Collecting fluency corrections for spoken learner English

Andrew Caines[1] **Emma Flint**[1] **Paula Buttery**[2]
[1] Department of Theoretical and Applied Linguistics
[2] Computer Laboratory
University of Cambridge, Cambridge, U.K.
{apc38|emf40|pjb48}@cam.ac.uk

Abstract

We present crowdsourced collection of error annotations for transcriptions of spoken learner English. Our emphasis in data collection is on *fluency* corrections, a more complete correction than has traditionally been aimed for in grammatical error correction research (GEC). Fluency corrections require improvements to the text, taking discourse and utterance level semantics into account: the result is a more naturalistic, holistic version of the original. We propose that this shifted emphasis be reflected in a new name for the task: 'holistic error correction' (HEC). We analyse crowdworker behaviour in HEC and conclude that the method is useful with certain amendments for future work.

1 Introduction

By convention, grammatical error detection and correction (GEC) systems depend on the availability of labelled training data in which tokens have been annotated with an error code and a correction. In (1) for example, taken from the open FCE subset of the Cambridge Learner Corpus (CLC) (Nicholls, 2003; Yannakoudakis et al., 2011), the original token 'waken' is coded as a 'TV' (verb tense) error and annotated with the correct token 'woken' on the right-hand side of the pipe.

(1) In the morning, you are <NS type="TV"> waken|woken </NS> up by a singing puppy.

Such efforts to annotate learner corpora are time-consuming and costly, but with sufficient quantities it is possible to train GEC systems to identify and correct errors in unseen texts. For example, 29 million tokens of the CLC have been error-annotated, of which the FCE is a publicly-available 500k token subset (Yannakoudakis et al., 2011). The Write & Improve[1] GEC system (W&I) has been built on these resources (Andersen et al., 2013), providing automated assessment and per-token error feedback. In common with other GEC systems, W&I prizes precision ahead of recall – so as to avoid false positive corrections being presented to the user.

Indeed the field of GEC as a whole adopts a conservative stance on error correction (hence preferring precision to recall in the well-established $F_{0.5}$ metric), is focused at the token level, and has tended to train separate classifiers for each error type (De Felice and Pulman, 2008; Tetreault et al., 2010; Dahlmeier and Ng, 2012), has adopted a machine translation approach (Brockett et al., 2006; Park and Levy, 2011; Yuan et al., 2016), or a hybrid of the two (Rozovskaya and Roth, 2016). Ease of correction varies by class of error, with Table 1 showing best-to-worst recall of the top-performing system for each error type in the CoNLL-2014 shared task on GEC of NUCLE data (Ng et al., 2014).

It is apparent that detection rates are relatively high for certain error types, namely issues of register, subject-verb agreement, determiner errors and noun number. We note that there are several error types in the lower half of Table 1 – such as sentence fragments, linking words, redundancy, unclear meaning and wrong collocations – which relate to fluency broadly defined. This indicates that these error types are harder to solve, or at least have not been worked on so much. Either way they require further attention.

Some notable blind-spots of the current GEC approach are found above the token level, in sentence and discourse level semantics and coher-

[1] https://writeandimprove.com

Code	Error	Training %	Recall %	System
Wtone	inappropriate register	1.3	81.8	AMU
SVA	subject-verb agreement	3.4	70.3	CUUI
ArtOrDet	article/determiner error	14.8	58.9	CUUI
Nn	noun number	8.4	58.7	AMU
Spar	parallelism	1.2	50.0	RAC
WOadv	adjective/adverb order	0.8	47.6	CAMB
Wform	word form	4.8	45.6	AMU
Mec	spelling & punctuation	7.0	43.5	RAC
Prep	preposition	5.4	38.3	CAMB
V0	missing verb	0.9	36.7	NARA
Vm	modal verb	1.0	35.9	RAC
Vform	verb form	3.2	27.6	NARA
Vt	verb tense	7.1	26.2	RAC
Sfrag	sentence fragment	0.6	25.0	UMC
Pform	pronoun form	0.4	22.6	CAMB
Trans	linking words	3.1	21.4	CAMB
Npos	possessive	0.5	20.0	NARA
Rloc–	redundancy	10.5	20.2	CAMB
Pref	pronoun reference	2.1	19.4	CAMB
Um	unclear meaning	2.6	15.8	PKU
Ssub	subordinate clause	0.8	15.4	NARA
Wci	wrong collocation	11.8	12.0	AMU
WOinc	word order	1.6	6.7	UMC
Others	miscellaneous	3.3	3.1	RAC
Cit	citation	1.5	0	_
Smod	dangling modifier	0.1	0	_
Srun	run on sentence	1.9	0	_
Wa	acronym	0.1	0	_

