of the dataset compared to written ones, the fact that the word representations are trained on written rather than spoken data, and the fairly different composition of error types found in CROWDED compared to equivalent written corpora. These factors, along with the possible need for extra pre-processing of the transcriptions to handle characteristic features of spoken language such as disfluencies, ellipsis and sections of unclear speech, may be explored in future work.

Another future improvement will be to augment the CROWDED Corpus with new data: new tasks, new languages, new recordings and annotation. Note that there are hundreds of CROWDED recordings which do not have the transcription updates and error annotation described here, so if funding allows there is unrealised potential in the existing data, besides adding to it with new data. Further insight may come from adapting the error typology to the spoken domain, in particular subdividing the majority 'other' error type into new types specific to spoken language, and then learning to better detect and correct these. We will then be able to improve upon the best GloVe_{300}+$\text{BERT}_{\text{BASE}}$ model with techniques for GED which are tailored to the kinds of errors found in spoken language.

Acknowledgements

This paper reports work supported by Cambridge Assessment, University of Cambridge. Additional funding was provided by the Cambridge Language Sciences Research Incubator Fund and the Isaac Newton Trust, and we are very grateful to Jane Walsh for her support. We thank the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research. Also, this work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service, provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council. We thank Dr Tim Polzehl and André Beyer of Crowdee for their help with the set up of our crowdsourcing jobs, as well as Dr Rogier van Dalen (now at Apple) and Prof Mark Gales of the ALTA Institute & Engineering Department, University of Cambridge for access to models and software for the merging of crowdsourced transcriptions. We thank the crowdworkers who, naturally, we do not know but who contributed hugely to this research; for instance taking great care over the transcription and annotation jobs, contacting us with queries, or sharing their voice recordings with us to form the corpus. Finally, we thank the three anonymous reviewers for their very helpful comments.

[6] https://www.ortolang.fr/market/corpora/ortolang-000913

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.

Claudia Baur, Cathy Chua, Johanna Gerlach, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2017. Overview of the 2017 Spoken CALL Shared Task. In *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*.

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2018. Overview of the 2018 Spoken CALL Shared Task. In *Proceedings of INTERSPEECH*.

Claudia Baur, Andrew Caines, Cathy Chua, Johanna Gerlach, Mengjie Qian, Manny Rayner, Martin Russell, Helmer Strik, and Xizi Wei. 2019. Overview of the 2019 Spoken CALL Shared Task. In *Proceedings of the 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE)*.

Samuel Bell, Helen Yannakoudakis, and Marek Rei. 2019. Context is key: Grammatical error detection with contextual word representations. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank LDC2012T13.

Brigitte Bigi. 2015. SPPAS – multi-lingual approaches to the automatic annotation of speech. *The Phonetician*, 111-112:54–69.

BNC Consortium. 2001. The British National Corpus, version 2 (BNC World).

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)*.

Christopher Bryant, Mariano Felice, Øistein 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 (BEA)*.

Andrew Caines, Christian Bentz, Calbert Graham, Tim Polzehl, and Paula Buttery. 2016. Crowdsourcing a multilingual speech corpus: recording, transcription and annotation of the CROWDED CORPUS. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Andrew Caines, Emma Flint, and Paula Buttery. 2017a. Collecting fluency corrections for spoken learner English. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Andrew Caines, Diane Nicholls, and Paula Buttery. 2017b. Annotating errors and disfluencies in transcriptions of speech. *University of Cambridge, Computer Laboratory*, UCAM-CL-TR-915.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson, 2020. *shiny: Web Application Framework for R*. R package version 1.4.0.2.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Proceedings of INTERSPEECH*.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Rachele De Felice and Stephen Pulman. 2008. A classifier-based approach to preposition and determiner error correction in L2 English. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*.

Bart Penning de Vries, Catia Cucchiarini, Stephen Bodnar, Helmer Strik, and Roeland van Hout. 2015. Spoken grammar practice and feedback in an ASR-based CALL system. *Computer Assisted Language Learning*, 28(6):550–576.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of Mechanical Turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*.

Keelan Evanini, Derrick Higgins, and Klaus Zechner. 2010. Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Jennifer Foster and Carl Vogel. 2004. Parsing ill-formed text using an error grammar. *Artificial Intelligence Review*, 21:269–291.

