Cross-corpora experiments of automatic proficiency assessment and error detection for spoken English

Stefano Bannò
Fondazione Bruno Kessler, Italy
Department of Cognitive Science,
University of Trento, Italy
sbanno@fbk.eu

Marco Matassoni
Fondazione Bruno Kessler, Italy
matasso@fbk.eu

Abstract

The growing demand for learning English as a second language has led to an increasing interest in automatic approaches for assessing spoken language proficiency. One of the most significant challenges in this field is the lack of publicly available annotated spoken data. Another common issue is the lack of consistency and coherence in human assessment. To tackle both problems, in this paper we address the task of automatically predicting the scores of spoken test responses of English-as-a-second-language learners by training neural models on written data and using the presence of grammatical errors as a feature, as they can be considered consistent indicators of proficiency through their distribution and frequency. Specifically, we train a feature extractor on EFCAMDAT, a large written corpus containing error annotations and proficiency levels assigned by human experts, in order to extract information related to grammatical errors and, in turn, we use the resulting model for inference on the CLC-FCE corpus, on the ICNALE corpus, and on the spoken section of the TLT-school corpus, a collection of proficiency tests taken by Italian students. The work investigates the impact of the feature extractor on spoken proficiency assessment as well as the written-to-spoken approach. We find that our error-based approach can be beneficial for assessing spoken proficiency. The results obtained on the considered datasets are discussed and evaluated with appropriate metrics.

1 Introduction

Automatic scoring of language proficiency is becoming a point of growing interest and importance in the field of second language (L2) assessment because the number of English-as-a-second-language (ESL) learners has been steadily increasing worldwide (Howson, 2013).

A common issue in this field is the lack of publicly available data specifically designed and annotated for automatic assessment, especially as regards spoken data. Another typical problem is the lack of consistency and coherence in human assessment, as it frequently relies on proficiency indicators that often have biases and are not clearly generalizable, therefore not easily transferable into automatic scoring systems (Zhang, 2013). Although L2 proficiency cannot be assessed on the mere basis of the presence of errors in learners' written and spoken productions, this aspect is highly consistent and plays a major role in language assessment by human experts (James, 2013). Nevertheless, to the best of our knowledge, the impact of errors on automatic spoken language assessment has not been thoroughly investigated yet, whereas other types of feature-based assessment have been more widely studied and explored (Crossley et al., 2015).

In this paper, we address the task of automatically predicting the scores of spoken responses of ESL learners leveraging written data and exploiting the presence of grammatical errors, thus tackling both the aforementioned problems: the issue related to the scarce availability of spoken data and the problem of inconsistency in human assessment.

In order to do so, we design a ranking of grammatical error gravity based on the frequency of each human-annotated error in the EF-Cambridge Open Language Database (EFCAMDAT), modelling it across 15 proficiency levels aligned with the CEFR (Common European Framework of Reference) levels ranging from A1 to C1 (Council of Europe, 2001); as our purpose is scoring spoken language proficiency, we discard spelling, punctuation and orthographic errors and we group errors into 5 categories.

Subsequently, we train a feature extraction model feeding the learners' texts of the EFCAMDAT as inputs and setting the 5 classes of errors as targets for our predictions and we use this model as an error feature extractor (EFEX) for inference on the Cambridge Learner Corpus - First Certificate

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 82 - 91
July 15, 2022 ©2022 Association for Computational Linguistics

in English (CLC-FCE) and on the International Corpus Network of Asian Learners of English (ICNALE), thus generating 5 labels corresponding to the aforementioned 5 classes of errors; then, we train a scoring model on the CLC-FCE injecting the 5 error labels generated by EFEX and we test it on the spoken annotated section of ICNALE.

Likewise, we use EFEX for inference on the TLT-school corpus. Subsequently, we train a scoring model on the written section of the corpus injecting the 5 error labels generated by EFEX and we test it on the spoken section. Figure 1 shows the proposed pipeline. Finally, we fine-tune our model on a small spoken subset.

The structure of the paper is as follows: in the next paragraphs, we briefly illustrate the theoretical framework and literature related to automatic scoring and assessment; in Section 2, we describe the data used in our experiments and our ranking of grammatical error gravity; in Section 3, we show the model architectures; in Section 4, we show the results of our experiments on the models; finally, in Section 5, we illustrate the conclusions of the study and reflect upon next steps.

