Leveraging the Intuitions of Lay People on Linguistic Complexity for Automatic Sentence Readability Assessment*

Ignatios Charalampidis[1,*], Xiaobin Chen[1]

[1]Hector Research Institute of Education Sciences and Psychology (University of Tübingen), Walter-Simon-Strasse 12, Tübingen, 72072, Germany

Abstract

Automatic readability assessment (ARA) is an invaluable tool for second language teachers, as it can be used to evaluate the difficulty of educational materials. Previous studies on ARA often used text-level linguistic complexity features to predict readability, while few studies have focused on the sentence level. This paper presents a study in which crowd-sourced evaluations of sentence readability were collected from non-expert native and non-native speakers of English. 50 participants were asked to evaluate 1000 sentences in English extracted from diverse sources for lexical, grammatical and overall sentence difficulty separately, and the resulting dataset of 10,000 evaluations (hereinafter referred to as SR-Crowd) is released for future research. Several machine learning (ML) models were trained on this data using linguistic and neural features. The best model relies on neural and linguistic features and it achieves a Quadratic Weighted Kappa (QWK) score of 0.7. Further, our findings suggest that readability perception differs among groups and individuals and a single ARA pipeline may not be enough to capture readability accurately for all of them. This research offers valuable insights that can be applied in the areas of evaluating language learning materials, content filtering, and text summarization, where accurate sentence-level readability assessment is essential.

Keywords

Second Language Acquisition, Language Learning, Computational Linguistics, Machine Learning, Automatic Readability Assessment, Neural Networks,

1. Introduction

In the field of second language acquisition (SLA), the Input Hypothesis, proposed by linguist Stephen Krashen, suggests that learners acquire language most effectively when exposed to input that is slightly beyond their current level of competence, often referred to as i+1, where i represents current competence and 1 represents new knowledge [1]. This approach emphasizes the importance of providing learners with comprehensible input that is slightly above their current level, allowing them to gradually expand their linguistic abilities. One way to achieve this in language learning is by providing materials whose readability is suitable for the learners' level.

Readability refers to the degree to which a piece of writing is understood and absorbed by its readers [2]. Greater readability reduces the effort required for reading and enhances the reading speed for all readers, especially for those who may struggle with comprehension [3]. This, in turn, boosts academic performance and leads to higher educational outcomes [4, 5, 6]. Therefore, foreign language teachers need to assess the readability of the text materials they provide to learners to make sure they add just enough new information to facilitate learning (+1 in Krashen's terms). In this paper we use the terms readability, complexity and difficulty loosely and interchangeably, even though we are well aware that technically they are not the same concept. Complexity hereinafter refers mostly to linguistic sophistication and is taken to mean the inverse of readability (the higher the complexity, the less readable a text is to read), while difficulty is used mostly in the context of overall cognitive effort required to read a text.

EvalLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy

*Corresponding author.

ignatios.charalampidis@uni-tuebingen.de (I. Charalampidis); xiaobin.chen@uni-tuebingen.de (X. Chen)
0000-0003-3158-0899 (X. Chen)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4006/paper9.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Though essential, readability assessment remains a laborious task for educators. This necessitated the creation of Automatic Readability Assessment (ARA) software and algorithms, whose goal is to evaluate the readability of a given text automatically. There are multiple ways to perform ARA, including simple formulas, machine learning (ML) and linguistic features, and more recently with the use of neural networks. Simple formulas rely on surface-level (or raw) features, such as the number of syllables and sentence length or the number of infrequent words in a text [7, 8, 9, 10]. In order to capture linguistic complexity more systematically and thoroughly other studies proposed the use of ML models for classification and a wide variety of linguistic features with high levels of success [11, 6]. Since there is high correlation between linguistic complexity and readability, linguistic features serve as a good proxy for readability and are therefore explored in this study.

ARA has been successfully applied on the document level, namely assessment of entire texts [12, 13, 14]. However, document-level readability suffers from lack of granularity; a text can be classified as a whole, but then no insight about its individual constituents is to be gained. This is problematic, for example, in cases where it is desirable to assess parts of a text in terms of complexity in order to simplify only those sentences that might not be readable enough for a particular audience (in this case students studying a foreign language). That is why research has shifted towards sentence-level ARA more recently [15, 16, 17]. To achieve sentence-level ARA previous research has leveraged judgments made by language experts or very proficient speakers and used them to train ML ARA models [18, 17, 15]. Nonetheless, previous studies relied on the judgments of a few annotators, who provided a single, holistic score of readability. For language learning purposes, it would be useful to tease apart the various factors that come into play when assessing readability, most notably vocabulary and grammar. That will provide insight into which features influence perception of readability the most, thereby leading to better ARA solutions.

In this study we aim to fill this gap by gathering sentence readability judgments from laypeople. Specifically, A corpus of 1000 sentences for human evaluation was curated [1]. The task was crowd-sourced and 50 participants (not language experts) were asked to assess random and diverse sentences on a scale of 1 to 100 for lexical, grammatical and overall difficulty separately. Then several ML classification and regression models were built and trained using the crowd's judgments as training data in an effort to discover which models perform best and how to best aggregate the intuitions of lay people into readability scores that can be used for training said models. We attempt to answer the following research questions:

1. Can lay people's intuitions be used to train ML models for ARA?
2. Which of the three criteria (lexical, grammatical and overall complexity) can be predicted best using ML?
3. Which ML models and features predict readability with the highest accuracy?

