

Developing an Automated Writing Placement System for ESL Learners

Helen Yannakoudakis^{1,2}, Øistein E. Andersen^{1,2}, Ardeshir Geranpayeh³, Ted Briscoe^{1,2,4},

Diane Nicholls⁴

¹ALTA Institute, department of Computer Science and Technology, University of
Cambridge

²iLexIR Ltd

³Cambridge English Language Assessment, University of Cambridge

⁴English Language iTutoring Ltd

Authors' Note

We are grateful to Gad Lim and Coreen Docherty, as well as the reviewers for their valuable contributions at various stages. We are also grateful to Cambridge Assessment for funding the ALTA Institute research. Further ALTA publications can be accessed here:

<http://www.wiki.cl.cam.ac.uk/rowiki/NaturalLanguage/ALTA>

Abstract

There are quite a few challenges in the development of an automated writing placement model for non-native English learners, among them the fact that exams that encompass the full range of language proficiency exhibited at different stages of learning are hard to design. However, acquisition of appropriate training data that are relevant to the task at hand is essential in the development of the model. Using the Cambridge Learner Corpus writing scores, which have been subsequently benchmarked to CEFR levels, we conceptualize the task as a supervised machine learning problem, and primarily focus on developing a generic writing model. Such an approach facilitates the modeling of truly consistent, internal marking criteria regardless of the prompt delivered, which has the additional advantage of requiring smaller dataset sizes and not necessarily requiring re-training or tuning for new tasks. The system is developed to predict someone's proficiency level on the CEFR scale, which allows learners to point to a specific standard of achievement. We furthermore integrate our model into Cambridge English Write & ImproveTM – a freely available, cloud-based tool that automatically provides diagnostic feedback to non-native English-language learners at different levels of granularity – and examine its use.

Developing an Automated Writing Placement System for ESL Learners

Introduction

The task of Automated Essay Scoring (AES) focuses on automatically assessing someone's writing competence and providing immediate feedback.¹ Learning to write a foreign language well requires a considerable amount of practice and appropriate feedback. AES systems provide a learning environment in which foreign language learners can practice their writing in an interactive manner and receive feedback (typically via multimedia presentations, such as text, graphics and sound) within a self-access and/or teacher-directed assessment context. Such systems can be a valid and useful supplement to writing instruction and can facilitate the language-learning process and promote learners' writing development (e.g., they reinforce the material that has been taught). Provision of feedback is a fundamental part of second language writing instruction, while prompt, accurate feedback is a vital component in the learning process that has been shown to increase learning efficiency (e.g., [Wang, Shang, and Briody, 2013](#)). Automating the free-text marking process² furthermore contributes to the reliability and consistency in the application of the assessment criteria, as well as to the reduced workload and costs inherent in the process of manual scoring.

To date, a number of AES systems for English-language learners have been developed as commercial products and/or deployed in classrooms. Examples of the earliest ones include e-Rater ([Attali & Burstein, 2006; Burstein, 2003](#)), an automated essay scoring system developed by Educational Testing Service (ETS), the first one to be deployed for operational scoring of high-stakes assessments in 1999; Criterion ([Burstein, Chodorow, & Leacock, 2003](#),

¹ Herein, we will use the terms 'essay' and 'text' interchangeably throughout.

² We note that 'marking' is another term used to refer to 'scoring'.

[2004](#)), a web-based writing assessment tool deployed in classrooms; and ESL Assistant ([Gamon et al., 2009](#)) for correction suggestions.

A system that automatically predicts someone's proficiency level, in conjunction with automated diagnostic feedback, has the advantage of allowing learners to reflect on their errors and simultaneously track their effect on their overall performance, thus facilitating self-assessment, self-tutoring and self-improvement through reflective use of feedback. Additionally, an attainment-level predictor³ can help to gauge a learner's readiness for proficiency-level certification / high-stakes assessment.

In this article, we present in detail the design, implementation and evaluation of an automated writing placement system for English-language learners that predicts someone's proficiency level based on benchmarks of language proficiency. Acquisition of appropriate data relevant to the task at hand is essential in the development of such a model. For our task, we require a corpus of prompts and corresponding answers representative of the full range of writing ability exhibited at different stages of learning, annotated on a common scale that spans the full ability range.

We devise a corpus based on the Cambridge English Language Assessment exams and their Common European Framework of Reference for Languages (CEFR) levels. Using the Cambridge Learner Corpus writing scores, which have been subsequently benchmarked to CEFR levels, we conceptualize the task as a *supervised machine learning* problem, and focus on developing a writing placement model that predicts someone's proficiency level by returning a CEFR level as the score, and therefore allows learners to point to an established standard of achievement. We furthermore integrate our model into Cambridge English Write & ImproveTM – a freely-available, cloud-based tool⁴ that automatically provides diagnostic feedback to non-native English-language learners at different levels of granularity ([Andersen, Yannakoudakis, Barker, & Parish, 2013](#)).

³ 'Attainment level' and 'proficiency level' will be used interchangeably throughout the article.

⁴ <https://writeandimprove.com>

The CEFR, the Cambridge English Exams and the Cambridge Learner Corpus

The Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) was developed in the late 1990s to provide ‘a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe’ ([Council of Europe, 2001](#)). The CEFR describes language ability on a scale consisting of six levels from A1 for beginners up to C2 for those who have mastered a language. Its use among language assessment users over the years, however, has had more far-reaching influences than was originally intended. The CEFR today is a well-known international standard for describing language ability, and is used around the world to describe learners’ language skills, with many language examination boards linking their exam results to the CEFR scale.

Cambridge English exams have a special link to the CEFR, and their embodiment of the scale goes back to the origins of the CEFR in the early 1980s. The Cambridge English First (previously First Certificate in English, FCE) exam was used as the corner stone of describing the B2 level in the CEFR development. The remaining Cambridge English Main Suite examinations were either developed to measure a language level corresponding to what was later defined as one of the CEFR levels or subsequently aligned to the scale after the publication of the CEFR in 2001: CPE (C2), CAE (C1), PET (B1) and KET (A2). Cambridge has a long history of collaboration with the Language Policy Unit of the Council of Europe in preparing various manuals for language test development linked to the CEFR. For more detailed discussion of the Cambridge English exams’ link and alignment to the CEFR, the reader is referred to the special issue of Research Notes (Issue 37, August 2009).

