Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5614–5623
Marseille, 11–16 May 2020

REPROLANG 2020: Automatic Proficiency Scoring of Czech, English, German, Italian, and Spanish learner essays

Andrew Caines & Paula Buttery
ALTA Institute & Computer Laboratory
University of Cambridge, 15 JJ Thomson Avenue, Cambridge U.K.
{andrew.caines,paula.buttery}@cl.cam.ac.uk

Abstract

We report on our attempts to reproduce the work described in Vajjala and Rama (2018), 'Experiments with universal CEFR classification', as part of REPROLANG 2020: this involves featured-based and neural approaches to essay scoring in Czech, German and Italian. Our results are broadly in line with those from the original paper, with some differences due to the stochastic nature of machine learning and programming language used. We correct an error in the reported metrics, introduce new baselines, apply the experiments to English and Spanish corpora, and generate adversarial data to test classifier robustness. We conclude that feature-based approaches perform better than neural network classifiers for text datasets of this size, though neural network modifications do bring performance closer to the best feature-based models.

Keywords: reproducibility, automated essay scoring, language proficiency, second language learning

1. Introduction

This paper relates to REPROLANG 2020, the Shared Task on the Reproduction of Research Results in Science and Technology of Language, specifically task D.2: Language Proficiency Scoring. The task involves repetition of the experiments described by Vajjala and Rama (2018) – namely automated essay scoring (AES) of texts written by learners of three languages contained in the MERLIN Corpus (Boyd et al., 2014): Czech, German and Italian.

Vajjala & Rama (V&R) set up their experiments as classification tasks, assigning each text to one of the six levels in the Common European Framework of Reference for Languages (CEFR). The reproduction exercise for REPROLANG involves three experiment types: **monolingual** classification in which both training and test data are the same language; **multilingual** classification in which training and test data come from all three languages; and **cross-lingual** classification in which German training data is used to score Czech and Italian test data. The primary metric for evaluation is weighted-F1.

We describe our attempts to reproduce V&R's experiments, report some differences in results and discuss why this might be, and present extensions to the work as well as proposals for future research. One extension involves additional languages – namely English and Spanish – in an attempt to further validate the idea of *universal* CEFR classification which V&R allude to in the title of their paper. Even with these additional languages, the representation of the typological variation found in the world's languages is limited, and therefore claims to 'universality' are premature. However, availability of CEFR-graded corpora is scarce, and a welcome future development would be further publication of such datasets for more languages.

2. Reproduction of the core experiments

V&R's experiments involved classification of essays into CEFR levels comparing feature-based and embedding inputs to machine learning algorithms.

2.1. Data

MERLIN contains 2287 texts[1] which have been proficiency scored and aligned with the CEFR scale from A1 (low) to C2 (high). There were no essays at level C2 – therefore C1 is the highest proficiency level found in MERLIN – and not all languages have instances of all CEFR levels between A1 and C1. V&R removed the 18 texts in the corpus without a CEFR level, along with texts from a CEFR level for which a language has fewer than 10 instances (see Table 1). In this way the original 2287 texts in MERLIN were reduced to 2267[2].

User	CEFR	CZ	DE	IT	Total
Basic	A1	0	57	29	86
	A2	188	306	381	875
Independent	B1	165	331	394	890
	B2	81	293	0	374
Proficient	C1	0	42	0	42
	C2	0	0	0	0
	Total	434	1029	804	2267

Table 1: Texts in the MERLIN corpus by CEFR level and target language, with a broad descriptor of the user type represented by the A*, B* and C* levels.

2.2. Feature extraction

V&R set out to use a range of linguistic features inspired by previous work in AES, on the assumption that they generalise well across languages. These are namely:

1. Word token and part-of-speech (PoS) tag n-grams (Yannakoudakis et al., 2011), from 1 to 5-gram se-

[1] Available from `https://merlin-platform.eu/C_data.php`.

[2] In Vajjala and Rama (2018) the count of texts is said to be 2286 which appears to be a miscount as there are 2267 files in the dataset they distribute with their repository.

quences. For cross-linguistic application, the Universal PoS tagset (UPoS) is used (Nivre et al., 2019).

2. Per-word dependency triples consisting of a dependency relation, PoS tag of the dependent and PoS tag of the head word. For instance, a subject noun in a sentence such as 'she saw', would be represented by the concatenated triple {nsubj,noun,verb} (Zesch and Horbach, 2018).
3. So-called *domain* features including document length (n.tokens per text), spelling and grammar error counts for German and Italian from the LanguageTool[3] (no such resource was found for Czech); in addition lexical density, lexical variation, and lexical diversity (Lu, 2012) which are defined below.
4. Task-specific word and character embeddings (Alikaniotis et al., 2016). In the monolingual experiments, only word embeddings are learned as 100-wide vectors. In the multilingual experiments, 32-wide word embeddings are concatenated with 16-wide character embeddings.
5. For multilingual experiments, the language of the text is supplied as an optional additional feature for feature-based approaches, or as an auxiliary learning task for the neural networks.