2. Related Work

ARA aims to automatically compute a score (either numerical or categorical) which shows how readable a text is. This can be achieved in a plethora of ways, including simple formulas, machine learning (ML) and linguistic features, and more recently with the use of neural networks. Simple formulas rely on surface-level (or raw) features, such as the number of syllables and sentence length or the number of infrequent words in a text [7, 8, 9, 10]. However, readability is influenced heavily by many factors, mostly grammatical and lexical, which raw text features alone cannot capture. In order to capture readability more systematically and thoroughly other studies proposed the use of ML models for classification and a wide variety of syntactic, lexical-morphological and semantic features [11, 6, 12] with high levels of success.

Several studies focused on document-level ARA, where the goal is to classify the readability level of an entire text. In [13, 19, 20] linguistic and surface-level features were successfully used to train ML

[1] All data and code will be made available soon at https://github.com/IgChar/sentence-readability-crowd

classification models, with reported accuracy reaching 93.3%, 98.12% and 80% respectively. Using a combination of linguistic features and the output of BERT-like [21] neural network models, the authors in [14] were able to achieve an almost perfect accuracy of 0.99 on the OneStopEnglish corpus [22], and an accuracy of 0.9 on WeeBit corpus [13].

These studies reported impressive results for document-level readability classification. However, for the purposes of SLA, a more robust and granular approach to ARA is necessary, especially if the end goal is text simplification and/or complexification of individual sentences. To accommodate this necessity other studies have focused on sentence-level ARA, thereby addressing the issue of lack of granular readability assessment. Various approaches and setups have been proposed for sentence-level ARA.

In [11, 16, 23] ARA was framed as a pairwise comparison between sentences and several linguistic features were used to determine which sentence is more difficult to read. All studies reported accuracy scores above 78%.

Other studies attempted to extract absolute readability scores for sentences without relying on comparisons. [15, 24, 18] achieved this by building sentence corpora and asking evaluators to assess sentence readability subjectively on a Likert scale. Then the authors used linguistic features to predict those scores, achieving high accuracy and Quadratic Weighted Kappa (QWK) scores [25] and revealing which features influence readability the most according to people's perception.

Most notably, [17] constructed the CEFR-based Sentence Profile (CEFR-SP), a collection of sentences classified by two language professionals into the six CEFR levels (A1, A2, B1, B2, C1, C2) . Unlike previous methods, they used BERT [21], for classification in two setups: 1) They extracted neural features from the sentences (embeddings) and built SVM and KNN classifiers with them and 2) they used BERT directly as a classifier. They then compared those methods and found that the BERT classifier outperformed the other methods with a QWK of 0.628 (rather high performance, indicating high agreement between ground truth and predicted values), but failed to accurately predict A1 and C levels, probably due to the dataset being unbalanced.

The integration of neural and linguistic features has also been explored for ARA on both document and sentence levels [26, 14]. [14] investigated two distinct approaches: 1) they combined sentences with linguistic features and fed the resulting array into a wide range of BERT-like neural networks for classification purposes and 2) they leveraged the output of these models as an additional feature in an XGBoost model, which was found to outperform other machine learning (ML) models in their experiments. Furthermore, the authors' hybrid approach, which utilized the output of BART-large [27], in conjunction with linguistic features, achieved an accuracy score of 0.729.

Previous studies have employed various approaches to ARA at the sentence level, leveraging linguistic, neural, and raw text features to achieve high assessment accuracy. However, there remain areas for further investigation. Specifically, previous research has primarily employed a single score or categorical value as a ground truth label to represent readability. Assessing people's perception of lexical and grammatical difficulty separately could provide more nuanced and fine-grained assessments of sentence readability. Furthermore, most prior work has relied on idealized sentences from official sources, which are often devoid of ungrammatical, ill-formed, or colloquial structures and vocabulary. As a result, systems trained on such data may struggle to generalize in contexts where more natural language is used. To fill these gaps 1) we recruited non-experts, both native and non-native speakers of English and explored how they perceive lexical, grammatical and overall sentence difficulty separately, 2) we used a diverse corpus of sentences covering more colloquial and possibly idiomatic language use and 3) we explored the efficacy of a variety of handcrafted linguistic and neural features and methods in producing reliable ARA models.

3. Study Design and Motivation

We theorize that complexity is not an inherent universal property of text, but rather one that depends on the receiver. Language experts perceive readability differently from second language (L2) learners and lay people with varying proficiency levels. What is more, while experts (e.g. foreign language teachers) may share a common understanding of linguistic complexity amongst themselves, lay people's individual differences, language proficiency level and lack of proper training may cause them to have vastly different opinions on complexity and readability, so much so that modeling a common perception of readability among this heterogeneous group using ML is impossible. To test this hypothesis we collected judgments of complexity from lay people with the purpose of training ARA models on them. The assumption is that if lay people's judgments on the same sentences are too diverse with a high deviation, than ML trained on this data will fail to converge. Conversely, if said models achieve an adequate performance, then this hints at the possibility that the collected data can be used to model some common perception or ground truth that the general crowd share. Further, we argue that extracting the ground truth from lay people's judgments is possible by leveraging the wisdom of the crowd [28], namely the idea that a large heterogeneous group of people can collectively make decisions and predictions that are often more accurate than those made by individuals or experts. This is based The aim of this study is to explore whether the intuitions of the crowd (lay people) can be modeled and captured using various ML methods and features, as well describe which methods resulted in higher performance.

