CEFR-Based Sentence Difficulty Annotation and Assessment

Yuki Arase[†] and **Satoru Uchida**[⋆] and **Tomoyuki Kajiwara**[◇]
[†]Graduate School of Information Science and Technology, Osaka University, Japan
[⋆]Faculty of Languages and Cultures, Kyushu University, Japan
[◇]Graduate School of Science and Engineering, Ehime University, Japan
arase@ist.osaka-u.ac.jp, uchida@flc.kyushu-u.ac.jp
kajiwara@cs.ehime-u.ac.jp

Abstract

Controllable text simplification is a crucial assistive technique for language learning and teaching. One of the primary factors hindering its advancement is the lack of a corpus annotated with sentence difficulty levels based on language ability descriptions. To address this problem, we created the CEFR-based Sentence Profile (CEFR-SP) corpus, containing 17k English sentences annotated with the levels based on the Common European Framework of Reference for Languages assigned by English-education professionals. In addition, we propose a sentence-level assessment model to handle unbalanced level distribution because the most basic and highly proficient sentences are naturally scarce. In the experiments in this study, our method achieved a macro-F1 score of 84.5% in the level assessment, thus outperforming strong baselines employed in readability assessment.

1 Introduction

Controllable text simplification, first proposed by Scarton and Specia (2018), is the automatic rewriting of sentences to make them comprehensible to a target audience with a specific proficiency level. Among its primary applications are providing reading assistance to language learners and helping teachers adjust the difficulty level of their teaching materials (Petersen and Ostendorf, 2007; Pellow and Eskenazi, 2014; Paetzold, 2016). The fine-grained control of output levels to match the linguistic ability of the readership is crucial for these educational applications.

While readability assessments have been actively studied (*e.g.*, in (Vajjala Balakrishna, 2015; Meng et al., 2020; Deutsch et al., 2020)), linking readability to language ability is difficult. Readability scores, such as the Flesch–Kincaid grade level (Kincaid et al., 1975), are intended for native speakers, not for language learners to whom very different considerations apply. Pilán et al. (2014) and Ozasa et al. (2007) revealed that readability metrics designed for L1 do not apply to L2 learners. Furthermore, readability definitions use documents rather than sentences, which are required by text simplification at the sentence-level, as their unit.

The lack of a corpus annotated by sentence difficulty level hinders the advancement of controllable text simplification. Previous studies (Scarton and Specia, 2018; Nishihara et al., 2019; Agrawal et al., 2021) necessarily used corpora annotated for readability rather than difficulty; furthermore, they assumed that all sentences in a document had the same readability (*i.e.*, the document level in Newsela (Xu et al., 2015)).

To solve these problems, we created a large-scale English corpus annotated by sentence difficulty levels based on the Common European Framework of Reference for Languages (CEFR),[1] the most widely used international standard describing learners' language ability. Our CEFR-based Sentence Profile (CEFR-SP) corpus adapts CEFR to sentence levels. A sentence is categorised as a certain level if a person with the corresponding CEFR-level can readily understand it. CEFR-SP provides CEFR levels for 17k sentences annotated by professionals with rich experience teaching English in higher education.

A major challenge in sentence-level assessment is the unbalanced distribution of levels: sentences at the basic (A1) and highly proficient (C2) levels are naturally scarce. To handle this, we propose a sentence-level assessment model with a macro-F1 score of 84.5%. We designed a metric-based classification method with a simple inductive bias that avoids overfitting to majority classes (Vinyals et al., 2016; Snell et al., 2017). Our method generates embeddings representing each CEFR-level and estimates a sentence's level based on its cosine similarity to these embeddings. Empirical results confirm that our method effectively copes with unbalanced

[1] https://www.coe.int/en/web/common-european-framework-reference-languages

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6206 - 6219
December 7-11, 2022

label distribution and outperforms the strong baselines employed in readability assessments.

This study makes two main contributions. First, we present the largest corpus to date of sentences annotated according to established language ability indicators. Second, we propose a sentence-level assessment model to handle unbalanced label distribution. CEFR-SP and sentence-level assessment codes are available[2] for future research at https://github.com/yukiar/CEFR-SP.

2 Related Work

Related studies have assessed text levels on different granularity (document and sentence) and level definitions (readability/complexity and CEFR).

2.1 Document-based Readability

Previous studies have assessed readability and created corpora with document readability annotations. WeeBit (Vajjala and Meurers, 2012), the OneStopEnglish corpus (Vajjala and Lučić, 2018), and Newsela provide manually written documents for various readability levels. Working with these annotated corpora, previous studies have used various linguistic and psycholinguistic features to develop models for assessing document-based readability (Heilman et al., 2007; Kate et al., 2010; Vajjala and Meurers, 2012; Xia et al., 2016; Vajjala and Lučić, 2018). Neural network-based approaches have proven to be better than feature-based models (Azpiazu and Pera, 2019; Meng et al., 2020; Imperial, 2021; Martinc et al., 2021). In particular, Deutsch et al. (2020) showed that pretrained language models outperform feature-based approaches, and the combination of linguistic features plays no role in performance gains.

2.2 Sentence-based Readability

Previous studies annotated sentences' *complexities* based on crowd workers' subjective perceptions. Stajner et al. (2017) used a 5-point scale to rate the complexity of sentences written by humans or generated by text simplification models. Brunato et al. (2018) used a 7-point scale for sentences extracted from the news sections of treebanks (McDonald et al., 2013). However, as Section 3.4 confirms, relating complexity to language ability descriptions is challenging. Naderi et al. (2019) annotated German sentence complexity based on language learners' subjective judgements. In contrast, the CEFR-level of a sentence should be judged *objectively* based on the understanding of language learners' skills. Hence, we presume that a sentence CEFR-level can be judged only by language education professionals based on their teaching experience. For sentence-based readability assessments, previous studies regarded all sentences in a document to have the same readability (Collins-Thompson and Callan, 2004; Dell'Orletta et al., 2011; Vajjala and Meurers, 2014; Ambati et al., 2016; Howcroft and Demberg, 2017). As we show in Section 3.4, this assumption hardly holds.