Features based on a syntactic parse – the PoS and dependency features – involved the use of UDPipe (Straka and Straková, 2017) trained on Universal Dependencies (UD) treebanks version 2.0 (Nivre et al., 2019).
The lexical domain features listed in point (4) above involve a distinction between lexical, open-class (OC) and functional, closed-class (CC) words: the former are namely adjectives, adverbs, interjections, nouns, proper nouns and verbs; the latter are adpositions, auxiliaries, conjunctions, determiners, numerals, particles, pronouns, symbols and 'other' in terms of UPoS tags. There is also a distinction between the number of unique words (types) and the count of all words (tokens) in a text.
Lexical density is thus the number of lexical tokens divided by the number of all tokens:

$$L_{dens} = \frac{OC_{tok}}{CC_{tok} + OC_{tok}} \quad (1)$$

The lexical variation for a given text is the number of lexical types over the number of lexical tokens:

$$L_{var} = \frac{OC_{typ}}{OC_{tok}} \quad (2)$$

Lexical diversity is the number of unique words in the text over the number of tokens in a text, also known as the type-token ratio:

$$L_{div} = TTR = \frac{CC_{typ} + OC_{typ}}{CC_{tok} + OC_{tok}} \quad (3)$$

In the monolingual experiments, the domain features were combined with the word n-grams, PoS n-grams and dependency triples: in order to do so the CEFR class probabilities for each text were first estimated based on the n-grams or triples alone, and these probabilities were concatenated to the domain features.

[3] https://languagetool.org

2.3. Classification experiments

CEFR grading can be treated as a regression problem if the CEFR scale is viewed as ranging from 0 (pre-A1) to 6 (C2). But the steps between levels are not necessarily equal, and it is also appropriate to treat CEFR grading as a classification problem as V&R do[4]. The task is therefore to automatically assign a CEFR level from A1 to C1 to the MERLIN essays, recalling that there are no essays at CEFR level C2 in the corpus.
V&R defined three experiment types: *monolingual, multilingual* and *cross-lingual*. In the first, **monolingual**, the language of the training and test data are the same. In the second, **multilingual**, all languages of the MERLIN corpus are mixed in as training and test data. In the third experiment, the **cross-lingual** one, German essays – because they are the most wide-ranging in terms of CEFR level – are used as training data in order to grade Czech and Italian test data in separate sub-experiments.
The features described in section 2.2. are not all suitable for each experiment type. For instance, document length is universally available, but word n-grams are not a suitable training feature for the cross-linguistic experiment because the train and test lexicons come from distinct languages (German and Czech or Italian respectively).
V&R use three classifier types for all experiments: logistic regression, random forests, and support vector machines. In addition, for the monolingual experiments they trained word embeddings as inputs to a neural network classifier, and for the multilingual experiments they trained both word and character embeddings. Such language-specific inputs are not appropriate for the cross-linguistic experiments, therefore V&R did not use neural nets in this case.
A summary of the three experimental settings is shown in Table 2. For all experiments except the cross-lingual ones, when the test set is defined, evaluation is done using ten-fold cross-validation and results are averaged across folds.

2.4. Reproduction

To reproduce V&R's experiments, we first ran the code available in the GitHub repository associated with their 2018 publication[5]. Secondly we re-implement their experiments by writing our own code, a task which we describe later in this section. Thirdly we extended their experiments as far as time allowed, described in section 3.

2.5. Running the experiments from V&R's codebase

V&R published a GitHub repository to go with their 2018 publication. It contains Python code as well as the original MERLIN texts and extracted features. It was not straightforward to start running the code: the readme file is sparse, there is no list of required libraries, and data files

[4] Nevertheless, below we do convert the scale to integers in order to report a useful error metric: root-mean-square error.

[5] https://github.com/nishkalavallabhi/UniversalCEFRScoring

Experiment	Train	Test	Features	Classifiers
Monolingual	CZ DE IT	CZ DE IT	doc.length, word n-grams, PoS n-grams, dep.triples, domain, word embeddings	LR, RF, SVM, NN
Multilingual	CZ,DE,IT	CZ,DE,IT	doc.length, word n-grams, PoS n-grams, dep.triples, domain, word and character embeddings	LR, RF, SVM, NN
Cross-lingual	DE DE	CZ IT	doc.length, PoS n-grams, dep.triples, domain	LR, RF, SVM

Table 2: Experiment types used by V&R: training languages, test languages, features and classifiers used. Key: CZ=Czech, DE=German, IT=Italian; LR=logistic regression, RF=random forests, SVM=support vector machine, NN=neural networks.

are called either through unexplained argument variables or via filepaths specific to the authors' machines.

The experiments are distributed across five scripts: one for the baseline classifier, another three for neural networks in multilingual and monolingual settings, and a fifth which contains the bulk of experiments named `IdeaPOC.py`. The various experimental settings have to be implemented by updating lines in the code – for instance, switching out the Czech for the German set of essays, or calling different feature types and setting language as an additional feature for the cross-lingual experiments. Having run the code, the resulting print statements are perhaps meaningful to the authors but are less so to outsiders. However, we stress that it is to the authors' credit that they released code accompanying their paper, in the spirit of open science and reproducibility, and also volunteered their work as one of the REPROLANG tasks.

We re-ran V&R's code once and got slightly different F1 scores to the ones reported in their 2018 paper. This demonstrates the effect of different computation settings as well as the random-ness of the machine learning process, particularly neural networks, despite V&R setting the values of random seeds in their code at various points. We re-ran the neural network scripts ten times, reporting mean scores in the results tables.

In addition we corrected the occasional mistaken use of macro-F1 calls in V&R's code (in their paper they state that they are reporting weighted-F1), indicating in the results tables where this applies. To be clear, macro-F1 is the mean of per-class F1 scores. Weighted-F1 assigns a weight to the per-class F1 scores according to the frequency of each class in the test set:

$$wF1 = \frac{F1_1 * w_1 + F1_2 * w_2 + \ldots + F1_N * w_N}{w_1 + w_2 + \ldots + w_N} \quad (4)$$

where $F1_1..F1_N$ are the F1 scores per class from 1 to N, and $w_1..w_N$ are the counts of instances for each class in the test set. Macro-F1 and weighted-F1 can give very different outcomes, as shown in the results tables below.