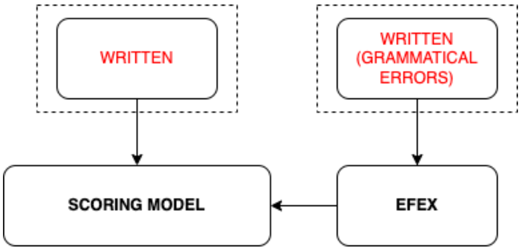

Figure 1: Diagram of the proposed training pipeline based on textual input (i.e the written train set). The scoring model is then used to predict proficiency scores on manual and ASR transcriptions (i.e. the spoken test set).

1.1 Theoretical framework

The origins of the field of L2 assessment date back to the influential work of Lado (1961), who believed that the problems of learning a new language could be predicted comparing the learners' native language and their target language, consistently with his structuralist perspective of language and contrastive linguistics. Language was taught - and thus assessed - as a set of distinct elements, starting from a contrastive analysis of sounds, grammar and vocabulary. As a result, errors play an important role in this construct. In response to and in continuation of contrastive analysis, at the end of the 1960s the work of Corder (1967) set the foundation for error analysis and considered the concept of error from a developmental perspective.

In the 1970s, the subsequent fundamental step in language testing and assessment was inspired by the forward-looking work on communicative competence by Hymes (1972), later refined and framed in the so-called communicative approach by Canale and Swain (1980). According to this approach, language is used to communicate meaning, which encompasses: grammatical knowledge, sociolinguistic competence, and strategic competence.

Around the 1990s, an approach theoretically rooted in the communicative approach, started to be developed and was later fixed in the Common European Framework of Reference (CEFR) (Council of Europe, 2001). Although it might seem that this approach privileges communication at the expense of formal correctness, errors still play a major role in assessing language proficiency (Pfingsthorn, 2013). Furthermore, Thewissen (2013) has shown that learner errors can be connected to CEFR proficiency levels and they can be considered as criterial features for each level, together with other linguistic features, as illustrated in Hawkins and Buttery (2010).

1.2 Reference to prior work

Deep learning techniques have brought significant improvements in the field of automatic scoring, for assessing both writing and speaking, such that end-to-end neural based approaches outperformed ETS's SpeechRater (Chen et al., 2018), one of the best known oral proficiency test engines (Xi et al., 2008). Specifically, transformer-based models have led to a remarkable improvement in tasks of predicting linguistic proficiency (Raina et al., 2020; Wang et al., 2021).

While grammatical error detection for speech assessment has been the focus of relatively few studies (Knill et al., 2019; Caines et al., 2020), grammatical errors have received more attention in the field of automatic essay scoring and are one of the features employed in Yannakoudakis et al. (2011) along with lexical, part-of-speech (POS) and syntactic features for automatically assessing ESL examination scripts, and they were found to be significant for enhancing the overall correlation between true scores and predicted ones. Gamon et al. (2013) uses Leacock and Chodorow (2003)'s findings on the influence of grammatical errors on

TOEFL (Test of English as a Foreign Language) scores for automatic essay scoring and feedback. Similarly, errors are a feature investigated in the work of Vajjala (2018), in which spelling and grammar errors are extracted by LanguageTool[1]. In this case, the error rate feature considered individually was found to have little impact on the classification performance. Similar experiments were conducted again by Vajjala and Rama (2018) with German, Czech and Italian, including errors as a feature. This work was reproduced by Caines and Buttery (2020), who applied such experiments also to English and Spanish corpora. Another research conducted on the CLC-FCE found that grammatical error detection highly influences essay scores (Cummins and Rei, 2018).

Recently, the work described by Ballier et al. (2019) has investigated the possibility of predicting CEFR proficiency levels based on manually annotated errors in the French and Spanish section of the EFCAMDAT corpus, but their study did not employ deep learning techniques. However, they identified that certain types of errors, such as punctuation, spelling and verb tense errors, are characteristic of specific CEFR proficiency levels. For our study, we reversed the process and we started from a ranking of error gravity across the CEFR proficiency levels.

Finally, some recent studies on automatic assessment of L2 proficiency have employed state-of-the-art models, combining associated auxiliary tasks (Craighead et al., 2020), none of which related to errors.