Table 1: Best recall by error type in the CoNLL-2014 shared task on GEC (Ng et al., 2014), including frequency of error type in the training data, and recall against gold-standard edits[3].

ence. Hence there has been a call for greater emphasis on *fluency* in error correction (Sakaguchi et al., 2016). We may think of fluency as encompassing the grammaticality-per-token focus of GEC thus far, with added layers of sentence and discourse level semantics and coherence. It is also more than just spoken fluency, which is a common usage of the term. Instead, it is a holistic notion of all-linguistic performance competence.

For example, in (2) we see the kind of sentence which in the GEC approach might only be corrected for the ungrammaticality of 'shorten', as in (3). But in fact the new version still lacks native-like fluency. The meaning is clear, a fact we can use to offer the fluent correction seen in (4).

(2) From this scope, social media has shorten our distance[4].

(3) From this scope, social media has shortened our distance.

(4) From this perspective, social media has shortened the distance between us.

Furthermore, in speech the problem is heightened by the fact that, relative to grammaticality, fluency is arguably of greater importance than it is in writing. In the immediate communication scenario of spontaneous conversation – the default setting for speech, though there are others – the signal is ephemeral and interlocutors are both forgiving of errors and adept at rapid repair (Clark and Schaefer, 1987; Cahn and Brennan, 1999; Branigan et al., 2007).

Except in classroom settings or when explicitly asked to do so, the listener rarely corrects or points out the speaker's grammatical errors. Instead she tends to signal understanding, offer signs of agreement or other emotional reaction, and seek clarification – all of which have been listed among the typical acts of 'alignment' in dialogue (Pickering and Garrod, 2004). She focuses more on the meaning of what is said, and the fluency of linguistic construction plays an important role in how successfully meaning is conveyed. We work with spoken data from learners, and the implication is that fluency takes on added importance in our view.

We therefore support the call for greater emphasis on fluency rather than grammaticality (Sakaguchi et al., 2016), propose that we represent that changed emphasis with a changed label for the field – 'holistic error correction' (HEC) is our

[4]Examples (2)–(4) from Sakaguchi et al (2016).

suggestion – and finally we present and evaluate a crowdsourcing method for fluency correction of transcriptions of spoken learner English. We analyse crowdworker behaviour in this task, discuss how the data can be used, and assess how the method can be improved in future work with a view to creating an open dataset of fluency annotations.

2 Crowdsourcing

Annotation of language corpora is an expensive process in both cost and time. And yet the labelling of corpora is highly desired as it opens the data up to further linguistic analysis and machine learning experiments. We describe our efforts to use *the crowd* for fast, low-cost annotation tasks and conclude as others have done before that, 'they can help' (Madnani et al., 2011) – the resultant annotations are good enough to be useful.

We engaged 120 crowdworkers through Prolific Academic[5] to provide fluency corrections for transcriptions of spoken learner English. A recent evaluation of Prolific Academic and two other widely-used crowdsourcing services, CrowdFlower and Amazon Mechanical Turk, reported favourable comparisons for Prolific in terms of both data quality and participant diversity (Peer et al., 2017). We recruited workers from Prolific on condition that they had an approval rating of 95% or more, that they reported English to be their first language, and that they were educated to at least U.K. GCSE level or equivalent (normally taken at 16 years).

This meant that the worker pool was reduced to 17,363 from an original pool of 23,973 at the time of recruitment (January 2017). Nevertheless recruitment proceeded at a rapid pace and all tasks had been completed within 24 hours of launch. Workers were paid £1 for what was estimated to be 10 minutes of work correcting 16 items (plus the two test items we put in to catch pathological contributions[6]). In fact our 120 workers spent an average of 16 minutes on the task (max=43 mins, min=7.2 mins, st.dev=7.6 mins). Workers declared themselves to be 45% female and 55% male and were in the age range 17–70 years (mean=33).

