Language Resources and Evaluation (2025) 59:2265–2296
https://doi.org/10.1007/s10579-024-09800-5

ORIGINAL PAPER

Automatic readability assessment for sentences: neural, hybrid and large language models

Fengkai Liu[1] · **Tan Jin**[2] · **John S. Y. Lee**[1]

Accepted: 6 December 2024 / Published online: 9 February 2025
© The Author(s), under exclusive licence to Springer Nature B.V. 2025, modified publication 2026

Abstract

Automatic readability assessment (ARA) aims to determine the cognitive load of a reader to comprehend a given text. ARA research has been mostly conducted at the text level, with numerous studies showing strong performance of neural and hybrid models. ARA at the sentence level, however, has received less attention, even though many applications in natural language processing (NLP) require assessment of the difficulty of individual sentences. This article compares the performance of neural models, hybrid models and large language models (LLMs) for sentence-level ARA, making three main contributions. First, we construct the first Chinese sentence-level ARA datasets, with nearly 70K sentences, to facilitate evaluation on Chinese. Second, we present the first experimental results on applying LLMs to sentence-level ARA. Finally, while previous work focused mostly on English data, we show that hybrid models outperform traditional classifiers, neural models, and LLMs in both English and Chinese data. The best hybrid model obtained state-of-the-art results on the Wall Street Journal dataset, surpassing the previous best result by almost 15% absolute. It also achieved competitive results on the CEFR-SP dataset. In detailed analyses, we identify the linguistic features that most significantly contributed to the performance of the hybrid model. We show that 10 linguistic features are correlated to readability across all datasets, and models trained on this reduced feature set achieve performance that rivals the full set. These results not only yield new insights into hybrid models for sentence-level ARA, but also set new benchmarks for future research in both English and Chinese.

Keywords Sentence readability assessment · Automatic readability assessment · Hybrid models · Large language models

Extended author information available on the last page of the article

1 Introduction

Text readability is defined as the cognitive load of a reader to comprehend a text (Martinc et al., 2021). Early research on automatic readability assessment (ARA) aimed to determine the difficulty level of a passage using lexical and syntactic features, e.g., readability formulas (Fry, 1968). Later work considered a larger range of linguistic features related to text comprehension to enhance accuracy, such as text cohesion and semantics (Nahatame, 2021). Recent research has shown growing interest in assessing the readability of individual sentences (Stajner et al., 2017; Brunato et al., 2018; Lu et al., 2020a; Schicchi et al., 2020), due to its applications in various downstream tasks in natural language processing (NLP). It is essential to difficulty-sensitive language generation tasks, such as generating pedagogical material and exercises (Pilán et al. 2014a). Sentence-level ARA also contributes to the field of explainable text simplification (Garbacea et al., 2021), by identifying sentences that are prime candidates for simplification, and evaluating sentence simplification (Yang et al., 2023). The challenge and necessity of sentence-level ARA as a standalone task have also been motivated by the substantial performance degradation when passage-level ARA models are used on individual sentences (Kilgarriff et al. 2008; Pilán et al. 2016).

Neural networks have been applied to the modeling of passage-level ARA (Filighera et al., 2019; Nadeem and Ostendorf, 2018), both as a classification or a ranking problem. They generally obtained better performance than traditional statistical classifiers (Todirascu et al., 2016). Recent research has demonstrated that hybrid models, which leverage both linguistically motivated features and predictions from neural models, can achieve superior results. Large language models (LLMs), such as LLaMa, ChatGPT (OpenAI, 2021) and other open-sourced models, have also been applied to passage-level ARA (Shi et al., 2023), although there are not yet consistent conclusions on their performance. As for sentence-level ARA, prior studies have appraised traditional classifiers grounded in linguistic features (Deutsch et al., 2020a) and machine learning methods (Brunato et al., 2018), but neural models have been reported to outperform these approaches (Arase et al., 2022; Schicchi et al., 2020). To the best of our knowledge, there has been only one paper on the use of hybrid model on sentence ARA (Liu and Lee, 2023). The evaluation in that study did not consider LLMs, and was limited to English and in-domain data. To fill this gap, we leverage LLMs for sentence-level ARA, analyzes cross-domain performance and conducts evaluation with both Chinese datasets and additional English datasets.