3.1. Data Collection

We collected evaluations from 50 participants for 1000 sentences in English and compiled a the Sentence Readability Crowd corpus (Sr-Crowd). A data collection tool was made in the form of a website where participants would log in in order to assess the complexity of SR-Crowd. Participants recruited via the platform www.prolific.com (hereinafter Prolific) were shown 200 random sentences each and they were asked to evaluate them in terms of grammatical complexity, lexical complexity and overall reading difficulty (henceforth referred to as V,G and O respectively) on a scale from 1 to 100, where 1 is the lowest difficulty and 100 is the highest, resulting in 10,000 evaluations in total, with each sentence being evaluated by 10 random participants. The 1-100 scale was selected because it provides more flexibility and leeway for assessors, enabling them to express their subjective judgments about sentence readability with greater precision than a traditional Likert scale would allow.

Prior to data collection each participant was presented with 15 example sentences and their respective evaluations done by linguists and language experts. These model evaluations act as ground-truth for the participants to learn from and calibrate their assessments accordingly. In other words, the participants were first primed on the examples and then they were asked to evaluate the difficulty of random sentences based on the examples they had seen. While the goal of the study was to examine the intuitions of the crowd, some example sentences and their respective indicative difficulty scores had to be presented to the participants in order to prevent them from overestimating or underestimating the complexity of sentences. We believe that showing participants a few examples is a middle-ground solution between training them to do readability and complexity assessment and letting them do the assessment without any criterion or standard whatsoever. The former would defeat the purpose of leveraging lay people's intuitions while the latter would render it more likely for participants to underperform in the task by judging sentences too quickly without thinking. By presenting some sentences and their respective indicative complexity scores as examples we minimize this probability while also avoiding to impose strict standards that would bias or cloud the participants' intuitive judgment.

Regarding the creation of SR-Crowd, many available corpora options were considered. The aim was to collect data from sources as diverse as possible to ensure that the results reflected the full range of English language usage. One dataset that fulfilled this criterion is the C4 corpus [29], which was specifically designed for training large language models. As it already included a broad range of documents and websites scraped from the internet, it was deemed suitable for this purpose. In order to promote diversity even more in our sentence corpus, we drew on a specially curated subset of the C4 corpus, dubbed Repset by Suzuki et al. (2023), [30]. This subset was created to aid in faster training of LLMs with less data and is therefore expected to contain diverse language and content. It was chosen for this study because it met these criteria. The sentences to be evaluated were randomly sampled from the Repset corpus and then randomly assigned to participants of the study, with the only criterion being sentence length. Each participant received a balanced mixture of short and long sentences, thereby increasing the likelihood that all participants got sentences of varying levels of readability, since sentence length correlates with perceived complexity. This allowed participants to assess the full spectrum of linguistic complexity.

Several ML classification and regression models were built and trained using the crowd's judgments as training data. In the following sections we describe the data we collected, how the crowd-sourced judgments for each sentence were aggregated into a single readability label for V, G and O, and which features and models were used for classification.

3.2. Data Description

The participants in this study came from a wide range of linguistic and proficiency backgrounds, including various native languages and levels of English proficiency. The sample was divided roughly evenly between individuals who were born and raised speaking English (27 participants) and those whose primary language at home was not English (23). As far as gender is concerned, 27 participants were male and 23 were female. Participants reported a total of 15 different native languages: English (en), Polish (pl), Isizulu (zu), Portuguese (pt), Greek (el), Spanish (es), Dutch (nl), Hungarian (hu), Italian (it), Urdu (ur), Xhosa (xh), Shona (sn), French (fr), German (de), Unknown (xx). Table 1 shows the distribution of speakers. As for the participants' self-reported proficiency levels, most participants reported a general proficiency of C1, followed by C2, B2 and finally B1. interestingly, some participants claimed that their native language is English, but reported a general proficiency of less than C2 level. Proficiency levels are summarized in Table 2.

Table 1
Count of Native Speakers per Language

Language	en	pl	zu	pt	el	es	nl	hu	it	ur	xh	sn	fr	de	xx
Count	17	6	5	4	3	3	2	2	1	1	1	1	1	1	1

Table 2
Count of Participants per General Proficiency Level

Proficiency Level(CEFR)	Count
C1	23
C2	16
B2	8
B1	3

Here a few statistical characteristics of SR-Crowd are mentioned. We report the descriptive statistics of the dataset as a whole, as well as those of each criterion separately (Vocabulary (V), Grammar (G) and Overall (O) sentence difficulty). Table 3 summarizes the mean, variance, std, range, kurtosis, skew and inter-quartile range for all participants for V,G,O. These values do not change significantly across the different criteria, which indicates that there is high correlation between V,G and O in natural language; sentences that are lexically complex often present grammatical complexity as well. Likewise the same values are reported across sentences in Table 4.

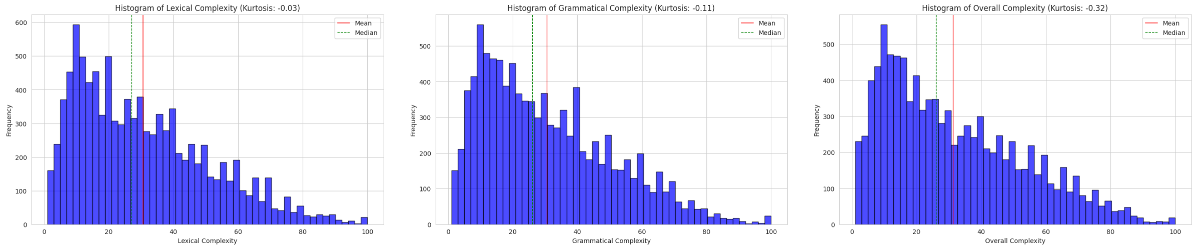
One property of the data is skewness to the left. Participants did not make use of the full range (1-100), but rather gravitated towards lower values. The mean and median of the dataset are located around 37, with values above 70 being very rare and values below 30 being more common. This is best illustrated in Figure 1, where the kurtosis values for all three axes are displayed. Apparently participants generally deemed sentences as simple and easy to understand, maybe partly because they were not trained or specifically instructed before data collection and probably due to their high language proficiency.