The *simplicity* of a sentence is one of the primary aspects in a text simplification evaluation, which is commonly judged by human. There are a few corpora annotated by the sentence simplicity for automatic quality estimation of text simplification (Štajner et al., 2016; Alva-Manchego et al., 2021). Nakamachi et al. (2020) applied a pretrained language model for estimating the sentence simplicity and used it to reward a reinforcement learning–based text simplification model. The sentence simplicity is distinctive from CEFR levels based on the established language ability descriptions.

2.3 CEFR-based Text Levels

Attempts have been made to establish criteria for CEFR-level assessments. For example, the English Profile (Salamoura and Saville, 2010) and CEFR-J (Ishii and Tono, 2018) projects relate English vocabulary and grammar to CEFR levels based on learner-written' and textbook corpora. Tools such as Text Inspector[3] and CVLA (Uchida and Negishi, 2018) endeavour to measure the level of English reading passages automatically. Xia et al. (2016) collected reading passages from Cambridge English Exams and predicted their CEFR levels using features proposed to assess readability. Rama and Vajjala (2021) demonstrated that Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) consistently achieved high accuracy for multilingual CEFR-level classification.

Although these micro- (*i.e.*, vocabulary and grammar) and macro-level (*i.e.*, passage-level) approaches have proven useful, few attempts have been made to assign CEFR levels at the *sentence* level, despite its importance in learning and teaching. Pilán et al. (2014) conducted a sentence-level assessment for Swedish based on CEFR; however,

[2]The licenses of the data sources are detailed in Ethics Statement section.

[3]https://textinspector.com/

they regarded document-based levels as sentence levels. Furthermore, their level assessment was as coarse as predicting either above B1 or not.

3 CEFR-SP Corpus

This section describes the design of the annotation procedure and discusses sentence-level profiles. CEFR describes language ability on a 6-point scale: A1 indicates the proficiency of beginners; A2, B1, B2, C1, and C2 indicates mastery of a language at the basic (A), independent (B), and proficient (C) levels. Because CEFR is skill-based, each level is defined by 'can-do' descriptors indicating what learners can do,[4] CEFR levels for sentences cannot be defined directly.

Therefore, we used a bottom-up approach, assigning CEFR levels to sentences based on the 'can-do' descriptors of reading skills under the definition that a sentence is, for example, at A1 level if it can be readily understood by A1-level learners. We hypothesise that with sufficient teaching experience and CEFR knowledge, it is possible to objectively determine at which level a learner can understand each sentence. We therefore carefully selected annotators with sufficient expertise through pilot and trial sessions.

3.1 Annotation Procedure

Pilot Study A pilot study was conducted to verify the hypothesis that sufficient teaching experience and CEFR knowledge will allow an objective evaluation of sentence levels. We recruited participants with three levels of expertise to label 228 sample sentences: an English-language education specialist with 12 years of teaching experience in higher education, a graduate student majoring in English education who is familiar with CEFR, and a group of three graduate students with various majors (natural language processing and ornithology) and no prior knowledge of CEFR or English-teaching experience. The results showed that the second expert had a high agreement rate with the first senior expert (Pearson correlation coefficient 0.74), whereas the members of the third group agreed less often with the senior expert (Pearson correlation coefficients: 0.45, 0.50, and 0.59). These results confirm that annotators with considerable experience and knowledge agree on the judgement of the CEFR levels of sentences.

Annotation Guidelines The annotators were familiarised with the annotation guidelines before beginning their work. The guidelines described the scales and 'can-do' descriptions of CEFR reading skills with example sentences of each level that were assessed by the expert. Importantly, the guidelines required the annotators to judge each sentence's level based on their English-teaching experience. Annotators were allowed to look in a dictionary to establish word levels but were instructed not to determine a sentence's level solely based on the levels of the words it contained.

Annotator Selection For formal annotation, we recruited eight annotators with diversified English-teaching experience. We then conducted a trial session in which the annotators were asked to label 100 samples extracted from the target corpora of formal annotation. These samples were labelled by the senior expert in the pilot study as references. Pearson correlation coefficients against the expert ranged from 0.59 to 0.77, roughly correlating with the participants' experience in English-teaching in terms of duration (years of teaching) and role (as private tutor or teacher in higher education). We finally selected two having high agreement rates (Pearson correlation coefficients: 0.75 and 0.73) and small average level-assignment differences (0.11 and 0.22) compared to the expert.[5] The annotation guidelines were finalised to provide example sentences with corresponding CEFR levels on which multiple annotators had agreed in the pilot and trial sessions.

3.2 Sentence Selection

Sentences were drawn from Newsela-Auto, Wiki-Auto, and the Sentence Corpus of Remedial English (SCoRE). Newsela-Auto and Wiki-Auto, created by Jiang et al. (2020), are specifically used for text simplification.[6] SCoRE (Chujo et al., 2015) was created for computer-assisted English learning, particularly for second language learners with lower-level proficiency. The sentences in SCoRE were carefully written by native English speakers, understanding the educational goals of each proficiency level; they include A-level sentences, which are scarce in text simplification corpora.

[4] https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52

[5] CEFR levels were converted into a 6-point scale.

[6] With the plan of expanding CEFR-SP to a parallel corpus in the future, we included parallel sentences. Note that our data-split policy (Section 5.1) ensures that highly similar sentences do NOT appear in training and validation/test sets.