This article makes three contributions to sentence-level ARA. First, we construct the first sentence-level ARA datasets for Chinese. With nearly 70K sentences covering different genres, these two large-scale datasets will facilitate training and evaluation in this field. Second, we are the first to apply large language models (LLMs) on sentence-level ARA, and report evaluation on zero-shot and few-shot performance. Third, building upon our previous work (Liu and Lee, 2023), we conduct more comprehensive experiments and analyses. Our expanded evaluation encompasses an additional English dataset based on CEFR (Arase et al., 2022), and an additional

language, namely Chinese. We show that hybrid models outperform traditional classifiers, neural models, and LLMs in both languages. The best hybrid model obtained state-of-the-art results on the Wall Street Journal dataset, surpassing the previous best result by almost 15% absolute. It also achieved competitive results on the CEFR-SP dataset. In a detailed statistical analysis on the hybrid models, we identify the linguistic features that most significantly contributed to their performance. Further, we show that 10 linguistic features are correlated to readability across all datasets, and models trained on this reduced feature set achieve performance that rivals the full set. These results not only demonstrate the effectiveness of hybrid models in sentence-level ARA, but also set new benchmarks for future research in both English and Chinese.

2 Literature review

This section describes the state-of-the-art in automatic readability assessment (ARA), with particular attention to sentence-level ARA.

2.1 Passage-level automatic readability assessment (ARA)

Readability formulas (Fry, 1968; Kincaid et al., 1975) and traditional assessment methods have primarily focused on discrete linguistic features presented by statistical measurements, such as word length, frequency, and sentence length (Collins-Thompson, 2008; Sung et al., 2015). Despite efforts to incorporate additional features such as cohesion and semantic information, few models have outperformed readability formulas (Todirascu et al., 2016). With advances in statistical Natural Language Processing (NLP), readability assessment began to be framed as a classification, regression or ranking task (Martinc et al., 2021), facilitated by manually constructed datasets such as Newsela (Xu et al., 2015), OneStopEnglish (Vajjala, S., and Lučić, I. 2018) and KRES (Logar et al., 2020). For new domains or languages, Grießhaber et al. (2020) investigated the use of domain adversarial training, which can prevent overfitting in low-resource and zero-resource contexts. The application of various word embedding methods and neural model architectures on these datasets has yielded improvement over traditional classifiers trained on linguistic features (Mikolov et al., 2013; Pennington et al., 2014). Recurrent neural networks (RNN), hierarchical attention networks (HAN), and transfer learning have obtained excellent performance on passage-level ARA (Cho et al., 2014; Yang et al., 2016; Vaswani et al., 2017). Azpiazu and Pera (2019) introduced a multiattentive RNN architecture designed to emphasize particular sections of a text for the purpose of determining its readability assessment. Mohammadi and Khasteh (2019) employed deep reinforcement learning on multilingual dataset, with a deep convolutional recurrent double dueling Q network (Wang et al., 2016), and achieved competitive performance on par with the state-of-the-art. For handling long texts, Deng et al. (2021) proposed a novel model called the attention-based BiLSTM fused

CNN with a gating mechanism (ABLG-CNN) and demonstrated commendable performance on a range of Chinese datasets.

An area of active recent research is the incorporation of linguistic features into neural models. Deutsch et al. (2020a) augmented deep learning models with 86 linguistically motivated features on the WeeBit (Vajjala and Meurers, 2012) and Newsela, but observed no significant improvement on passage-level ARA. Filighera et al. (2019) reported only slight enhancements, while Lee et al. (2021) noted substantial improvement with Random Forest and RoBERTa. Imperial (2021) explored the concatenation of linguistic features with sentence embeddings derived from BERT's hidden layers. Liu et al. (2020) introduced Tensor Graph Convolutional Networks (TensorGCN), a novel method designed to incorporate semantic, syntactic, and sequential contextual data into diverse graph types for the purpose of text classification. This approach outperformed the Text GCN model alone (Yao et al., 2019). A burgeoning field of study involves evaluating the efficacy of newly introduced Large Language Models (LLMs), like ChatGPT, on text classification tasks due to their demonstrated excellence in a wide array of language-related tasks. Shi et al. (2023) employed ChatGPT for text classification by transforming structural knowledge into a graph, but observed a performance decline compared to prior results. Sun et al. (2023) introduced Clue And Reasoning Prompting (CARP) strategy, whose performance is comparable to supervised neural models.

2.2 Sentence-level automatic readability assessment (ARA)

Fry (1990) noted the unreliability of readability formulas when applied to shorter passages, making it necessary to develop other ARA approaches for short texts. Pilán et al. (2016) also observed a reduction of approximately 7% accuracy when ARA models are trained with sentences in the same manner as passages. Sentence-level ARA was recognized as a distinct task from passage-level ARA and was applied to the selection of appropriate vocabulary example sentences (Pilán et al. 2014a). This task warrants comprehensive investigation given its applicability not only in Natural Language Processing (NLP) downstream tasks, such as the development of sentence simplification evaluation metrics (Yang et al., 2023), the interpretation of text simplification (Garbacea et al., 2021), and the facilitation of controllable text simplification (Arase et al., 2022), but also in Computer-Assisted Language Learning (CALL) applications, such as generating sentence-level multiple choice exercises (Volodina et al., 2014), selecting appropriate sentences for unfamiliar words (Segler, 2007), and facilitating pedagogy materials and exercises (Pilán et al. 2014a).