The data were language learner monologues from Cambridge English Language Assessment Business Language Testing Service (BULATS) oral exams[7]. The learners were prompted to discuss business topic scenarios and allowed to talk for up to a minute at a time. Recordings were transcribed by two different workers from the Amazon Mechanical Turk crowdsourcing service and subsequently combined into a single transcript by finding the best path through a word network constructed out of the two transcript versions, using automated speech recognition (van Dalen et al., 2015). This method is of course not error-free: van Dalen *et al* report a word error rate of 28% on a 55k token test set.

The learners' first languages (L1s) were Arabic, Dutch, French, Polish, Thai and Vietnamese (Table 2), and their proficiency was judged by two examiners such that they could be placed on the CEFR scale (Common European Framework of Reference for Languages) as shown in Table 3.

Whilst the learners are fairly well-balanced by L1 in terms of both speaker numbers and token counts, it is clear that there's a skew towards the middle ranks of the CEFR scale – namely, A2 to C1 – with fewer A1 learners and only two C2 level learners. As would be expected, the token-to-speaker ratio rises with increasing proficiency: thus there are more tokens for each proficiency level (excepting C2), even where speaker numbers do not go up.

We prepared a web application using R Shiny and shinyapps hosting (R Core Team, 2017; Chang et al., 2016; Allaire, 2016). We named it 'Correcting English' and directed crowdworkers to it from Prolific Academic. If necessary, transcriptions were divided into 'speech-units' (Moore et al., 2016) – analogous to the sentence in writing – and presented speech-unit by speech-unit (SU). Workers were greeted with a welcome page explaining that they would be shown transcriptions of spoken learner English, that the learners were talking about business topics, and that they could expect to see mistakes.

Workers were asked to make corrections so that, "it sounds like something you would expect to hear or produce yourself in English". Whether the target should be the proficiency of a native speaker or a high proficiency learner is a fraught ques-

[5] http://www.prolific.ac

[6] These were the straightforward grammatical errors in, 'The currency of the USA be dhollars', and, 'The capital of the UK are Londoin', where we could pattern match for the corrections we expected. The absence of such corrections warned us to check the worker's whole contribution and judge whether to reject it and refuse payment.

[7] http://www.bulats.org

L1	Speakers	Tokens	SUs
Arabic	40	12,181	425
Dutch	33	11,549	396
French	37	11,716	383
Polish	40	9729	393
Thai	37	10,207	414
Vietnamese	39	9858	361
Total	226	65,240	2372

Table 2: L1 of speakers in the BULATS corpus: number of tokens and speech-units per group.

CEFR	Speakers	Tokens	SUs
A1	38	4553	325
A2	48	9584	451
B1	48	14,766	520
B2	48	16,854	509
C1	42	17,749	541
C2	2	624	26
Total	226	65,240	2372

Table 3: CEFR proficiency level of speakers in the BULATS corpus: number of tokens and speech-units per group.

tion in second language acquisition research, so we avoid reference to any such target and instead ask the worker to envisage how they might express the information contained in the SU. We intended that this gave the worker a concrete standard of English to aim for, and we assume that they are native speakers in any case, since we filtered for that in the recruitment stage. Moreover it encourages them to think about how they would *speak* the same thought, the intention being that this would lead them to think more about fluency than about grammaticality. We added that they should make as many changes as necessary, echoing Sakaguchi and colleagues' instruction for 'fluency edits' as opposed to 'minimal edits' (2016).

On the annotation page, workers were also able to view the context of a learner's response: that is, a summary of the 'prompt' to which they had responded. They could opt to skip the given transcription if they could not make any sense of it (and it would be replaced with another: such a move did not 'run down' the 18 required annotations). They could indicate with a tick-box that the transcription needed no correction. And they could grade their own confidence in their judgements, from 'not sure' to 'very sure' with 'quite sure' in between. A screenshot of a Correcting English page is given in Figure 1.

Once the worker completed 18 annotations (the 16 BULATS items and 2 test items) they were redirected to Prolific Academic and we were required to approve or reject their submission. In total we approved 120 submissions.