Cambridge English in collaboration with Cambridge University Press established the Cambridge Learner Corpus (CLC) in 1993. The CLC was developed to inform both test development and publishing activities focusing on learners. It contains examinee responses

from the written component of Cambridge's English examinations, initially at intermediate (B) and higher proficiency (C) CEFR levels. Scripts are keyed in exactly as examinees have written them (i.e., with errors intact) to form a computer-readable version. These scripts are accompanied by comprehensive candidate information and score data so that the corpus can be searched by variables such as age, gender, native language (L1) or score/grade achieved on the written component. A unique feature of the CLC is the error-tagging system (see [Nicholls, 2003](#)). The corpus is searchable through proprietary software either for particular types of error, or lexically through a concordancer, collocation search or frequency word lists. The CLC, which grows by 2-3 million words annually, is one of the few learner corpora to have been compiled using only examination scripts selected according to examinees' L1 and directly linked to the CEFR levels. The CLC is increasingly being used for research on providing automatic feedback to language learners.

Previous work

There is a large body of literature regarding automated assessment systems for non-native English text in general, designed to assess various aspects of writing – such as linguistic accuracy, content relevance, coherence and so on – and/or provide formative feedback and facilitate writing instruction ([Andersen et al., 2013](#); [Attali & Burstein, 2006](#); [Azab, Hokamp, & Mihalcea, 2015](#); [Briscoe, Medlock, & Andersen, 2010](#); [Burstein, Chodorow, & Leacock, 2003](#); [Burstein et al., 2004](#); [Burstein, Marcu, & Knight, 2003](#); [Chang & Chang, 2015](#); [Dickinson, Kübler, & Meyer, 2012](#); [Higgins, Burstein, & Attali, 2006](#); [Kakkonen, Myller, & Sutinen, 2004](#); [Landauer, Laham, & Foltz, 2003](#); [Macdonald, Frase, Gingrich, & Keenan, 1982](#); [Page, 1968](#); [Soyer, Topic, Stenetorp, & Aizawa, 2015](#)). Existing systems, overviews of which have been published in various studies ([Dikli, 2006](#); [Shermis & Hammer, 2012](#); [Valenti, Neri, & Cucchiarelli, 2003](#); [Williamson, 2009](#)), involve a wide range of techniques from dimensionality reduction over matrices of terms through to extraction of linguistically deeper features such as types of syntactic constructions and specific error types (e.g., non-agreement of subject and main verb). However, few attempts have been made to automatically predict the proficiency level of learners' writing on the CEFR scale.

[Dickinson et al. \(2012\)](#) describe a system for predicting the level of Hebrew-language learners based on placement exam exercises in which the learners have to order words into a grammatical sentence. [Hancke and Meurers \(2013\)](#) investigate CEFR classification of short essays in German and focus on identifying aspects of learner language that characterize the different levels. [Vajjala and Lõo \(2014\)](#) develop a model for predicting a learner's CEFR language proficiency in Estonian, while the systems by [Pilán and Volodina \(2016\)](#), [Pilán, Volodina, and Zesch \(2016\)](#), [Volodina, Pilán, and Alfter \(2016\)](#) are based on a corpus of second language (L2) Swedish learner texts that have been manually linked to the CEFR levels.

[Alexopoulou, Yannakoudakis, and Salamoura \(2013\)](#) and [Yannakoudakis, Briscoe, and Alexopoulou \(2012\)](#) developed a tool that visualizes the internal ‘marking criteria’ of an automated model that predicts whether a learner of English has shown evidence of having attained the upper-intermediate (B2) CEFR level. The authors’ aim is to facilitate the linguistic interpretation of highly discriminative linguistic features and delimit the boundaries of the B2 level, with an overall goal of facilitating the creation of a set of ‘descriptors’ (i.e., features) that are linked to the different CEFR levels. [Andersen et al. \(2013\)](#) describe a self-assessment and tutoring system that automatically provides feedback to non-native speakers of English at various levels of granularity within a free-text response to a prompt. One of their models predicts the overall linguistic quality of a text on a scoring scale that is benchmarked at the upper-intermediate (B2) CEFR proficiency level. The scoring scale is calibrated so that performances beneath and beyond the B2 level can be roughly estimated (to the extent that the linguistic constructions elicited by the prompts can reflect the varying proficiency levels), and can therefore roughly estimate someone’s proficiency level as being far below, just below, around or above an upper intermediate level.

Method

Our goal is to automatically predict the proficiency level of learners’ writing using the CEFR scale as the basis for the scoring algorithm. In the following sections, we describe our approach to the task, details of the system developed, and the evaluation methodology adopted.

Instruments

Based on the Cambridge English exams, we devised a representative corpus consisting of texts and their CEFR levels. More specifically, we employed a range of prompts, benchmarked at different proficiency levels and covering the full spectrum of language proficiency. We used the responses to these tasks to develop an attainment level predictor for writing. However, different prompts are designed to assess attainment of different levels and

are therefore calibrated to different underlying scoring scales and rubrics. As a result, establishing relationships across levels and determining how performance on one exam relates to performance on another is a complex and challenging endeavor; however, a reliable mapping of the scoring scales to a common underlying standard is fundamental for the development of a proficiency-level predictor. Cambridge English exams have been calibrated with explicit links across levels that make it possible to know how a performance would have been evaluated at a different level. We note, however, that this is bounded by the maximum elicited performance at a level: for example, the CEFR B1 scale would be capped at CEFR B2 (for more details on this, see, e.g., [Lim, 2012](#)).

Our model's scoring scale stretches across the full CEFR proficiency continuum and is based on the Cambridge English (CE) Main Suite exams that have been benchmarked at the different CEFR levels: Cambridge English Key (previously Key English Test, KET) for A2, Cambridge English Preliminary (previously Preliminary English Test, PET) for B1, Cambridge English First (FCE) for B2, Cambridge English Advanced (previously Certificate in Advanced English, CAE) for C1 and Cambridge English Proficiency (previously Certificate of Proficiency in English, CPE) for C2.⁵ Each of these exams is marked on a scale from 0 to 5 (where 3 is the lowest pass score), and has been calibrated with explicit links across the levels. As detailed in Table 1, our model predicts a score on a scale from 0 to 13. The passing scores on each CE exam overlap across the adjacent proficiency levels, and there is an unambiguous and conceptually straightforward mapping between the 0–13 scoring scale and the CEFR levels (where levels A1– and C2+ represent performances beneath and beyond the A1 and C2 level respectively).