The difficulty level can also be affected by external factors, such as discourse and readers' knowledge of a topic. For example, consider the sentence 'The white house announced his return.' Though it is simple in terms of wording and grammar, understanding it requires the knowledge that 'the white house' is an organisation name and the resolution of the coreference of 'he (his)' from outside the sentence. We consider comprehension of anaphora and cultural and factual knowledge to be different aspects of language proficiency. The dependence on external factors makes the sentence-level assessment ill-formed. To minimise the effect of outside factors, we selected *stand-alone* sentences for annotation, that is, sentences comprehensible independent of their surrounding context.

Thus, we selected the first sentences in paragraphs to avoid requiring coreference resolution. We excluded sentences with named entities (although dates, times, country names, and numeral expressions were allowed), quotations, and brackets. Appendix A describes the complete heuristics for sentence selection. We conducted several rounds of manual checks by observing a few hundred samples to finalise the heuristics of the sentence selection.

After filtering, we randomly sampled 5—30 word sentences to obtain 8.5k sentences each from Newsela-Auto and Wiki-Auto and 3.0k sentences from SCoRE (excluding the 100 sentences used in the trial session). Note that we excluded sentences from the Newsela-Auto test set so that CEFR-SP can be employed in training text simplification models in the future.

3.3 Sentence Profile

The two annotators independently supplied 40k labels for the 20k sentences. They assigned the same level to 37.6% sentences and levels with one grade difference to 50.8% sentences, which resulted in 88.4% sentences with levels within one grade difference. Given that many sentences are likely to have intermediate levels of difficulty, we regarded both assignments as correct if they differed by only one; thus, for example, the same sentence could be labelled as both B1 and B2. This left us with 27,841 labels for 17,676 unique sentences. Table 1 shows example sentences sampled from CEFR-SP.

Table 2 shows the number of sentences per level, average sentence length (number of words), and distribution (%) of lexical levels computed on the content words in the 27,841 labelled sentences.[7] We used the CEFR-J Wordlist[8], which assigns A1 to B2 levels to pairs of lemmas and part-of-speech tags. This allowed us to determine word levels without word sense disambiguation.[9] The content words in sentences were matched with the CEFR-J wordlist using their lemmas and part-of-speech tags. The frequency of each lexical level was computed by dividing the count of words with that level by the number of all content words at each sentence-level. We excluded function words, assuming that they had less effect on the sentence-level.

As expected, sentences in the A1 and C2 levels

A1	She had a beautiful necklace around her neck.
A2	Some experts say the classes should be changed.
B1	Historically there have also been negative consequences.
B2	Alligators are generally timid towards humans and tend to walk or swim away if one approaches.
C1	The metal-carbon bond in organometallic compounds is generally highly covalent.
C2	In the past, non-photosynthetic plants were mistakenly thought to get food by breaking down organic matter in a manner similar to saprotrophic fungi.

Table 1: Example sentences for each CEFR-level

	Num.	Length	Lexical level			
			A1	A2	B1	B2
A1	771	7.7	66.3	15.2	4.8	1.3
A2	4,775	10.9	54.6	18.2	10.1	3.2
B1	11,274	15.2	41.7	20.1	15.5	5.9
B2	8,283	18.0	31.9	19.1	17.8	7.9
C1	2,490	19.0	23.7	16.9	17.3	8.5
C2	248	19.2	16.5	15.2	16.3	6.8

Table 2: Distribution of sentence lengths and lexical levels of content words (%) in CEFR-SP

[7] We used Stanza (Qi et al., 2020) version 1.3.0 for preprocessing.

[8] CEFR-J Wordlist Version 1.6 http://www.cefr-j.org/data/CEFRJ_wordlist_ver1.6.zip

[9] Another possible lexicon is English Vocabulary Profile (EVP; http://www.englishprofile.org/wordlists). Although EVP provides C-level words, it requires word sense disambiguation to determine the level of a word, which hinders precise word-level estimation.

	Newsela								
	2	3	4	5	6	7	8	9-10	11-12
A1	12	16	35	20	7	4	3	0	2
A2	41	148	602	446	243	172	65	24	59
B1	30	187	1,155	1,302	1,015	969	475	290	442
B2	3	37	315	607	615	709	483	322	555
C1	0	2	23	51	59	89	91	58	176
C2	0	0	0	1	1	5	2	3	6

Table 3: Confusion matrix between CEFR and Newsela-Auto levels: grade levels scatter across CEFR levels.

	Length	Lexical level			
		A1	A2	B1	B2
Lv.1	8.8	22.3	10.9	7.7	6.3
Lv.2	13.4	17.7	13.4	9.9	6.7
Lv.3	21.8	17.2	13.7	11.5	7.3
Lv.4	26.8	16.5	14.2	12.3	8.9
Lv.5	27.1	16.4	9.9	12.0	7.3

Table 4: Distribution of sentence lengths and lexical levels of content words (%) in the sentence complexity dataset created by Brunato et al. (2018)

were particularly scarce. Sentence lengths are not proportional to CEFR levels; A level-sentences are short, whereas B-level sentences and above are similar in length. In contrast, the distribution of lexical levels shows a roughly positive correlation to sentence levels; A1-level words appear significantly more frequently in lower-level sentences, and B1 and B2 words in higher-level ones. A2-level words form an exception, appearing most frequently in the intermediate levels of A2 to B2.

3.4 Comparison with Existing Corpora

Table 3 shows the confusion matrix between CEFR levels and Newsela-Auto grade levels assembled using sentences extracted from Newsela-Auto. Newsela assigns readability levels using Lexile and converts them into a K–12 grade level.[10] Newsela-Auto assigns the grade level of the document to all the sentences contained in it. The Newsela-Auto levels scatter across CEFR levels, indicating that document-based readability levels do not agree with sentence-based CEFR language ability.