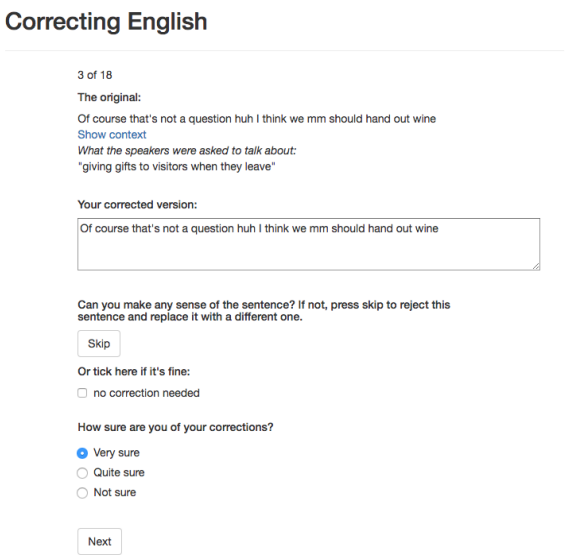

Figure 1: Screenshot from the Correcting English web application for crowdsourcing fluency corrections of spoken learner English: note that the original speech-unit is reproduced verbatim in the correction text-box, ready for the crowdworker to edit (or not).

3 Results

The BULATS dataset is different to those previously submitted for crowdsourced error annotation, to the best of our knowledge, in that it is *spoken* data and it is *learner* English. In all, 1507 unique SUs were selected at random and presented to crowdworkers for annotation, representing 63.5% of the 2372 SUs in the corpus. Workers made a total of 5706 judgements, excluding the test items.

3.1 Skipped speech-units

The majority of judgements were 'skip' moves to reject the presented SU. Overall workers skipped almost two-and-a-half SUs for every one they annotated (Table 4).

We found that variation in proficiency level explains the SU skip rate to some extent. The ratio of skipped to annotated SUs decreases from 5.8:1 to 1.5:1 from level A1 to C1, indicating that workers were more willing to annotate SUs uttered by higher proficiency speakers. There is a non-significant correlation between the percen and the grade assigned to the recording ($r = -0.182, p <$

CEFR	Skips	Annotations	Skip:Annotation	Unique SUs	Corpus %
A1	507	87	5.8	46	14.2
A2	832	238	3.5	117	25.9
B1	948	387	2.4	162	31.2
B2	837	359	2.3	164	32.2
C1	870	582	1.5	232	42.9
C2	23	18	1.3	6	23.1
Total	4017	1671	2.4	727	30.6

Table 4: CEFR proficiency level of speakers in the BULATS corpus: number of tokens and speech-units per group.

$0.001, df = 1155$). As a consequence our corpus of annotations is skewed towards higher proficiency levels (ignoring the small C2 subset for now), with almost half of the C1 SUs in our corpus being annotated at least once, in contrast to just one-sixth of A1 SUs (Table 4).

Of the SUs presented to crowdworkers, 348 were never skipped (Table 5). Recall that the skip action was intended for workers to indicate that they could make no sense of the speech-unit, and therefore could not reasonably be expected to correct it. Of the skipped SUs, 282 were skipped once only. Since linguistic intuitions are highly subjective, we put these aside as singular opinions on the SUs while we wait for a second opinion. Therefore we have 877 SUs which have been skipped two or more times, and we pay attention to this subset in some way.

Skips	SUs	Skips	SUs
0	348	9	27
1	282	10	9
2	259	11	7
3	194	12	8
4	128	13	5
5	97	14	3
6	59	15	5
7	48	16	2
8	24	18	2

Table 5: Number of skips per speech-unit in the BULATS corpus.

Examples of highly-skipped SUs include the following:

(5) A lot of coaching ment mentor.

(6) Ah we work very very well together ah we uh very close we can share lots of things er we also have time to uh sit down and talk about how school is developing and ah whether we are doing the right things together or not.

(7) Uh so I think I think location of facility is where the is good to store it to store.

In (5) the SU is too short, disfluent and lacking in a main verb to make any sense of. In contrast (6) is very long, peppered with filled pauses ('ah', 'uh', 'er'), and made up of several main clauses run on to one another in a chain. Both are difficult to make sense of for different reasons. Both were spoken by learners of CEFR level C1, whereas in (7) the level is B1 and the difficulty in interpretation perhaps stems more from the low proficiency level of the speaker.