Another challenge in the development of the model is that of task biases: that is, different prompts have different topics and registers. Correlations between topic, genre and level can confound the model, as it may end up predicting proficiency levels primarily based on topics and/or register⁶ (the latter going beyond the use of topic words) rather than writing competence.⁷ One way to ameliorate this problem is to utilize prompts of the same and/or similar topic and register and their responses to eliminate such biases in the data to

⁵ <http://www.cambridgeenglish.org/exams/general-and-business-english/>

⁶ 'Register' refers to the level of language formality.

⁷ We note that similar issues have been observed in the task of automatic Native Language (L1) Identification, where topic biases may lead the model into learning how to discriminate between topics rather than L1s (for more details see, e.g., Brooke and Hirst, 2012).

the extent possible. An example set of such prompts is presented in Table [2](#), where we can see that the exact same tasks stretch across two adjacent levels (i.e., A1 / A2, B1 / B2, and C1 / C2), with similar ones used for levels further apart. As new prompts are added to the system, and maintenance of a balanced dataset becomes more expensive, (re-)tuning the model across different tasks should allow for a more robust approach.

Insert Table [1](#) here

Insert Table [2](#) here

Corpus Our final corpus (i.e., collection of written texts) consists of 2,312 texts (a text refers to an individual’s written response) and their CEFR scores (assigned by a human expert)⁸ written by 2,312 distinct English learners from a number of different native languages. The resulting set is representative of the full set of responses and of all CEFR levels from A1– to C2+. A large number of learners are teenagers at the B level. On average, there are around 200 words per text.⁹ We randomly select 260 of these texts to test the performance of our final model (i.e., we do not use this subset during model development). The remaining texts are used to develop our model (see details in the following section). The CEFR level distribution of the data can be seen in Figure [1](#).

Procedure

We approach the task of automatically predicting someone’s CEFR proficiency level as a Supervised Machine Learning (SML) problem. SML refers to a class of Artificial Intelligence algorithms that enable machines to *learn* or otherwise acquire knowledge from data that have been annotated with what we are trying to predict (referred to as *training data*), and subsequently make inferences about new, unseen and unannotated data. In our case, the training data consist of the set of texts written in response to prompts benchmarked at the various CEFR levels, and annotated by human experts with a CEFR score in the 0–13 scale. The training set is used to generate a set of *features* per text (along with the text’s *target* variable, that is, the CEFR score that we are trying to predict). Features refer to measurable properties of text that can be automatically calculated and used to “explain” the

⁸ Texts are scored directly on the full CEFR scoring scale.

⁹ We note that prompts (and responses) are administered digitally, and responses are not timed.

mapping between a text and a CEFR score (see the following section for more details on the features we employ).

The learning algorithm then tries to learn the function, called the *hypothesis*, that best describes the relationship between the features and the target variables. This is otherwise known as an input–output mapping function that can map input features to a target variable (i.e., map input features to a CEFR score). Once the model is trained (i.e., the hypothesis function is learned), we can apply it to unseen *test* examples (i.e., ones whose target variables are unknown to the model), and measure its performance by comparing the predicted scores with those assigned by human experts on the same data.

In its basic form, a SML algorithm can perform *classification* by learning a linear threshold function that can discriminate data points of two categories (e.g., pass from fail essays). Here, we use the training data to learn a *ranking* function: a ranking SML model seeks to identify an optimal ordering of the data directly, and outputs a score for each data point / essay, from which a global ordering of the data is constructed ([Joachims, 2002](#)). The intuition behind the use of a ranking model in AES is that this class of SML models can directly exploit the partial ranking implicit in any discrete scoring scale.

The principal advantage of applying ranking to the AES task is that it allows us to explicitly model the relationships between texts (i.e., that some texts are “better” than others) and the gradation of text quality, without necessarily having to specify numerical scores or introduce a pass/fail boundary. This is a more appropriate model for scoring on a discrete ordinal or symbolic scale (such as CEFR levels) and one which exploits the labeling information in the training data efficiently and directly ([Briscoe, Medlock, & Andersen, 2010](#); [Yannakoudakis, Briscoe, & Medlock, 2011](#)).

Concretely, given our set of training samples / feature vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ and a ranking $<_r$ such that the relation $\mathbf{x}_i <_r \mathbf{x}_j$ holds if and only if \mathbf{x}_i should be ranked ahead of \mathbf{x}_j , a ranking model computes a weight vector \mathbf{w} that maximizes the number of correctly ranked pairs of training samples, as formalized by the following constraints on *pairwise difference vectors* (i.e., the difference between feature vectors \mathbf{x}_i and \mathbf{x}_j):

$$\forall(\mathbf{x}_i <_r \mathbf{x}_j): \mathbf{w} \cdot (\mathbf{x}_i - \mathbf{x}_j) > 0$$

\mathbf{w} represents the parameters of the model that are learned during training and that define the hypothesis function. The parameters are chosen so that we get the best possible correlation with

the human-assigned scores in the training data, and thus predict CEFR scores as accurately as possible.¹⁰

The output from the training procedure is \mathbf{w} and, now, given a test essay \mathbf{x} , predictions are made by computing the dot product $\mathbf{w} \cdot \mathbf{x}$. The predicted ranks can then be converted to CEFR scores using, for example, linear regression.

We use the data described in the previous section to train our ranking model. Specifically, the CEFR level distribution in the training and test sets is presented in Figure 1 (left and right respectively). The test set contains 260 texts and their CEFR scores, and is not used during model development, while the rest of the 2,052 texts are used for training and tuning of the model.

Insert Figure 1 here

Feature space

In order to facilitate learning of the model parameters, the input data should be represented appropriately with the most relevant set of features possible. We focus on developing a model that assesses general writing competence and is not topic-/genre-specific. Such an approach facilitates the modeling of truly internally consistent “marking criteria” regardless of the prompt delivered. Systems that measure English competence directly are easier and faster to deploy, since they are more likely to be re-usable and generalize better across different genres than topic-specific models. Topic-specific models, on the other hand, are not immediately usable when new tasks are added, since the model cannot be applied until a sufficient number of manually annotated responses have been collected for a specific prompt.¹¹

We parse the data using the Robust Accurate Statistical Parsing (RASP) system with the standard tokenization and sentence boundary detection modules ([Briscoe, Carroll, & Watson, 2006](#)) in order to broaden the space of candidate features suitable for the task, and utilize linguistic features that carry task-independent information to the extent possible. More specifically, the model uses the following feature types:

¹⁰ For more details on the ranking algorithm we employ, the reader is referred to the work by [Briscoe, Medlock, & Andersen, 2010](#).