Early investigations into sentence readability primarily pursued binary classification or pairwise difficulty prediction. Ambati et al. (2016) implemented an incremental CCG parser to determine the simpler sentence within a pair. Schumacher et al. (2016) leveraged a collection of lexical and grammatical features to predict pairwise relative sentence complexity and found that contextual information improved performance. A hybrid algorithm combining rule-based and statistical classifiers yielded a 71% accuracy rate in the binary classification of texts for learning Swedish as a

foreign language (Pilán et al. 2014a). Statistical classifiers achieved 66% accuracy on an English dataset derived from Wikipedia and Simple Wikipedia (Vajjala and Meurers, 2014) and between 78.9% and 83.7% on an Italian dataset (Dell'Orletta et al., 2014). Bosco et al. (2021) presented DeepEva, which leverages Treetagger annotation tool and LSTM layers to classify the complexity of Italian and English sentences as easy or hard to understand. Several studies have also explored multi-way sentence-level ARA using conventional machine learning methods. Brunato et al. (2018) developed an SVM linear regression model that incorporates a range of surface, morphological and syntactic features. The model achieved 59.1% and 60% accuracy on Italian and an English datasets, respectively, graded on a 7-point scale. Sentence length and nominal modification were identified as significant correlates of sentence difficulty. A Bayesian Ridge Regression Model, trained on a variety of linguistic features including syntax, lexical, morphology and cohesion, attained a high level of correlation with human evaluations of German sentence difficulty (Weiss and Meurers, 2022). A classifier trained on features extracted from the phrase complexity level of n-grams achieved a 0.66 weighted F-score on an English dataset graded on a 5-point scale (Stajner et al., 2017). A classifier for Chinese sentences, based on vocabulary and grammar points, reached 31.92% accuracy on 10-way classification (Lu et al., 2020a).

There have been two notable studies that have implemented neural models for sentence-level ARA. Schicchi et al. (2020) demonstrated that an RNN-based architecture surpassed the performance of Vec2Read (Mikolov et al., 2013). Arase et al. (2022) reported that the BERT-base model outperformed conventional machine learning classifiers on their annotated CEFR-based sentence difficulty dataset. In the first study on hybrid models for sentence-level ARA, Liu and Lee (2023) showed that they outperformed neural models. This preliminary study was limited to two relatively small English datasets. Furthermore, a more in-depth corpus analysis of linguistic features should be undertaken to enhance our understanding of sentence readability assessment. This paper seeks to address this gap by comparing hybrid models, neural models, and traditional classifiers on larger datasets across different text genres and two languages. The performance of large language models will also be evaluated.

3 Data

This section introduces the datasets used in our study. These included three publicly available English corpora (Sect. 3.1), and two newly constructed Chinese corpora (Sect. 3.2).

3.1 English datasets

3.1.1 CEFR-based sentence profile (CEFR-SP)

This corpus (Arase et al., 2022) is composed of 10k English sentences, sourced from the Wiki-Auto dataset (Jiang et al., 2020), and the Sentence Corpus of Remedial

Table 1 Statistics on the CEFR-SP dataset

Level	# sentences	%	Sent. Length	
			Mean	SD
1	124	1.24	8.52	3.18
2	1,271	12.70	8.98	3.08
3	3,305	33.04	12.99	4.79
4	3,300	33.29	16.95	5.63
5	1,744	17.43	18.22	5.97
6	230	2.30	18.84	5.84
Total	10004	100.00	14.79	6.02

English (SCoRE) (Chujo et al., 2015) .[1] These sentences have been graded by professionals in English language education according to the Common European Framework of Reference (CEFR) for Languages scale, which spans from level 1 to level 6 (with level 1 corresponding to A1 and level 6 to C2). Detailed statistics are shown in Table 1.

3.1.2 Wall street journal (WSJ)

This corpus (Brunato et al., 2018) comprises 1,200 sentences drawn from the Wall Street Journal (Nivre et al., 2007). Each sentence was annotated by 20 native speakers on a difficult scale from 1 ("very easy") to 7 ("very difficulty") . Our evaluation focuses on sentences that garnered a consensus on their difficulty rating from at least 14, 15, 16, or 17 out of the 20 raters. These sentences were shown to produce superior performance in the original study Brunato et al. (2018) .[2] Detailed statistics can be found in Table 2.