The weighted-F1 measures from re-running V&R's code are reported alongside the original F1 scores given in their 2018 publication in Tables 3 (monolingual experiments), 4 (multilingual experiments), and 5 (cross-lingual experiments).

2.6. Re-implementing V&R's experiments

As a further exercise in the reproducibility of V&R's work, we wrote our own codebase to implement the experiments they describe and also make it available as a GitHub repository[6]. Whereas V&R wrote in Python and mainly used `scikit-learn` functions (Pedregosa et al., 2011), we wrote in R and mainly use `caret` functions (R Core Team, 2019; Kuhn, 2019). For neural networks we both used the Keras interface to TensorFlow (Chollet and others, 2015; Abadi et al., 2016).

For the most part we were able to reproduce the results reported in V&R, once all experiments were evaluated with weighted-F1. There are slightly different outcomes in terms of best models, which we discuss, and overall our results are superior to V&R's but at the cost of a marked deterioration in speed.

For instance, we ran V&R's `IdeaPOC` script (which executes most of the feature-based experiments for all languages) within an hour whereas on the same machine our monolingual feature-based experiments for one language took several hours, while the larger multilingual experiments took a few days to run. In accordance with the proposal that computational budgets should be taken into account (Dodge et al., 2019), we acknowledge that the `scikit-learn` approach to machine learning is superior to `caret` in that performance is only slightly lower for much reduced execution time.

2.6.1. Pre-processing

Since V&R include the processed MERLIN data in the repository, it cannot have been high priority to walk others through the pre-processing steps required to get from the original texts to extracted features. For this reason there is some missing information but the problems were easily remedied and we created a pull request with our updates to their pre-processing script on GitHub.

As a first step we transform the corpus texts from the original files which include various metadata to new files which only include the students' essays, while keeping count to ensure that all language-CEFR groups have at least ten text instances. This step is largely based on V&R's original script, but with modification so that file filtering is handled at this stage (rather than removed in a posthoc step).

[6] `https://github.com/cainesap/CEFRgrader`

Features	V&R 2018			V&R re-run 2019			REPROLANG 2020		
	DE	IT	CZ	DE	IT	CZ	DE	IT	CZ
†Zero rule							0.157	0.322	0.262
†Probabilistic							0.288	0.459	0.376
*Document length	0.497	0.578^{L}	0.587^{L}	0.616^{L}	0.800^{L}	0.596^{L}	0.643^{L}	0.815^{L}	0.597^{L}
Word n-grams	0.666	0.827	0.721	0.590	0.800	0.728	0.666	0.823	0.721^{S}
PoS n-grams	0.663	0.825	0.699	**0.659**	0.801	0.679	0.672	0.806	0.704
Dependency triples	0.663	0.813	0.704	0.637	**0.808**	0.707	0.666	0.800	0.679
Domain features	0.533^{L}	0.653^{L}	0.663	0.625	0.807	0.663	0.691^{L}	0.812^{L}	0.648^{L}
Word n-grams + Domain	**0.686**	**0.837**	**0.734**	0.638	0.793	0.721	0.700^{L}	**0.838**	**0.729**
PoS n-grams + Domain	**0.686**	0.816	0.709	0.653	0.792	0.687	0.690	0.823^{L}	0.702^{S}
Dep. triples + Domain	0.682	0.806	0.712	0.637	0.782	**0.730**	**0.702^{L}**	0.821	0.693
†Word embeddings	0.646^{N}	0.794^{N}	0.625^{N}	0.602^{N}	0.771^{N}	0.623^{N}	0.382^{N}	0.616^{N}	0.399^{N}
+ 300w embeddings				0.606^{N}	0.786^{N}	0.621^{N}	0.492^{N}	0.717^{N}	0.368^{N}
+ Adam				0.620^{N}	0.773^{N}	0.658^{N}	0.650^{N}	0.819^{N}	0.682^{N}

Table 3: Monolingual CEFR classification experiments with German (DE), Italian (IT) and Czech (CZ) texts (Table 2 in V&R). Asterisks indicate that the original V&R 2018 values are macro-F1 whereas they should be weighted-F1. A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

2.6.2. Feature extraction

We followed V&R's feature extraction methods closely, using UDPipe for R (Wijffels, 2019) and parsing models pretrained on UD 2.0, querying the LanguageTool jar for German and Italian texts, and sampling ten test-folds stratified over CEFR levels. Data handling is largely carried out with 'tidyverse' and 'tidytext' methods (Wickham et al., 2019; Silge and Robinson, 2016).

We re-scale all feature values to between 0 and 1, a normalisation step we do not believe V&R carried out, having inspected their code. In addition, in an attempt to mitigate the problem of slow training, we pruned the least frequent features iteratively until our feature matrices were less than or equal to 1000 columns wide, based on document frequency[7]. V&R set a fixed minimum document frequency of 10 for feature inclusion, while our threshold is thus variable, sometimes being much higher than 10.

For the neural network classifiers, input texts were tokenized and a minimum *corpus* frequency[8] of 15 is set for inclusion (matching V&R's threshold) otherwise words are set as out-of-vocabulary. We also ensured all text sequences were 400 tokens long for word embeddings or 2000 characters for the character embeddings, either by clipping long texts or padding short ones, per V&R.