¹¹ We note however that separate models can be developed to assess prompt relevance and detect off-topic responses ([Cummins, Yannakoudakis, & Briscoe, 2016](#)).

- i. Character sequences (Chars)
- ii. Parts of Speech sequences (PoS)
- iii. Hybrid word and Parts of Speech sequences (Word/PoS)
- iv. Phrase structure rules (PS rules)
- v. Errors and error rate (Error rate)

The feature types above are mainly motivated by the fact that (sub-)lexical and grammatical properties should be highly predictive for our task. We use contiguous sequences of characters of up to length 3, extracted from individual words. For example, for the word “home”, we extract the character sequences “h”, “o”, “m”, “e”, “ho”, “om”, “me”, “hom”, and “ome”. PoS sequences of up to length 3 are extracted using the RASP tagger, which uses the CLAWS¹² tagset. An example PoS sequence is “VM RR”, which denotes a modal auxiliary followed by a general adverb, as in “could clearly”. Additionally, hybrid sequences of words and PoS of up to length 3 are used, in which open class words are replaced with their PoS tag; for example, “the NN1”, in which the determiner is followed by a singular common noun, as in “the girl”.

Based on the most likely parse for each sentence, we extract the rule names from the Phrase Structure (PS) tree. RASP’s rule names encode detailed information about the grammatical constructions found. For example, “V1/modal_bse/+” denotes a verb phrase consisting of a modal auxiliary head followed by an (optional) adverbial phrase, followed by a verb phrase headed by a verb with base inflection. Moreover, rule names explicitly represent information about peripheral or rare constructions (e.g., a sentence with preposed prepositional phrase with adjectival complement, as in “for better or worse, he left”), as well as about fragmentary and likely extra-grammatical sequences (e.g., a text unit consisting of 2 or more sub-analyses that cannot be combined using any rule in the grammar). Therefore, we believe that many (longer-distance) grammatical constructions and errors found in texts can be implicitly captured by features automatically derived from RASP PS rule names.¹³

Although the data we use contain information about the linguistic errors committed (Nicholls, 2003), we estimate an error rate in a way that does not require manually error-annotated data, and therefore allows our model to be applied directly to plain un-

¹² <http://ucrel.lancs.ac.uk/claws/>

¹³ For details of the rule name taxonomy and automatic generation of rule variants and their associated names, the reader is referred to Briscoe (2006).

annotated text once trained. In order to estimate the error-rate, we use a large background corpus of correct English containing more than 2 billion words ([Ferraresi, Zanchetta, Baroni, & Bernardini, 2008](#)) and identify those word sequences in the input data (of length 3) that are not present in the background corpus. Additionally, we automatically generate error rules from the Cambridge Learner Corpus (CLC) ([Nicholls, 2003](#)) by detecting word sequences that have been manually annotated as incorrect at least ninety per cent of the times they occur. This way, rules can be extracted from the existing error annotation in the CLC, obviating the need for manually constructed mal-rules.¹⁴ We also extend our set of error rules with classes of incorrect but plausible derivational and inflectional morphology based on a machine-readable dictionary (examples of the automatically generated error rules are presented in Table [3](#)). Using this set of error rules, we then generate another error-rate feature that the assessment model can utilize.

We note at this point that the term ‘feature type’ refers to a category of features (e.g., RASP PS rule names), while the term ‘features’ refers to instances of a feature type (e.g., RASP PS rule names refers to one feature type, but has over 1K unique instances).

Insert Table [3](#) here

Results

We measure the quality of the predicted scores (i.e., the output from the ranking model) by calculating Pearson’s product-moment (r) and Spearman’s rank (ρ) correlation coefficient against the scores assigned by a human expert on the same test dataset (the texts are marked directly on the full CEFR scale). Pearson’s correlation describes the extent to which the human-assigned and the predicted scores co-vary relative to the degree to which they vary independently. Spearman’s correlation assesses the strength of a monotonic relation between the two, and measures the quality of the ranking of the predicted scores produced by the model. Additionally, we calculate quadratic weighted κ ([Cohen, 1960, 1968](#)), a measure of agreement between the human-assigned and the predicted CEFR scores that is adjusted for chance agreement. For κ , we use linear regression to convert the predicted scores to real-valued scores in the 0—13 CEFR range, which are then rounded to the nearest score on the CEFR scale (referred to as the ‘predicted CEFR scores’).

¹⁴ A term used in computational linguistics to refer to rules capturing (common) grammatical errors.

In Table 4, we can see the performance of the model on the test set when incrementally adding more feature types to facilitate learning of an accurate CEFR level predictor. We can see that each of the feature types improves the model’s performance with respect to all evaluation measures. Our final model has a Pearson r of 0.765 and a Spearman ρ of 0.773, while a κ of 0.738 indicates high agreement between the predicted CEFR scores and those assigned by humans (the standard error of κ is 0.026).

As mentioned in the previous section, we aim at developing a model that assesses general linguistic competence based on task-independent feature types to the extent possible. Such types should be able to generalize better, as they are less biased by task idiosyncrasies. To further investigate this, we develop a CEFR-level model that uses feature types similar to the ones described earlier, but now instead of character sequences, it utilizes word sequences, and open class words are not replaced with their PoS tag;¹⁵ therefore, this model can capture topic directly. In Table 4 (last row), we can see the performance of this model on the test set, referred to as “topic-dependent features”. Relying on such features substantially decreases performance, further supporting our hypothesis that our approach generalizes better.

Insert Table 4 here

In order to assess the independent as opposed to the order-dependent additive contribution of the feature types to the overall performance of the system, we conducted a number of ablation tests. An ablation test consists of removing one type from the system at a time and re-evaluating the resulting model on the test set. Table 5 presents ablation results on the test set. We can see that all types have a positive effect on performance, while the error rate and PS rules appear to have the greatest impact, as removing either of these gives the largest decreases in both correlation and agreement.