2 Datasets and setup

2.1 EFCAMDAT

Firstly, we use the EFCAMDAT corpus (Geertzen et al., 2014) that comprises L2 learners' scripts annotated with their respective score on a scale from 0 to 100, their proficiency level from 1 to 16 (mapped to CEFR levels from A1 to C2) and partially error-tagged by human experts. As our work investigates the efficacy of errors as features, we only use the error-tagged section of the EFCAMDAT Cleaned Subcorpus (Shatz, 2020), consisting of 498,208 scripts ranging from proficiency level 1 to 15 (i.e. from A1 to C1), which we divided into training and test set. The error tagset of the corpus consists of 24 types of errors, of which we discarded 7 related to spelling, punctuation and orthographic errors, as they would be of no use for assessing speech (see Table 1). As a preliminary analysis, we computed the KL-Divergence between the distribution of the 17 error labels counts across CEFR proficiency levels in EFCAMDAT. The labels were converted into a smoothed distribution, by applying add-one smoothing. The symmetric KL-Divergence was then calculated. Therefore, for error type t_i for proficiency level L_k:

Code	Meaning	Code	Meaning
XC	change from x to y	NSW	no such word
AG	agreement	PH	phraseology
AR	article	PL	plural
D	delete	PO	possessive
PS	part of speech	PR	prepositions
EX	expression of idiom	SI	singular
IS	insert	VT	verb tense
MW	missing word	WC	word choice
WO	word order		

Table 1: EFCAMDAT error tagset without codes related to spelling, punctuation and orthographic errors.

$$\mathrm{P}(t_\mathrm{i}|L_\mathrm{k}) = \frac{\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1}{\sum_{j=1}^{N}(\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k}) + 1)}$$

where $\mathrm{cnt}(t_\mathrm{i}, L_\mathrm{k})$ is the number of occurrences for a given label in a given grade.

The symmetric KL Divergence was subsequently calculated across proficiency levels:

$$\mathrm{KL}(L_\mathrm{k}|L_\mathrm{l}) = \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}\right)\right) + \left(\sum_{i=1}^{N}\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})\log\left(\frac{\mathrm{P}(t_\mathrm{i}|L_\mathrm{l})}{\mathrm{P}(t_\mathrm{i}|L_\mathrm{k})}\right)\right)$$

Table 2 reports the symmetric KL-Divergence between distributions of counts from all the 17 error labels across CEFR proficiency levels. It appears that we can consider errors as criterial features of linguistic proficiency, as there are differences in the distributions of grammatical errors across proficiency levels, to which we can correlate differences in their frequency.

[1] https://languagetool.org/

	A1	**A2**	**B1**	**B2**	**C1**
A1	0.0	0.055	0.065	0.085	0.066
A2	0.055	0.0	0.013	0.029	0.028
B1	0.065	0.013	0.0	0.005	0.009
B2	0.085	0.029	0.005	0.0	0.010
C1	0.066	0.028	0.009	0.010	0.0

Table 2: Symmetric KL Divergence between distributions of counts from all 17 error labels in EFCAMDAT.

2.2 Ranking of error gravity

In light of this, we analyzed the frequency of each type of error across the 15 proficiency levels of the corpus. We calculated it dividing the sum of all the occurrences of a given type of error in a given proficiency level by the number of texts assigned to a given proficiency level. We then decided to design a ranking of error gravity for each type of error in relation to each proficiency level, by introducing a negative bias in the error count when this amounts to 0:

$$b_t = \begin{cases} -1 & 0.1 \leq F_{t,L} < 0.2 \\ -2 & 0.2 \leq F_{t,L} < 0.3 \\ \dots & \\ -9 & 0.9 \leq F_{t,L} < 1.0 \end{cases}$$

where $F_{t,L}$ is the normalized frequency of error type t at proficiency level L; e.g. if $F_{AR,1}$ is 0.2, all the occurrences of error AR at level 1 reporting 0 errors are replaced by -2. The rationale behind this idea is to "award" learners who have not made a frequent error in their proficiency level. Subsequently, in order to avoid having a too sparse representation, we grouped the 17 types of errors into 5 classes of errors: verb tense (VT), lexis and use of words (LUW), prepositions, articles, possessives and part of speech (PAP), agreement (AG) and generic errors (GE), as shown in Table 3. We divided each of the 5 error counts by the word count, in order to weigh also the text length. Finally, the error count in each level is normalized on a scale from 0 to 1.

Before applying our ranking of error gravity and introducing the negative bias, we also calculated the averaged error rates (i.e. the number of errors divided by the number of words times 100) of each of the 5 classes and of their sum for each proficiency level (see Table 4). In the VT class, the increase of the error rate at A2 can be explained by the fact that A1 learners generally use a smaller variety of tenses. As a result, they tend to make fewer verb tense errors.