2.6.3. Classification

Like V&R, we trained three classifiers for each feature-based approach: multinomial logistic regression, random forests, and linear support vector machines. The first two are much slower than SVMs, though the latter in general perform more poorly with the exception of Czech monolingual classification, which displays a few oddities compared to German and Italian (Table 3).

Neural networks were implemented with Keras, following V&R in learning a 100-wide word embedding for monolingual experiments, or a 32-wide word embedding concatenated to a 16-wide character embedding for multilingual experiments. In monolingual experiments a batch size of 32 was used for a maximum of 10 epochs, with categorical cross-entropy loss and the AdaDelta optimisation algorithm (Zeiler, 2012). Finally a softmax layer emits class probabilities.

In the multilingual experiments, dropout is applied to the embedding layer outputs at a rate of 0.25, and the set-up is otherwise the same except for a batch size of 128 and a maximum of 8 epochs. For the multilingual experiments with the language of the text as an auxiliary objective, language prediction is weighted at 0.5.

As a final modification we introduce new baselines to put the feature-based approaches in context. V&R portrayed document length as a baseline approach but it can often be an informative feature so we instead evaluate two naive baselines: a *zero rule* classifier which predicts the most frequent class seen in training, and a probabilistic classifier which assigns CEFR levels to a text based on the distribution of classes found in the training data. As part of reviewing the role of the document length classifier, we also report a 'document length plus language' multilingual experiment in Table 4 which V&R did not do.

2.6.4. Evaluation

In keeping with V&R's experiments and suitably for imbalanced multi-class classification, weighted-F1 is our primary metric. We report two additional metrics for a more comprehensive portrayal of model performance: root-mean-square error (**RMSE**) and percent-within-one-level (**within1**). The former is a common metric for regression evaluation and involves first transforming the CEFR levels to integers from 1 to 6. RMSE thus indicates the average distance of model predictions from the true labels:

[7]The count of documents in which feature f appears.

[8]The count of word w in all texts in the corpus.

Features	V&R 2018		V&R re-run 2019		REPROLANG 2020	
	lang(-)	lang(+)	lang(-)	lang(+)	lang(-)	lang(+)
†Zero rule					0.221	n/a
†Probabilistic					0.334	n/a
*Document length	0.428^{L}	n/a	0.574^{L}	n/a	0.600^{L}	0.709^{L}
Word n-grams	0.721	0.719	0.606	0.607	**0.740**	**0.736**
PoS n-grams	**0.726**	**0.724**	**0.680**	**0.681**	0.732	0.731
Dependency triples	0.703	0.693	0.651	0.653	0.710	0.716
Domain features	0.449^{L}	0.471^{L}	0.597	0.647	0.698	0.726^{L}
†Word + Char embeddings	0.693^{N}	0.689^{N}	0.659^{N}	0.657^{N}	0.391^{N}	0.401^{N}
+ 300w embeddings			0.666^{N}	0.648^{N}	0.486^{N}	0.482^{N}
+ Adam			0.667^{N}	0.662^{N}	0.724^{N}	0.725^{N}

Table 4: Multilingual CEFR classification experiments with and without language information as a feature or auxiliary task (Table 3 in V&R); lang(+) indicates that the language of the text is used as an additional feature or as an auxiliary objective for neural networks, lang(-) indicates the absence of such a feature or task. Asterisks indicate that the original V&R 2018 values are macro-F1 whereas they should be weighted-F1. A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

Features	V&R 2018		V&R re-run 2019		REPROLANG 2020	
	Test:IT	Test:CZ	Test:IT	Test:CZ	Test:IT	Test:CZ
†Zero rule					0.322	0.210
†Probabilistic					0.368	0.326
Document length	0.553^{L}	0.487^{L}	0.553^{L}	0.487^{L}	0.595^{L}	0.339
PoS n-grams	**0.758**	0.649	**0.751**	**0.680**	**0.689**	0.377^{S}
Dependency triples	0.624	**0.653**	0.601	0.665	0.591	**0.387**
Domain features	0.630^{L}	0.475	0.575	0.476	0.614^{S}	0.339

Table 5: Cross-lingual CEFR classification experiments with German training data and Italian (IT) or Czech (CZ) test data (Table 4 in V&R). A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}} \quad (5)$$

where t is a single text and T represents the number of texts in the test corpus.

Within1 is a measure which indicates AES stability in terms of a proportion of 'damaging' errors which are more than one CEFR level from the true label. The underlying intuition is that a model which assigns a B2 text two levels away at A1 or C2 has made an error which is more than twice as grave as one which assigns the text to level A2 or C1 (within one level). It can be thought of as tolerable error in contrast to being out-by-two, which can have severe consequences for the student if they are under-scored (for example, affecting employment or education prospects), or conversely for the testing organisation if they drastically over-score students who are not of that proficiency level, thereby causing themselves reputational damage.

2.6.5. Results

We report results in Table 3 for monolingual experiments, Table 4 for multilingual experiments, and Table 5 for cross-lingual experiments. These correspond to Tables 2, 3 and 4 in V&R's 2018 publication.