Insert Table 5 here

As a further analysis of model performance, we also examined performance per CEFR level by looking at how far the predicted CEFR scores are from the human-assigned scores. This allows us to identify the levels it can predict more accurately. Specifically, we found that disagreements between the predicted and the human scores are higher for the A1 and C2 levels. On the other hand, the model has high agreement with the human scores on the rest of

¹⁵ As model performance is affected by feature interactions, this model is further tuned to achieve the best possible generalizable result.

the levels, and particularly with B1 and B2. This can be explained by the CEFR level distribution in the training data: the more training data we have for a level, the better the performance of the model for that particular level. When looking at the distribution of predicted CEFR scores against human-assigned scores on the test data (Figure 1, right), we can see that the model tends to over-predict B1, B2 and C1 labels, and under-predict the rest. As mentioned earlier, the model uses a training set of 2,052 texts, and we expect that more training data will further improve its performance. The mean of the human-assigned scores on the test set is 6.604 (standard deviation: 2.312), and the mean of the system-predicted CEFR scores is 6.627 (standard deviation: 1.840). In Table 1, we can see that a score of 6 represents a high B1 level.

Analysis

A manual inspection of the feature space reveals the specific features that are found to be highly predictive of CEFR levels. Below we can see a subset of those features, together with the weight assigned to them by the model, and example instantiations. A negative weight indicates that the feature contributes negatively to overall performance, whereas a positive weight indicates that it contributes positively to overall performance. For example, the character sequence “cas” has a negative weight (this could, for example, be in the context of a misspelling, such as “ocasion”), whereas the use of the modal “would” has a positive weight. We note, however, that the model uses a combination of a large set of features in order to make a prediction.

Feature	Weight	Example
AP/a1	5.99959	<i>very cleverly</i>
Apostrophes	3.60846	<i>yesterday's weather</i>
PP/p1	3.60492	<i>in the mornings</i>
NP/det_n1	2.26438	<i>the film</i>
would	1.65683	<i>would rather go</i>
T/frag	-1.94559	<i>but know she knew</i>
cas	-1.21774	<i>ocasion</i>
VBDZ_RR	1.19119	<i>was truly demanding</i>
VB0_VVN	1.14131	<i>be involved in</i>

ily	-1.13626	<i>easily</i>
PPIS2_VM_VV0	-0.939866	<i>as we can see</i>

Other features include the use of common nouns, prepositions, leading coordinators (e.g., “both” in “both . . . and . . .”), the infinitive marker “to”, spelling errors and noun-agreement errors. More analysis, however, is needed in order to link highly predictive features to *Reference Level Descriptions* (RLDs)¹⁶ that distinguish between the different CEFR levels in terms of various functions that non-native learners can perform as they gradually master the English language. [Alexopoulou et al. \(2013\)](#); [Yannakoudakis et al. \(2012\)](#) propose such a framework and examine the internal workings of a B2-level assessment model via visual presentations, link them to RLDs for B2, and identify novel and data-driven Second Language Acquisition hypotheses about developmental aspects of learner grammars. However, this is beyond the scope of this paper.

Writing Tool

We make the automated CEFR level predictor publicly available as a web browser-based cloud service for learners of English to practice their writing and receive feedback about their CEFR level as their writing progresses. Specifically, we integrate the placement model in Cambridge English Write & ImproveTM, and present a summary of usage behavior and statistics.

Cambridge English Write & ImproveTM

Cambridge English Write & ImproveTM (W&I hereafter) is a free, online tool¹⁷ that automatically provides diagnostic feedback to non-native English-language learners at different levels of granularity ([Andersen et al., 2013](#)): an overall assessment of their proficiency; an assessment for each individual sentence by highlighting sentences requiring more work, where darker shades indicate more problematic regions; and diagnostic word-

¹⁶ <http://englishprofile.org/the-cefr/reference-level-descriptions>

¹⁷ <https://writeandimprove.com>

level feedback on local issues, such as spelling, word choice and agreement errors. Learners can choose from a range of tasks, write or upload their text, save their work at any time and/or submit their writing for feedback. Then they can try again and use the feedback to improve their writing.

The previous version of W&I ([Andersen et al., 2013](#)) was aimed at assessing attainment of the upper-intermediate level in English (CEFR level B2). We develop a range of prompts and topics to write about at the different CEFR proficiency levels that reflect what a learner might face at a Cambridge English exam, and add them to W&I (Figure [2](#)).

Insert Figure [2](#) here

We employ the assessment model described in the previous sections and extend W&I to provide an overall assessment of someone's proficiency based on the full spectrum of language proficiency as defined in the CEFR. Specifically, given a piece of text, we assign an overall score for someone's writing on a scale from A1 to C2. This is illustrated in Figure [3](#): given a learner text on the left, the system provides feedback to the learner in a separate column on the right. At the top right, we can see the learner's proficiency level (A2), followed by W&I's topic relevance feedback in the middle, and sentence-level and local word-level diagnostic feedback at the bottom right. Additionally, on the same page, a progress graph is presented to the learner, which shows the CEFR level progression for up to the previous five checks of their text (Figure [4](#)). In this example, the learner has progressed within the B level, but has not quite reached the B2 level yet.¹⁸

The assessment time is around 15 seconds,¹⁹ which facilitates incremental and exploratory editing of a text to improve it, giving the learners the ability to try out different ways of correcting a problematic turn of phrase.

Insert Figure [3](#) here

Insert Figure [4](#) here

¹⁸ We note at this point that W&I is already context-aware and assesses whether responses are relevant to what was asked ([Cummins et al., 2016](#)), though a description of this component is beyond the scope of this paper.

¹⁹ 'Assessment time' refers to the overall time needed by the system to automatically process input text and provide feedback to the learner (see Figure 3).