Roland Fried, Karen Schettlinger, and Matthias Borowski, 2019. *robfilter: Robust Time Series Filters*. R package version 4.1.2.

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92)*.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2004. The NICT JLE Corpus: Exploiting the language learners' speech database for research and education. *International Journal of The Computer, the Internet and Management*, 12(2):119–125.

Sudhanshu Kasewa, Pontus Stenetorp, and Sebastian Riedel. 2018. Wronging a right: Generating better errors to improve grammatical error detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kate Knill, Mark Gales, Potsawee Manakul, and Andrew Caines. 2019. Automatic grammatical error detection of non-native spoken learner English. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Ann Lee and James Glass. 2012. Sentence detection using multiple annotations. In *Proceedings of INTERSPEECH 2012*. International Speech Communication Association.

Kyusong Lee, Seonghan Ryu, Paul Hongsuck Seo, Seokhwan Kim, and Gary Geunbae Lee. 2014. Grammatical error correction based on learner comprehension model in oral conversation. In *Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, 1907.11692.

Oier Lopez de Lacalle and Eneko Agirre. 2015. Crowdsourced word sense annotations and difficult words and examples. In *Proceedings of the 11th International Conference on Computational Semantics*.

Yiting Lu, Mark Gales, Kate Knill, Potsawee Manakul, Linlin Wang, and Yu Wang. 2019. Impact of ASR performance on spoken grammatical error detection. In *Proceedings of INTERSPEECH*.

Nitin Madnani, Joel Tetreault, Martin Chodorow, and Alla Rozovskaya. 2011. They can help: using crowdsourcing to improve the evaluation of grammatical error detection systems. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2016. Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.

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 *Eighteenth Conference on Computational Natural Language Learning, Proceedings of the Shared Task*. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics 2003 conference; UCREL technical paper number 16*. Lancaster University.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Gabrielė Aleksandravičiūtė, Lene Antonsen, Katya Aplonova, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, John Bauer, Sandra Bellato, Kepa Bengoetxea, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Johannes Heinecke, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Takumi Ikeda, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Andre Kaasen, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Arne Köhn, Kamil Kopacewicz, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Lucia Lam, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Yuan Li, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Tomohiko Morioka, Shinsuke Mori, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adédayọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Daria Petrova, Slav Petrov, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Jack Rueter, Shoval Sadde, Benoît Sagot, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Takaaki Tanaka, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2019. Universal Dependencies 2.4. http://hdl.handle.net/11234/1-2988.

Stefan Palan and Christian Schitter. 2018. Prolific.ac – a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Gabriele Paolacci and Jesse Chandler. 2014. Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

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

Manny Rayner, Ian Frank, Cathy Chua, Nikos Tsourakis, and Pierrette Bouillon. 2011. For a fistful of dollars: Using crowd-sourcing to evaluate a spoken language CALL application. In *Proceedings of the Fourth ISCA Workshop on Speech and Language Technology in Education (SLaTE)*.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Marek Rei and Helen Yannakoudakis. 2017. Auxiliary objectives for neural error detection models. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Joel Tetreault and Martin Chodorow. 2008. Native judgments of non-native usage: experiments in preposition error detection. In *Proceedings of the workshop on Human Judgments in Computational Linguistics, COLING*.

Rogier van Dalen, Kate Knill, Pirros Tsiakoulis, and Mark Gales. 2015. Improving multiple-crowd-sourced transcriptions using a speech recogniser. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers.

Mark van der Loo. 2014. The stringdist package for approximate string matching. *The R Journal*, 6:111–122.

Yu Wang, Mark Gales, Katherine Knill, Konstantinos Kyriakopoulos, Andrey Malinin, Rogier van Dalen, and Mohammad Rashid. 2018. Towards automatic assessment of spontaneous spoken English. *Speech Communication*, 104:47–56.

Jan Wijffels, 2019. *udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' NLP Toolkit*. R package version 0.8.3.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv*, 1910.03771.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. *arXiv*, 1812.06280v3.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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*.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. CWIG3G2 – complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv*, 1212.5701.

