Human and Automated CEFR-based Grading of Short Answers

Anaïs Tack[1,2,4a] **Thomas François**[1,4b] **Sophie Roekhaut**[3] **Cédrick Fairon**[1]
[1] CENTAL, UCL, Place Blaise Pascal 1, B-1348 Louvain-la-Neuve, Belgium
[2] ITEC, imec, KU Leuven Kulak, Etienne Sabbelaan 51, B-8500 Kortrijk, Belgium
[3] ALTISSIA International, Place de l'Université 16, B-1348 Louvain-la-Neuve, Belgium
[4] F.R.S.-FNRS [a] Research Fellow, [b] Postdoctoral Researcher
anais.tack@uclouvain.be thomas.francois@uclouvain.be
sroekhaut@altissia.com cedrick.fairon@uclouvain.be

Abstract

This paper is concerned with the task of automatically assessing the written proficiency level of non-native (L2) learners of English. Drawing on previous research on automated L2 writing assessment following the Common European Framework of Reference for Languages (CEFR), we investigate the possibilities and difficulties of deriving the CEFR level from short answers to open-ended questions, which has not yet been subjected to numerous studies up to date.

The object of our study is twofold: to examine the intricacy involved with both human and automated CEFR-based grading of short answers. On the one hand, we describe the compilation of a learner corpus of short answers graded with CEFR levels by three certified Cambridge examiners. We mainly observe that, although the shortness of the answers is reported as undermining a clear-cut evaluation, the length of the answer does not necessarily correlate with inter-examiner disagreement. On the other hand, we explore the development of a soft-voting system for the automated CEFR-based grading of short answers and draw tentative conclusions about its use in a computer-assisted testing (CAT) setting.

1 Introduction

The recent years have seen a growth of interest in Automated Writing Evaluation (AWE) for levelling non-native (L2) writing proficiency. Among the variety of assessment scales used, a number of studies have focused on levelling the writing proficiency following the Common European Framework of Reference (CEFR) (Council of Europe, 2001) through a combination of machine learning techniques and linguistic complexity features (Vajjala and Lõo, 2014; Volodina et al., 2016a; Pilán et al., 2016). One of the often cited benefits for using such assistive systems is that they could increase the effectiveness of large-scale testing procedures where a large panel of examiners are grading a mass of responses in a short period of time.

One application that comes to mind is the validation of the required writing skills of a large group of university students. In this scenario, implementing an expert-only testing procedure is costly for two reasons. On the one hand, a sufficiently large panel of experts evaluating the same text is needed to guarantee the validity of the evaluation. On the other hand, the large number of students who are participating in the programme makes the procedure even more time-consuming. Integrating an automated evaluator in the panel of examiners could therefore contribute to an increase in effectiveness of the evaluation procedure.

The present study takes part in a broader project which very aim is to research the possibility of using a computer-assisted setting for evaluating the level of written proficiency in English of non-native university students. The main idea of the project is to validate whether the students have the writing skills matching the CEFR descriptors of the proficiency level in which they have been placed. As a follow-up to a more general placement test, the students are queried to write an original short answer (ranging from 30 to 200 words) to an open-ended question, on the basis of

Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 169–179
Copenhagen, Denmark, September 8, 2017.

which a panel of examiners validate or adapt the CEFR level resulting from the global evaluation. In this context, we investigated the possibilities of partially automatising the short answer evaluation procedure, which is the general subject of the current paper.

The paper is structured as follows. After a brief review of the previous work on automated grading and the CEFR (Section 2), we will introduce our work on (i) the collection of a CEFR-graded learner corpus of short answers (Section 3) and (ii) the development of an automated grading system through ensemble learning (Section 4). In Section 5, we will compare the human and automated grading of short answers.

2 Background

2.1 Learner Writing Proficiency

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) is one of the most commonly used scale for measuring the proficiency of L2 users, dividing them into three groups: the basic (levels A1 and A2), independent (levels B1 and B2) and proficient users (levels C1 and C2). For various dimensions of proficiency (i.e. speaking, writing, etc.), it lists 'can-do' descriptors that can be used to assign a level to a learner. Although these criteria have been widely used in L2 teaching and research, studies have also stressed the need for more empirical research on how the different levels are linked with particular aspects of L2 proficiency (Hulstijn, 2007) (f.i. writing proficiency). Indeed, it is important to evaluate the learners' writing proficiency regardless of their overall L2 proficiency, since there is no proof that the overall CEFR level is necessarily transferred to the various dimensions composing L2 proficiency.

Over more than the past two decades, the most indispensable resource for gaining empirical insight into learner writing proficiency has been the learner corpus (Granger, 2009), as shown by the continuous emergence of written and spoken corpora available for numerous target languages and discourse types. For English in particular, the *International Corpus of Learner English* (ICLE) (Granger et al., 2009) and the *Cambridge Learner Corpus* (CLC) have been the go-to standard. Moreover, the recent years have also seen an increasing availability of learner corpora aligned with the CEFR (Boyd et al., 2014; Vajjala and Lõo, 2014; Volodina et al., 2016b), including the subsets of the CLC used by the English Profile (Salamoura and Saville, 2010).