3.1.3 OneStopEnglish (OSE)

This corpus (Vajjala, S., and Lučić, I. 2018) consists of aligned texts graded at three reading grades: beginner, intermediate, and advanced. There are 189 original texts, each with three different versions tailored to match these respective reading levels, which amount to 19,904 sentences across the 567 text variants.[3].

Instead of assigning a uniform difficulty level to all sentences within a text (Pilán et al. 2014b), we determined the difficulty of each individual sentence based on its revision, or lack thereof. Within the intermediate-level texts, 10.21% of the sentences are identical to those in the beginner texts; similarly, 18.76% of sentences in the advanced texts are copied verbatim from one of the simpler versions. These sentences were tagged with the lowest reading level at which they are found. The

[1] We were not able to access the Newsela-Auto dataset via the public repository at https://github.com/yukiar/CEFR-SP. Our results will be based on the aforementioned datasets only.

[2] None of these sentences was graded at level 6 or 7.

[3] The segmentation of sentences was conducted utilizing the NLTK toolkit (Bird et al., 2009)

Table 2 Statistics on the WSJ dataset

Level	# sentences	%	Sent. Length		Level	# sentences	%	Sent. Length	
			Mean	SD				Mean	SD
WSJ14					WSJ15				
1	69	10.62	10.43	1.42	1	60	12.63	10.50	1.51
2	262	40.31	14.51	4.66	2	237	49.89	14.05	4.36
3	203	31.23	25.00	6.17	3	113	23.79	24.60	6.30
4	96	14.77	30.70	5.49	4	51	10.74	30.59	4.96
5	20	3.08	31.50	8.77	5	14	2.95	31.43	6.91
Total	650	100.00	20.27	8.77	Total	475	100.00	18.40	8.40
WSJ16					WSJ17				
1	54	15.30	10.46	1.46	1	47	20.01	10.53	1.56
2	211	59.78	13.75	3.94	2	161	68.80	13.40	3.72
3	61	17.28	24.90	6.65	3	20	8.55	22.45	7.78
4	23	6.52	31.74	4.42	4	5	2.14	32.00	4.47
5	4	1.13	32.50	5.00	5	1	0.42	35.00	0.00
Total	353	100.00	16.56	7.68	Total	234	100.00	14.09	5.74

remaining sentences were assigned the same difficulty level as the text. Since these sentences were revised by the editors, that implies that they could not be of a lower difficulty level. Detailed statistics are shown in Table 3.

These three English corpora originated from the domain of text simplification, which aims to assist non-native learners and individuals with readability disorders. The OSE corpus, derived from a parallel corpus for text simplification, employed human text-level revision to assign sentence readability labels. Different from OSE, the sentences in the CEFR-SP and WSJ corpora were individually annotated by experts on a difficulty scale. Although the CEFR-SP and WSJ corpora share a common annotation scheme, they differ significantly in terms of text genre. The CEFR-SP corpus comprises numerous texts sourced from Wikipedia, which adopt an expository writing format and may incorporate evidence-based descriptions to convey knowledge and explain concepts. In contrast, the WSJ corpus consists of newspaper materials. The variety of annotation schemes and text genres can facilitate a comprehensive evaluation of our proposed models.

Table 3 Statistics on the OSE dataset

Version	# sentences	%	Sent. Length	
			Mean	SD
Begin.	4,834	33.98	18.77	8.60
Inter.	4,759	33.46	22.44	10.24
Advan.	4,631	32.58	25.91	12.24
Total	14224	100.00	22.31	10.84

3.2 Chinese datasets

We constructed the first publicly available Chinese datasets for sentence-level ARA. While Chinese ARA datasets have been developed at the passage level (Qiu et al., 2021; Zeng et al., 2022), none is publicly available at the sentence level.

3.2.1 Chinese L1 corpus (CLC)

We constructed a sentence-level dataset for ARA in Chinese, targeting native learners (L1). This dataset was derived from a collection of passages sourced from textbooks used in Chinese language education for native speakers (Cheng et al., 2019). To ensure a well-distributed dataset, we randomly selected passages from each grade such that there are about 1,000 sentences representing each grade, from Grade 3 through 12. This procedure yielded 9,700 sentences. Detailed statistics are shown in Table 5.