How can we make use of the information in crowdworkers' skipping actions? We could interpret them as judgements as to the futility of attempting automatic correction on these units. For example, we could choose to exclude those SUs which have been skipped on at least two of the occasions they have been presented to crowdworkers. These SUs would constitute a 'nonsensical' portion of the corpus which (for now) we might deem too hard to automatically correct, as it is not possible to infer what the speaker intended to say. With the proposed threshold, 282 SUs would have to be set aside – or, 38.8% of the 727 SUs in the current dataset.

The implication for HEC evaluation is that we are only judging system performance against those SUs which we can reasonably expect to be corrected. The implication for computer-assisted language learning (CALL) applications is that if such an utterance were automatically detected, the system could ask the learner to clarify what they said or ask them to try again, rather than attempting a correction and damaging the system's reputation through nonsensical corrections to nonsensical SUs. However, it is apparent that many SUs would be trimmed through this method and with the proposed threshold. Is this a sensible approach? We leave this as a matter for debate, and welcome feedback in this regard.

3.2 Corrected speech-units

In terms of annotations then, 727 (30.6%) of the 2372 SUs in the corpus were annotated at least once (Table 4). If all 120 crowdworkers had submitted 16 SU annotations of suitable quality, it would give us a corpus of 1920 annotated SUs. However, in a quality control stage we removed 249 units due to poor contributions by workers, thereby losing just over one-eighth of the total submissions and leaving us with 1671 remaining annotations. Data loss of 13% seems a reasonable amount to allow for in designing a crowdsourcing study, and certainly we never expected a 100% success rate in terms of data quality.

These 1671 remaining annotations represent 727 unique SUs. Thus we have approximately two annotations for each SU on average. How can we assess what changes crowdworkers made to the original texts? Firstly we note that on the whole SUs were shortened in correction: the mean character difference between the original and corrected SU is -9.2 characters, while the median was -4 characters.

Self-reported confidence levels are generally high: workers rated their confidence level as 'very sure' or 'quite sure' for 85% of their annotations. We could choose to exclude the remaining 15% of annotations of which the workers declared themselves unsure. This would reduce the 1671 annotations to 1425 and the number of included SUs from 727 to 632. That would be the conservative approach, and probably the decision one would take before training a HEC system. Nevertheless we can use this information in evaluating HEC outputs, weighting scoring so that hypotheses measured against gold-standard fluency edits (of which the worker is at least quite sure) are valued more highly than those measured against silver-standard edits (the 'not sure' annotations).

Moreover, confidence level tends to be lower the greater the character difference between original and corrected SUs: in Figure 2 we see that the character difference values are more widely spread around the zero mark for the lower confidence levels, 'not sure' and 'quite sure'. For 'very sure' on the other hand, there is a peak of character differences around the zero mark, suggesting that no change has been made in the majority of cases. This indicates that crowdworkers tended to feel unsure when they *took action*: whether this is a property of the dataset or human nature is a matter for further investigation. It could also be that where no change was needed, the worker felt no need to change the confidence level from its default setting ('very sure'). Thus in future work we will consider alternative methods of collecting confidence ratings: either with larger scales or an interface other than radio buttons.

Another indicator of the changes made by the crowdworkers comes from lexical diversity scores: the mean type-token ratio (TTR) of the original SUs is 0.872 (st.dev=0.114), whereas mean TTR of the corrected SUs is 0.915 (st.dev=0.089). This overall increase in diversity suggests that one way in which workers 'improved' the SUs was to make them more expressive in terms of vocabulary use.

Of the 727 SUs annotated by crowdworkers, 433 were annotated at least twice. For all pairwise comparisons within a set of SU annotations we measured identical corrections, like Sakaguchi and colleagues (2016) on the basis that interannotator agreement is difficult to operationalise and arguably an inappropriate measure for error annotation (Bryant and Ng, 2015). Having made 7676 comparisons in this way, we find that 14.8% of error corrections are identical, a figure close to the 15.3% reported for the 'expert' annotators in Sakaguchi et al's study (and well above the 5.9% for the 'non-expert' crowdworkers).