Table 4 shows the distribution of sentence lengths and lexical levels of content words (%) in the sentence complexity corpus of Brunato et al. (2018). This corpus rated sentence complexity on a 7-point scale, with 1 indicating 'very simple' and 7 indicating 'very complex'. Based on this paper, we extracted sentences having degrees of agreement greater than or equal to 10 and determined their levels as rounded means of assigned levels. We found that no sentences were assigned levels higher than 5, which means this corpus lacks sentences at the most complex levels. In contrast, CEFR-SP provides C-level sentences, which are considered the most complex.

The distributions in Table 4 are distinct from those in our corpus (Table 2). Although Brunato et al. (2018) reported that sentence length shows a clear correlation with complexity level, this was not true for our sentences of level B1 or higher. In Table 4, the distribution of each lexical level across complexity levels was relatively uniform. In contrast, CEFR-SP showed a positive correlation between sentence and lexical levels. The results suggest that the standards of our CEFR-level annotations based on formal language ability descriptions were significantly different from the annotators' subjective perception of complexity.

4 Sentence-Level Assessment

We propose a sentence-level assessment model robust to imbalances in label distribution.

4.1 Problem Definition

CEFR levels are ordinal: *e.g.*, the B2 level is higher than the B1 level. It might therefore seem natural to model the level assessment as a regression problem. However, the gaps between the levels can be nonuniform, making the interpretation of regression outputs difficult; for example, we cannot decide whether an output of 0.7 corresponds to A1 or A2 (Heilman et al., 2008; François, 2009). Therefore, we model CEFR-level assessment as a multiclass classification problem.[11]

Given a training corpus with N labelled samples $\{(x_0, y_0), (x_1, y_1), \cdots, (x_{N-1}, y_{N-1})\}$, where x_i is a sentence and $y_i \in \{0, 1, \cdots, J-1\}$ indicates the index of the corresponding level, we train a classifier that classifies an input sentence into J classes; $J = 6$ in CEFR. For brevity, we do not distinguish between a level and its index hereafter.

[10] https://support.newsela.com/article/grade-to-lexile-conversion/

[11] Moreover, a classification model was superior to a regression model in our preliminary experiments.

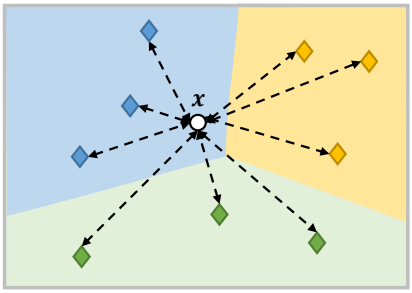

Figure 1: How sentence-level is estimated by measuring similarities to level embeddings (represented by ♢).

4.2 Background: Metric-based Method

Table 2 empirically shows that the distribution of sentence levels is unbalanced; the most basic and highly proficient sentences are the least common. An unbalanced label distribution leads to overfitting major classes and ignoring minor ones; for educational applications, such infrequent levels cannot be dismissed.

Therefore, we propose a sentence-level assessment model that is robust against label imbalance. We use a metric-based approach (Vinyals et al., 2016; Snell et al., 2017; Ye and Ling, 2019; Sun et al., 2019) that classifies samples based on distances in a vector space, thereby avoiding overfitting by virtue of the simple inductive bias of a classifier. The metric-based approach has been studied for few-shot classification, where unlabelled sentences are classified by the embedding distances between labelled and unlabelled samples. In contrast, we explicitly learn embeddings representing CEFR levels (hereafter referred to as *prototypes*) and predict sentence levels using cosine similarity.

4.3 Metric-based Level Assessment

We assume that representing a CEFR-level by a single vector may be insufficient; allowing multiple prototypes improves the expressiveness of level representation. We generate K prototypes for each CEFR-level, *i.e.*, KJ prototypes in total, constituting a prototype matrix $\boldsymbol{C} \in \mathbb{R}^{KJ \times d}$. The k-th prototype of the i-th CEFR-level $\boldsymbol{c}_i^k \in \mathbb{R}^d$ has the same dimension d as the sentence embedding. We assume that the similarity between the input sentence embedding and prototype measures the likelihood that the sentence has the corresponding label, as shown in Figure 1.

We employ a pretrained masked language model (MLM) to encode a sentence. Specifically, we encode an input sentence with m tokens $x = \{w_0, w_1, \cdots, w_{m-1}\}$ using MLM to obtain the hidden outputs of each token[12]

$$\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{m-1} = \text{MLM}(w_0, w_1, \cdots, w_{m-1}),$$

where $\boldsymbol{h}_i \in \mathbb{R}^d$. We generate a sentence embedding $\boldsymbol{x} \in \mathbb{R}^d$ by mean pooling these token embeddings (Reimers and Gurevych, 2019):

$$\boldsymbol{x} = \text{MeanPool}(\boldsymbol{h}_0, \boldsymbol{h}_1, \cdots, \boldsymbol{h}_{m-1}). \quad (1)$$

Finally, we compute the distribution p over the levels for $\boldsymbol{x}$ using softmax considering similarities to the prototypes:

$$p(y = j|\boldsymbol{x}) = \frac{\exp(\text{CosSim}(\boldsymbol{x}, \boldsymbol{c}_j))}{\Sigma_j \exp(\text{CosSim}(\boldsymbol{x}, \boldsymbol{c}_j))},$$

where $\text{CosSim}(\cdot, \cdot)$ calculates cosine similarity. When a level has multiple prototypes $K > 1$, we compute the mean of the cosine similarities:

$$\text{CosSim}(\boldsymbol{x}, \boldsymbol{c}_j) = \frac{\Sigma_k \text{CosSim}(\boldsymbol{x}, \boldsymbol{c}_j^k)}{K}.$$

4.4 Loss Weighting

The entire model, including MLM, is trained to minimise cross-entropy loss. For further alleviation of the unbalanced label distribution, loss weighting is applied according to the multinomial distribution of the level frequency (Conneau and Lample, 2019).

$$p_i = \frac{q_i^\alpha}{\Sigma_{i=0}^{J-1} q_i^\alpha}, \quad (2)$$

where q_i is the frequency of level i in the training set, and $\alpha \in [0, 1]$ controls the weight strength. A small alpha gives large weights to infrequent labels.