Due to the often indistinct demarcations between educational grades, assigning a precise grade or level to an individual Chinese sentence can be difficult. To mitigate the subjectiveness, instead of directly judging the level of each sentence (Arase et al., 2022; Brunato et al., 2018), we asked annotators to label the relative difficulty of each sentence with respect to the rest of the passage:

In-grade The sentence is at the expected level of difficulty given the grade of the passage;
Below-grade The sentence is easier than the in-grade sentences;
Above-grade The sentence is harder than the in-grade sentences.

Table 4 shows three example sentences taken from a Grade 3 passage. The top sentence was judged to be typical, with multiple subjects from *tangmu* 'Thomas' to *hua* 'speech' and *shengyin* 'voice'. The middle sentence has a simpler structure, with only subject *shaonian* 'youth'. The bottom sentence is more difficult to read due to the complex modifiers for the noun *xinyin* 'trust' and for the noun *xiyue* 'joy'.

We recruited two native speakers of Mandarin to perform the annotation. One annotator was a certified Chinese-language teacher with over 26 years of experience, and the other was an undergraduate student majoring in linguistics. On average, they identified 7.90% of sentences as being more complex (above-grade) and 29.07% as less complex (below-grade) relative to the passage's grade. The two annotators reached a consensus on 82.72% of the sentences, with a Cohen's Kappa coefficient of 0.61, indicating a substantial level of agreement (Landis and Koch, 1977). They reconciled the label differences through discussion to reach the final labels.

As depicted in Fig. 1a, the proportion of above-grade sentences tends to decrease with increasing grade levels. In most grades (except Grade 3), more sentences were classified as below-grade than above-grade. The percentage of below-grade sentences fluctuated across grades, ranging from 41.26 to 12.17%. Only the 6,106 sentences that were annotated as *in-grade* --- and were hence tagged with the

Table 4 Example sentences taken from a Grade 3 passage, annotated with their difficulty relative to Grade 3

Sentence	Difficulty
汤姆开头有点吞吞吐吐，渐渐地，话越来越多，声音也越来越大，越来越自然了。 'Thomas stuttered at first, but gradually his speech got longer, his voice grew louder and more natural.'	in-grade
少年连连摆手，用不太标准的中国话说："不，不要钱。 'The youth waved his hands and said with an accent, "No, I don't want money." '	below-grade
在那里，我们得到的是人与人之间的信任和被信任的喜悦。 'There, what we get is the trust between people and the joy of being trusted.'	above-grade

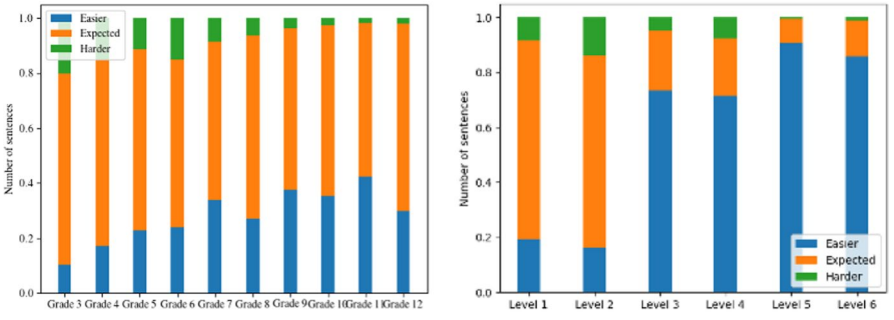
Reaction conditions: **1** (0.5 mmol), CuPT (5 mol%), KOH (25 mol%), DMSO (1 mL), 25 °C, air; isolated yield

[a]CuPT (10 mol%).

corresponding grade of the passage --- were employed for model training. Detailed statistics are shown in Table 5.

3.2.2 Teaching chinese to speakers of other languages (TCSOL) corpus

In addition to our work on the first language (L1) dataset, we developed a sentence-level Chinese dataset tailored for non-native language learners, utilizing passages from textbooks designed for teaching Chinese to non-native speakers (Bo et al., 2019). Each passage within these textbooks was categorized from Level 1 through Level 6, according to the standards of the Ministry of Education of the People's Republic of China in 'International Curriculum for Chinese Language Education' (Hanban/Confucius Institute Headquarters, 2008). We randomly selected 500 passages from each level, resulting in a compilation of 69424 sentences. For the annotation of the sentence-level ARA within the TCSOL corpus, we engaged twelve teaching assistants with considerable experience in instructing Chinese as a foreign language. The annotation went through three stages: familiarization, adjudication, and leveling.



(a) CLC corpus. (b) TCSOL corpus.