Errors	**Class**
VT	VT
NSW + PH + EX + MW + WC + WO	LUW
AR + PO + PR + PS	PAP
AG + PL + SI	AG
D + IS + XC	GE

Table 3: The 5 error classes we used for our study.

Furthermore, we performed ANOVA on each of the 5 classes and we always obtained significant p-values (<0.05), thus finding that there are significant differences between proficiency levels in terms of errors.

	mean (%)				
	A1	**A2**	**B1**	**B2**	**C1**
LUW	3.67	3.10	2.69	1.96	1.58
PAP	1.63	1.42	1.20	0.99	0.70
AG	0.99	0.49	0.47	0.36	0.31
GE	2.00	1.67	1.29	0.95	0.80
VT	0.31	0.43	0.41	0.36	0.19
total	8.62	7.13	6.08	4.63	3.59

Table 4: Averaged error rate of each error class and their sum across proficiency levels.

2.3 ICNALE

In order to test our approach, we consider ICNALE (Ishikawa), a publicly available dataset [2] comprising written and spoken responses of ESL learners ranging from A2 to B2 and partially of native speakers. The CEFR levels were assigned prior to collecting the data, as the ICNALE team required all the learners to take an L2 vocabulary size test and to present their scores in English proficiency tests such as TOEFL, TOEIC, IELTS, etc. On the basis of these two scores, the learners were classified into proficiency levels. Only a small section of dialogues and essays has been scored by human experts so far and has been included in the ICNALE Global Rating Archives (Ishikawa, 2020): it currently includes the assessments and scores (on a scale from 0 to 100) of 140 dialogues and 140 essays by 40 human raters. Since not all the dialogues and essays were previously assigned a proficiency level, for our experiments we selected only the ones classified into CEFR levels and scored by human experts, and we also considered the scored texts

[2] http://language.sakura.ne.jp/icnale/download.html

and speeches of native speakers, therefore reducing the written section to 121 essays and the spoken section to 116 dialogues, of which we considered only the learners' utterances. Out of the 40 raters involved in the project, we only selected the native speakers with more than 5 years of experience in ESL teaching and assessment, i.e. 4 raters for the written section and 3 raters for the spoken section. We set the average of these scores as targets. Details about average and standard deviation of the raters' scores can be found in Ishikawa (2020).

2.4 CLC-FCE

Due to the limited amount of annotated data in the ICNALE corpus, we train our models on the CLC-FCE corpus, a publicly available dataset [3], containing the scripts of an English language exam aimed at around B2 level of the CEFR, which is also the highest level of the ICNALE corpus. Its 1244 exam scripts include responses to two different prompts asking the test-takers to write a short answer (e.g. a letter, an article, a report, a short story) and range from 200 to 400 words on average. Each answer has been error-tagged and annotated by human experts with a mark. Note that we eliminated the answers that did not report a score. More information about the dataset can be found in Yannakoudakis et al. (2011).

2.5 TLT-school

In Trentino, an autonomous region in northern Italy, the linguistic competence of Italian students have been assessed over years through proficiency tests in both English and German (Gretter et al., 2020), involving about 3000 students ranging from 9 to 16 years old, belonging to four different school grade levels (5^{th}, 8^{th}, 10^{th}, 11^{th}) and three proficiency levels (A1, A2, B1). Since our experiments are conducted only on the B1 section of the English written and spoken parts of the corpus, we will not describe the section concerning the texts and utterances of the German section, as their analysis goes beyond the scope of this paper.

The written section consists of 895 answers to 2 question prompts. Test-takers are asked two questions: the first one requires them to write a blog entry in which they have to describe what happened during the day and to talk about their plans for the rest of the week, while the second one asks them to write an email to a friend who broke an object borrowed from them. The spoken section is composed of 442 responses to 7 small talk questions about everyday life situations. It is worth mentioning that some answers are characterized by a number of issues (e.g. presence of words belonging to multiple languages or presence of off-topic answers). We decided not to eliminate these answers from the data used in the experiments, but we removed the empty responses.

As regards the speech transcriptions, we eliminated the annotations related to spontaneous speech phenomena such as hesitations and fragments of words etc. Detailed information about the manual transcriptions and other aspects of the corpus can be found in Gretter et al. (2020).