Appendix A: corpus statistics & GED experiment results

		FCE	All speakers	Native speakers	Learners
Edit type	Missing	21.0%	13.9%	12.8%	14.8%
	Replacement	64.4%	47.9%	44.3%	51.0%
	Unnecessary	11.5%	38.2%	42.9%	34.2%
Error type	Adjective	1.4%	0.8%	0.5%	1.0%
	Adjective:form	0.3%	0.06%	0.1%	0.02%
	Adverb	1.9%	1.5%	1.5%	1.8%
	Conjunction	0.7%	1.3%	1.1%	1.5%
	Contraction	0.3%	0.4%	0.6%	0.4%
	Determiner	10.9%	4.0%	2.6%	5.2%
	Morphology	1.9%	0.6%	0.4%	0.7%
	Noun	4.6%	5.8%	5.5%	6.1%
	Noun:inflection	0.5%	0.01%	0.0%	0.02%
	Noun:number	3.3%	1.0%	1.0%	1.3%
	Noun:possessive	0.5%	0.1%	0.2%	0.02%
	Orthography	2.9%	3.0%	3.4%	2.9%
	Other	13.3%	61.0%	66.3%	56.4%
	Particle	0.3%	0.5%	0.6%	0.4%
	Preposition	11.2%	2.9%	2.0%	3.7%
	Pronoun	3.5%	1.2%	1.0%	1.4%
	Punctuation	9.7%	8.7%	8.8%	8.7%
	Spelling	9.6%	0.3%	0.3%	0.4%
	Verb	7.0%	3.1%	2.1%	3.9%
	Verb:form	3.6%	0.4%	0.4%	0.6%
	Verb:inflection	0.2%	0.01%	0.0%	0.02%
	Verb:subj-verb-agr	1.5%	0.3%	0.2%	0.4%
	Verb:tense	6.0%	1.1%	0.6%	1.7%
	Word order	1.8%	1.2%	0.8%	1.6%
Corpus stats	N.speakers	-	76	30	46
	Recordings	-	383	140	243
	Texts/Transcripts	1244	1108	406	702
	Word tokens	531,416	39,726	20,814	18,912
	Total edits	52,671	8454	3926	4528

Table 2: The proportional distribution of error types determined by Prolific crowdworkers for a subset of the CROWDED Corpus (for a full description of error types see Bryant et al. (2017)); followed by descriptive corpus statistics. Proportions supplied for the FCE Corpus for comparison, from Bryant et al. (2019).

Pretrained	**Contextual**	**Optimizer**	**LR**	**Batch**	**P**	**R**	**F$_{0.5}$**	**Hours**
GloVe$_{300}$	n/a	AdaDelta	1.0	32	.386 (.115)	.111 (.058)	.227 (.063)	**1.18**
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.488 (.070)	.258 (.059)	**.405 (.040)**	2.52
GloVe$_{300}$	BERT$_{\mathrm{LG}_{cased}}$	AdaDelta	1.0	32	.446 (.079)	.244 (.066))	.373 (.046)	2.76
GloVe$_{300}$	RoBERTa$_{\mathrm{LG}}$	AdaDelta	1.0	32	.446 (.067)	.221 (.062)	.360 (.049)	2.73
GloVe$_{300}$	BERT$_{\mathrm{LG}_{uncased}}$	AdaDelta	1.0	32	.432 (.073)	.207 (.056)	.342 (.041)	4.64
GloVe$_{300}$	ELMo	AdaDelta	1.0	32	.402 (.092)	.186 (.059)	.308 (.039)	5.86
GloVe$_{300}$	Transfo-XL	AdaDelta	1.0	32	.394 (.085)	.157 (.050)	.287 (.047)	5.11
GloVe$_{300}$	FLAIR	AdaDelta	1.0	32	.392 (.093)	.166 (.075)	.286 (.060)	4.88
GloVe$_{300}$	XLNet	AdaDelta	1.0	32	.394 (.059)	.146 (.045)	.284 (.035)	3.34
GloVe$_{300}$	GPT2$_{\mathrm{LG}}$	AdaDelta	1.0	32	.408 (.106)	.144 (.066)	.270 (050)	3.92
Wiki2vec	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.489 (.087)	.254 (.069)	.399 (.042)	2.11
GloVe$_{400}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.480 (.073)	.256 (.068)	.398 (.039)	2.13
GloVe$_{200}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.480 (.071)	.253 (.061)	.484 (.039)	2.11
FastText$_{\mathrm{CC}}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.482 (.071)	.249 (.058)	.397 (.039)	2.42
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	Adam	.001	32	**.491 (.073)**	.239 (.052)	.397 (.043)	2.23
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	.001	32	.194 (.059)	.354 (.287)	.191 (.063)	1.27
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	64	**.491 (.077)**	.251 (.056)	.403 (.040)	1.93
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	16	.471 (.073)	**.260 (.055)**	.396 (.041)	2.21
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.481 (.073)	.255 (.063)	.399 (.044)	2.88
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.483 (.077)	.255 (.067)	.400 (.043)	2.78
GloVe$_{300}$	BERT$_{\mathrm{BASE}_{cased}}$	AdaDelta	1.0	32	.485 (.078)	.250 (.060)	.398 (.040)	2.27