Drawing on these developments, many studies have aimed at identifying the linguistic variables that are indicative (or *criterial*) of a particular L2 proficiency level (Díaz-Negrillo et al., 2013) and in particular those that are predictive of qualitative L2 writing (Crossley and McNamara, 2011; Vajjala, 2017). As a result, we know lexical complexity features, such as lexical diversity, word familiarity, meaningfulness and imageability, to be good predictors of L2 writing. As for the criterial features that apply specifically to the CEFR, important advances have been made in the context of the English Profile with the creation of a valuable inventory of structural patterns and learner errors (Hawkins and Buttery, 2010).

2.2 Automated Learner Writing Assessment

The advances made towards developing error-annotated and human-graded learner corpora (such as the CLC), as well as understanding the features underlying L2 proficiency, have subsequently furthered the development of systems for automated learner writing assessment, which include intelligent writing assistants (e.g. Andersen et al., 2013) and automated scoring systems (e.g. Yannakoudakis et al., 2011). In the case of automated scoring, two kinds of systems are generally distinguished, viz. automated essay grading (AEG) and automated short answer grading (ASAG)[1], depending on the length and type of texts as well as the kind of scoring method used. However, Burrows et al. (2015, p. 66) observe that '[t]he difference between these types can be fuzzy'.

Essay grading, on the one hand, is concerned with the evaluation of the quality or proficiency – often by means of a standard scale – of writings spanning several paragraphs or pages. In the context of L2 essay grading, a number of recently developed systems have achieved promising results with a wide range of complexity features and machine learning techniques for English, using the Cambridge English Scale (Yannakoudakis et al., 2011)[2] or the TOEFL scale (Vajjala,

[1] For a more extensive overview of the fields, see Shermis and Burstein (2013) and Burrows et al. (2015).

[2] The scores of the scale are aligned with the CEFR. The Cambridge English First (FCE) corpus used by Yannakoudakis et al. (2011) is known to correlate with a B1/B2 level.

2017). Other CEFR-based grading systems have been developed for German (Hancke and Meurers, 2013), Estonian (Vajjala and Lõo, 2014) and Swedish (Pilán et al., 2016).

The specificity of short answer grading, on the other hand, is the fact that it deals with 'objective questions' and length-restricted answers ranging 'between one phrase and one paragraph' (Burrows et al., 2015, p. 61). Its goal is to evaluate the learner responses as regards their correctness with respect to the initial question. The adequacy of the answer is thus compared to a model answer and graded either on a pass/fail basis or along a scale of correctness, using a range of concept and pattern matching techniques, alignment-based evaluation metrics (e.g. BLEU) or machine learning algorithms. In the context of L2 short answer grading, we mainly find systems developed for evaluating responses to reading comprehension questions, such as the CoMiC systems developed for English and German (Meurers et al., 2011).

The writing task underlying the current study can be situated between the extreme ends of essay and short answer grading presented above, aiming at assessing the CEFR level associated with short texts. On the one hand, the task is based on a series of questions (e.g."What is the best book you ever read?") which are more open-ended than the objective questions generally used in ASAG. On the other hand, contrary to essay writing, the task aims to assess writing proficiency based on a shorter display of writing, by adding more restrictions on the length of the answers (approximately one paragraph, or between 30 and 200 words).

3 A Corpus of Short Answers Graded per CEFR Level

In the context of the writing proficiency test we introduced in Section 1, we conducted a pilot study for collecting a CEFR-graded learner corpus that was representative of the task at hand.

3.1 Design

CEFR levels We defined a pool of questions (Table 1) that were used for querying the students' based on the result of the placement test. We will refer to the CEFR level defined by the placement test as the *initial proficiency level*. Note that although we defined the same set of questions for both the advanced C1 and C2 levels, hence grouping them in a common C level, we decided to keep the original six-level distinction in the graded learner corpus in order to ensure the reusability of the collected data.

level	min. words	topics
A1	30	(A) family (B) daily habits (C) hobbies
A2	60	(A) holiday memories (B) birthday invitation (C) lifetime goals
B1	80	(A) book reading (B) spending 1 million euros (C) blog writing
B2	100	(A) improve the environment (B) enjoy work or earn money (C) study abroad
C	150	(A) social networks (B) leading a healthy life (C) living in the public eye

Table 1: Question types per initial CEFR level

Question types The questions were all open-ended questions intended to trigger as wide a range of answers as possible. In order to vary the range of topics targeted by each question, we defined a pool of three different topics per initial level, which were construed bearing the CEFR guidelines in mind.

Length During the corpus collection procedure, each question trigger was followed by an indication of the minimal word limit required for submitting an answer. We mainly targeted answers ranging from 30 words at the A1 levels to 150 words at the C levels.