As for the automatic speech recognition (ASR) output text, its word error rate is 35.9% on the whole spoken test data, whereas it amounts to 41.13% for the B1 subset we used in our experiments; acoustic and language models are described in Gretter et al. (2019).

The total score ranges from 0 to 8 in the written section and from 0 to 12 in the spoken section and consists of the sum of the subscores assigned by human experts for each specific proficiency indicator assigned by the human raters (i.e. fulfillment, formal correctness and lexical complexity, cohesion, and narrative and descriptive competences for writing; and relevance, formal correctness, lexical complexity, pronunciation, fluency, and communicative competence for speaking). For each indicator human raters could choose 0, 1 or 2 points. Since every utterance was scored by only one expert, it was not possible to evaluate any kind of agreement among experts. Note that the CEFR levels were assigned before the tests and should be considered as expected proficiency levels, whereas the test scores are effectively representing each learner's performance in the exam. Table 6 shows the number of answers and word counts of the TLT-school spoken test set across test scores.

3 Model architectures

We build our models using a BERT architecture (Devlin et al., 2018) in the version provided by the HuggingFace Transformer Library (Wolf et al., 2019) (*bert-base-uncased*). In both the feature extractor and the scoring models BERT layers are frozen.

[3] https://ilexir.co.uk/datasets/index.html

	ICNALE		CLC	TLT	
	Wr	Sp		Wr	Sp
Train	-	-	2122	594	345
Dev	-	-	160	-	-
Test	121	116	194	301	97
Avg. len	225	186	192	103	28
Max. len	302	455	462	279	221
Min. len	179	23	72	1	1
Score	0-100	0-100	1-40	0-8	0-12

Table 5: Statistics (number of answers and word counts) for the three test sets: ICNALE (Written and Spoken), CLC-FCE, TLT-school (Written and Spoken).

Score	Samples	Min. len	Max. len	Avg. len
0-3	27	1	100	11.18
3-6	23	9	85	22.00
6-9	14	11	51	27.07
9-12	33	20	196	55.57

Table 6: Statistics (number of answers and word counts) for the TLT-school spoken test set across test scores.

3.1 Feature extractor

Specifically, EFEX takes a sequence of token embeddings i.e. of the answers provided by the learners $[x_1, ..., x_n]$, as inputs and predicts the 'biased' estimate (see formula in section 2.1) of error rate of each class of error, i.e. VT, LUW, PAP, AG and GE. Each rate is calculated by a final dense layer and the model uses mean squared error (MSE) as the loss function. For the GE and LUW outputs we add one and two extra dense layers respectively. We used Adam optimizer (Kingma and Ba, 2014) with learning rate of 8e-6, batch size set at 16, validation split at 0.1, and we trained our models for 60 epochs. Figure 2 shows the architecture of EFEX.

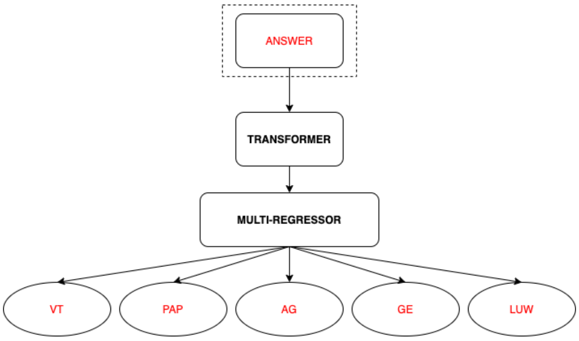

Figure 2: EFEX model architecture.

3.2 Scoring models

Before testing the impact of the labels generated by EFEX, we run several experiments on the selected datasets using our simple baseline scoring models, which take only a sequence of token embeddings, i.e. of the answers provided by the test-takers $[x_1, ..., x_n]$, as inputs and predict the total score of each answer normalized on a scale from -1 to 1. The EFEX-enriched models take the answers as inputs combined with a 5-dimensional vector, i.e. the number of classes of errors generated by EFEX, and have the same outputs as the baselines, as shown in Figure 3.

In both the baseline models and the EFEX-enriched models, the scores are calculated by a final dense layer and the model employs MSE as the loss function. The structure and hyper-parameters of the models are shown in Table 7. For the evaluation we consider two metrics: MSE and Pearson's correlation coefficient (PCC) between the true scores and the predicted ones.