Table 3: CROWDED Corpus GED with variations of pre-trained and contextual word representations, optimizer, learning rate (LR) and batch size. Precision, recall and F$_{0.5}$ averaged over 10-fold cross-validation executed 10 times (standard deviations in brackets), with average time per run in hours.

Model	Learned as	P	R	$F_{0.5}$	Hours
GloVe$_{300}$	n/a	.386 (.115)	.111 (.058)	.227 (.063)	**1.18**
+Lemma	discrete.feat:50w	.373 (.099)	.123 (.061)	.239 (.057)	1.80
+UPOS	discrete.feat:16w	.385 (.115)	.113 (.055)	.232 (.067)	1.46
+XPOS	discrete.feat:50w	.377 (.111)	.114 (.063)	.226 (.063)	1.33
+DepRel	discrete.feat:39w	.391 (.115)	.112 (.059)	.233 (.063)	1.76
+HeadToken	real.feat:normed	.388 (.120)	.098 (.052)	.213 (.064)	1.39
+ComplexWord	discrete.feat:1w	.378 (.111)	.119 (.076)	.231 (.061)	1.45
GloVe$_{300}$+BERT$_{\mathrm{BASE}}$	n/a	**.488 (.070)**	.258 (.059)	**.405 (.040)**	2.52
+Lemma	discrete.feat:50w	.455 (.080)	.233 (.053)	.372 (.043)	4.84
+UPOS	discrete.feat:16w	.450 (.076)	.235 (.057)	.372 (.046)	5.12
+XPOS	discrete.feat:42w	.457 (.078)	.250 (.069)	.380 (.046)	2.13
+DepRel	discrete.feat:39w	.449 (.078)	.257 (.061)	.380 (.042)	2.74
+HeadToken	real.feat:unnormed	.456 (.079)	.229 (.057)	.370 (.045)	5.06
+ComplexWord	discrete.feat:1	.452 (.074)	.260 (.067)	.383 (.046)	2.27
+Lemma	discrete.target:wt1.0	.450 (.077)	.236 (.056)	.371 (.041)	4.49
+Lemma	discrete.target:wt0.1	.447 (.074)	.241 (.057)	.372 (.042)	4.42
+Lemma	discrete.target:wt0.01	.456 (.078)	.231 (.060)	.373 (.045)	4.73
+XPOS	discrete.target:wt1.0	.454 (.073)	.226 (.053)	.369 (.042)	5.98
+XPOS	discrete.target:wt0.1	.458 (.085)	.233 (.061)	.371 (.040)	5.93
+XPOS	discrete.target:wt0.01	.453 (.081)	.230 (.057)	.368 (.044)	5.93
+DepRel	discrete.target:wt1.0	.457 (.070)	.231 (.055)	.373 (.041)	4.71
+DepRel	discrete.target:wt0.1	.454 (.081)	.232 (.058)	.369 (.042)	4.86
+DepRel	discrete.target:wt0.01	.451 (.075)	.237 (.060)	.372 (.043)	5.26
+HeadToken	real.target:wt1.0	.454 (.079)	.236 (.058)	.373 (.039)	5.01
+HeadToken	real.target:wt0.1	.447 (.077)	.241 (.058)	.372 (.044)	4.98
+HeadToken	real.target:wt0.01	.454 (.077)	.236 (.055)	.373 (.043)	4.92
+ComplexWord	discrete.target:wt1.0	.449 (.076)	.242 (.063)	.373 (.044)	5.05
+ComplexWord	discrete.target:wt0.1	.452 (.078)	.244 (.055)	.377 (.043)	4.89
+ComplexWord	discrete.target:wt0.01	.454 (.084)	.228 (.055)	.368 (047)	4.93
+ErrorType	discrete.target:wt1.0	.453 (.075)	.237 (.063)	.372 (.042)	4.80
+ErrorType	discrete.target:wt0.1	.449 (.078)	.232 (.058)	.368 (.040)	5.35
+ErrorType	discrete.target:wt0.01	.448 (.073)	.235 (.051)	.371 (.042)	5.24
+CrowdEDfreqs	real.feat:normed	.444 (.073)	**.264 (.066)**	.381 (.042)	2.62
+BNCfreqs	real.feat:normed	.458 (.067)	.249 (.064)	.383 (.043)	2.62
+1BWfreqs	real.feat:normed	.452 (.076)	.231 (.052)	.371 (.039)	5.99
+WordDuration	real.feat:normed	.453 (.083)	.232 (.055)	.370 (.042)	3.68
+PauseDuration	real.feat:normed	.456 (.077)	.230 (.051)	.373 (.046)	4.54
+Pitch	real.feat:normed	.450 (.079)	.233 (.055)	.370 (.046)	4.15
+Energy	real.feat:normed	.449 (.077)	.237 (.055)	.372 (.047)	4.54