3.2 Collection

To collect a corpus of short answers, we conducted an on-line survey where each participant answered a question based on the CEFR level of the course in which they were enrolled. Each question was chosen in a circular fashion from the pool of questions previously defined. The minimal word limit of each answer was controlled so as to only allow a submission when the minimal word limit had been reached. After having submitted a valid answer, the students also responded to a short sociological questionnaire and were given the opportunity to enter in a raffle as a reward for participating.

We targeted learners coming from two different learning environments. On the one hand, we contacted participants who were enrolled in an e-learning platform. Their initial level was defined based on the CEFR level of the course they were

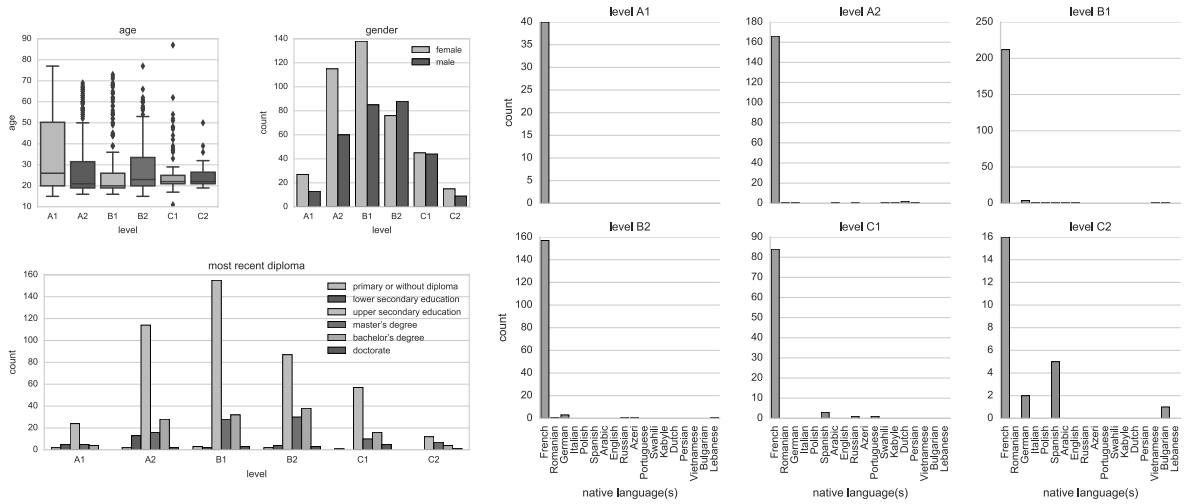

Figure 1: Sociological variables of the participants.

(a) original

	question			
initial level	A	B	C	all
A1	14	8	17	39
A2	65	50	60	175
B1	75	88	60	223
B2	42	66	54	162
C1	40	27	22	89
C2	11	8	5	24
all	247	247	218	712

(b) resampled

	question			
initial level	A	B	C	all
A1	14	7	17	38
A2	35	19	18	72
B1	19	19	18	56
B2	18	19	19	56
C1	20	17	16	53
C2	11	8	5	24
all	117	89	93	299

Table 2: The number of answers collected per initial level and per question type

following after having completed a general proficiency placement test with vocabulary, grammar, reading and listening exercises. On the other hand, we also contacted a group of participants enrolled in university-level English language classrooms targeting a particular CEFR level.

In all, we collected a total of 712 responses (Table 2). Based on the responses given in the questionnaire (Figure 1), we can observe that the majority of the participants were French-speaking learners of English studying at the bachelor's and master's level (all disciplines included).

3.3 Grading

The data used in this study contains a sample of the learner responses graded (i) according to their initial level and (ii) according to their *assessed proficiency level* as evaluated by majority voting[3] of a panel of three certified CEFR-expert Cambridge examiners. We will referred to them as examiners $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ respectively.

Before assessing the written proficiency level of the learner responses, we decided to keep the dataset as balanced as possible. Indeed, as we observe from the number of responses per initial CEFR level (Table 2b), there is an important difference between the number of texts collected for the beginner (A1) and advanced levels (C1 and C2) and the number of texts collected for the intermediate levels (A2, B1 and B2). We therefore performed a stratified random sampling of the data to balance the number of texts per initial level and question type, (i) by randomly selecting an equivalent number of texts per individual level ($\pm$ 25 texts) and (ii) by randomly supplying additional texts per grouped levels A, B, and C (60, 62, and 28 texts respectively) with the aim of having as similar a distribution per group as possible. As a result, a sample of 299 texts was used for the remainder of the study.

The panel of examiners used an on-line evaluation interface for grading. The examiners were prompted with the initial question and submitted

[3]In cases without agreement, the assessed level was derived by taking the nearest integer of the mean of the votes. These cases were then manually verified taking the hesitations observed in the examiners' comments into account.