Table 3
Mean of descriptive statistics for all criteria (V,G,O) (Across Users)

Criterion	mean	variance	std	range	kurtosis	skew	q1	median	q3	iqr
Lexical (V)	30.2	231.83	14.49	74.24	2.30	1.00	20.18	28.35	39.20	19.02
Grammatical (G)	30.45	231.76	14.39	72.80	2.27	0.95	19.93	28.44	39.13	19.20
Overall (O)	31.02	247.46	14.74	72.34	1.97	0.92	20.26	28.94	39.88	19.61

Table 4
Mean of descriptive statistics for all criteria (V,G,O) (Across Sentences)

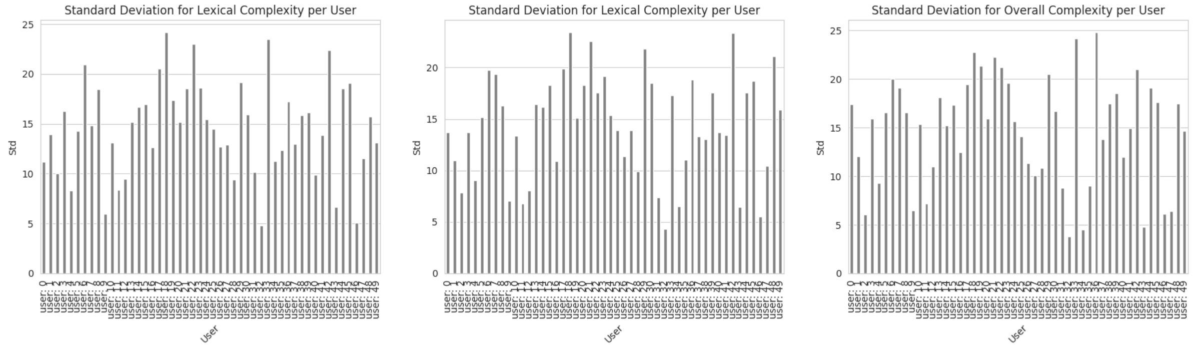
Criterion	mean	variance	std	range	kurtosis	skew	q1	median	q3	iqr
Lexical (V)	30.55	336.16	17.63	55.84	-0.58	0.47	17.27	28.04	41.45	24.18
Grammatical (G)	30.45	326.16	17.39	54.94	-0.62	0.44	17.41	28.40	41.31	23.90
Overall (O)	31.23	367.79	18.52	57.98	-0.64	0.45	17.05	28.84	42.92	25.87



(a) Lexical Complexity (b) Grammatical Complexity (c) Overall Complexity

Figure 1: Kurtosis Plots for Lexical, Grammatical and Overall Complexity

To reduce the effects of variability, all evaluations that deviated by 2 std above or below the mean were not considered during model building. This can be further illustrated in Figure 2 which displays the std for all axes per participant. Some participants gave consistently low scores for most sentences, which is unlikely to reflect the true intuition of the crowd, since one would expect at least some of those sentences to be considerably more difficult to read and understand than others. A low std among evaluations of different sentences by the same participant is an indication that said participant underperformed.



(a) Lexical Difficulty (b) Grammatical Difficulty (c) Overall Difficulty

Figure 2: STD Plots for Lexical, Grammatical and Overall Complexity (per Participant). Some participants seem to have too little variation in their assessments

Finally, we hypothesize that participants who are native speakers of English may have evaluated most sentences as easy to understand because they are very proficient themselves. Here the differences between the two groups (native vs non-native speakers of English) are examined. Any participant who has a std below the threshold of 10 is shown in Table 5. Most participants who showed little variation among their assessments are indeed English speaking. Except for participant with userid 13, all others' assessments are on the lower end of the scale, as evident by the mean values for the three axes.

Table 5
Participants with STD Lower than 10

userid	proficiency	language	mean_std	lexical_mean	gram_mean	overall_mean	lex_std	gram_std	overall_std
3	c1	Hungarian	7.97	15.00	14.50	13.00	10.01	7.87	6.04
5	c1	English	8.90	15.00	17.00	18.00	8.31	9.06	9.32
10	c2	Greek	6.51	8.00	9.00	8.00	5.98	7.05	6.50
12	c2	Portuguese	7.47	12.00	13.00	16.00	8.38	6.81	7.21
13	c1	English	9.48	45.00	31.00	57.00	9.45	8.03	10.98
32	c2	Polish	8.79	11.00	10.00	11.00	10.16	7.39	8.84
33	c2	English	4.32	5.00	5.00	5.00	4.83	4.34	3.79
35	c2	English	7.43	18.50	11.00	10.00	11.24	6.53	4.52
44	c1	Isizulu	5.98	4.00	5.00	5.00	6.69	6.45	4.79
47	c2	English	5.59	14.00	15.00	14.00	5.08	5.55	6.13
48	c2	English	9.45	20.00	10.00	5.00	11.52	10.42	6.41

This discrepancy between the two groups can be further illustrated in the box plots in Figure 3. Native speakers had a tendency to evaluate sentences as less complex than the non-native group did. Native speakers' mean was generally lower than the non-native group's and their std also. This may indicate a fundamental difference between the two groups which would prove that readability is subjective and it should be modeled based on who is perceiving it.