In the left-most columns we repeat their reported results. In the middle columns we report from re-running their Python code, with corrections to weighted-F1 rather than macro-F1 where appropriate (marked by asterisks). In the right-hand columns we show the results from our re-implementation of their experiments. Like V&R we report results from the random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

In terms of results reproduction from V&R's original **monolingual** experiments, one difference involves the reporting of weighted-F1 rather than macro-F1 for document length and domain features. This leads to higher scores, visible in the middle 'V&R re-run' columns of Table 3, because the classifiers trained on these features are performing better for the larger CEFR groups in the corpus. However, we also found in our 2019 re-run that most measures are slightly down on the 2018 results. A notable exception is the Czech subset with dependency triples and domain features which became the leading model for that language in our re-run, whereas in the 2018 results the domain and word n-gram model was best for all languages. In the re-run the leading models are PoS n-grams for German, dependency triples for Italian, and dependency triples

Features	Weighted-F1			RMSE			Within1		
	DE	IT	CZ	DE	IT	CZ	DE	IT	CZ
†Zero rule	0.157	0.322	0.262	2.19	1.56	1.06	35.3	51.0	81.3
†Probabilistic	0.288	0.459	0.376	1.38	0.801	1.04	71.3	96.6	84.7
Document length	0.643^{L}	0.815^{L}	0.597^{L}	0.597^{L}	0.419^{L}	0.624^{L}	99.1^{L}	$\mathbf{100^{L}}$	99.3^{L}
Word n-grams	0.666	0.823	0.721^{S}	0.565	0.404	0.503^{S}	99.6	**100**	$\mathbf{99.8^{S}}$
PoS n-grams	0.672	0.806	0.704	0.558	1.07	0.547	**99.8**	91.7	**99.8**
Dependency triples	0.666	0.800	0.679	0.569	1.07	0.574	99.5	91.7	99.3
Domain features	0.691^{L}	0.812^{L}	0.648^{L}	0.561^{L}	0.420^{L}	0.596^{L}	99.4^{L}	$\mathbf{100^{L}}$	99.1^{L}
Word n-grams + Domain	0.700^{L}	**0.838**	**0.729**	$\mathbf{0.546^{L}}$	**0.390**	**0.524**	99.6^{L}	**100**	**99.8**
PoS n-grams + Domain	0.690	0.823^{L}	0.702^{S}	**0.546**	0.411^{L}	0.556^{S}	**99.8**	$\mathbf{100^{L}}$	99.5^{S}
Dep. triples + Domain	$\mathbf{0.702^{L}}$	0.821	0.693	0.547^{L}	0.416	0.564	99.6^{L}	**100**	99.5
†Word embeddings	0.382^{N}	0.616^{N}	0.399^{N}	0.978^{N}	1.34^{N}	0.944^{N}	88.6^{N}	71.1^{N}	87.1^{N}
+ 300w embeddings	0.492^{N}	0.717^{N}	0.368^{N}	1.46^{N}	1.25^{N}	0.958^{N}	64.3^{N}	79.3^{N}	86.5^{N}
+ Adam	0.650^{N}	0.819^{N}	0.682^{N}	0.584^{N}	0.411^{N}	0.569^{N}	99.7^{N}	$\mathbf{100^{N}}$	99.6^{N}

Table 6: Monolingual CEFR classification experiments with German (DE), Italian (IT) and Czech (CZ) texts: REPROLANG 2020 code re-implementation of V&R, weighted-F1, RMSE and 'within1' percentage. A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks). These notations also apply to the following two tables below.

Features	Weighted-F1		RMSE		Within1	
	lang(-)	lang(+)	lang(-)	lang(+)	lang(-)	lang(+)
†Zero rule	0.221	n/a	1.93	n/a	42.4	n/a
†Probabilistic	0.334	n/a	1.19	n/a	79.9	n/a
Document length	0.600^{L}	0.709^{L}	0.648^{L}	0.540^{L}	98.5^{L}	99.6^{L}
Word n-grams	**0.740**	**0.736**	**0.501**	**0.505**	99.7	99.7
PoS n-grams	0.732	0.731	0.514	0.511	99.7	**99.8**
Dependency triples	0.710	0.716	0.547	0.534	99.3	99.6
Domain features	0.698	0.726^{L}	0.558	0.525	99.4	99.7
†Word + Char embeddings	0.391^{N}	0.401^{N}	0.963^{N}	0.941^{N}	88.7^{N}	89.9^{N}
+ 300w embeddings	0.486^{N}	0.482^{N}	1.50^{N}	1.52^{N}	62.7^{N}	62.2^{N}
+ Adam	0.724^{N}	0.725^{N}	0.521^{N}	0.520^{N}	$\mathbf{99.8^{N}}$	$\mathbf{99.8^{N}}$

Table 7: Multilingual CEFR classification experiments with and without language information as a feature or auxiliary task: REPROLANG 2020 code re-implementation of V&R, weighted-F1, RMSE and 'within1' percentage.

Features	Weighted-F1		RMSE		Within1	
	Test:IT	Test:CZ	Test:IT	Test:CZ	Test:IT	Test:CZ
†Zero rule	0.322	0.210	1.56	1.06	51.0	81.3
†Probabilistic	0.368	0.326	1.25	1.71	76.9	58.6
Document length	0.595^{L}	0.339	0.736^{L}	1.54^{L}	98.0^{L}	61.1^{L}
PoS n-grams	**0.689**	0.377^{S}	**0.632**	1.34^{S}	**98.5**	72.4^{S}
Dependency triples	0.591	**0.387**	0.731	**0.952**	98.1	**89.2**
Domain features	0.614^{S}	0.339	0.730^{S}	0.979	97.8^{S}	88.7

Table 8: Cross-lingual CEFR classification experiments with German training data and Italian (IT) or Czech (CZ) test data: REPROLANG 2020 code re-implementation of V&R, weighted-F1, RMSE and 'within1' percentage.

plus domain features for Czech.