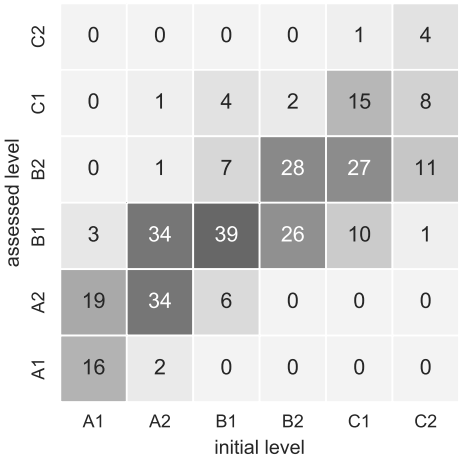

Figure 2: A comparison of the distribution between the initial and assessed CEFR levels.

answer, but did not receive any indication of the initial question level. They were then asked to evaluate the proficiency level of the answer based on the CEFR scale (ranging from A1 to C2), which they could turn back to and review as much as possible. The examiners could also flag the text as "Impossible to evaluate" in case they were, for whatever reason, unable to derive its proficiency level. Finally, they were also given the option of adding a comment to provide further details and justifications of their choice.

Figure 2 shows the number of texts distributed per initial and assessed levels. We observe that particularly the initial B1 answers were assessed as being indicative of a B1 written proficiency level (70%), whereas the initial C1 and C2 levels seem to have been relatively overestimated with only 28% and 17% of them assessed as having the C1 and C2 levels respectively.

4 A Soft-Voting CEFR-based Grader

In this section, we describe the general architecture of the system developed for the automated grading of the collected learner texts on a 5-point scale (A1, A2, B1, B2 and C). We decided to collapse the C1 and C2 levels into one C label for two reasons. First, although the small number of observatons that received an assessed C2 level (N=5) was considered insufficient, we did not want them to be discarded. Second, the original test setup on which this study was based did not aim to make a distinction between these assessed levels.

Features As preprocessing step to feature extraction, we used the Stanford CoreNLP suite (Manning et al., 2014) for performing tokenisation, lemmatisation, part-of-speech tagging, constituency and dependency parsing as well as coreferential resolution.

We defined a feature set of 18 different families, counting 695 individual feature configurations. We included a number of traditional readability features (François and Fairon, 2012; Vajjala and Lõo, 2014), including lexical features (word length, number of syllables, lexical frequency from SUBTLEX (Brysbaert and New, 2009), lexical likelihood based on Simple-Good Turing Smoothing (Gale and Sampson, 1995), lexical variation, lexical sophistication and part-of-speech tag ratios), syntactic features (sentence length and constituency tree structural patterns), WordNet-based (Fellbaum, 1998) and discursive features (synonyms, number of referential expressions and degree of content overlap), as well as a number of psycholinguistic norms (age of acquisition, imageability, familiarity, etc.) extracted from the MRC database (Wilson, 1988). We also included additional features for L2 complexity such as the types of (shallow) spelling and grammar errors as well as corpus-driven criterial features based on the English Profile (Hawkins and Buttery, 2010).

We should note that, contrary to previous work on Swedish L2 essay grading where the learner texts were normalised for error correction (Pilán et al., 2016), we only included error-based features without performing any error normalisation – apart from sentence segmentation errors and run-on sentences in particular – as preprocessing step to feature extraction. The error-based features were computed based on a noisy channel spelling correction (Kernighan et al., 1990) and handcrafted orthographic and syntactic (constituency- and dependency-based) patterns.

By means of a Spearman rank correlation test and a randomised logistic regression stability selection procedure on the entire sample, we found a set of 29 features to be of significant importance for the task at hand (Table 3 on the next page). This procedure was then reapplied on each of the model training folds before model fitting during nested cross-validation (cf. *infra*). Not surprisingly, we find that the most informative predictors of writing proficiency are the lexical ones

family	*feature*	μ_{A1}	μ_{A2}	μ_{B1}	μ_{B2}	μ_{C}	ρ
AoA	*Bristol lem*	−0.8	−0.8	−0.1	0.7	0.6	0.57***
	Kup lem	−1.1	−0.7	−0.2	0.7	0.8	0.62***
CEFR	*B1*	−0.7	−0.6	−0.2	0.7	0.7	0.53***
Disc	*global content overlap*	−1.0	−0.8	−0.2	0.7	0.9	0.73***
	global noun overlap	−0.9	−0.5	−0.2	0.6	0.7	0.56***
GrCorr	*missing subject*	−0.8	−0.7	−0.2	0.7	0.8	0.63***
LexFreq	*all mean*	−0.6	−0.4	0.3	0.0	−0.1	0.13*
	all mean $\mathcal{L}$	−0.8	−0.4	0.3	0.0	−0.1	0.18**
	grammatical 75P $\mathcal{L}$	−0.6	−0.6	0.3	0.2	0.0	0.22***
LexLike	*all* $\mathcal{L}$	0.0	0.2	0.3	−0.3	−0.6	−0.29***
LexVar	*adjective UberIndex* $\mathcal{L}$	−1.5	−0.9	−0.1	0.8	1.0	0.75***
	all UberIndex	−1.9	−0.9	0.0	0.8	1.0	0.78***
	modifier LogTTR	−2.2	−0.7	0.0	0.7	0.8	0.74***
	modifier SquaredTTR	−1.1	−0.8	−0.2	0.7	1.1	0.72***
	modifier UberIndex	−1.7	−0.9	−0.1	0.8	1.0	0.78***
	verb1 LogTTR	−2.1	−0.7	0.1	0.6	0.8	0.71***
	verb1 SquaredTTR	−1.4	−0.8	−0.1	0.6	1.2	0.71***
	verb1 UberIndex	−1.9	−0.9	0.0	0.7	1.0	0.77***
	verb2 UberIndex	−1.9	−0.9	0.0	0.7	1.0	0.78***
POSTag	*noun : grammatical*	1.3	0.2	−0.3	0.1	−0.1	−0.12*
	noun : preposition	1.8	0.3	−0.2	−0.2	−0.3	−0.26***
	determiner : noun	−0.8	−0.3	0.3	0.0	−0.1	0.14*
	grammatical : noun	−1.0	−0.1	0.3	−0.2	−0.1	0.12*
	lexical : grammatical	0.4	−0.1	−0.4	0.3	0.3	0.19***
	nominal : preposition	1.8	0.3	−0.2	−0.2	−0.3	−0.33***
	past part. : wh pron.	−0.4	−0.4	−0.2	0.2	0.9	0.43***
SentLen	*median*	−0.8	−0.5	−0.1	0.4	0.9	0.52***
WordLen	*mean*	−0.8	−0.6	−0.3	0.7	0.7	0.54***
	proportion 5 letters	−0.3	−0.4	−0.3	0.5	0.6	0.40***