Usage

Current users have been submitting their work for assessment more than once, which suggests that the system is being used in an iterative fashion as envisaged. From a sample size of 517,285 texts, the distribution of all scores across the system has a mean of 5.22 (which corresponds to the lower end of the B1 band) and a standard deviation of 1.8 (which corresponds to 0.9 CEFR levels). Users tend to select beginner prompts, and therefore generally shorter and less-sophisticated texts are over-represented.²⁰ Looking at the statistics on how the score develops from earlier to later submissions to the same prompt by a given user, we observe a positive effect. For example, on a sample size of 75,968 and when considering the score difference between the first and second submissions, the mean difference in score between all consecutive submissions is 0.310 with a standard deviation of 0.605. Further analyzing this per CEFR level of the first submission, we observe that it is harder for C-level users to increase their score. A more thorough investigation might show that the degree of improvement depends on variables such as the initial level, the genre elicited by the prompt, the proportion of sentences that have been modified, the increase (or decrease) in length, and the time between submissions. However, a detailed discussion of the usefulness and usability of the online system is beyond the scope of this paper.²¹

²⁰ Users are free to select from a set of prompts provided (or be directed by teachers towards specific tasks).

²¹ We intend to publish the detailed results of user-based evaluation experiments in the near future.

Conclusions

In this paper, we presented in detail the design, implementation and evaluation of an automated writing placement system for English-language learners that predicts someone's proficiency level on the full spectrum of the CEFR scale. We conceptualized the task as a supervised machine learning problem, and focused on developing a model that assesses general linguistic competence. We identified textual feature types that were highly predictive of someone's CEFR level, and performed ablation studies to investigate their contribution to overall performance. When compared against human-assigned CEFR scores, the model achieves both high correlation and agreement.

We further integrated our model into Cambridge English Write & ImproveTM (W&I) – a web-based tool that automatically provides diagnostic feedback to non-native English-language learners at different levels of granularity – and made it publicly available at no cost. The previous version of W&I was aimed at assessing attainment of the B2 level. In the current version, learners are assessed on the full CEFR scale and can point to an established standard of achievement. As part of future work, we plan to run further experiments to measure long-term learning effects.

References

- Alexopoulou, T., Yannakoudakis, H., & Salamoura, A. (2013). Classifying intermediate learner english: a data-driven approach to learner corpora. *Twenty Years of Learner Corpus Research: Looking back, Moving ahead, Corpora and Language in Use—Proceedings, 1*, 11–23.
- Andersen, Ø. E., Yannakoudakis, H., Barker, F., & Parish, T. (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* (pp. 32–41).
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-Rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3), 1–30.
- Azab, M., Hokamp, C., & Mihalcea, R. (2015). Using word semantics to assist english as a second language learners. In *Proceedings of the Human Language Technology conference of the North American chapter of the Association for Computational Linguistics* (pp. 116–120).
- Briscoe, T. (2006). *An introduction to tag sequence grammars and the RASP system parser* (Tech. Rep. No. UCAM-CL-TR-662). University of Cambridge, Computer Laboratory.
- Briscoe, T., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *ACL-COLING '06 Interactive Presentation Session* (pp. 77–80).
- Briscoe, T., Medlock, B., & Andersen, Ø. E. (2010). *Automated assessment of ESOL free text examinations* (Tech. Rep. No. UCAM-CL-TR-790). University of Cambridge, Computer Laboratory.
- Brooke, J., & Hirst, G. (2012). Robust, lexicalized native language identification. In *Proceedings of COLING 2012 Technical Papers* (pp. 391–408).
- Burstein, J. (2003). The e-Rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Lawrence Erlbaum Associates.

- Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence* (pp. 3–10). American Association for Artificial Intelligence.
- Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25(3), 27–36.
- Burstein, J., Marcu, D., & Knight, K. (2003). Finding the write stuff: Automatic identification of discourse structure in student essays. *Intelligent Systems, IEEE*, 18(1), 32–39.
- Chang, J., & Chang, J. S. (2015). WriteAhead2: Mining lexical grammar patterns for assisted writing. In *Proceedings of the human language technology conference of the north American chapter of the association for computational linguistics* (pp. 106–110).
- Cohen, J. (1960, April). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 4(70), 213–220.
- Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge University Press.
Retrieved from <http://books.google.co.uk/books?id=PygQ8Gk4k4YC>
- Cummins, R., Yannakoudakis, H., & Briscoe, T. (2016). Unsupervised modeling of topical relevance in L2 learner text. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 95–104).
- Dickinson, M., Kübler, S., & Meyer, A. (2012). Predicting learner levels for online exercises of Hebrew. In *Proceedings of the seventh workshop on innovative use of NLP for building educational applications* (pp. 95–104). Association for Computational Linguistics.
- Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).
- Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In S. Evert, A. Kilgarriff, & S. Sharoff (Eds.), *Proceedings of the 4th web as corpus workshop*.

- Gamon, M., Leacock, C., Brockett, C., Dolan, W. B., Gao, J., Belenko, D., & Klementiev, A. (2009). Using statistical techniques and web search to correct ESL errors. *Calico Journal*, 26(3), 491–511.
- Hancke, J., & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the learner corpus research (LCR) conference*.
- Higgins, D., Burstein, J., & Attali, Y. (2006, May). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2), 145 – 159.
- Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM conference on knowledge discovery and data mining* (pp. 133–142).
- Kakkonen, T., Myller, N., & Sutinen, E. (2004). Semi-automatic evaluation features in computer-assisted essay assessment. In *Proceedings of the 7th IASTED international conference on computers and advanced technology in education* (pp. 456 – 461).
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112).
- Lim, G. S. (2012). Developing and validating a mark scheme for writing. In H. Khalifa & F. Barker (Eds.), *Research notes* (Vol. 49, pp. 6–10). Cambridge ESOL.
- Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The writer's workbench: Computer aids for text analysis. *Educational psychologist*, 17(3), 172–179.
- Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the corpus linguistics 2003 conference* (pp. 572–581).
- Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 14(2), 210–225.
- Pilán, I., & Volodina, E. (2016). Classification of language proficiency levels in Swedish learners' texts. In *Proceedings of Swedish language technology conference*.
- Pilán, I., Volodina, E., & Zesch, T. (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), Osaka, Japan. Association for Computational Linguistics.