We also report translation edit rate (TER) – a measure of the number of edits needed to transform one text into another, where an edit is an insertion, deletion, substitution, or phrasal shift, and where TER is expressed as edits per token (Snover et al., 2006).

In Table 6 we selected a speech-unit from the BULATS corpus along with two crowdsourced corrections. In the first correction, minimal edits have been made to make the SU more acceptable in grammatical terms (*that's* $\rightarrow$ *is*, *a the* $\rightarrow$ *a*, *are* $\rightarrow$ *is*). In the second version the correction is more holistic, even with punctuation (which was not called for), and the resulting SU is fluent. This latter type of correction is the one we seek, though it's clear from this example that not all corrections were done in a holistic way. One method to determine the success of crowdsourcing fluency edits would be to sample and rate corrections for fluency. We will incorporate this approach into further inspection of speech-units and the way they were corrected in future work.

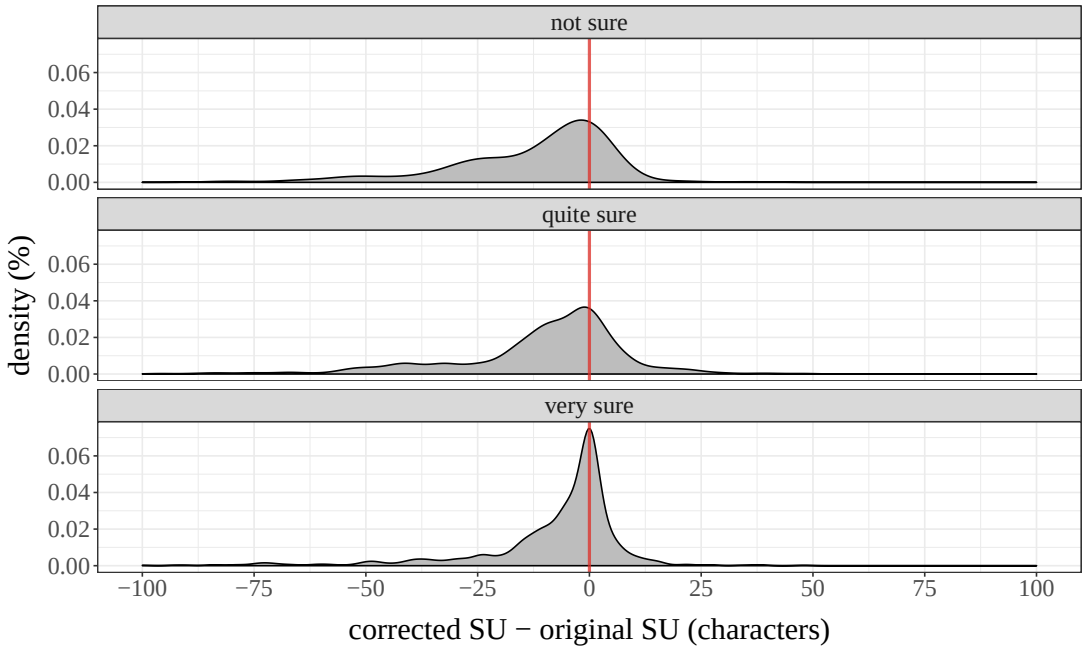

Figure 2: Density plot of the difference between corrected SU and original SU in characters, by crowdworkers' self-reported confidence level.

Version	Speech-unit	TER
original	I think in a newspaper that's an option and a the reference from a past employer are very important	0
corrected.1	I think in a newspaper is an option and a reference from a past employer is very important	3/19
corrected.2	I think that when advertising in a newspaper that's an option and also asking for a reference from a past employer is very important	10/19

Table 6: Example crowdsourced corrections for a speech-unit from the BULATS corpus.

In Figure 3 we show that for each CEFR level, firstly the proportion of SUs marked 'fine', or in need of no correction, tends to increase with increasing proficiency, and secondly mean TER scores for each SU rise from levels A1 to B1, and then fall again to C1 and C2. We hypothesise that the reason for this is that learners become more 'adventurous' in the linguistic constructions they attempt to use as they move from the A1 and A2 proficiency levels to B1 and B2. Thus their speech-units become in need of *more* correction, despite their improving capability with English. Part of their development into C1 and C2 level speakers is to become more accurate with the more complex construction types; hence SUs are in *less* need of correction. This is a 'U-shaped' developmental trajectory previously observed in language acquisition (Gershkoff-Stowe and Thelen, 2004).