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3: Features selected through a Spearman rank correlation test and a stability selection procedure. All features are standardised to a Gaussian scale and their average is reported per assessed level. Lemma-based indices are marked with $\mathcal{L}$.

and in particular lexical diversity features, which is in line with previous studies (Crossley and McNamara, 2011; Hancke and Meurers, 2013; Vajjala and Lõo, 2014; Pilán et al., 2016). Furthermore, we find that the sentence length and word length, as well as the average age of acquisition of the words used by the learners display a strong positive correlation with the assessed CEFR level. We also observe that the frequent use of B1 criterial feature patterns are indicative of the learner writings from the B2 levels onwards. One surprising observation, however, can be drawn from the apparent positive correlation of lexical frequencies. This could be explained by the fact that beginners (A1 and A2) quite commonly display a use of L1 interference in their texts – as can be seen in the use of the French *caractères* ("characters") in Figure 3 – which are subsequently tagged as foreign (infrequent) words.

Model Figure 3 illustrates the model architecture used for the automated CEFR-based grading of a short answer (initial A1 level and assessed A2 level). Our system used the *Scikit-learn* library (Pedregosa et al., 2011) for training an ensemble learning approach via a soft-voting classifier integrating a panel of five traditional models: a

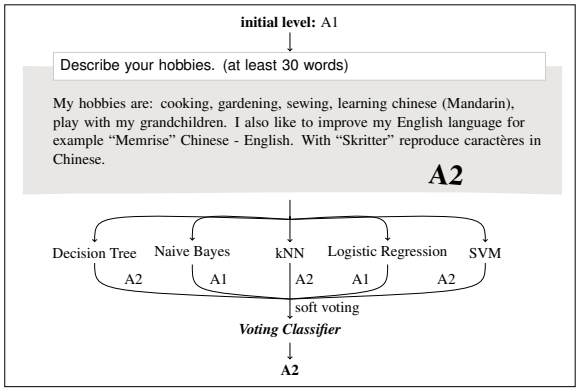

Figure 3: Example of the ensemble learning approach to the automated scoring of short answers.

Gaussian Naive Bayes classifier, a CART Decision Tree, a kNN classifier, a one-vs.-rest (OvR) Logistic Regressor and a OvR polynomial LibSVM Support Vector Machine. The system was developed via a nested cross-validation procedure and its hyperparameters were optimised via a two-stage model selection procedure on the training fold, performing a 10-fold grid search on the individual models first and then on the ensemble method.

5 Results

5.1 Expert Grading

Reliability To measure the inter-rater reliability of the assessed proficiency levels, we use Krippendorff's α with interval metric.[4] Krippendorff's statistic suggests a strong agreement (α = .81; $.80 < \alpha < .90$) between our examiners, which ensures the reliability of the CEFR-labelled corpus. The strong agreement is also reflected by the fact that all three examiners gave the same proficiency level (i.e. perfect agreement) to 44% of the texts and that for 50% of the texts at least one pair of examiners gave the same proficiency level (Table 4 on the following page). Only for 6% of the texts do they seem to not agree at all. Furthermore, the high agreement score for the interval metric indicates that, in the cases where our examiners did not perfectly agree on the target proficiency level, the distance between the given levels was not large.

[4] Although we could have used Fleiss' κ for comparing the CEFR categories, we decided to use the former because it enables us not only to compare the assessed levels on a *scale*, but also to properly deal with missing values (cf. *supra*, "Impossible to evaluate") instead of discarding them as would Fleiss' κ. We should also note that only two missing values were observed for one examiner.

		initial level						
agreement	**%**	A1	A2	B1	B2	C1	C2	all
perfect	**43.8**	19	29	28	25	22	8	131
partial	**50.2**	17	41	27	27	26	12	150
no	**6.0**	2	2	1	4	5	4	18
		38	72	56	56	53	24	299

Table 4: Inter-examiner agreement scores.

