Speak & Improve Corpus 2025: an L2 English Speech Corpus for Language Assessment and Feedback

Kate Knill[*1,3], *Diane Nicholls*[*2], *Mark J.F. Gales*[1,3], *Mengjie Qian*[1], *Pawel Stroinski*[2]

[1]ALTA Institute/MIL Lab, Dept. of Engineering, University of Cambridge, UK
[2]Cambridge University Press and Assessment, UK
[3]Enhanced Speech Technology Ltd, UK

kmk1001@cam.ac.uk, diane.nicholls@cambridge.org

Abstract

We introduce the *Speak & Improve Corpus 2025*, a dataset of L2 learner English data with holistic scores and language error annotation, collected from open (spontaneous) speaking tests on the Speak & Improve learning platform. The aim of the corpus release is to address a major challenge to developing L2 spoken language processing systems, the lack of publicly available data with high-quality annotations. It is being made available for non-commercial use on the ELiT website. In designing this corpus we have sought to make it cover a wide-range of speaker attributes, from their L1 to their speaking ability, as well as providing manual annotations. This enables a range of language-learning tasks to be examined, such as assessing speaking proficiency or providing feedback on grammatical errors in a learner's speech. Additionally the data supports research into the underlying technology required for these tasks including automatic speech recognition (ASR) of low resource L2 learner English, disfluency detection or spoken grammatical error correction (GEC). The corpus consists of around 315 hours of L2 English learners audio with holistic scores, and a subset of audio annotated with transcriptions and error labels.

Index Terms: L2 speech, non-native speech, automatic speech recognition, spoken grammar error correction, language assessment and feedback

1. Introduction

Spoken language assessment and feedback are crucial components of language learning. Developing robust and inclusive automated tools for these tasks remains a significant challenge. One important limitation to developing systems is the lack of high-quality labelled data. To address this issue and encourage research in these areas, we are distributing a new annotated corpus of spoken L2 (second language) learner English, collected from speaking tests performed on the Speak & Improve (S&I) learning platform [1]. The *Speak & Improve Corpus 2025* consists of around 315 hours of recordings of L2 English learner speech on open speaking tasks, annotated with English speaking proficiency scores. In addition a subset of the data, around 55 hours, was manually transcribed, including disfluencies, and grammatically error corrected phrases [2]. The S&I platform was used by speakers from around the globe yielding a wide range of L1 (first language) backgrounds. The proficiency levels of the speakers range from Elementary (A2) to Advanced (C1) on the CEFR scale [3]. This breadth of speaker attributes and abilities is crucial for the development of inclusive language learning technologies. We believe this is the most comprehensive L2 English learner public corpus released to date.

Data in the corpus was collected from practice tests done on the S&I web application. Learners take a multi-level, monologic (computer prompt:human response), English speaking practice test. This is similar in format to exams such as Cambridge University Press & Assessment's Linguaskill Speaking Test, Pearson's PTE and Duolingo's English Test. Learners undertake a series of open speaking tasks[1] each of which aims to examine a different aspect of their speaking ability. Scores are given for each task, or part, and over the complete test. The latter takes into account the candidates skills at use of language resource, coherence, hesitation/extent and task achievement as well as their pronunciation/intelligibility. S&I's open speaking nature means that the learners are better able to demonstrate their competency in spoken English than is the case for read speech tests. This does mean, however, that the text of what the learner said is unknown. Additionally, their speech is relatively spontaneous so includes natural speaking effects such as disfluencies (filler words, false starts, repetitions) and not necessarily grammatically correct sentences from a writing perspective. This is in addition to grammatical errors that they might make related to their level of proficiency in English. The range of proficiency levels and L1s will cause a wide variety of pronunciation errors relating to the learner's L1 accent and whether they know how to pronounce an English word.

The S&I Corpus 2025 supports research on a wide range of language learning and assessment tasks including Spoken Language Assessment (SLA), Spoken Grammatical Error Correction (SGEC), and Spoken Grammatical Error Correction Feedback (SGECF). This will be explored by participants in the Speak & Improve Challenge 2025 [4] who will have access to a pre-release of the corpus December 2024-March 2025, prior to the full public release of the corpus in April 2025.

Related work, the Speak & Improve application, annotation of the data and the corpus selection and distribution are described in following sections.

[1] The S&I speaking test includes a read aloud component but this is not included in this corpus.