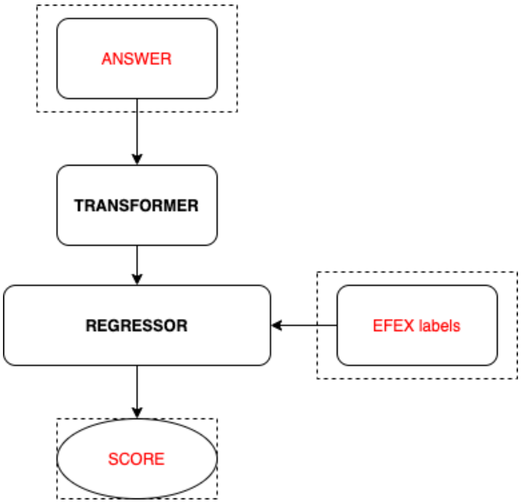

Figure 3: Scoring model architecture.

4 Experiments and results

4.1 CLC-FCE to ICNALE

We run a series of experiments starting from training EFEX on the EFCAMDAT dataset, setting VT, PAP, AG, GE and LUW as our prediction targets, feeding only the input text. We tested EFEX on the EFCAMDAT test set and we obtained significant results when comparing the true labels with the predicted ones in terms of PCC (see Table 8).

Secondly, we run the scorer on ICNALE (see Table 9); since we do not have enough ICNALE data for a proper training, we train our models on

	TLT	CLC/ICNALE
Max. seq. len.	256	512
Learning rate	9e-6	2e-6
Epochs	60 (120)	60 (150)
Batch size	32	16
1st Dense layer	768 - relu	768 - relu
Dropout	0.2	0.2
2nd Dense layer	128 - relu	64 - relu
Dropout	0.2	0.2
Output layer	1	1

Table 7: Model architectures and hyperparameters. The number of epochs in brackets refers to the EFEX-enriched model.

	PCC
LUW	0.796
PAP	0.862
AG	0.868
GE	0.831
VT	0.876

Table 8: EFEX performance in terms of PCC on EF-CAMDAT.

the CLC-FCE. Considering that we test our models trained on the CLC-FCE directly on out-of-domain data without fine-tuning, we achieve overall interesting results. In this case, the performance of the EFEX-enriched model is slightly lower than the baseline when tested on the scores of the ICNALE written set, but still better in terms of PCC when used for predicting the scores of the spoken set.

ICNALE	**Written**		**Spoken**	
Model	MSE	PCC	MSE	PCC
CLC baseline	0.201	0.719	0.121	0.614
+ EFEX labels	0.254	0.709	0.134	**0.625**

Table 9: Results on the ICNALE test dataset (MSE and PCC).

4.2 TLT-school - Written to spoken

Finally, we run our experiments on the TLT-school, training our baseline on the written training set and testing it on the spoken test set. We follow the same steps with our EFEX-enriched model and we gain a higher performance when predicting the spoken scores both using the manual transcriptions and the ASR output text, as shown in Table 10. Additionally, we fine-tune our model on the spoken training set for 2 epochs reducing the learning rate to 2e-6 and we obtain our best performance, reaching a PCC of 0.764 on the manual transcriptions.

	TLT - Spoken			
	Man. transcr.		**ASR**	
	MSE	PCC	MSE	PCC
Baseline	0.555	0.734	0.793	0.605
+ fine-tuning	0.488	0.741	0.715	0.609
+ EFEX labels	0.468	0.759	0.688	0.638
+ fine-tuning	**0.400**	**0.764**	**0.606**	**0.642**

Table 10: Results on the TLT test dataset (MSE and PCC): baseline; baseline + fine-tuning; baseline + EFEX labels; baseline + EFEX labels + fine-tuning.

Also the results on the ASR output appear to be enhanced by fine-tuning, as we obtain a PCC of 0.642. Fine-tuning the baseline without additional features reaches a PCC of 0.741 on the manual transcriptions and of 0.609 on the ASR. We find that the EFEX-enriched model achieves higher results across both metrics.

Furthermore, we continue our analysis comparing the performance of the baseline and the EFEX-enriched model across test scores. Figure 4 shows the MSE variation across 4 ranges of scores, i.e. 0-3, 3-6, 6-9, 9-12. It can be observed that the MSE is always lower for the EFEX-enriched model except in the range of scores between 0 and 3 on both the manual transcriptions and ASR output text, for which the EFEX-enriched model shows a minute increase of the MSE. Such difference is probably due to the fact that, in this specific range of scores, learners' answers, in addition to having lower quality, are also shorter on average (about 11 words), as shown in Table 6. As the score increases, the word average rises to 56 for scores between 9 and 12. Fewer words also means fewer and less variety of errors. Therefore, EFEX might be introducing some information that is not needed for answers with lower scores.