In our re-implementation of V&R's experiments (the 'REPROLANG 2020' columns), the weighted-F1 measures are very close to V&R's 2018 results, and on the whole are slightly better with the exception again of Czech which oscillates above and below the original results. The best models involve combinations of domain and word or dependency features, supporting V&R's original experiments.

A notable issue involves the word embeddings and neural network classifier: our new results are much worse than V&R's, both from their 2018 paper and from our re-run of their code. Having closely followed their method and checked parameter settings from their code, we are not sure why this should be other than the fundamental difference in

Keras environments (Python for V&R, R for us). To further investigate the problem with the neural network experiments, we increased the output width of the embedding layer from 100 to 300, and used the Adam optimiser (Kingma and Ba, 2015) as opposed to AdaDelta. We applied these same changes to V&R's code and include the results as extra rows in Table 3. We find that with these two updates our results improve markedly and in fact surpass the equivalent scores from V&R's modified code. This points to a problem with the R Keras AdaDelta call, but it requires further investigation to confirm this supposition.

Again, for the **multilingual** experiments, our 2019 re-run of V&R's code leads to lower weighted-F1 scores in general, but the same conclusion in terms of the best approach: PoS n-grams. In our re-implementation we again see higher performance overall and a different conclusion as to the best model: in our case word n-grams. The addition of language as a feature or auxiliary task has little effect for the most part, with the notable exception of document length. We encounter the same puzzle in the performance of neural network classifiers as seen in the monolingual experiments. Again, wider outputs from the embedding layer make a small improvement, and a switch to Adam optimiser makes a large improvement such that performance surpasses V&R's.

In the **cross-lingual** experiments, we find broadly similar results between V&R's original paper and the re-run of their code. The results from our re-implementation are slightly down on V&R's results for German, and markedly inferior for Czech (nevertheless above our baselines). The experimental set-up is the same, so we are not sure why this should be. However, the idea of cross-lingual CEFR classification is an interesting toy experiment but perhaps not a setting one would encounter, or wish to implement, in the wild, unless there were a very good reason to do so – namely CEFR scoring based on *very* similar languages, which German, Italian and Czech are not. It is linguistically telling that the best features for cross-lingual classification are PoS n-grams and dependency triples, indicating that there are at least some similarities in the cross-linguistic morpho-syntactic patterns of development shown by language learners as their proficiency improves.

In Tables 6, 7, 8 we show a broader range of performance metrics from our re-implementation of V&R's work: left-hand columns show weighted-F1, the middle columns show root-mean-square error (RMSE) with CEFR treated as a numeric scale, and the right-hand columns show the percentage of texts scored within 1 level of their true CEFR level. These show that the models which perform best on weighted-F1 tend to be best on RMSE and within1, with a few exceptions – notably the neural networks with Adam. The better Italian scores indicate that this is the least challenging of the three languages, with most texts at 2 of its 3 levels (Table 1), whereas German has more levels (5) and Czech has a more balanced distribution across 3 levels.

3. Extensions

We ran the following extensions to V&R's experiments, involving more languages, adversarial data, and further work on the neural network classifiers. Due to time limitations we ran only the baselines, single feature and neural network models (not the domain+ feature combinations) for extensions 1 and 2.

3.1. Extension 1: Adding new languages

User	CEFR	EN	ES
Basic	A1	585	69
	A2	845	387
Independent	B1	631	312
	B2	469	363
Proficient	C1	483	406
	C2	287	237
	Total	3300	1774

Table 9: Texts in the Write & Improve (EN: English) and CEDEL2 (ES: Spanish) corpora by CEFR level.

Features	DE	IT	CZ	EN	ES
†Zero rule	.156	.344	.410	.262	.210
†Proba.	.277	.494	.498	.355	.340
Doc.len.	$.65^{L}$	$.81^{L}$	$.60^{L}$	.321	$.31^{L}$
Words	.668	**.827**	$.71^{L}$	.367	**.477**
PoS	.660	.815	**.724**	$\mathbf{.37^{L}}$	.410
Dep.trips	.642	.788	$.66^{S}$	.343	.417
Domain	$\mathbf{.69^{L}}$	.807	$.65^{L}$	$\mathbf{.37^{L}}$	.415
†Embeds	$.65^{N}$	$.82^{N}$	$.67^{N}$	$.36^{N}$	$.41^{N}$

Table 10: Extension 1. Monolingual 6-level CEFR classification experiments with five languages. A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

We sought out corpora of CEFR-labelled essays in languages other than those used so far. Thanks to the BEA Shared Task 2019 (Bryant et al., 2019) we were aware of the Write & Improve public set of 3300 essays, and thanks to work by del Río (2019) we became aware of the Spanish CEDEL2 corpora (Lozano and Mendikoetxea, 2013) which contains 1774 essays. Note that CEDEL2 essays are mapped to CEFR levels from University of Wisconsin placement test scores following the schema developed by Crístobal Lozano. Counts of essays at each CEFR level are shown in Table 9.

Note that the English and Spanish corpora span all 6 CEFR levels from A1 to C2, while both are larger than the MERLIN language subsets. The new corpora have many more C-level texts than are found in MERLIN: thus we can seek to answer two questions – whether the same features which work for the languages seen so far also work in other languages, and whether those features generalise to language at a more advanced level.