2. Related Work

While there are other publicly available datasets, none match the breadth and diversity of the Speak & Improve dataset for L2 English assessment and feedback. For instance, the International Corpus Network of Asian Learners of English (ICNALE) [5] focuses exclusively on Asian L1s and only features CEFR levels from A2 to B2. The latter are pre-assigned based on the learners' previously obtained language certificates rather than based on their actual speech. Similarly, the CLES corpus [6] only features L1 French speakers whose CEFR ranges from B1 to B2. The LearnerVoice dataset [7] presented at Interspeech 2024 contains audio and annotations on grammatical errors and disfluencies, but is not publicly available yet, although its authors are planning to release it. It is restricted to L1 Korean speakers. Other L2 corpora such as speechocean762 [8] and L2-ARCTIC [9], only focus on pronunciation assessment of read speech. Existing English datasets for automatic speech recognition (ASR) training and evaluation of L2 learner speech, such as the ETLT Italian-L1 dataset [10], also tend to focus on single L1 groups, limiting their generalisability. The version of the ETLT corpus with annotated analytic and holistic scores [11] has not been made public.

No current public dataset provides comprehensive support for spoken language assessment or spoken grammatical error correction/feedback at the scale of Speak & Improve. Datasets such as the NICT-JLE [12] and KISTEC [13] support disfluency detection, but only the former has annotations on grammatical error corrections. In both cases, the only available data consists of manual text transcriptions. The respective audio recordings are not available.

While written GEC is an established area of study with five shared tasks in the last 15 years (i.e., the HOO 2011 Pilot Shared Task [14], the CoNLL-2013 Shared Task [15], the CoNLL2014 Shared Task [16], the BEA-2019 Shared Task [17], and MULTIGEC-2025 [18]), one of the key innovations of this corpus is that it is the first to provide audio with grammatical error corrections to enable more complete research on spoken GEC. The complexities of spoken GEC, such as handling disfluencies, varied accents, and spontaneous speech patterns, make this task significantly different from written GEC, requiring new approaches and innovation.

The Speak & Improve Challenge 2025 demonstrates how the S&I Corpus can be used for spoken language assessment and spoken grammatical error correction feedback, as well as for producing underlying technology such as automatic speech recognition and disfluency detection and grammatical error correction. For written assessment, the Automated Student Assessment Prize (ASAP) competition for essay scoring was organised in 2012 by the Hewlett Foundation [19], and focused on L1 learners. There have been other spoken language assessment shared tasks but they have generally targeted specific aspects such as pronunciation, grammar, or semantic meaning [20, 21, 22].

3. Speak & Improve

Speak & Improve (S&I) is a research project from the University of Cambridge in association with Cambridge University Press & Assessment (CUP&A) and English Language iTutoring Ltd (ELiT). The S&I web application has been developed by ELiT with automated speaking assessment technology provided by Enhanced Speech Technology Ltd, developed through technology transfer from the ALTA Institute [1]. Always available and free, users can interact with S&I through many different devices including laptops, tablets and mobile phones, as it is based on the browser. Learners are able to practice their English speaking and improve their confidence on a wide range of communicative speaking tasks. S&I is designed for all proficiency levels, from basic beginner through independent intermediate to proficient learners; on the internationally-recognised CEFR [3] scale from below A1 to C1 and above.

The S&I Corpus 2025 is selected from data collected by Speak & Improve version 1. There were 1.7 million users of this version from across the globe from its launch in December 2018 to its retirement in September 2024. Over 18.4 million answers were submitted in total during that time. Learners were offered one of five complete practice tests to perform. They could choose to do the whole test, which takes around 10-15 minutes, or to stop after one or more parts of the test. The test structure is based on the Linguaskill Speaking test [23] which is designed to cover a variety of communicative speaking functions [23]. A single test consists of 5 parts, always presented to a user in the same order:

- **Part 1: Interview**.
  The learner is asked 8 questions about themselves e.g. ”Who in your family are you most similar to?”. They are given 10 seconds of speaking time for the first 4 questions, and 20 for the second 4. The first two questions which may contain personal identity information are not marked and are not included in this corpus.
- **Part 2: Read Aloud**.
  The learner must read aloud 8 sentences.
- **Part 3: Long Turn 1: Give your opinion**.
  The learner has 1 minute to give their opinion on a specific topic using 3 questions to guide them.
- **Part 4: Long Turn 2: Give a presentation about a graphic**.
  The learner has 1 minute to describe a process depicted in a diagram.
- **Part 5: Communication Activity: Answer questions about a topic**.
  The learner has to respond to 5 questions relating to an overall topic, each for up to 20 seconds.