Fig. 1 Distribution of sentences annotated as in-grade, above-grade and below-grade at each grade

Table 5 Statistics on the CLC dataset

Grade	# sentences	%	Sent. Length	
			Mean	SD
3	565	68.32	24.49	12.13
4	683	70.41	26.75	12.88
5	516	65.15	28.70	14.73
6	582	63.82	32.04	15.78
7	535	58.09	33.31	18.36
8	612	66.74	37.04	19.62
9	604	56.19	37.59	22.34
10	604	56.19	37.59	22.34
11	892	56.10	30.74	16.92
12	651	66.91	48.47	27.71
Total	6,106	62.95	33.67	19.65

During the familiarization stage, the annotators were given a subset of passage samples to code as a preliminary exercise. This activity was designed to acquaint them with the content and context of the materials. Following this initial coding session, a coding scheme was established based on the collective input of the twelve annotators. Their discussions were further informed by the insights of a seasoned expert who had dedicated 30 years to teaching Chinese to foreign students. The annotation approach was similar to that used for the CLC corpus. Each sentence within the passages was assessed and designated as above-grade, below-grade, or in-grade. Differences in the labels assigned by the annotators were then resolved through discussion until a consensus was reached for the final labels. This collaborative and iterative process ensured that the annotations were both accurate and reflective of a shared understanding of sentence difficulty levels within the context of teaching Chinese to second language learners.

In the adjudication stage, twelve annotators were organized into three groups. Each group was tasked with annotating a set of 180 texts and subsequently cross-checking the levels assigned to each sentence. For any given text within the 180 assigned to a group, if the proportion of disagreement among the annotators reached or exceeded 10%, this discrepancy would trigger a third round of coding by a corpus linguist. The mean level information from the sentences was then utilized to refine the initially agreed-upon annotation schemes. They discussed the disagreements and worked together to establish a final, agreed-upon annotation scheme. The level of agreement among the annotators during this stage was quantified by a Pearson Coefficient Correlation of 0.58 and a Cohen's Kappa of 0.46, as cited by Landis and Koch (1977). These values indicate a moderate level of inter-annotator agreement. In the third and final stage, leveling, the twelve annotators proceeded to independently annotate the sentences in the remaining 2,400 texts. This was a substantial task, with each annotator responsible for labeling 200 texts.

Within all proficiency levels of the TCSOL corpus, the number of sentences deemed below-grade surpassed those considered above-grade. This pattern, which

Table 6 Statistics on the TCSOL dataset

Level	# sentences	%	Sent. Length	
			Mean	SD
1	7,863	72.31	16.30	21.24
2	6,826	69.95	19.45	10.95
3	2,602	21.95	29.41	17.29
4	2,604	20.93	29.78	15.72
5	1,176	8.81	45.66	26.04
6	1,467	13.22	36.97	23.98
Total	22538	32.46	23.20	19.95

is visually represented in Fig. 1b, mirrors that of the CLC corpus. Generally, as the proficiency level increases, there is a notable decrease in the proportion of sentences classified as above-grade. Conversely, the frequency of below-grade sentences demonstrates significant variability, ranging from a high of 90.55% to a low of 16.24%. For model training, only the 22,538 sentences that were identified as *in-grade* were utilized. Detailed statistics are shown in Table 6.

Despite the similarity in annotation scheme between the CLC and TCSOL corpora, the fundamental difference in their intended audience has an impact on the classification of sentence difficulty. The source texts in the CLC corpus are exemplar texts in Chinese language education for native speakers, whereas those in the TCSOL corpus are specifically designed for non-native Chinese language learners. These two datasets will help test the ability of our model on sentences intended for both native and non-native speakers.

3.3 Linguistic feature analysis

To investigate the relationship between linguistic features and sentence readability, linguistic features were extracted from the sentences in our datasets. For English, 255 features were extracted from LingFeat[4]; for Chinese, 221 features were extracted from ChiLingFeat[5,6] Five linguistic categories are represented in these features, namely: (a) Advanced Semantic, which aims to evaluate the intricacy of meaning structures; (b) Discourse, which assesses the coherence and cohesion of texts; (c) Syntactic, which gauges the intricacy of grammar and overall structure; (d) Lexico Semantic, intended to measure the difficulty of individual words and phrases; and (e) Shallow Traditional, encompassing conventional features and formulas used in assessing text readability.

To refine our feature set, we employed the Variance Threshold method from scikit-learn on the development set, setting the threshold at 0.8 to eliminate

[4] https://github.com/brucewlee/lingfeat

[5] https://github.com/ffliu6/ChiLingFeat

[6] ChiLingFeat was developed based on LingFeat and most of linguistic features share the same names.