4.5 Prototype Initialisation

The experiments established that the initialisation of prototypes affects the training stability, as the prototypes are learned from scratch. Therefore, the prototypes have consistent values set during initialisation to stabilise model training. Assuming that common characteristics of the same level of sentences are reflected in their embeddings, we use the mean of sentence embeddings in Equation (1): $\hat{\boldsymbol{c}}_i = \text{MeanPool}(\boldsymbol{x}_0^i, \boldsymbol{x}_1^i, \cdots, \boldsymbol{x}_{n-1}^i)$, where $\boldsymbol{x}_k^i$ is the k-th sentence embedding of level i and n is the number of sentences at level i in the training set.

[12] In practice, MLM may attach special tokens such as `[CLS]` and `[SEP]` to an input, which are omitted for brevity.

	A1	A2	B1	B2	C1	C2
Train	535	3,646	8,996	6,636	1,908	100
Validation	125	568	1,130	821	290	74
Test	111	561	1,148	826	292	74

Table 5: Distribution of sentence levels in training, validation, and test sets

Because a level is allowed to have multiple prototypes, an initialisation vector is generated for the k-th prototype at level i, $\hat{\boldsymbol{c}}_i^k \in \hat{\boldsymbol{C}}$, by adding Gaussian noise with mean $\mu = 0$ and variance σ^2 set to 5% of that computed on all elements in $\hat{\boldsymbol{c}}_0, \hat{\boldsymbol{c}}_1, \ldots, \hat{\boldsymbol{c}}_{J-1}$:

$$\hat{\boldsymbol{c}}_i^k = \hat{\boldsymbol{c}}_i + \mathcal{N}(\mu, \sigma^2).$$

Finally, expecting these prototypes to capture the distinctive features of different levels, we orthogonalise the matrix $\hat{\boldsymbol{C}}$ and set the initial values of the prototype matrix $\boldsymbol{C}$.

5 Evaluation

In this section, the proposed level assessment model is evaluated using the CEFR-SP corpus.

5.1 Corpus Splitting

We split CEFR-SP into three sets: approximately 80% for training, 10% for validation, and 10% for the test set, as shown in Table 5. We adjusted the number of sentences for infrequent levels to preserve a reasonable number of test and validation cases.[13] In corpus splitting, we ensured that highly similar sentences did not appear in both the training and validation/test sets, as detailed in Appendix B.

A sentence in CEFR-SP may have as many as two levels, both assignments being regarded as equally reliable. Therefore, the predictions during training, validating, and testing were assumed correct if they matched either of the annotated labels.

5.2 Evaluation Metrics

The ability to predict *all* levels correctly is important for educational applications. As the distribution of levels was unbalanced, the models were evaluated using macro-F1 to penalise models that ignored minor classes. In addition, because CEFR levels are ordinal, the models were also evaluated using the quadratic weighted kappa (Bakeman and Gottman, 1997).

To reduce the dependence of performance fluctuation on initialisation seeds, the experiments were conducted 12 times with randomly selected seeds. We then discarded the best and worst results and reported a mean macro-F1 score and kappa value with a 95% confidence interval.

5.3 Setting

We used BERT-Base, cased model (Devlin et al., 2019) as the pretrained MLM to encode sentences in the models that were compared.[14] Specifically, we used the outputs of the 11-th layer, which performed strongly. K, the number of prototypes of the proposed method, was set to 3 to maximise the average macro-F1 of the validation set in the 1–10 range.

Comparison Because of the roughly positive correlation between the word and sentence levels (Section 3.3), we implemented a bag-of-words (BoW)[15] classifier using support vector machines (Cortes and Vapnik, 1995) as the naive baseline. Moreover, as a simpler variant of metric-based classification method, we implemented a k-nearest neighbour (kNN) (Fix and Hodges, 1989) classifier. We used mean-pooled token embeddings of freezed BERT as features and the cosine distance for distance computation. The size of k was set to 6 which marked the highest macro-F1 on the development set.

As the state-of-the-art baseline, we used a BERT-based classifier that outperforms conventional linguistic-feature-based classifiers in predicting passage-level readability (Deutsch et al., 2020) and CEFR levels (Rama and Vajjala, 2021), as well as on simple and complex binary classification (Garbacea et al., 2021) of the WikiLarge corpus (Zhang and Lapata, 2017). The proposed model was compared with these baselines with or without loss weighting.

Ablation Study We investigated the effect of K with an ablation study. We also implemented variations of the proposed method without loss weighting and initialisation based on sentence embeddings. The former method achieved its maximum

[13] We tentatively used the higher level among the two annotated labels for assigning a sentence into either the training, validation, or test sets.

[14] In a preliminary experiment, we compared BERT, RoBERTa (Liu et al., 2019), and Sentence-BERT (Reimers and Gurevych, 2019) with different configurations and confirmed that there was no significant difference between them. Therefore, we decided to use the standard BERT-Base.

[15] Word-level features performed much worse and were omitted in this experiment.