At the end of each part the user receives an auto-marker assessment in the form of a CEFR-like level, along with an estimate of how confident the auto-marker is that the predicted score is a true reflection of the user's speaking level. Users who complete the test are awarded an indicative CEFR grade for the whole test. Unlike the Linguaskill test where restrictions are put on the length of thinking time test takers have before responding, the user can take as much time as they would like to prepare before answering each question. The questions for the tests were written exclusively for S&I and do not form part of any CUP&A test.

Since the test takes some time many users drop out before the completion of the full test. Whilst only complete recordings are used for the test sets in the S&I Corpus, some recordings from partially completed tests are included in the corpus as part of the data for training spoken language assessment systems. Version 2 of S&I has addressed this issue by offering users the chance to practise a variety of specific speaking skills in addition to doing complete tests.

4. Corpus

This section describes the S&I Corpus 2025 including the data annotation, selection and distribution. The first release of the corpus is split into three data sets: Train, Development (Dev) and Evaluation (Eval). These sets are being made available as a pre-release for the Speak & Improve Challenge 2025 [4].

4.1. Data Annotation

The S&I data was annotated by ELiT human annotators using their bespoke annotation tool. They use a three stage approach as shown in Figure 1. This section describes the key points of the annotation with respect to the corpus. Further details can be found in [2].

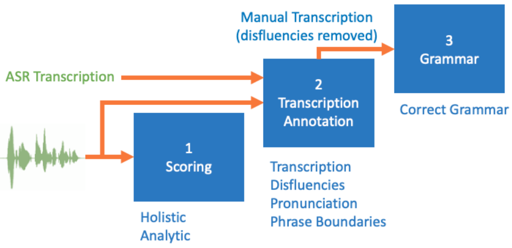

Figure 1: *Annotation phases.*

Phase 1: Scoring. Utterances are initially given an audio quality score (range 3-5, utterances with an audio quality less than 3 were removed). Each test part, comprising one or more utterances, is then awarded an holistic score (scoring range 1-6). This holistic score summarises the annotator's overall/aggregated impression of the speaker's performance across all the utterances in the given part. Parts receiving a holistic score equivalent to CEFR level A2 or higher are considered of sufficient quality and passed to the next annotation phase. Parts scored as A1 or below, in addition to those with poor audio quality, lack of meaningful responses, or inappropriate content (e.g. foul language), are excluded from further annotation and from consideration for the S& I Corpus 2025.

Phase 2: Transcription Annotation. The second annotation phase generates the manual transcriptions of the audio. For efficiency reasons, these transcriptions were derived by correcting the output of an Automatic Speech Recognition (ASR) system[2]. The goal of this stage is to produce a detailed transcription that accurately reflects exactly what the learner said - including all language errors, hesitations, false starts and repairs and code-switching. In addition annotators mark phrase boundaries when they occur within an utterance. Word-level pronunciation errors are also indicated[3]. Where annotators were unable to identify a word, or words, in the audio these are tagged as unknown words. Where it is clear that the user is saying a proper noun which is foreign to British English these words were tagged as such to avoid the annotator having to guess what the user had said. Similarly code-words, where the user switched into speaking another language, were tagged. The associated phrases containing unknown, foreign proper nouns and/or code-switched words are excluded from the final annotation stage[4]. Tables 1, 2 and 3 show the marks and tags that annotators can apply in phase 2 and are provided with the corpus transcriptions.

Mark	Description
backchannel	Speaker spoke to themselves
disfluency	Word is part of a false start or repeated disfluent sequence
partial	Only part of the word was spoken
pronunciation	Lexical pronunciation error made

Table 1: *Phase 2 annotation marks attached to a transcribed word.*

Phase 3: Error Annotation. The final annotation phase focuses on correcting learner language errors. Phase 3 annotators are provided with texts only, to avoid any conflict of interpretation or second guessing by them as to whether the phase 2 annotator was correct. This phase aims to generate an accurate transcription of the learner's intended speech. Prior to starting annotation, fluent phrase-level transcriptions are made from the disfluent utterance-level transcriptions output by phase 2. This is done by first splitting the utterance into phrases using the phase 2 phrase tags. Since it will not be possible to say what the correct grammar should be if only a partial phrase was spoken and/or it contains words that the annotators were unable to transcribe at phase 2, phrases containing any of these are excluded from error annotation at this point; i.e. phrases tagged as incomplete, or including words tagged as unknown, foreign proper nouns and/or code-switched. Fluent transcriptions of the remaining phrases are then created by removing hesitations, partial words, false starts and repetitions from the manual disfluent transcription; these are marked as a disfluency or partial at phase 2.