4 Related work

Our work relates to previous attempts to collect error annotations through crowdsourcing (Tetreault et al., 2010; Madnani et al., 2011), which have concluded in its favour on the whole. Moreover we focus on *fluent* error corrections, as Sakaguchi and colleagues do (2016). Note also that crowdworkers were engaged in speech transcription, which is itself an established practice (Snow et al., 2008; Novotney and Callison-Burch, 2010).

Within second language acquisition research, we are focused on the *fluency* part of the well-established 'complexity accuracy fluency' framework (Housen and Kuiken, 2009). In future work we intend to turn to the complexity and accuracy dimensions as well. The framework gives us a useful way to consider automated assessment and feedback for language learners.

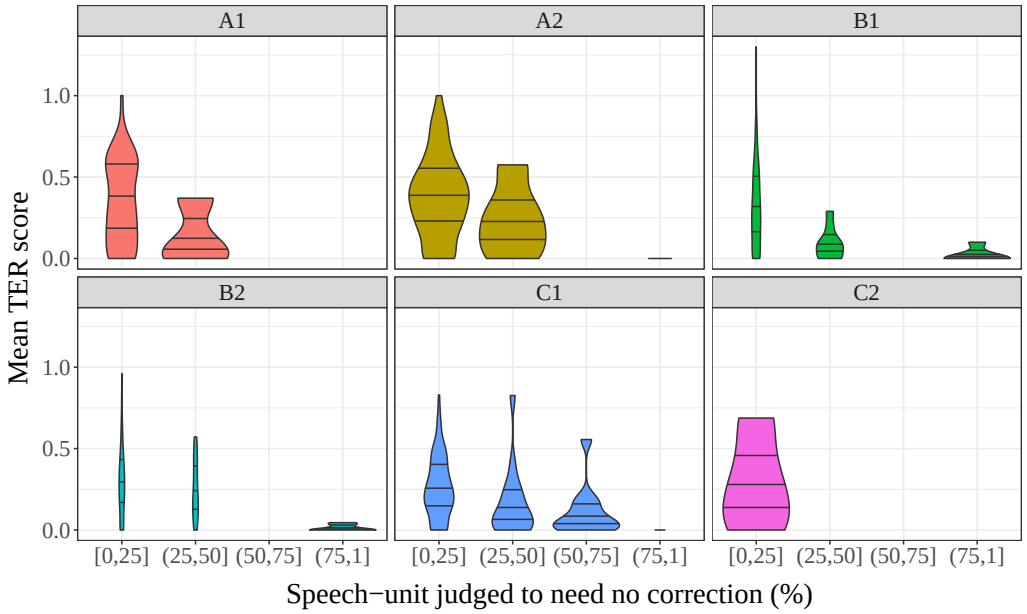

Figure 3: Proportion of SUs marked 'fine' by crowdworkers x Mean TER score for each CEFR level (width of 'violins' indicates density; horizontal lines mark first, second and third quartiles).

5 Conclusion and future work

In this paper we have presented our efforts to crowdsource fluency corrections of spoken learner English. We found that crowdworkers were tentative in applying corrections to SUs, more so for low CEFRs. When they did attempt to correct SUs though, we did find an overall decrease in SU length, an increase in lexical diversity, and TER scores which suggest U-shaped edit quantities by proficiency level.

Further evaluation of annotation quality remains to be carried out, including fluency ratings of the corrected versions. Also in future work we intend to repeat this work on an open dataset, such as the CrowdED Corpus (Caines et al., 2016), so that the resulting annotations can be released to others. Currently the BULATS corpus is not openly available.