- Shermis, M. D., & Hammer, B. (2012). Contrasting state-of-the-art automated scoring of essays: analysis. In *Annual national council on measurement in education meeting* (pp. 1–54). http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf.
- Soyer, H., Topic, G., Stenetorp, P., & Aizawa, A. (2015). Crovewa: Cross-lingual vector-based writing assistance. In *Proceedings of the Human Language Technology conference of the North American chapter of the Association for Computational Linguistics* (pp. 91–95).
- Vajjala, S., & Lõo, K. (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 / Linköping Electronic Conference Proceedings 107: 113–127*.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 3–118.
- Volodina, E., Pilán, I., & Alfter, D. (2016). Classification of Swedish learner essays by CEFR levels. *CALL communities and culture—short papers from EUROCALL 2016*, 456.
- Wang, Y.-J., Shang, H.-F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*, 26(3), 234–257.
- Williamson, D. M. (2009). A Framework for Implementing Automated Scoring. In *Proceedings of the annual meeting of the American educational research association and the national council on measurement in education* (pp. 1 – 39). San Diego, CA.
- Yannakoudakis, H., Briscoe, T., & Alexopoulou, T. (2012). Automating second language acquisition research: integrating information visualisation and machine learning. In *Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH* (pp. 35–43).
- Yannakoudakis, H., Briscoe, T., & 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* (pp. 180–189).

	A1–	A1	A2	B1	B2	C1	C2	C2+
KET	0	1 2	3 4	5				
PET				3 4	5			
FCE					3 4	5		
CAE						3 4	5	
CPE							3 4	5
Overall scale	0	1 2	3 4	5 6	7 8	9 10	11 12	13

Table 1. Overall scoring scale that covers the full CEFR spectrum, based on the Cambridge English Main Suite exams. A score of 5 in the 0—13 scale represents a low B1, while a score of 6 represents a high B1.

Level	Prompt
A1 / A2	Describe the kinds of movies you like and why.
B1 / B2	Describe the kinds of movies you like and why. Why do you think people watch movies?
C1 / C2	Describe the kinds of movies you like and why. Why do you think people watch movies? To what extent will other media fulfil these functions in the future?
A1 / A2	Describe two websites that you use. What do you like about them?
B1 / B2	Describe two websites that you use. What do you like about them? How can these websites be improved?
C1 / C2	Describe two websites that you use. What do you like about them? How have these websites changed people and society, for better or worse?

Table 2. Prompts overlapping in topic and register across the CEFR levels.

Sequence	Error	Correction
he] want [to	Verb agreement	wants
to] thanks [all	Verb form	thank
are] to [old	Spelling confusion	too
's] interesting [place	Missing determiner	an+
is] need [to	Missing determiner	a+
of] whole	Missing determiner	the+
This [why	Missing verb	+is
few] absence	Noun agreement	absences
listening] at	Replace preposition	to
beloveds	Countability	beloved
disappointement	Spelling	disappointment
singed	Verb inflection	sang

Table 3. Examples of automatically generated error rules (first and second column) along with the proposed corrections (last column): square brackets indicate left and right context, while ‘+’ indicates that a word needs to be inserted. For example, if “want” is preceded by “he” and followed by “to” (first row), we are confident we have a verb agreement error.

Feature type	Pearson r	Spearman ρ	Cohen's κ
Chars	0.659	0.666	0.643
+Word/PoS	0.733	0.735	0.695
+PS rules	0.741	0.738	0.711
+Error rate	0.763	0.771	0.733
+PoS	0.765	0.773	0.738
Topic-dependent features	0.714	0.728	0.671

Table 4. Performance of the model on the test set (consisting of 260 texts) when incrementally adding more feature types to facilitate learning of an accurate CEFR level predictor. “Topic-dependent features” refers to a CEFR level predictor model that uses a feature set that can directly capture aspects of topic.

Ablated feature type	Pearson r	Spearman ρ	Cohen's κ
none	0.765	0.773	0.738
Chars	0.745	0.760	0.724
Word/PoS	0.763	0.770	0.739
PS rules	0.742	0.745	0.700
Error rate	0.740	0.738	0.704
PoS	0.763	0.771	0.733

Table 5. *Ablation*
tests on the test set (consisting of 260 texts) showing the independent contribution of each feature type to the performance of the CEFR model.

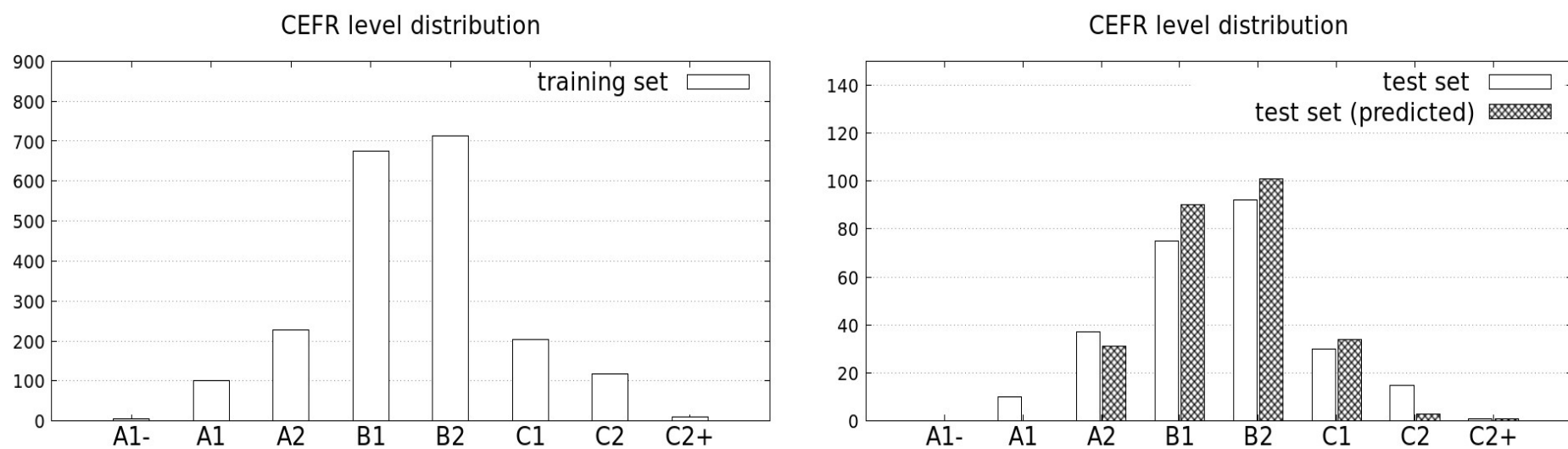


Figure 1. CEFR level distribution in the training and test set according to human experts ('training set' and 'test set' respectively) along with the model's predicted CEFR levels on the test set on the right ['test set (predicted)']. The y-axes refer to the total number of texts assigned to a specific CEFR level (please note that the y-axes are on different scales). The gaps between 'test set' and 'test set (predicted)' for each of the CEFR levels represent the difference between human scores and system scores (e.g., the system tends to overpredict B1). The mean score in the 'test set' is 6.604 (standard deviation: 2.312), and the mean score in 'test set (predicted)' is 6.627 (standard deviation: 1.840). In Table 1 we can see that a score of 6 represents a high B1 level.