[2] The ASR system used at this stage is not related to the Whisper models used in the distributed baseline pipeline for the Speak & Improve Challenge 2025.

[3] For these pronunciation errors annotators were asked to ignore accent errors, and focus on lexical errors. For more details see [2].

[4] These phrases are also excluded from ASR evaluation in the Speak & Improve Challenge 2025.

Tag	Description
hesitation	Speaker made a sound indicating a hesitation e.g. um, er
code-switch	Word is code-switched into language other than English
foreign-proper-noun	Foreign proper noun
unknown	Annotator was unable to determine what the word spoken was

Table 2: *Phase 2 word level transcription tags. The word in the transcript matches the tag i.e. is not a transliteration of what was said. Phase 3 annotators may tag a word as unknown if they are unsure how to correct the word.*

Tag	Description
speech-unit-incomplete	Partial phrase/speech unit
speech-unit-statement	Phrase boundary ending with a full stop (period)
speech-unit-question	Phrase boundary ending with a question

Table 3: *Phase 2 phrase level transcription tags. Phase 3 annotators may adjust phrase boundaries but these utterances were excluded from the corpus.*

For example, *"i am sti- still care about %hesitation% the the environment"* will become *"i am still care about the environment"*. Using this fluent transcription the annotators generated the grammatically corrected transcription, e.g. *"i still care about the environment"*. By comparing the fluent transcriptions to these corrected transcriptions it is possible to obtain a set of reference edits that can be fed back to learners. Phase 3 annotators can mark a word with an error tag where the annotator knows the phase 2 annotator overlooked something or created a typo. They may also amend the phrase boundaries. Utterances with these edits were excluded from inclusion in the corpus for ease of processing.

4.2. Data distribution and selection

The data for the S&I Corpus 2025 was selected from tests recorded on version 1 of the S&I platform from 2019 to 2024. Learners were guided to do a full practice test. All data underwent Phase 1 holistic scoring at the test part level. A subset of the data received manual transcription annotation at Phase 2. The fluent phrases from Phase 2 were then passed to Phase 3 for grammatical error annotation. Annotators at phase 1 identified responses with poor audio recording quality and/or aberrant/malpractice responses. These were excluded from consideration for the corpus.

The Dev and Eval test sets were first defined to consist of full test submissions only i.e. where the learner has completed the test and all parts were successfully annotated at each phase of interest. 300 fully annotated test submissions were selected for each of the Dev and Eval test sets. These were extended for spoken language assessment (SLA) with a further (approximately) 150 submissions which had undergone phase 1 holistic proficiency scoring only. The submissions were selected so that the grade distribution was as even as possible but also reflects the nature of the learners using the platform. Table 4 shows the grade distribution for the Dev set. As can be seen the majority are in the CEFR range B1-B2+. The Eval and Train sets have similar grade distributions. The scores in Table 4 correspond to the marks awarded for a CEFR grade level on a part part basis. The overall score is the average of the scores across all the test parts so has a finer gradation. For this corpus the test is considered to consist of 4 parts. There was insufficient L1 and speaker data provided to incorporate this information into the data set selection. Note, this does mean that a speaker could appear in both the training and test sets. Table 5 shows the statistics for the Dev and Eval sets.

CEFR Grade	Score	% Data
A2	2.0	2.1
A2+	2.5	5.7
B1	3.0	18.3
B1+	3.5	25.3
B2	4.0	25.1
B2+	4.5	18.3
C1	5.0	5.0
C1+	5.5	0.2

Table 4: *Distribution of holistic CEFR grades for Dev set. Eval and Train sets have similar distributions.*

The Train set is a combination of data from full submissions and from individual parts, and is a mix of fully annotated and SLA annotated only data. The data was selected to be approximately equal across parts for the full data set, as shown in Table 6.