One option for future annotations is to offer the original and corrected speech-units in parallel corpus format for machine translation approaches to error correction (Brockett et al., 2006; Park and Levy, 2011; Susanto et al., 2014; Junczys-Dowmunt and Grundkiewicz, 2016; Yuan et al., 2016), and with automatically aligned error annotations at the token level for classifier and rule-based approaches – the format used for GEC so far, as in the FCE and NUCLE datasets (Yannakoudakis et al., 2011; Dahlmeier et al., 2013). This would be in line with the call by Sakaguchi and colleagues for new annotated corpora for HEC research (Sakaguchi et al., 2016). We believe that whole sentence or speech-unit corrections lend themselves well to the recent emergence of neural network MT systems for error correction, since these are essentially sequence-to-sequence translations (Yuan and Briscoe, 2016). The challenge would be to build a sufficiently large training corpus for NMT: crowdsourcing would seem to be a fast and good-enough data collection method. Moreover, a hybrid MT-classifier system (Rozovskaya and Roth, 2016) may suit the goal of automated feedback, whereby the learner can be informed of detected errors and how to avoid them.

In any future data collection we need to install controls against crowdworkers' tendency to annotate higher proficiency items in preference to lower proficiency items. For example, we could remove the facility for skipping items, or there could be only a limited facility to do so (since we do find this information useful too). We could also present more context than the prompt alone – for example, the preceding and following speech-units. Finally, we will further investigate correction behaviours: to what extent crowdworkers followed our request to consider *spoken* English as the model, rather than written norms, and to what extent they aimed for holistic fluency corrections rather than minimal grammatical edits.

Acknowledgments

This paper reports on research supported by Cambridge English, University of Cambridge. We are grateful to our colleagues Kate Knill, Calbert Graham and Russell Moore. The second author received funding to pay crowdworkers from Sidney Sussex College, Cambridge, and the Department of Theoretical & Applied Linguistics at the University of Cambridge. We thank the three reviewers for their very helpful comments and have attempted to improve the paper in line with their suggestions.

References

JJ Allaire. 2016. *rsconnect: Deployment Interface for R Markdown Documents and Shiny Applications*. R package version 0.5.

Øistein Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Holly P. Branigan, Martin J. Pickering, Janet F. McLean, and Alexandra A. Cleland. 2007. Syntactic alignment and participant role in dialogue. *Cognition* 104:163–197.

Chris Brockett, William B. Dolan, and Michael Gamon. 2006. Correcting ESL errors using phrasal SMT techniques. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Christopher Bryant and Hwee Tou Ng. 2015. How far are we from fully automatic high quality grammatical error correction? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Janet Cahn and Susan Brennan. 1999. A psychological model of grounding and repair in dialog. In *Proceedings, AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*.

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 2016)*.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2016. *shiny: Web Application Framework for R*. R package version 0.14.2.

Herbert Clark and Edward Schaefer. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes* 2:19–41.

Daniel Dahlmeier and Hwee Tou Ng. 2012. A beam-search decoder for grammatical error correction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: the NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

Rachele De Felice and Stephen G. 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 2008)*.

Lisa Gershkoff-Stowe and Esther Thelen. 2004. U-shaped changes in behavior: A dynamic systems perspective. *Journal of Cognition and Development* 5(1):11–36.

Alex Housen and Folkert Kuiken. 2009. Complexity, fluency, and accuracy in second language acquisition. *Applied Linguistics* 30:461–473.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

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.

Russell Moore, Andrew Caines, Calbert Graham, and Paula Buttery. 2016. Automated speech-unit delimitation in spoken learner English. In *Proceedings of COLING*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*.

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.

Scott Novotney and Chris Callison-Burch. 2010. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Y. Albert Park and Roger Levy. 2011. Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

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.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27:169–190.

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Alla Rozovskaya and Dan Roth. 2016. Grammatical error correction: Machine translation and classifiers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Keisuke Sakaguchi, Courtney Napoles, Matt Post, and Joel Tetreault. 2016. Reassessing the goals of grammatical error correction: fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics* 4:169–182.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of Translation Edit Rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*.

Raymond Hendy Susanto, Peter Phandi, and Hwee Tou Ng. 2014. System combination for grammatical error correction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Joel Tetreault, Elena Filatova, and Martin Chodorow. 2010. Rethinking grammatical error annotation and evaluation with the Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*.

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)*.

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-Volume 1*.

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zheng Yuan, Ted Briscoe, and Mariano Felice. 2016. Candidate re-ranking for SMT-based grammatical error correction. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*.