Welcome, Helen
Sign out

W&I workbooks

W&I Beginner

W&I Intermediate

W&I Advanced

My workbooks

Create a workbook

Join a workbook

Progress

My writing

My activity & awards

My account

W&I workbooks

W&I Beginner

Open

- An email: A school friend**
A friend from your class, Cristina, had an accident last week.
Now, her doctor says she must stay at home for two weeks.
- A paragraph: A place you like**
Your teacher wants you to write about a place you like. Write about
 - where the place is
- A postcard: My town**
You receive this postcard from your English pen friend, Joe.
Here is a postcard of my town. You can see it's very nice. Please send me
- A paragraph: What I can see through the window**
Describe what you can see through the window where you are now, or from your bedroom at home. Write about:

W&I Intermediate

Open

- A report: Write and Improve**
Your English teacher told you about Write and Improve and said you should use it to practise and improve your English writing. Your teacher has asked you to write **a report** about your experience with
- An information sheet: How to succeed at job interviews**
You have recently had some job interviews so your college careers office has asked you to write a leaflet giving advice on how to succeed at job interviews.
- An article (for a newspaper or magazine): Someone you admire**
The local newspaper in the community where you live is planning to publish articles about famous people from different countries. This might be a film star, a politician or another person from the past or
- An opinion essay: Learning a new language**
Is learning to speak a foreign language as important these days as it was in the past?

Figure 2. W&I front-end with a range of topics to write about at different proficiency levels, along with a range of prompts that reflect what a learner might face at a Cambridge English exam.

W&I Intermediate

A report: Making a video

Your English class is going to make a short video about daily life at your school. Your teacher has asked you to write a report suggesting which lessons should be filmed **and** why.

Write your **report**.

Start again 

 Saved

My teacher has asked me to write a report about which lessons should be filmed.

I am think it will be good idea to filming the English class with our teacher who is Mrs Roberts. All of students love Mrs Roberts and we enjoyn this class alot. We could show the students all having nice time and doing activits. People who watch the film will see that it is a fun. They will want come to our school if they see this class.

Then, we can show them the restaurant where we having the lunch. This will show them you can to have good food in this schools' restaurant. And they

118 words entered. For this task you should enter between 140 and 190 words. Try to write more.

Check again →

Task help

 History

 Help

Level
A2

Images

Feedback

Changes



Very good! Your new level is A2. Your writing is improving! Use the feedback and think about ways to improve more. Make changes. Click Check again when you are ready. Good luck!

Did you write about the question? (5 is best)

0

1





2

3

4

5

My teacher has asked me to write a report about which lessons should be filmed.

I am think it will be  good idea to filming the English class with our teacher who is Mrs Roberts. All of students love Mrs Roberts and we enjoyn this class  alot. We could show the students all having nice time and doing  activits. People who watch the film will see that it is a fun. They will want  come to our school if they see this class.





Then, we can show them the restaurant where we  having  the lunch. This will show them you can to have good food in  this schools' restaurant. And they will see the students are happy  to.

Figure 3. W&I: learners can examine the automated feedback and revise their piece of writing. Given a learner text on the left, the system provides feedback in a separate column on the right: at the top right corner, we can see the learner's proficiency level (A2), followed by W&I's topic relevance feedback (middle), and sentence and local diagnostic feedback (bottom right).

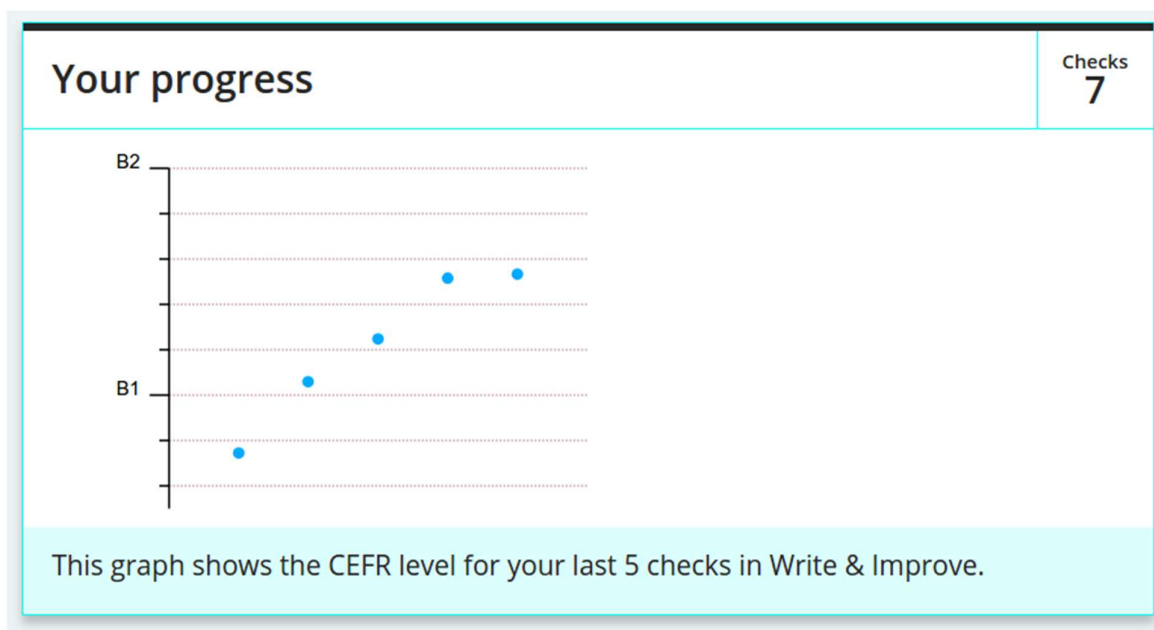


Figure 4. Progress graph indicating the CEFR level progression for up to the previous five checks of the text.