Data Set	No. of Submissions	No. of Utterances	No. of Hours			No. of Words	
			Trans	GEC	SLA	Transcript	GEC
Dev	438	5616	22.9	20.8	35.3	140k	105k
Eval	442	5642	22.7	20.4	35.4	140k	104k

Table 5: *Data set statistics for Dev and Eval set.*

Part	No. of Submissions	No. of Utterances	No. of Hours			No. of Words	
			Trans	GEC	SLA	Transcript	GEC
1	3068	18072	7.3	3.9	70.3	47k	22k
3	3060	3060	5.8	2.9	47.0	37k	14k
4	3005	3005	5.8	1.5	45.5	36k	7k
5	3085	15353	9.4	4.8	81.4	60k	24k
Total	6640	39490	28.2	13.0	244.2	170k	65k

Table 6: *Data set statistics for train set. 1742 submissions are complete for SLA scoring at an overall level.*

4.3. Corpus contents

The corpus contents are listed in Table 7. Annotation markup and tags are described in Tables 1, 2 and 3. All annotation was manually derived except for the time alignments provided. These were generated using the HTK HVite tool [24] adapted to do lattice-based forced alignment with an L2 English acoustic model.

Item	Description
Audio files	16kHz, single channel recordings. (flac)
Audio file lists	List of file IDs and their corresponding audio file. Correspond to data sets and subsets for specific tasks (tsv)
Transcript annotations GEC annotations	Marked up manual (disfluent) transcriptions with File IDs, question prompt and timing information. (json) As for manual transcript but the grammatical error corrected transcript. (json)
SLA marks	Holistic marks in the range 2-5.5 corresponding to A2-C1+ CEFR-like grades on a per-part and overall submission level (tsv)
STM transcriptions	Reference STMs for disfluent, fluent and GEC transcripts at the phrase level for use in NIST sclite scoring. Contains automatically aligned phrase start and end times. Category information included: audio quality (utterance level); grade (part); part number (part). (STM)

Table 7: *Material provided with the S&I Corpus.*

4.4. Use of the Corpus with LLMs

We encourage the reader to consider the problem of ‘data leakage’ with regard to LLMs: whereby NLP/SLP datasets are leaked into LLM training datasets via commercial APIs [25]. We therefore ask that the Corpus is only passed to locally-stored LLMs, such as might be downloaded from the Hugging Face Transformers library [26], or to commercial LLMs in such a way that the data is not retained for model training purposes.

4.5. How to obtain the Corpus

From December 2024-March 2025 the Speak & Improve Corpus 2025 will be made available to participants in the Speak & Improve Challenge 2025 in connection with the ISCA SLaTE Workshop 2025, and will then be released for non-commercial academic research purposes. Access to the corpus may be obtained by visiting the ELiT website, completing the form and agreeing to the licence when available. Future versions of the Corpus will be made available on the same website. Queries about the corpus should be made to support@speakandimprove.com.

5. Conclusions and Future Work

In this paper we describe the Speak & Improve Corpus 2025. This corpus of L2 learner English speaking provides a comprehensive resource for researchers interested in spoken language assessment and/or feedback.

The S&I Corpus 2025 consists of audio and manual annotations of open speaking test submissions performed on the Speak & Improve learning platform. We provide around 315 hours of speech data from L2 English speakers at A2 to C1 CEFR levels, split into train, dev and eval data sets. There are around 950 fully annotated test submissions and the equivalent of a further 2500+ test submissions with manual speaking proficiency assessment scores.

6. Acknowledgements

This paper reports on research supported by Cambridge University Press & Assessment (CUP&A), a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

Thanks to the team at ELiT who have created and run the Speak & Improve learning platform, the ELiT annotation tool and the hosting and establishment of this corpus. In particular, Paul Ricketts, Scott Thomas, Jon Phillips and Alex Watkinson. The ELiT humannotator team provided the annotations, with Gloria George and David Barnett providing phase 2 and 3 annotations. Annabelle Pinnington and her team in Cambridge University Press & Assessment's Multi-Level Testing team have provided all the question prompts for Speak & Improve. Amanda Cowan and Jing Xu at Cambridge English have supported the development of this corpus and helped with all the paperwork.

Thanks to the ALTA Spoken Language Processing Technology team at Cambridge University Engineering Department for helping to setup the Speak & Improve Challenge 2025 and being beta-testers for the corpus, particularly Stefano Bannò, Penny Karanasou and Siyuan Tang.

The authors marked with* made equal contributions to this paper.

7. References

[1] D. Nicholls, K. M. Knill, M. J. F. Gales, A. Ragni, and P. Ricketts, "Speak & Improve: L2 English Speaking Practice Tool," in *INTERSPEECH 2023*, 2023, pp. 3669–3670.