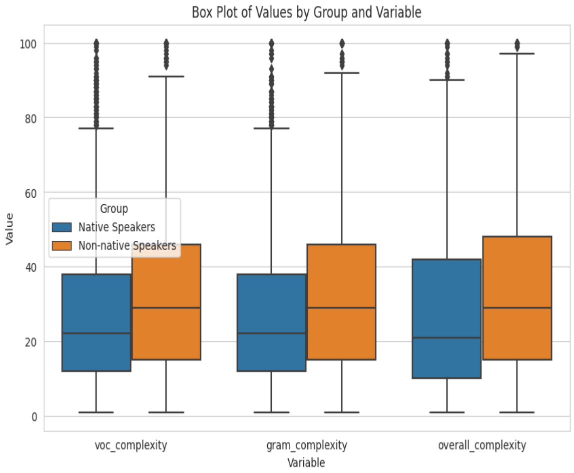


Figure 3: Native vs Non-native Evaluations Box Plot

3.3. Data Preparation and Scaling

Since the data collected is in the range of 1 to 100, any model trained on it would have to produce a continuous variable in the same range as an ARA score. This means that ARA scoring is framed as a regression problem, which in turn means that the MSE loss should be used for the training and evaluation of the models presented in this study. However, for the sake of comparison with other baselines that frame ARA as a classification problem and for future studies that will aim to compare results among methods, this study reports classification scores, namely QWK, accuracy and Macro F1. In order to achieve this comparability, our data had to be quantized into discrete classes from 0 to 5, reflecting CEFR levels. That way more representative scores can be reported that are compatible with the results of other studies (e.g. [17, 31]).

Turning a range of 1 to 100 into a 6-tier scale requires thresholding to optimal values. One approach would be to discretize based on 7 cutoff values separated linearly, effectively splitting the 1 to 100 range into 6 equally spaced bins. However, we noticed that the mean of evaluations is not centered around 50, but around 37. Moreover, most participants refrained from scoring higher than 67 on average, which means that higher scores are rare. To address this imbalance, we used thresholds obtained thusly: First we averaged all values per sentence per criterion and then calculated equal length bins starting from the minimum of these averages up to the maximum. That way the data was split into 6 bins with an unequal amount of sentences. We chose this method because it splits the data into a distribution similar to that of real texts. That is to say, sentences of lower-to-medium complexity (A2, B1 and B2) are more frequent, while very simple or very challenging sentences (A1, C1 and C2) are less frequently encountered. For brevity we show the cutoff values and the number of sentences per bin only for the Overall Difficulty criterion (O) in Table 6.

Table 6
Label distribution for SR-Crowd after thresholding

	0(A1)	1(A2)	2(B1)	3(B2)	4(C1)	5(C2)
Thresholds (O)	6.4	16.58	26.76	36.95	47.13	57.31
Number of Sentences	48	291	396	193	64	8

Using the first method of thresholding yields a Gaussian distribution of data, with A1 and C2 having the least amount of sentences and B1 having the most. Another factor we experimented with is averaging multiple assessments into a single score for each criterion (V, G and O) for each sentence. Since every sentence was evaluated by multiple participants, we had to decide how to average those judgments into a single score that is most representative of the sentence's difficulty. Concretely, we considered 1) averaging values per sentence per criterion first and then thresholding (hereinafter referred to as AVG THRESHOLD) and 2) thresholding to discrete levels first and then taking majority vote per sentence per criterion (referred to as MAJORITY VOTING). In both cases outliers were ignored. It was found that AVG THRESHOLD yielded the best results, so we chose to run our experiments using ground truth labels extracted with this method.

4. Model Building and Criterial Features

Several models were trained on the collected data and a total of four general pipeline variations were built: 1) Traditional ML models trained solely on linguistic features as input, 2) traditional ML models trained on neural embeddings as input, 3) BERT-like models fine-tuned for classification on SR-Crowd, 4) A pipeline that combines BERT-like models' scores & embeddings and linguistic features with a traditional ML model as the backbone. As far as traditional ML modeling is concerned an XGBoost model was used. Preliminary experiments showed that XGBoost achieves better performance than other methods (SVM, Random Forests), so we conducted our main experiments with it. The neural models tested are a) MiniLM [32], a lightweight sentence transformer, b) BGE-M3 [33], a multilingual BERT transformer, DeBERTa-v3-large [34], a state-of-the-art Transformer often used for classification with peak performance, and ModernBERT [35], a modernized bidirectional encoder-only Transformer capable of encoding long contexts.

4.1. ML Models with Linguistic Complexity Features

Pipeline#1 is a classification setting that involves criterial feature extraction as done in [24, 15]. The CTAP platform [36] was used for the extraction of linguistic complexity features. CTAP extracts a comprehensive list of 576 syntactic, lexical and morphological complexity features which can be used to classify texts in terms of linguistic complexity. Since complexity is highly correlated with readability, those features serve as a good proxy of readability. This pipeline achieves a QWK score of 0.5 across all three criteria. More specifically, XGBoost was used with 576 CTAP features as input and the aggregated participants' judgments as output.