		A1	A2	B1	B2	C1	C2	Average	Weighted κ
BoW	w/o lossW	0.0	69.7	76.3	66.4	34.7	0.0	41.2	0.354 ± 0.000
		44.2	64.9	73.0	69.6	53.8	8.0	52.3	0.429 ± 0.000
kNN		1.5 ± 1.4	75.2 ± 0.7	81.8 ± 0.4	66.4 ± 0.6	8.1 ± 2.6	0.0 ± 0.0	38.8 ± 0.4	0.373 ± 0.004
BERT	w/o lossW	12.8 ± 9.4	$\mathbf{83.6 \pm 0.3}$	87.0 ± 1.1	$\mathbf{86.7 \pm 1.2}$	82.9 ± 1.5	76.8 ± 5.5	71.7 ± 1.7	0.592 ± 0.012
		72.7 ± 3.9	82.7 ± 1.1	85.5 ± 0.9	86.4 ± 0.7	84.9 ± 1.2	83.6 ± 3.0	82.5 ± 0.9	0.609 ± 0.014
Proposed	w/o lossW	12.0 ± 13.4	$\mathbf{83.6 \pm 0.4}$	$\mathbf{87.8 \pm 1.2}$	86.3 ± 1.4	83.0 ± 0.9	0.0 ± 0.0	58.7 ± 1.8	0.595 ± 0.013
	w/o init	76.1 ± 1.5	80.5 ± 1.4	84.7 ± 1.4	85.7 ± 1.3	85.3 ± 1.2	88.1 ± 2.1	83.3 ± 0.9	$\mathbf{0.628 \pm 0.017}$
		$\mathbf{78.0 \pm 1.3}$	81.4 ± 0.9	86.5 ± 1.1	85.9 ± 0.8	$\mathbf{85.4 \pm 1.3}$	$\mathbf{89.7 \pm 1.6}$	$\mathbf{84.5 \pm 0.7}$	$\mathbf{0.628 \pm 0.010}$

Table 6: Macro-F1 scores (%) per level and quadratic weighted kappa values measured on the CEFR-SP test set; 'w/o lossW' indicates a model without loss weights and 'w/o init' indicates a model without initialisation using sentence embeddings. The proposed method (last row) preserves high F1 scores at the infrequent A1 and C2 levels and the best quadratic weighted kappa value.

validation macro-F1 score when $K = 1$. The latter method used the same settings as the proposed method, except for prototype initialisation; it initialised the prototype embeddings using a normal distribution $\mathcal{N}(0, 1)$.

5.4 Implementation Details

The classifier layer of the BERT baselines comprised a linear layer with weights $W \in \mathbb{R}^{d \times J}$ and a 10% dropout to the input sentence embedding. Other conditions remained the same as those of the proposed method. We input a sentence embedding computed by Equation (1) and calculated the standard classification loss of cross-entropy. Loss weights were computed by Equation (2).

All models were implemented using the PyTorch, Lightning, Transformers (Wolf et al., 2020), and scikit-learn libraries.[16] The neural network models were trained on an NVIDIA Tesla V100 GPU using an AdamW (Loshchilov and Hutter, 2019) optimiser with a batch size of 128. The training was stopped early, with 10 patience epochs and a minimum delta of $1.0e-5$ based on the average macro-F1 score of all levels measured on the validation set. The loss weighting factor α and other hyperparameters were tuned using Optuna (Akiba et al., 2019). For the proposed method and BoW and BERT baselines, α values were set to 0.2, 0.3, and 0.4, respectively. The complete hyperparameter settings are described in Appendix C.

5.5 Results

Table 6 shows the CEFR-SP test set results by means of macro-F1 scores (%) per level and quadratic weighted kappa values with 95% confidence intervals. As in previous studies, the BERT-based classifiers outperformed the BoW baselines.

[16] Hyperlinks to these libraries are listed in Appendix D.

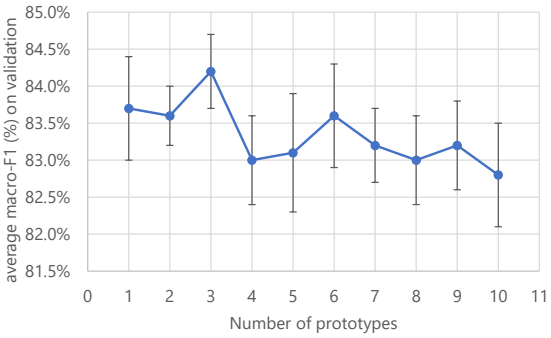

Figure 2: Effects of number of prototypes: average macro-F1 scores (%) measured on validation set

This result confirms that words and their levels, despite their importance, are not solely responsible for determining sentence levels. The kNN classifier showed higher macro-F1 scores than BoW without loss weighting on A2 and B1 because of the powerful BERT embeddings. However, it failed to identify A1 and C levels, which indicates the significance of addressing unbalanced label distribution.

The proposed method (last row) had the highest F1 scores for infrequent levels, *i.e.*, A1 and C2, but a slightly reduced performance for the more common levels. We consider this acceptable, considering the method's capability to assess infrequent levels. Overall, the proposed method achieved the highest average macro-F1 score (84.5%) and quadratic weighted kappa value (0.628).

Effects of Loss Weighting While loss weighting is highly effective in alleviating the effects of unbalanced label distribution on all models, it is more critical for the proposed method. Exclusion of loss weighting overlooks the A1 and C2 levels, as is clear from the sixth row of Table 6. Confusion matrices confirmed that A1 and C2 sentences were misclassified to their adjacent levels.

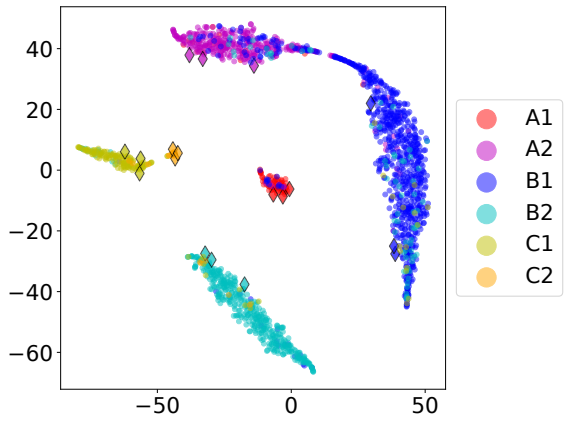

Figure 3: Visualisation of prototypes (represented by ♢) and sentence embeddings of the proposed method.