We ran experiments for this extension in the same way as the core ones (the reproduction of V&R's work) with the following exceptions: (1) we used the most recent UDPipe

models, those trained on UD 2.4 treebanks; (2) for feature-based approaches we reduced the maximum number of features from 1000 to 400 (to ensure that feature matrices are narrower than they are high, with documents arranged in rows and features in columns), seeking out efficiencies in training without detriment to performance.

Results are reported in Table 10, and we find that the models perform more poorly on the English and Spanish texts. We do not propose that this is to do with the languages themselves, but rather to properties of the new corpora, especially as models unaffected by pre-processing steps (document length and neural networks) show deterioration similar to the models requiring parse information. The baselines indicate that classification is harder for these corpora than for the Italian and Czech subcorpora but easier than for German, and yet performance is worse: this is a matter for further investigation, but it does seem that the presence of all 6 CEFR levels and more advanced level texts presents a greater challenge for automatic scoring. We note that performance on the MERLIN languages is not affected by the reduction in features from 1000 to 400.

3.2. Extension 2: Adversarial data

In extension 2 we question how robust the various classification models are to adversarial data: that is, by rogue texts which would undermine the reliability of CEFR classifiers if they are not detected as such. Some have suggested that auto-markers are vulnerable to such 'attacks' (Hockly, 2018; Yoon et al., 2018) and so it is important to consider how our models react to unexpected inputs.

We use the English W&I corpus as our starting point and introduce four types of adversarial data, all of which we label as spam (i.e. the task is now 7-way classification, with the spam label alongside the 6 CEFR levels):

1. Randomly scrambling the order of word tokens in 100 randomly selected English texts in the W&I corpus, resulting in nonsense texts;
2. Randomly selecting 25 texts from each of the Czech, German and Italian corpora and adding them to the training and test sets: *seen other languages*;
3. Randomly selecting 25 Spanish texts from CEDEL2 and adding them to the test set only: *unseen other languages*;
4. Randomly extracting 100 English texts from the W&I corpus, clipping the first sentence or first 10 word tokens (whichever is shorter), providing those words as a prompt to the 355m parameter GPT-2 language model (Radford et al., 2019), and requesting a response length to replace the omitted words.

We report results in Table 11, with overall weighted-F1 for the extended English corpus (EN) and F1 scores for each adversarial text type on spam detection. It does seem that scrambled texts based on genuine learner essays are hard to detect, as are texts in other languages which have been seen during training. These two outcomes suggest that deeper syntactic and semantic representations are needed, while foreign language texts should be detected by a first-step filter rather than added to training data. This conclusion is borne out by the strong performance on unseen other language texts. Finally the GPT-2 texts are also successfully detected, presumably because the more complex vocabulary generated by GPT-2 is distinctive compared to genuine learner essays.

Features	EN	Scr	Seen	Unseen	GPT-2
†Zero rule	.100				
†Proba.	.174				
Doc.len.	.255	.131	.148	.084	.165
Words	.390	0.0	0.0	**1.0**	.808
PoS	.379	.022	.029	**1.0**	**.831**
Dep.trips	.368	.022	.029	**1.0**	.816
Domain	.377	.039	.193	**1.0**	.857
†Embeds	**.41**^{N}	.06^{N}	**1.0**^{N}	**1.0**^{N}	.73^{N}

Table 11: Extension 2. Monolingual 7-level CEFR classification experiments with English (EN) and adversarial texts (Scr: scrambled English texts, Seen: other languages seen in training, Unseen: other languages not seen in training, GPT-2: English texts modified by GPT-2). A text dagger indicates that the score is the mean of 10 runs. Cells in bold highlight the best performing model for each column. Results are from random forest classifiers unless indicated by a superscript character (L for logistic regression, S for support vector machines, N for neural networks).

3.3. Extension 3: Neural network modifications

It is apparent in the core experiments that for the MERLIN datasets feature-based classification models outperform the neural networks. However, it is not clear whether modification to the neural networks implemented by V&R might yet outperform the feature-based approaches. We trialled several modifications including:

NN1. reducing minimum word frequency from 15 to 10;

NN2. increasing the fixed-length text size from 400 to 500 word tokens;

NN3. omitting out-of-vocabulary words, rather than representing them with a generic word index and weights vector;

NN4. introducing a random 20% validation split to the training set when fitting the model;

NN5. regularisation measures such as L2 at 0.001, dropout at 50%, a smaller batch size of 16 rather than 32, and allowing more epochs: 40 rather than 10;

NN6. the use of pre-trained fastText embeddings learned from Wikipedia in each language (Bojanowski et al., 2017).

Results are compared with the V&R model with Adam optimiser which performed best in the core experiments: model NN0 in Table 12. Some of the modifications move performance close to the best feature-based models, while those thought to prevent over-fitting (NN4, NN5) do not improve performance on these datasets but might help the models generalise to other data. Thus, improvement in neural network scoring of these corpora remains a matter for further investigation.

NN model	DE	IT	CZ	EN	ES
NN0	.650	.819	**.682**	.358	.413
NN1	.655	.817	.661	.360	**.419**
NN2	.651	.816	.673	.358	.410
NN3	**.671**	**.826**	.627	**.361**	.412
NN4	.644	.812	.622	.316	.358
NN5	.633	.799	.655	.306	.368
NN6	.665	.813	.639	.343	.372

Table 12: Extension 3. Monolingual 6-level CEFR classification experiments with German (DE), Italian (IT), Czech (CZ), English (EN) and Spanish (ES) texts: neural network modifications. All scores are the mean of 10 runs. Cells in bold highlight the best performing model for each column.