As can be seen in Table 7, XGBoost with CTAP features a relatively low QWK and moderate accuracy and F1 scores. It is worth noting that XGBoost model's predictions tend to center around the value of 2,

Table 7
Performance metrics for XGBoost + CTAP Features

CTAP + XGB	Accuracy	Macro F1	QWK
Lexical	0.488	0.47	0.552
Grammatical	0.51	0.485	0.56
Overall	0.508	0.485	0.485

which is the most common level found in the SR-Crowd after thresholding the participants' assessments. This may make the model not generalize well for more diverse datasets.

4.2. ML Models with Embeddings as Features

Neural embeddings are often used for classification, clustering and retrieval tasks. To test their efficacy in classification of sentence difficulty we extracted embeddings using three state-of-the-art models: 1) jina-embeddings-v3 [37], b) OpenAI's text-embeddings-3-large and c) Gemma 2b-it [38] [2]. After the embeddings were extracted they were fed into an XGBoost model.

As can be seen in Table 8, results were very poor, indicating that neural embeddings alone do not capture readability accurately. Even though F1 score and Accuracy are around 0.4, QWK is universally too low, indicating that the model produces almost random predictions centered around the most common level found in the dataset. This could be attributed to the fact that neural embeddings are optimized to retain semantic information, not information about grammatical or lexical complexity.

Table 8
ML model performance with embeddings as features for Lexical (V), Grammatical (G) and Overall (O) difficulty

Model	Criterion	Accuracy	Macro F1	QWK
Jina SVC	V	0.4	0.3	0.2
	G	0.42	0.321	0.09
	O	0.475	0.396	0.256
Gemma SVC	V	0.375	0.35	0.239
	G	0.405	0.375	0.26
	O	0.405	0.405	0.312
OpenAI SVC	V	0.44	0.375	0.274
	G	0.41	0.347	0.197
	O	0.395	0.35	0.29

4.3. BERT Model Training

Pipeline#2 involved the use of BERT-like models directly as classifiers by attaching a classification head on the Transformer model, without extracting features or using other ML models. Every BERT-like model was trained using the adam optimizer in a 3-fold cross-validation setting with different seeds for train-test splitting. For training efficiency the larger BERT models were trained using parameter efficient LoRAs [39], which greatly reduce training time while performing comparably to normal training, as was found in our preliminary experiments. Only MiniLM was trained normally without the use of LoRAs. All models were trained with a batch size of 64 (factoring in gradient accumulation steps).

Contrary to [17], we found that treating ARA as a regression problem and then thresholding to optimal cutoff values (similar to the method used to aggregate assessments described above) worked best. This works because CEFR levels (unlike other categorical labels) are ordinal values, which means that each level indicates higher complexity than the previous one. As for thresholding, optimal cutoff values were found using Scipy's minimize function in Python, which finds optimal parameters for a given function. We also tried the Python library Optuna, which discovers optimal hyperparameters for

[2] the Gemma implementation can be found here https://huggingface.co/trapoom555/Gemma-2B-Text-Embedding-cft

ML models, but found Scipy's minimize to work decently for the purposes of this study. In Table 9 we show an overview of the models' performance.

Table 9
ML model performance with embeddings as features for Lexical (V), Grammatical (G) and Overall (O) difficulty

Model	Criterion	Accuracy	Macro F1	QWK
MiniLM	V	0.39	0.26	0.65
	G	0.445	0.288	0.647
	O	0.445	0.33	0.66
BGE-M3	V	0.48	0.35	0.632
	G	0.47	0.33	0.64
	O	0.46	0.31	0.623
DeBERTaV3	V	0.42	0.324	0.651
	G	0.435	0.267	0.63
	O	0.5	0.34	0.664
ModernBERT	V	0.505	0.394	0.662
	G	0.55	0.53	0.685
	O	0.47	0.29	0.672

In terms of QWK all models seem to peak at around 0.65. However, bigger models produce values in the full range of 0 to 5, whereas MiniLM produced values centered around 2 and 3, as these were most common in the dataset. ModernBERT performed best with an accuracy of 0.55 for grammatical difficulty prediction, while being the most efficient and fast to train and run inference with. DeBERTa shows promise and with better hyperparameter tuning it could reach a higher performance, but the computational cost and training time make it an unappealing choice for future experiments.

4.4. Combining Linguistic and Neural Features with BERT scores

Previous studies have demonstrated that combining linguistic and neural features can lead to better performance [26, 14, 40]. To test whether this holds true for SR-Crowd, pipeline#3 combined all features extracted with the previously described methods into a single traditional ML model. ModernBERT's score combined with CTAP features achieved the highest overall accuracy of 0.48, with the grammatical criterion prediction performing best at 0.48 accuracy and 0.7 QWK. The Macro F1 score of 0.364 indicates a reasonable balance across classes, while the QWK score of 0.7 reflects a strong level of agreement between predicted and actual levels. This suggests that the combination of ModernBERT and CTAP effectively leveraged the strengths of both methods.

On the other hand, combining neural and linguistic features in an XGBoost classifier did not yield any better results for SR-Crowd. The model performed the same as or even worse than the baseline that uses CTAP features alone. What is more, predictions are unbalanced with levels 1-3 being overrepresented.

Finally, pipeline#4 combines BERT scores from ModernBERT, CTAP features and embeddings from embedding models, but does not yield better results. Table 10 shows an overview of the results.