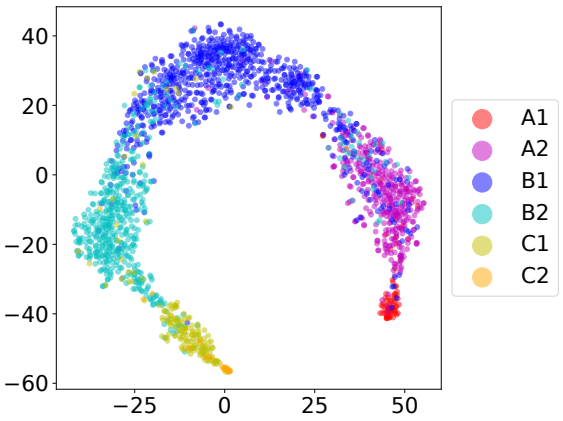

Figure 4: Visualisation of sentence embeddings of BERT baseline with loss weighting.

Effects of Initialisation The seventh row of Table 6 presents the results for the proposed method without initialisation using sentence embeddings. This method tended to have larger confidence intervals than the proposed model. Moreover, we observed that it fell into an undesired solution that overlooked A1 and C2 levels depending on initialisation seeds, as reflected in lower macro-F1 scores. These results confirm that our initialisation was effective for training stabilisation.

Effects of Number of Prototypes Figure 2 shows the average macro-F1 scores with 95% confidence intervals measured on the validation set when the number of prototypes in the proposed method changed from 1 to 10. The average macro-F1 score initially improved as the number of prototypes increased; it peaked at three, and then gradually decreased. This trend empirically confirms the effectiveness of multiple prototypes and shows that a relatively small number of prototypes is sufficient for CEFR-SP.

Visualisation Figure 3 plots the sentence embeddings generated by the proposed method, and Figure 4 those generated by the BERT baseline with loss weighting. The gold levels are colour-coded; for the proposed method, the prototypes are indicated by diamond markers. We used T-SNE (van der Maaten and Hinton, 2008) for visualisation, setting the perplexity to 30 and number of iterations to 5k to ensure convergence.

The class boundaries were not clear in the embeddings of the baseline. In contrast, the embeddings of the proposed method formed clear clusters by level owing to the metric-based classification; this improved the interpretability. When assessing the level of a new sentence, the cosine similarity to each prototype indicates whether the assessment result is high-confidence, *i.e.*, prototypes of a single level exhibit significantly high cosine similarity to the sentence, or ambiguous, *i.e.*, multiple levels exhibit competitive cosine similarities.

6 Summary and Future Work

In this study, we introduced CEFR-SP, the first English sentence corpus annotated with CEFR levels. The carefully designed annotation procedure involved recruiting experts with strong backgrounds in English education to ensure the reliability of the assigned labels. CEFR-SP allows the development of an automatic sentence-level assessment model. The proposed method can handle unbalanced level distributions using a metric-based classification.

Our future work will involve collecting parallel sentences of CEFR-SP to make it directly applicable for training text simplification models. We will also develop controllable text simplification models based on reinforcement learning: the proposed level assessment model will be employed to reward the generation of lower-level sentences.

Limitations

Because of severe space constraints, we have reported only the lexical profile of CEFR-SP. We will present its syntactic and psycholinguistic features and analyse it from an educational perspective in a future publication. Moreover, CEFR-SP is not directly applicable to train controllable text simplification models that require parallel sentences with different levels. Therefore, we are currently expanding CEFR-SP to make it parallel through the

manual rewriting of sentences in the corpus. Our sentence-level assessment model helps this process. We can complement sentences of scarce levels by adding additional rewriting tasks.

We suspect that the proposed method is directly applicable to other label-imbalanced classification problems. The empirical investigation of this is out of the scope of the present paper and is left for future work.

Ethics Statement

Ethics in Annotation Process The sentences in CEFR-SP were sampled from Newsela-Auto (news articles), Wiki-Auto (Wikipedia articles), and SCoRE (sentences written for an educational application of an academic project). We believe them to be free from harmful content that insults annotators.

We contracted with a commercial company that provides data annotation services for academia, including the management of annotators. We paid annotators $0.50 per sentence, *i.e.*, approximately $44/h. This was significantly higher than the minimum wage in the place where this study was conducted, reflecting our respect for the expertise required.

License Compliance We comply with the licenses of the original data sources of CEFR-SP. Specifically, we separate CEFR-SP sentences by data source and distribute them with the same license as the original datasets from which they were sampled.

Wiki-Auto CC BY-SA 3.0

SCoRE CC BY-NC-SA 4.0

Newsela-Auto We ask people first to obtain Newsela corpus (`https://newsela.com/data/`) and then contact us, following the distribution policy of Newsela-Auto.

For the reproducibility of the study, the training-, validation-, and test-set splits are maintained.

Acknowledgements

We appreciate the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported by JSPS KAKENHI Grant Number JP21H03564.

References

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 3757–3769.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2623–2631.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental CCG parser. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1051–1057.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association of Computational Linguistics (TACL)*, 7:421–436.

Roger Bakeman and John M. Gottman. 1997. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press.