Put differently, the examiners tended to disagree more on adjacent proficiency levels (such as B1 and B2) than between levels at the extreme ends of the scale (such as A1 and C2).

Grading difficulty and disagreement Although we observe a strong human-human agreement (HHA) between the three examiners, we also noted their comments with respect to the difficulty of the task of assigning a CEFR level to a very short text. Indeed, for the A1 and A2 levels (counting minimally 30 and 60 words respectively) they frequently reported needing more context to correctly assess the proficiency level, in particular for those texts that displayed "no errors" and were written in "mainly accurate English". This is illustrated in the few texts where the initial A2 level seemed to have been underestimated in favour of a B2 or C1 level. We were therefore interested in examining what characteristics define the texts that were difficult to grade.

We measured the difficulty of grading a text on the basis of the per-item observed disagreement $D^{\alpha}_{o_i}$ on the label x given by coder c on item i (5.1.1). We derived this measure by decomposing Krippendorff's formula for the observed disagreement D^{α}_{o} (Artstein and Poesio, 2008, pp. 564-7) , which amounts to two times the per-item empirical variance s^2_i.

$$D^{\alpha}_{o_i} = \frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{m=1}^{\mathbf{c}} \sum_{n=1}^{\mathbf{c}} \delta_{interval}(x_{ic_m}, x_{ic_n}) = 2s^2_i \quad (5.1.1)$$

Interestingly, we find that, although the examiners reported having difficulties evaluating the CEFR level of the shortest answers, the length of the answer was not significantly correlated with the amount of per-item disagreement (Pearson's r = .04; p = .455) In fact, Pearson's r as well as the number of agreeing or disagreeing cases per initial level (Table 4) show that the annotators tended to disagree more on the longer ones, as most of the texts where no agreement was observed were concerned with the initial level ranging from the C1 to C2 levels (min. 150 words).

	acc.	adj. acc.	$F_{1\ macro}$	RMSE	α
Soft Voting	.530 ± .115	.978 ± .040	.495 ± .142	.721 ± .124	.757
Decision Tree	.504 ± .103	.946 ± .053	.438 ± .126	.802 ± .125	.713
kNN	.500 ± .084	.972 ± .047	.403 ± .107	.758 ± .104	.690
Logistic Regression	.462 ± .138	.958 ± .044	.422 ± .142	.807 ± .120	.717
Naive Bayes	.486 ± .117	.952 ± .047	.487 ± .132	.802 ± .173	.742
SVM	.496 ± .129	.977 ± .041	.451 ± .135	.750 ± .164	.737
baseline (prior)	.378 ± .013	.824 ± .017	.110 ± .003	1.072 ± 0.031	–.010
baseline (stratified)	.282 ± .041	.606 ± .024	.201 ± .046	1.524 ± 0.057	–.161
baseline (random)	.191 ± .015	.484 ± .027	.163 ± .020	1.930 ± 0.058	–.131

Table 5: Performance of the system compared to a set of baselines on 10-fold cross-validation.

Multiple semipartial Spearman correlation tests were then carried out as a way of investigating which complexity features might be characteristic of the per-item grading difficulty D (as previously defined by $D^{\alpha}_{o_i}$), while controlling for text length L (in number of words). We observed a number of significant effects with a small set of lexical features, such as the overall lexical diversity ($r_{D(X.L)}$ = .142; $p < .05$), the variation in use of modifiers ($r_{D(X.L)}$ = .183; $p < .01$) and adjectives ($r_{D(X.L)}$ = .182; $p < .01$), as well as the average lexical likelihood ($r_{D(X.L)}$ = -.151; $p < .01$).

5.2 Automated Scoring

Performance The voting classifier described in Section 4 achieves a good human-system agreement[5] (HSA) (α = .76, .67 $< \alpha <$ .80) with respect to the answers' assessed CEFR level obtained by majority voting (Table 5). Although our system did not surpass the strong HHA ceiling we observed earlier (which amounts to α = .82 when using a 5-point scale), the HSA of our ensemble method still outperformed the HSA of its individual classifiers. What is more, in cases where there is a human-system *dis*agreement, we find that the output mainly differs by an adjacent level, leading to an adjacent accuracy of 98% and an RMSE of .7 on a scale of five (A1, A2, B1, B2 and C).

A Friedman test with a post-hoc Holm correction was then carried out as a means of comparing the performance of our voting classifier with respect to the models it is composed of as well as to the most performant baseline. Our system achieved a significant gain in perform-

[5]The α values were computed by aggregating the predictions on all 10 test folds and by comparing them to the true labels obtained after majority voting.

ance (RMSE) with respect to a prior baseline[6] (F_F = 4.865, $p < .01$, k = 6, α = .05). Although the test did not reveal any other significant gain beyond the one observed over the baseline, we find that the system's performance is comparable to previous work for Swedish CEFR-based essay grading where an F_1 of .438 is attained on original (not error-normalised) learner texts (Pilán et al., 2016).