[2] K. M. Knill, D. Nicholls, M. Gales, P. Stroinski, and A. Watkinson, "Annotation of L2 English Speech for Developing and Evaluating End-to-End Spoken Grammatical Error Correction," in *9th Workshop on Speech and Language Technology in Education (SLaTE)*, 2023, pp. 146–150.

[3] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[4] M. Qian, K. Knill1, S. Banno, S. Tang, P. Karanasou, and M. J. G. andDiane Nicholls, "Speak & Improve Challenge 2025: Tasks and Baseline Systems," 2024.

[5] S. Ishikawa, *The ICNALE guide: An introduction to a learner corpus study on Asian learners' L2 English*. Routledge, 2023.

[6] S. Coulange, M.-H. Fries, M. Masperi, and S. Rossato, "A corpus of spontaneous L2 English speech for real-situation speaking assessment," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 293–297.

[7] H. Kim, J. Myung, S. Kim, S. Lee, D. Kang, and J. Kim, "LearnerVoice: A Dataset of Non-Native English Learners' Spontaneous Speech," in *Interspeech 2024*, 2024, pp. 2325–2329.

[8] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, "speechocean762: An Open-Source Non-Native English Speech Corpus for Pronunciation Assessment," in *Interspeech 2021*, 2021, pp. 3710–3714.

[9] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A Non-native English Speech Corpus," in *Interspeech 2018*, 2018, pp. 2783–2787.

[10] R. Gretter, M. Matassoni, D. Falavigna, A. Misra, C. Leong, K. Knill, and L. Wang, "ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech," in *Interspeech 2021*, 2021, pp. 3845–3849.

[11] R. Gretter, M. Matassoni, S. Bannò, and F. Daniele, "TLT-school: a Corpus of Non Native Children Speech," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 378–385.

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

[13] K. Kanzawa, Y. Kobayashi, J. Lee, H. Mitsunaga, M. Mori, , and Y. Tanaka, "The KIT Speaking Test Corpus (KISTEC)," https://kitstcorpus.jp/, 2022.

[14] R. Dale and A. Kilgarriff, "Helping our own: The HOO 2011 pilot shared task," in *Proceedings of the 13th European Workshop on Natural Language Generation*, 2011, pp. 242–249.

[15] H. T. Ng, S. M. Wu, Y. Wu, C. Hadiwinoto, and J. Tetreault, "The CoNLL-2013 Shared Task on Grammatical Error Correction," in *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*. Sofia, Bulgaria: Association for Computational Linguistics, 2013, pp. 1–12. [Online]. Available: https://aclanthology.org/W13-3601

[16] H. T. Ng, S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant, "The CoNLL-2014 shared task on grammatical error correction," in *Proceedings of the eighteenth conference on computational natural language learning: shared task*, 2014, pp. 1–14.

[17] C. Bryant, M. Felice, Ø. E. Andersen, and T. Briscoe, "The BEA-2019 shared task on grammatical error correction," in *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, 2019, pp. 52–75.

[18] "Shared task on Multilingual Grammatical Error Correction 2025 ," https://www.aclweb.org/portal/content/shared-task-multilingual-grammatical-error-correction-2025, 2024.

[19] B. Hamner, J. Morgan, lynnvandev, M. Shermis, and T. V. Ark, "The Hewlett Foundation: Automated Essay Scoring," https://kaggle.com/competitions/asap-aes, 2012, kaggle.

[20] C. Baur, C. Chua, J. Gerlach, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2017 Spoken CALL Shared Task," in *7th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2017)*, 2017, pp. 71–78.

[21] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Qian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2018 Spoken CALL Shared Task," in *Interspeech 2018*, 2018, pp. 2354–2358.

[22] ——, “Overview of the 2019 Spoken CALL Shared Task,” in *8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 1–5.

[23] J. Xu, M. Brenchley, E. Jones, A. Pinnington, T. Benjamin, K. Knill, G. Seal-Coon, M. Robinson, and A. Geranpayeh, “Linguaskill Building a validity argument for the Speaking test,” *Linguaskill Research Reports, UCLES, Tech. Rep*, 2020.

[24] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.4.1).* University of Cambridge, 2009. [Online]. Available: http://htk.eng.cam.ac.uk

[25] S. Balloccu, P. Schmidtová, M. Lango, and O. Dusek, “Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian’s, Malta: Association for Computational Linguistics, Mar. 2024, pp. 67–93. [Online]. Available: https://aclanthology.org/2024.eacl-long.5

[26] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6