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? Do you agree? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2690–2699.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial EFL learners. In Agnieszka Leńko-Szymańska and Alex Boulton, editors, *Multiple Affordances of Language Corpora for Data-driven Learning*, Studies in Corpus Linguistics 69, pages 109–128. John Benjamins.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 193–200.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. READ–IT: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 1–17.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Evelyn Fix and J. L. Hodges, Jr. 1989. Discriminatory analysis - Nonparametric discrimination: Consistency properties. *International Statistical Review*, 57(3):238–247.

Thomas François. 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the Student Research Workshop at EACL*, pages 19–27.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1086–1097.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 71–79.

David M. Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 958–968.

Joseph Marvin Imperial. 2021. BERT embeddings for automatic readability assessment. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 611–618.

Yasutake Ishii and Yukio Tono. 2018. Investigating Japanese EFL learners' overuse/underuse of English grammar categories and their relevance to CEFR levels. In *Proceedings of Asia Pacific Corpus Linguistics Conference*, pages 160–165.

Patrick Jacob and Alexandra Uitdenbogerd. 2019. Readability of Twitter tweets for second language learners. In *Proceedings of the Annual Workshop of the Australasian Language Technology Association*, pages 19–27.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7943–7960.

Rohit Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 546–554.

Peter J. Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical Report 56, Institute for Simulation and Training.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, 1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 92–97.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. ReadNet: A hierarchical transformer framework for web article readability analysis. In *Proceedings of European Conference on IR Research (ECIR)*, pages 33–49.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for German language. *arXiv*, 1904.07733.

Akifumi Nakamachi, Tomoyuki Kajiwara, and Yuki Arase. 2020. Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity. In *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 153–159.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266.

Toshiaki Ozasa, George R. S. Weir, and Masayasu Fukui. 2007. Measuring readability for Japanese learners of English. In *Proceedings of the Conference of Pan-Pacific Association of Applied Linguistics (PAAL)*, pages 122–125.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.

David Pellow and Maxine Eskenazi. 2014. An open corpus of everyday documents for simplification tasks. In *Proceedings of the Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Proceedings of Speech and Language Technology in Education (SLaTE)*, pages 69–72.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 174–184.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 101–108.

Taraka Rama and Sowmya Vajjala. 2021. Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling? *arXiv*, 2102.12971.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

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*, Eurosla Monographs Series 1, pages 101–132. Eurosla.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 712–718.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Sanja Stajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4096–4102.

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification. In *Proceedings of Workshop & Shared Task on Quality Assessment for Text Simplification (QATS)*, pages 22–31.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 476–485.

Satoru Uchida and Masashi Negishi. 2018. Assigning CEFR-J levels to English texts based on textual features. In *Proceedings of Proceedings of Asia Pacific Corpus Linguistics Conference*, pages 463–467.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 163–173.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 288–297.

Sowmya Vajjala Balakrishna. 2015. *Analyzing text complexity and text simplification: Connecting linguistics, processing and educational applications*. Ph.D. thesis, University of Tübingen.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 38–45.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 12–22.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2872–2881.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 584–594.

A Details of Sentence Selection

Dependence on external factors makes the sentence-level-assessment problem ill-formed. This phenomenon was noticed in (Jacob and Uitdenbogerd, 2019): linguistic features that are typically well-correlated with document readability were poorly correlated with it in tweets, which inevitably depend on external factors. To avoid this problem, we carefully selected stand-alone sentences for annotation.

For Wiki-Auto, we excluded the first paragraphs of an article to avoid dictionary-definition-like sentences, *e.g.*, 'X is the capital of country Y'. While we excluded sentences containing named entities recognised by Stanza, we allowed named entities of types of DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, and CARDINAL, as well as those in a list that we manually prepared containing names of well-known regions, countries, and cities (*e.g.*, Europe, France, and Paris), and common personal names (*e.g.*, William). Finally, we regularised spellings to the American forms using the localspelling library.[17]

B Details of Corpus Splitting

First, we computed the cosine distances between all pairs of sentence embeddings obtained using a pretrained Sentence-BERT model (Reimers and Gurevych, 2019).[18] Next, the average cosine distance for each sentence was calculated. The sentences were then allocated to the test, validation, and training sets according to the descending order of their average cosine distances. Thus, sentences with the least similarity to other sentences were allocated to the test and validation sets, and the rest to the training set.

C Hyperparameter Settings

For all models, the loss weighting factor α was searched in the range $[0.1, 1.0]$ with 0.1 interval. For neural network models, the learning rate was searched in the range $[1e-5, 7e-5]$ with $1e-5$ interval. For the BoW baseline using support vector machines, the kernel was chosen from linear or radial basis function networks, and the regularisation parameter γ was searched in the range $[0.01, 100]$ by loguniform sampling of 40 points. Table 7 presents the hyperparameter settings of the proposed and BERT baseline models, Table 8 those of the BoW baseline.

D Hyperlinks to Libraries

Here we list hyperlinks to the libraries used in implementation.

PyTorch `https://pytorch.org/`

Lightning `https://www.pytorchlightning.ai/`

[17] `https://github.com/fastdatascience/localspelling`

[18] Specifically, we used `all-mpnet-base-v2`, which had the highest performance at `https://www.sbert.net/docs/pretrained_models.html`.

		Learning Rate	α
BERT baseline	w/o lossW	$6.0e-5$	–
		$3.0e-5$	0.4
Proposed	w/o lossW	$3.0e-5$	–
	w/o init	$1.0e-5$	0.2
		$1.0e-5$	0.2

Table 7: Hyperparameter settings of the proposed and BERT baseline models

		Kernel	γ	α
BoW	w/o lossW	linear	4.6	–
		linear	0.7	0.3

Table 8: Hyper-parameter settings of the Bag-of-Words baseline

Transformers https://huggingface.co/docs/transformers/index

scikit-learn https://scikit-learn.org/

Optuna https://optuna.readthedocs.io/