Nevertheless, we do observe a difficulty of attaining a perfect HSA with the system's accuracy peaking at 53%. Even though this result may seem inferior to previous CEFR-based essay grading systems (Vajjala and Lõo, 2014; Volodina et al., 2016a), we should note that the data sets used in these studies were slightly different from our data set and mainly included longer texts graded on either a 4-point scale (A2, B1, B2 and C1) (Vajjala and Lõo, 2014; Pilán et al., 2016) or a 5-point scale (A1, A2, B1, B2 and C1) (Volodina et al., 2016a). Furthermore, we should also note the parallel between the difficulty of deriving the exact CEFR level from the answers and the difficulty experienced by our human raters of achieving a perfect agreement (43.8%) (see Table 4 on the previous page).

However, linking the length of the answers with the per-item human-system disagreement (cf. formula 5.1.1 on the preceding page), we observe yet again a non-significant correlation between both (Pearson's r = .07; p = .22). Thus, it seems that, similarly to the expert graders, our system did not particularly have a difficulty grading the shortest answers. In addition, the system did not have any particular difficulties in correctly predicting the lowest CEFR levels either (Figure 4).

For enhancing our automated CEFR-based scoring of short answers, the two following options could be explored. First, we could explore the possibility of pinpointing and resolving the difficulties involved with attaining a high HHA and HSA using more high-level learner features indicative of the advanced CEFR levels. Second, similarly to Pilán et al. (2016), we could examine the effect of applying (automatic) learner error normalisation on the system's performance, provided that the applied normalisation technique is accurate enough for correctly dealing with learner language. However, we should note that the absence of error normalisation did not seem to have impacted the grading accuracy of the A1 and A2 levels (see Figure 4) where the presence of errors is known to be particularly prevalent (Hulstijn, 2007).

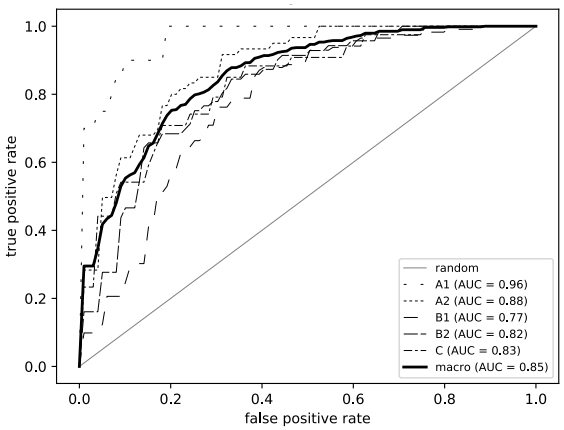

Figure 4: Receiver operating characteristic for the voting classifier.

Computer-assisted testing simulation To explore the possibility of using the system in a computer-assisted setting, we simulated the reliability of replacing one of the three examiners by our system. Table 6 on the next page shows the performance scores and reliability coefficients of all possible configurations using a panel of three examiners where we replaced one examiner with a soft-voting short answer grader which was retrained on the examiner's evaluations.

The good agreement scores for Krippendorff's α[7] enable us to draw tentative conclusions as to the possibility of using the system in a panel of examiners. Replacing one examiner by our system could therefore be possible, but the simulation did not reveal any configurations (α = .75 on average) that topped the strong agreement of having three human examiners (α = .82 when using a 5-point scale).

Interestingly, we also observed that the best results were achieved when training the system on examiner $\mathcal{Z}$, who could be typed as being neither too "demanding" nor too "lenient" compared to the other examiners (Table 7 on the following page). To perform this comparison, we ranked the exam-

[6]The prior baseline predicts the class with the maximum prior probability, which is the B1 level (113 out of 299 observations; Figure 2 on page 5). The stratified baseline gives random predictions based on the class distribution as observed on the training set.

[7]As before, the α values were computed by aggregating the predictions on all 10 test folds, but now comparing them to the individual labels given by the two other examiners.

trained on	$\mathcal{X}$	$\mathcal{Y}$	$\mathcal{Z}$	avg.
acc.	.51	.37	.56	.48
adj. acc.	.97	.84	.99	.93
F_1	.48	.33	.52	.44
RMSE	.77	1.05	.67	.83
agreeing with (%)	$\begin{pmatrix}\mathcal{Y}\\ \mathcal{Z}\end{pmatrix}$	$\begin{pmatrix}\mathcal{X}\\ \mathcal{Z}\end{pmatrix}$	$\begin{pmatrix}\mathcal{X}\\ \mathcal{Y}\end{pmatrix}$	
perfect	33.78	31.42	34.11	33.10
partial	60.54	61.49	60.87	60.97
HHA	47.51	53.85	42.86	48.07
HSA	52.49	46.15	57.14	51.93
no	5.69	7.09	5.02	5.93
Krippendorff's α	.76	.74	.75	.75

Table 6: Reliability of replacing one examiner with the system. The partial agreement scores are further broken down into percentages per human-human agreement (HHA) and human-system agreement (HSA).

examiner	average rank
$\mathcal{X}$	1.81
$\mathcal{Z}$	1.96
$\mathcal{Y}$	2.23

rank 1: gave the lowest level ("demanding")
rank 2: gave neither one, or all scores tied
rank 3: gave the highest level ("lenient")

Table 7: Comparative ranking of the examiners according to their evaluations.

iners according to their evaluation for each text and used 'average' ranking for tied labels (i.e. for perfect or pairwise agreement).

