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.

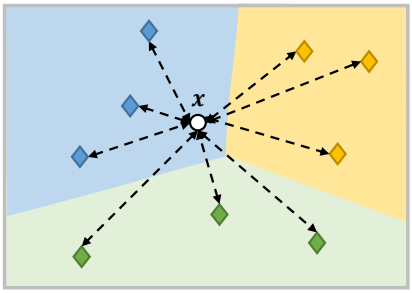

Figure 1: How sentence-level is estimated by measuring similarities to level embeddings (represented by ♢).

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.