Table 7 Corpus-invariant features, i.e. features found to be significant across all datasets and readability levels

Category	Abbre.	Correl.	Description
LxSem	1. MTLDTTR_S	0.207	Measure of Textual Lexical Diversity (default TTR = 0.72)
	2. SquaAvV_S	0.054	(Unique Adverbs**2)/total Adverbs (Squared AdVerb Variation-1)
Synta	3. ra_VeCoT_C	0.220	Ratio of Verb POS count to Coordinating Conjunction count
	4. ra_VeSuT_C	0.138	Ratio of Verb POS count to Subordinating Conjunction count
	5. ra_NoSuT_C	0.146	Ratio of Noun POS count to Subordinating Conjunction count
	6. ra_AvVeT_C	0.096	Ratio of Adverb POS count to Verb POS count
	7. ra_AvCoT_C	0.167	Ratio of Adverb POS count to Coordinating Conjunction count
	8. ra_NoVeT_C	0.199	Ratio of Noun POS count to Verb POS count
Disco	9. to_UEnti_C	0.264	Total number of unique Entities
	10. to_EntiM_C	0.268	Total number of Entities Mentions counts

features with low variance. Following this, we executed a one-way analysis of variance (ANOVA) on the remaining features to identify the linguistic characteristics with significant variances across different levels of sentence readability. In this analysis, readability levels were treated as the independent variable, while the linguistic features were the dependent variables. To meet the prerequisite of a normal distribution for the application of ANOVA, we identified and excluded outliers. Following Shiffler's (1988) recommendation, sentences that deviated by more than three standard deviations from the mean for any linguistic feature were considered outliers. This step ensured the integrity and validity of the ANOVA results.

Significant linguistic features were identified across different levels in various datasets .[7]In the CEFR-SP dataset, we identified 95 linguistic features that differed across readability levels. The OSE dataset revealed a total of 160 significant features, while analysis of the WSJ datasets yielded 147 significant features in WSJ14, 146 in WSJ15, 137 in WSJ16, and 115 in WSJ17. Within the CLC dataset, we discovered 105 significant linguistic features, and the TCSOL dataset presented with 30 such features. These results underscore the diverse linguistic attributes and trends that manifest across various datasets and readability levels.

Importantly, ten features were found to be significant across the sentence readability levels in all datasets. These features will henceforth be referred to as the "*corpus-invariant features*", and all other features as "*corpus-specific features*". Table 7 presents the corpus-invariant features and their Spearman correlation coefficients.[8] One of these ten features is the total number of Entities Mentions. Consider the following example sentence, which mentions twelve entities:

[7] $p < .001$

[8] We computed the correlation coefficients for each corpus-invariant feature across all datasets and then determined the average of these values to gain insights into the general trends and relationships between features and readability levels.

Aetna Life Casualty_ORG gained 1 3/8_CARDINAL to 61 1/8_CARDINAL , Cigna_GPE advanced 7/8_CARDINAL to 64 3/4_CARDINAL , Travelers_ORG added 1/2_CARDINAL to 40 3/8_CARDINAL and American International Group_ORG rose 3 1/8_CARDINAL to 107 3/4_CARDINAL .

As agreed by the annotators, it is difficult for readers to comprehend a sentence with so many entities.

4 Models

4.1 Baseline 1: Linguistic model

We employed the scikit-learn implementation of Random Forests (RF) and XGBoost (XGB) algorithms (Pedregosa et al., 2011). We selected linguistic features that exhibited notable differences across various levels (see Sect. 3.3). Following Lu et al. (2020b), we trained statistical classifiers on these linguistic features as well as bag-of-word features.

4.2 Baseline 2: Neural model

Transformer-based neural models have demonstrated impressive performance in many natural language processing tasks. For the English datasets (Sect. 3.1), we fine-tuned BERT (Devlin et al., 2019), BART (Lewis et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020) into an ARA classifier[9], using the pre-trained versions released by Huggingface (Wolf et al., 2019). We used the base versions of all of the above. For BART, RoBERTa and ELECTRA, we also used the large versions.

For the Chinese datasets (Sect. 3.2), we fine-tuned BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), along with MacBERT (Cui et al., 2020) and LERT (Cui et al., 2022) into an ARA classifier. We used the base versions of all of the above. For RoBERTa, MacBERT and LERT, we also used the large versions. All models were downloaded from Huggingface (Wolf et al., 2019).

4.3 Hybrid models

We implemented three hybrid models. The following model incorporates linguistic features into a neural model:

[9] We used the Adam algorithm (Kingma and Ba, 2015) for optimization. The epoch for each training is 10, and set the maximum word embedding size as 128. This setting is the same on Chinese datasets.