Specifically, the error distribution for the lowest range might be less informative, as can be inferred from the Frobenius norm values of the EFEX vectors for each score range shown in Table 11.

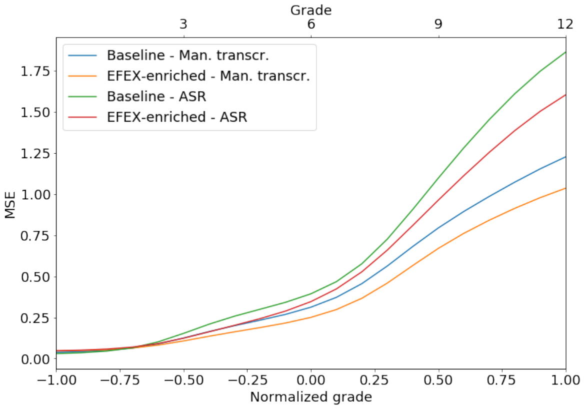

Figure 4: MSE variation across scores on manual transcriptions and ASR output text.

	Norm	
Score range	Man. transcr.	ASR
0-3	1.786	1.780
3-6	2.386	2.540
6-9	2.022	2.090
9-12	4.011	3.986

Table 11: Frobenius norm values of EFEX vectors across score ranges.

5 Conclusions and future work

In this work we presented a promising approach to automatic proficiency assessment of spoken responses based on the presence of errors across proficiency levels, extracted with an error feature extractor that we developed using a BERT-based architecture. Furthermore, we proposed to use models previously trained on written data in order to tackle the problem related to limited availability of spoken data. First, we tried our error-based approach on some publicly available datasets, training our models on the CLC-FCE and testing them on the ICNALE. In this case, our EFEX-enriched model managed to modestly improve the prediction of the dialogues scores in terms of PCC. Specifically for this experiment, one also has to consider the difference in domain and scoring metrics between the two corpora, albeit they are approximately around the same proficiency levels.

Subsequently, we discovered that the use of EFEX labels shows a more interesting improvement in scoring the spoken section of TLT-school after training our models on written data, suggesting that these additional features can mitigate the impact of ASR errors and some typical phenomena of the spoken modality. An example drawn from the data could be the following: *"in fact when a person does a lot of movement and moves a lot and goes out in the in the nature then his his body is in more healthy"*. The repetitions ’in the’ and ’his’ as well as what appears to be a wrongly inserted preposition ’in’ would be considered actual errors if they occurred in written productions, but not necessarily so in spoken texts.

Our assumption is that BERT models, as they are trained on large written corpora, already possess written grammatical knowledge and are sensitive to grammatical violations to a certain extent. Therefore, when evaluating written proficiency, they do not need to be warned with explicit indications with regard to errors, but error-related features can be beneficial to understand and decode the typical phenomena of oral language and learn spoken and conversational grammar. Considering that in spoken responses the scoring module could take advantage of a distinction of errors made by the speaker or introduced by ASR (Knill et al., 2019), we assume that there is still room for improvement in the approaches that detect errors as additional features.

Further work should be undertaken starting from the first step of our pipeline, i.e. the error feature extractor, since, despite the good results shown in Table 8, we can still improve it and analyse its effectiveness in various ways, e.g. by rearranging the error classes and remapping the ranking of error gravity.

Considering that we removed spontaneous speech phenomena such as hesitations and fragments of words from the data for our experiments, we envisage a combination of the approach presented in this paper and the use of error-related features derived from audio recordings, such as phonological errors as well as repetitions and other types of disfluency.

Moreover, we plan to investigate the impact of models trained on written data and tested on spoken data also for other CEFR levels. Finally, we acknowledge that the presence of errors cannot be the only feature to be taken into account when assessing L2 proficiency at higher levels, but, if properly weighted and balanced with other proficiency indicators, it might improve consistency and objectivity in assessment.

References

N. Ballier, T. Gaillat, A. Simpkin, B. Stearns, M. Bouyé, and M. Zarrouk. 2019. A supervised learning model for the automatic assessment of language levels based on learner errors. In *EC-TEL 2019 14th European Conference on Technology Enhanced Learning*, pages 1–13.