Table 4: CROWDED Corpus GED with additional features and auxiliary objectives. Precision, recall and $F_{0.5}$ averaged over 10-fold cross-validation executed 10 times (standard deviations in brackets), with average time per run in hours. Features are marked as discrete or real-valued, with discrete features encoded as one-hot vectors and their width noted, and real-valued features may be normalised to values between 0 and 1. Auxiliary objectives are marked as discrete or real-valued targets, with varying weights.

Model	**CROWDED injection**	**P**	**R**	**$F_{0.5}$**	**Hours**
GloVe$_{300}$	outset	.386 (.115)	.111 (.058)	.227 (.063)	1.18
+FCE	none	.180 (.)	.042 (.)	.108 (.)	2.46
+FCE	outset*	.409 (.046)	.133 (.067)	.273 (.032)	25.3
+FCE	fine-tuning	.386 (.038)	.169 (.047)	.303 (.042)	1.21
+W&I	none	.411 (.)	.089 (.)	.238 (.)	2.79
+W&I	outset*	.434 (.008)	.154 (.041)	.313 (.038)	32.5
+W&I	fine-tuning	.392 (.076)	.183 (.067)	.305 (.048)	0.85
+JFLEG	none	.312 (.)	.128 (.)	.242 (.)	**0.06**
+JFLEG	outset	.405 (.147)	.130 (.072)	.231 (.078)	1.80
+JFLEG	fine-tuning	.358 (.072)	.136 (.079)	.243 (.051)	0.91
GloVe$_{300}$+BERT$_{\text{BASE}}$	outset	**.488 (.070)**	.258 (.059)	**.405 (.040)**	2.52
+FCE	none	.427 (.)	.205 (.)	.351 (.)	8.34
+FCE	outset*	.460 (.056)	.218 (.037)	.371 (.031)	82.7
+FCE	fine-tuning	.460 (.082)	.214 (.044)	.365 (.042)	4.93
+W&I	none	.462 (.)	.211 (.)	.373 (.)	13.2
+W&I	outset*	.467 (.049)	.234 (.037)	.386 (.036)	91.2
+W&I	fine-tuning	.464 (.064)	.224 (.044)	.375 (.036)	5.23
+JFLEG	none	.365 (.)	**.402 (.)**	.372 (.)	**0.38**
+JFLEG	outset	.450 (.075)	.258 (.061)	.382 (.042)	5.31
+JFLEG	fine-tuning	.458 (.083)	.225 (.050)	.371 (.046)	3.80

Table 5: CROWDED Corpus GED with additional training data. Precision, recall and $F_{0.5}$ averaged over 10-fold cross-validation executed 10 times (standard deviations in brackets; dots indicate that there was only one run and hence no standard deviation), with average time per run in hours. Point of introduction of CROWDED training data is noted as 'none' (no CROWDED data used in training), 'outset' (mixed with written corpus) or 'fine-tuning' (used to update the weights of a pre-trained model). *Asterisked experiments were run with 4 random seeds rather than 10 due to their long duration.