4. Conclusion

Overall this has been a thought-provoking exercise, one which provides many ideas for future work and underlines the benefit of reproducibility in research. The sharing of open code repositories, datasets and publications, accompanied by thorough documentation and checkpointing are positive developments that can aid the field as a whole, with research viewed as a joint enterprise in which we build on each other's work. In this task we have been able to broadly imitate V&R's original findings, concluding that feature combination is the most robust approach to essay scoring, as confirmed in other work (Zechner et al., 2009; Yannakoudakis et al., 2018).

5. Acknowledgements

This paper reports on research supported by Cambridge Assessment, University of Cambridge. We thank Christopher Bryant, Diane Nicholls, Cristóbal Lozano and Iria del Río Gayo for assistance with corpora, the NVIDIA Corporation for the donation of the Titan X Pascal GPU used in this research, and the anonymous reviewers for very helpful feedback. We also thank the REPROLANG organisers, and Vajjala & Rama for volunteering their work for this shared task.

6. Bibliographical References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*.

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schone, K., Stindlov, B., and Vettori, C. (2014). The MERLIN Corpus: Learner language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

Bryant, C., Felice, M., Andersen, Ø., and Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Chollet, F. et al. (2015). Keras.

del Río, I. (2019). Linguistic features and proficiency classification in L2 Spanish and L2Portuguese. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL)*.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., and Smith, N. A. (2019). Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Hockly, N. (2018). Automated writing evaluation. *ELT Journal*, 73(1):82–88.

Kingma, D. P. and Ba, J. L. (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Kuhn, M., (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Lozano, C. and Mendikoetxea, A. (2013). Learner corpora and SLA: the design and collection of CEDEL2. In Ana Díaz-Negrillo, et al., editors, *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam: John Benjamins.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96:190–208.

Nivre, J., Abrams, M., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Basmov, V., Bauer, J., Bellato, S., Bengoetxea, K., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cetin, S., Chalub, F., Choi, J., Cho, Y., Chun, J., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Davidson, E., de Marneffe, M.-C., de Paiva, V., Diaz de Ilarraza, A., Dickerson, C., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eli, M., Elkahky, A., Ephrem, B., Erjavec, T., Etienne, A., Farkas, R., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerdes, K., Ginter, F., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saave-

dra, B., Grioni, M., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Habash, N., Hajič, J., Hajič jr., J., Hà Mỹ, L., Han, N.-R., Harris, K., Haug, D., Heinecke, J., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Hwang, J., Ikeda, T., Ion, R., Irimia, E., Ishola, Ọ., Jelínek, T., Johannsen, A., Jørgensen, F., Kaşıkara, H., Kaasen, A., Kahane, S., Kanayama, H., Kanerva, J., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Köhn, A., Kopacewicz, K., Kotsyba, N., Kovalevskaitė, J., Krek, S., Kwak, S., Laippala, V., Lambertino, L., Lam, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Lê H`ông, P., Lenci, A., Lertpradit, S., Leung, H., Li, C. Y., Li, J., Li, K., Lim, K., Li, Y., Ljubešić, N., Loginova, O., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsumoto, Y., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Misirpashayeva, M., Missilä, A., Mititelu, C., Miyao, Y., Montemagni, S., More, A., Moreno Romero, L., Mori, K. S., Morioka, T., Mori, S., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Murawaki, Y., Müürisep, K., Nainwani, P., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nguy˜ên Thị, L., Nguy˜ên Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nurmi, H., Ojala, S., Olúòkun, A., Omura, M., Osenova, P., Östling, R., Øvrelid, L., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perrier, G., Petrova, D., Petrov, S., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Rääbis, A., Rademaker, A., Ramasamy, L., Rama, T., Ramisch, C., Ravishankar, V., Real, L., Reddy, S., Rehm, G., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Romanenko, M., Rosa, R., Rovati, D., RoÈca, V., Rudina, O., Rueter, J., Sadde, S., Sagot, B., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Särg, D., Saulīte, B., Sawanakunanon, Y., Schneider, N., Schuster, S., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shohibussirri, M., Sichinava, D., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Smith, A., Soares-Bastos, I., Spadine, C., Stella, A., Straka, M., Strnadová, J., Suhr, A., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tanaka, T., Tellier, I., Thomas, G., Torga, L., Trosterud, T., Trukhina, A., Tsarfaty, R., Tyers, F., Uematsu, S., Urešová, Z., Uria, L., Uszkoreit, H., Vajjala, S., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yu, Z., Žabokrtský, Z., Zeldes, A., Zeman, D., Zhang, M., and Zhu, H. (2019). Universal dependencies 2.4.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R Core Team, (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Silge, J. and Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. *The Open Journal*, 1(3).

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

Vajjala, S. and Rama, T. (2018). Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., FranÃ§ois, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., MÃ$\frac{1}{4}$ller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

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

Yannakoudakis, H., Briscoe, T., and Medlock, B. (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*.

Yannakoudakis, H., Andersen, Ø. E., Geranpayeh, A., Briscoe, T., and Nicholls, D. (2018). Developing an automated writing placement system for esl learners. *Applied Measurement in Education*, 31:251–267.

Yoon, S.-Y., Cahill, A., Loukina, A., Zechner, K., Riordan, B., and Madnani, N. (2018). Atypical inputs in educational applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*.

Zechner, K., Higgins, D., Xi, X., and Williamson, D. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51:883–895.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Zesch, T. and Horbach, A. (2018). ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*.