Table 10
ML model performance with embeddings + CTAP features

Model	Criterion	Accuracy	Macro F1	QWK
ModernBERT (score) + CTAP	V	0.47	0.41	0.703
	G	0.48	0.364	0.7
	O	0.43	0.36	0.69
CTAP + text-embedding-3-large	V	0.491	0.455	0.5
	G	0.547	0.51	0.55
	O	0.52	0.484	0.525
ModernBERT (score) + CTAP + text-embedding-3-large	V	0.43	0.42	0.65
	G	0.46	0.43	0.67
	O	0.425	0.424	0.68

Overall, the results indicate that combinations of models, such as ModernBERT with CTAP features and XGBoost, tend to yield better performance metrics compared to single approaches. Further, it is hereby proven that fine-tuning these models can be done cheaply and efficiently by harnessing newer methods (e.g. LoRA) thereby dramatically decreasing training and inference costs.

5. Conclusions

This study presented a new dataset of sentence difficulty evaluations performed by non-experts in the field of linguistics or teaching. The dataset was analysed and several ML models were trained on it for sentence-based ARA. It was found that participants had a tendency to evaluate sentences as easy to understand, with the average of their evaluations being 37 out of 100. This tendency was found to be slightly more pronounced for Native-speakers. Furthermore, the std across participants for the same sentences was rather high, indicating that agreement was low among participants assigned to evaluate those sentences. Since evaluations were in a range of 1-100 and because we want our models and methods to be comparable to those in related works, we had to convert this scale to one between 0 and 5, roughly representing CEFR levels, as other works do. Then ground truth labels were extracted by first removing outliers and then averaging assessments for V, G and O separately. as it was found to lead to higher convergence in the trained models.

As regards mapping the 1-100 range of assessments to the CEFR scale (0-5, A1-C2), it was found that using AVG THRESHOLD as described in section 3.3 resulted in splitting of the data into bins roughly representing a Gaussian distribution; most sentences were classified as levels 2-3 while fewer sentences were classified as levels 0, 1, 4 and 5. This is commensurate with our expectations, as we empirically assumed that sentences sampled from random lay texts would follow this distribution. This assumption, however, should be tested more thoroughly in the future using corpus analysis. In any case, our method of aggregating the crowd's judgments to produce ground truth labels and train our models on them proved to be significantly more effective than MAJORITY VOTING, as models trained on these aggregated judgments had a much higher QWK score.

As for modeling, BERT-like models performed moderately well in sentence difficulty prediction, and definitely better than ML models trained on linguistic complexity features from CTAP alone. However, the combination of BERT regression scores and CTAP features using an XGBoost model produced the best outcome with a QWK of 0.7 for grammatical difficulty prediction and an average score of 0.67 for V and O. This indicates that the best pipeline, which combines linguistic features with a regression score from ModernBERT, is not significantly better or worse predicting any of the three criteria (V,G and O). Further, our statistical analysis shows that participants' judgments on these criteria may be codependent; A sentence that is lexically difficult may influence lay people's perception of grammar and overall difficulty as well.

Finally, the best model's performance is moderately high, indicating that readability is highly subjective and the readability judgments of one group of people may not reflect those of others due to individual differences. To develop effective ARA models, it is crucial to train them on the unique intuitions and perceptual norms of specific target audience groups, rather than relying solely on expert opinions or those of highly proficient speakers. Examining the divide between expert and non-expert judgments is a goal for a future study. However, using the intuitions of the crowd as training data seems to lead to generalizable ARA models that capture the general crowd's perception.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. D. Krashen, Principles and Practice in Second Language Acquisition, Pergamon Press, Oxford, 1982.

[2] E. Dale, J. S. Chall, A formula for predicting readability: Instructions, Educational Research Bulletin 27 (1948) 37–54.

[3] W. H. DuBay, The Principles of Readability, Impact Information, Costa Mesa, CA, 2004.

[4] H. A. E. Mesmer, Tools for Matching Readers to Texts: Research-Based Practices, Guilford Press, New York, 2007.

[5] K. E. Stanovich, Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy, Journal of education 189 (2009) 23–55. doi:10.1177/0022057409189001-204.

[6] Y.-T. Sung, W.-C. Lin, S. B. Dyson, K.-E. Chang, Y.-C. Chen, Leveling l2 texts through readability: Combining multilevel linguistic features with the cefr, The Modern Language Journal 99 (2015) 371–391. doi:https://doi.org/10.1111/modl.12213.

[7] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (1975).

[8] J. S. Chall, E. Dale, Readability Revisited: The New Dale-Chall Readability Formula, Brookline Books, Cambridge, MA, 1995.

[9] D. R. Smith, et al., The Lexile Scale in Theory and Practice. Final Report, Technical Report ERIC ED307577, MetaMetrics, Inc., Washington, DC, 1989. Available from ERIC: https://eric.ed.gov/?id=ED307577.

[10] W. Dubay, Unlocking language: The classic readability studies, Professional Communication, IEEE Transactions on 51 (2009) 416–417. doi:10.1109/TPC.2008.2007872.

[11] S. Vajjala, D. Meurers, Assessing the relative reading level of sentence pairs for text simplification, in: S. Wintner, S. Goldwater, S. Riezler (Eds.), Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 288–297. URL: https://aclanthology.org/E14-1031/. doi:10.3115/v1/E14-1031.

[12] I. Pilán, E. Volodina, R. Johansson, Rule-based and machine learning approaches for second language sentence-level readability, in: J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 174–184. URL: https://aclanthology.org/W14-1821/. doi:10.3115/v1/W14-1821.