Moreover, it appears that training the system on examiner $\mathcal{Z}$ even bettered the performance of the voting classifier trained on the data labelled by the entire panel of examiners (see Table 5 on page 7). However, for future endeavours, we argue that we should not solely rely on such idiosyncratic evaluations merely because they enhance a system's performance – however appealing that may be – and that we should therefore continue to use the labelled data obtained via majority voting.

6 Conclusion

In this paper, we compared human and automated scoring of short answers using the Common European Framework of Reference (CEFR). For this purpose, we compiled a learner corpus of short answers, written by non-native learners of English and evaluated by a panel of three certified Cambridge examiners, and which will be made available for non-commercial use. Furthermore, we developed a soft-voting CEFR-based classifier based on a set of traditional linguistic complexity features as well as some more specific L2 complexity features.

We obtained positive results, although more work is needed to further examine the difficulties involved with predicting the CEFR written proficiency level from short texts. Indeed, our findings showed that the shortness of the answer is not necessarily correlated with the amount of human-human or human-system disagreement. Yet, our results were inconclusive as to what indicators could explain the difficulty of grading a short answer according to the CEFR scale.

We therefore propose to continue investigating the influence of more advanced L2 complexity features on explaining the intricacy involved with the current task. As regards our system, we propose to examine the impact of error normalisation on its performance. Finally, other aspects associated with the task still remain to be considered as well, such as the replication of the results to other target L2 languages as well as to groups with more diverse L1 backgrounds.

Acknowledgements

This work received the financial support of the ALTISSIA e-learning company. We would also like to thank the *Centres de Langues* (CLL) and the *Institut des langues vivantes* (ILV) of Louvain-la-Neuve, Belgium for their indispensable contributions to the corpus collection and annotation.

References

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Adriane Boyd, Jirka Hana, Lionel Nicolas, Detmar Meurers, Katrin Wisniewski, Andrea Abel, Karin Schöne, Barbora Štindlová, and Chiara Vettori. 2014. The MERLIN corpus: Learner Language and the CEFR. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1281–88, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Council of Europe. 2001. *Common European Framework of Reference for Languages*. Cambridge University Press, Cambridge, UK.

Scott A. Crossley and Danielle S. McNamara. 2011. Understanding expert ratings of essay quality: CohMetrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2-3):170–191.

Ana Díaz-Negrillo, Nicolas Ballier, and Paul Thompson, editors. 2013. *Automatic Treatment and Analysis of Learner Corpus Data*, volume 59 of *Studies in Corpus Linguistics*. John Benjamins Publishing Company, Amsterdam.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Thomas François and Cédrick Fairon. 2012. An "AI Readability" Formula for French As a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, Stroudsburg, PA, USA. Association for Computational Linguistics.

William A. Gale and Geoffrey Sampson. 1995. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237.

Sylviane Granger. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Karin Aijmer, editor, *Studies in Corpus Linguistics*, volume 33, pages 13–332. John Benjamins Publishing Company, Amsterdam.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English. Version 2. Handbook and CD-ROM*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research 2013, Book of Abstracts*, pages 54–56, Bergen, Norway.

John A. Hawkins and Paula Buttery. 2010. Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, 1(01):1–23.

Jan H. Hulstijn. 2007. The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency. *The Modern Language Journal*, 91(4):663–667.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A Spelling Correction Program Based on a Noisy Channel Model. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, pages 205–210, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David Mcclosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland USA. Association for Computational Linguistics.

Detmar Meurers, Ramon Ziai, Niels Ott, and Janina Kopp. 2011. Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 2101–2111, Osaka, Japan. Association for Computational Linguistics.

Angeliki Salamoura and Nick Saville. 2010. Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In Inge Bartning, Maisa Martin, and Ineke Vedder, editors, *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, number 1 in Eurosla Monographs Series, pages 101–132. European Second Language Association.

Mark D. Shermis and Jill Burstein, editors. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, New York, NY.

Sowmya Vajjala. 2017. Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features. *International Journal of Artificial Intelligence in Education*, pages 1–27.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, NEALT Proceedings Series 22, pages 113–127.

Elena Volodina, Ildikó Pilán, and David Alfter. 2016a. Classification of Swedish learner essays by CEFR levels. In *CALL communities and culture – short papers from EUROCALL 2016*, pages 456–461, Limassol, Cyprus. European Association for Computer Assisted Language Learning.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016b. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Wilson. 1988. MRC Psycholinguistic Database: Machine Usable Dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.