A. Caines, C. Bentz, K. Knill, M. Rei, and P. Buttery. 2020. Grammatical error detection in transcriptions of spoken English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2144–2162.

A. Caines and P. Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 5614–5623.

M. Canale and M. Swain. 1980. Theoretical bases for communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1):1–47.

L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian. 2018. End-to-end neural network based automated speech scoring. In *IEEE International Conference on Acoustics Speech and Signal Processing*.

S. P. Corder. 1967. The significance of learner's errors. *International Review of Applied Linguistics in Language Teaching*, V(1):161–170.

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

H. Craighead, A. Caines, P. Buttery, and H. 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*, pages 2258–2269.

S.A. Crossley, K. Kyle, and D.S. McNamara. 2015. To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *The Journal of Writing Assessment*, 9(1):1–19.

R. Cummins and M. Rei. 2018. Neural multi-task learning in automated assessment. page arXiv:1801.06830.

J. Devlin, M. Chang, L. Kenton, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

M. Gamon, M. Chodorow, C. Leacock, and J. Tetreault. 2013. Grammatical error detection in automatic essay scoring and feedback. In M.D. Shermis and J.C. Burstein, editors, *Handbook of Automated Essay Evaluation*, chapter 15, pages 251–266. Routledge, New York.

J. Geertzen, T. Alexopolou, and A. Korhonen. 2014. Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). In *Proceedings of the 2012 Second Language Research Forum*, pages 240–254.

R. Gretter, M. Matassoni, K. Allgaier, S. Tchistiakova, and D. Falavigna. 2019. Automatic assessment of spoken language proficiency of non-native children. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna. 2020. TLT-school: a corpus of non native children speech. In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.

J.A. Hawkins and P. Buttery. 2010. Criterial features in learner corpora: Theories and illustrations. *English Profile Journal*, 1(1):1–23.

P. Howson. 2013. *The English effect*. British Council, London.

D. Hymes. 1972. On communicative competence. In J. Pride J. Holmes, editor, *Sociolinguistics: Selected Readings*, pages 269–293. Penguin, Harmondsworth.

S. Ishikawa. A new horizon in learner corpus studies: The aim of the ICNALE project. In *Corpora and language technologies in teaching, learning and research*. University of Strathclyde Press.

S. Ishikawa. 2020. Aim of the ICNALE GRA project: Global collaboration to collect ratings of asian learners' l2 english essays and speeches from an ELF perspective. *Learner Corpus Studies in Asia and the World*, 5:121–144.

C. James. 2013. *Errors in language learning and use: Exploring error analysis*. Routledge.

D.P. Kingma and J. Ba. 2014. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*.

K. Knill, M. Gales, P. Manakul, and A. Caines. 2019. Automatic grammatical error detection of non-native spoken learner English. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 8127–8131.

R. Lado. 1961. *Language testing: the construction and use of foreign language tests*. Longman, London.

C. Leacock and M. Chodorow. 2003. Automated grammatical error detection. In M.D. Shermis and J.C. Burstein, editors, *Automated essay scoring: a cross-disciplinary perspective*, pages 195–207. Lawrence Erlbaum Associates, Mahwah, NJ.

J. Pfingsthorn. 2013. *Variability in learner errors as a reflection of the CLT paradigm shift*. Frankfurt am Main.

V. Raina, M.J.F. Gales, and K.M. Knill. 2020. Universal adversarial attacks on spoken language assessment systems. In *Interspeech 2020*, pages 3855–3859.

I. Shatz. 2020. Refining and modifying the EFCAMDAT. *International Journal of Learner Corpus Research*, 6(2):220–223.

J. Thewissen. 2013. Capturing l2 accuracy developmental patterns: Insights from an error-tagged efl learner corpus. *The Modern Language Journal*, 97(1):77–101.

S. Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28:79–105.

S. Vajjala and T. Rama. 2018. Experiments with universal cefr classification. In *Proceedings of 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153.

X. Wang, K. Evanini, Y. Qian, and M. Mulholland. 2021. Automated scoring of spontaneous speech from young learners of english using transformers. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 705–712.

T. Wolf et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. page arXiv:1910.03771.

X. Xi, D. Higgins, K. Zechner, and D.M. Williamson. 2008. Automated scoring of spontaneous speech using SpeechRater SM v1.0.

H. Yannakoudakis, T. Briscoe, and B. 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, pages 180–189.

M. Zhang. 2013. Contrasting automated and human scoring of essays. *R&D Connections*, (21):1–11.