[13] S. Vajjala, D. Meurers, On improving the accuracy of readability classification using insights from second language acquisition, in: J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 163–173. URL: https://aclanthology.org/W12-2019/.

[14] B. W. Lee, Y. S. Jang, J. Lee, Pushing on text readability assessment: A transformer meets handcrafted linguistic features, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10669–10686. URL: https://aclanthology.org/2021.emnlp-main.834/. doi:10.18653/v1/2021.emnlp-main.834.

[15] D. Brunato, L. De Mattei, F. Dell'Orletta, B. Iavarone, G. Venturi, Is this sentence difficult? do you agree?, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2690–2699. URL: https://aclanthology.org/D18-1289/. doi:10.18653/v1/D18-1289.

[16] B. R. Ambati, S. Reddy, M. Steedman, Assessing relative sentence complexity using an incremental CCG parser, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp.

1051–1057. URL: https://aclanthology.org/N16-1120/. doi:10.18653/v1/N16-1120.

[17] Y. Arase, S. Uchida, T. Kajiwara, CEFR-based sentence difficulty annotation and assessment, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6206–6219. URL: https://aclanthology.org/2022.emnlp-main.416/. doi:10.18653/v1/2022.emnlp-main.416.

[18] L. Seiffe, F. Kallel, S. Möller, B. Naderi, R. Roller, Subjective text complexity assessment for German, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 707–714. URL: https://aclanthology.org/2022.lrec-1.74/.

[19] F. Dell'Orletta, S. Montemagni, G. Venturi, READ–IT: Assessing readability of Italian texts with a view to text simplification, in: N. Alm (Ed.), Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 73–83. URL: https://aclanthology.org/W11-2308/.

[20] M. Xia, E. Kochmar, T. Briscoe, Text readability assessment for second language learners, in: J. Tetreault, J. Burstein, C. Leacock, H. Yannakoudakis (Eds.), Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, San Diego, CA, 2016, pp. 12–22. URL: https://aclanthology.org/W16-0502/. doi:10.18653/v1/W16-0502.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[22] S. Vajjala, I. Lučić, OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification, in: J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, H. Yannakoudakis (Eds.), Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 297–304. URL: https://aclanthology.org/W18-0535/. doi:10.18653/v1/W18-0535.

[23] F. Dell'Orletta, M. Wieling, G. Venturi, A. Cimino, S. Montemagni, Assessing the readability of sentences: Which corpora and features?, in: J. Tetreault, J. Burstein, C. Leacock (Eds.), Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 163–173. URL: https://aclanthology.org/W14-1820/. doi:10.3115/v1/W14-1820.

[24] S. Stajner, S. P. Ponzetto, H. Stuckenschmidt, Automatic assessment of absolute sentence complexity, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 4096–4102. URL: https://doi.org/10.24963/ijcai.2017/572. doi:10.24963/ijcai.2017/572.

[25] R. Bakeman, J. M. Gottman, Observing interaction: An introduction to sequential analysis, Cambridge university press, 1997.

[26] T. Deutsch, M. Jasbi, S. Shieber, Linguistic features for readability assessment, in: J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, T. Zesch (Eds.), Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 1–17. URL: https://aclanthology.org/2020.bea-1.1/. doi:10.18653/v1/2020.bea-1.1.

[27] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703/. doi:10.

18653/v1/2020.acl-main.703.

[28] J. Surowiecki, The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations, Doubleday, New York, 2004.

[29] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[30] J. Suzuki, H. Zen, H. Kazawa, Extracting representative subset from extensive text data for training pre-trained language models, Information Processing & Management 60 (2023) 103249. URL: https://www.sciencedirect.com/science/article/pii/S0306457322003508. doi:https://doi.org/10.1016/j.ipm.2022.103249.

[31] T. Naous, M. J. Ryan, A. Lavrouk, M. Chandra, W. Xu, ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 12230–12266. URL: https://aclanthology.org/2024.emnlp-main.682. doi:10.18653/v1/2024.emnlp-main.682.

[32] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. doi:10.48550/arXiv.2002.10957.

[33] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. URL: https://aclanthology.org/2024.findings-acl.137/. doi:10.18653/v1/2024.findings-acl.137.

[34] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. doi:10.48550/arXiv.2111.09543.

[35] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[36] X. Chen, D. Meurers, CTAP: A web-based tool supporting automatic complexity analysis, in: D. Brunato, F. Dell'Orletta, G. Venturi, T. François, P. Blache (Eds.), Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119. URL: https://aclanthology.org/W16-4113/.

[37] S. Sturua, I. Mohr, M. K. Akram, M. Günther, B. Wang, M. Krimmel, F. Wang, G. Mastrapas, A. Koukounas, A. Koukounas, N. Wang, H. Xiao, jina-embeddings-v3: Multilingual embeddings with task lora, 2024. URL: https://arxiv.org/abs/2409.10173. arXiv:2409.10173.

[38] G. Team, Gemma (2024). URL: https://www.kaggle.com/m/3301. doi:10.34740/KAGGLE/M/3301.

[39] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: ICLR 2022, 2022. URL: https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

[40] F. Liu, J. Lee, Hybrid models for sentence readability assessment, in: E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. Tack, V. Yaneva, Z. Yuan, T. Zesch (Eds.), Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 448–454. URL: https://aclanthology.org/2023.bea-1.37/. doi:10.18653/v1/2023.bea-1.37.