Concatenated Model Similar to Song et al. (2021), the input to the model consists of the input sentence $w_1 w_2 \ldots w_n$ concatenated with the linguistic features $f_1, f_2 \ldots f_n$, in the format "`[CLS]` $w_1 w_2 \ldots w_n$ `[SEP]` $f_1 f_2 \ldots f_n$".

The following two models wrap the linguistic features and neural model output in a non-neural statistical classifier:

Hard Label Following Deutsch et al. (2020b), the grade of the sentence, as predicted by the Neural Model (Sect. 4.2), serves as an additional feature in the statistical classifier (Sect. 4.1).
Soft Labels Following Lee et al. (2021), the probability of each grade, as predicted by the Neural Model (Sect. 4.2), serve as additional features alongside the linguistic features in the statistical classifier (Sect. 4.1).

4.4 Large language models (LLMs)

Large language models (LLMs), such as the OpenAI tool "Chat Generative Pre-Trained Transformer (ChatGPT)" (OpenAI, 2021), have showcased exceptional capabilities in comprehending and generating textual content. We deployed a range of prominent LLMs and carried out experiments with zero and few examples to investigate their ARA performance.

ChatGLM2-6B [10] is the second iteration of the open-source, bilingual (Chinese-English) chatbot model, ChatGLM-6B ,[11] with stronger performance and the capacity to consider more extensive contextual information. Utilizing a training approach akin to that of ChatGPT, ChatGLM-6B is specifically fine-tuned to optimize its performance for the Chinese language.

GPT-3.5-turbo-0613 [12] is the snapshot of GPT−3.5-turbo from 13 Jun 2023 and the replaced model for GPT−3.5-turbo-0301. This model will not receive updates. Our selection of this stable release is to guarantee that our results can be consistently reproduced in future research.

GPT-4–0613 [13] is the snapshot of gpt-4 from June 13th 2023. Despite OpenAI's assertion that the performance gap between GPT-4 and GPT−3.5 is marginal across numerous basic tasks, we proceeded to evaluate both models to determine whether GPT-4 could achieve superior results specifically in sentence-level ARA task.

Meta-Llama-3-8B[14] is the third generation of the open-sourced and foundation language model Llama.[15] This model has obtained the state-of-art performance on a wide range of industry benchmarks and outperformed its second generation.

[10] https://github.com/THUDM/ChatGLM2-6B
[11] https://github.com/THUDM/ChatGLM-6B
[12] https://platform.openai.com/docs/models/gpt-3-5
[13] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[14] https://huggingface.co/meta-llama/Meta-Llama-3-8B
[15] https://huggingface.co/docs/transformers/model_doc/llama

Table 8 Sample sentences of several corpus-invariant features. *Min. Level* stands for the minimum sentence readability level in the dataset and *Max. Level* represents the maximum one. *Sample sentence* presents the selected sample in its corresponded level

Zero-shot
You are a sentence classifier and your task is to assess the readability level of the given sentence into a number from {Min. Level} to {Max. Level}, in which {Min. Level} denotes easiest and {Max. Level} denotes hardest. You should directly output the assessed level only. You answer should be a number. Do not output any explanation. Sentence: {} Level:
你是一个句子分类器，你的任务是将给定句子的可读性级别评估为{Min. Level}到{Max. Level}之间的数字，其中{Min. Level}表示最容易，{Max. Level}表示最难。 你应该只直接输出评估的级别。你的答案应该是一个数字。不输出任何解释。 句子：{} 等级：
Few-shot
You are a sentence classifier and your task is to assess the readability level of the given sentence into a number from {Min. Level} to {Max. Level}, in which {Min. Level} denotes easiest and {Max. Level} denotes hardest. You should directly output the assessed level only. You answer should be a number. Do not output any explanation. Sentence: {Sample sentence} Level: {Min. Level} … Sentence: {Sample sentence} Level: {Max. Level} Sentence: {} Level:
你是一个句子分类器，你的任务是将给定句子的可读性级别评估为{Min. Level}到{Max. Level}之间的数字，其中{Min. Level}表示最容易，{Max. Level}表示最难。 你应该只直接输出评估的级别。你的答案应该是一个数字。不输出任何解释。 句子：{Sample sentence} 等级：{Min. Level} … 句子：{Sample sentence} 等级：{Max. Level} 句子：{} 等级：

Reaction conditions: **1** (0.5 mmol), CuPT (5 mol%), KOH (25 mol%), DMSO (1 mL), 25 °C, air; isolated yield

[a]CuPT (10 mol%).

Recent studies involving zero- and few-shot learning with pre-trained language models have yielded impressive results, particularly in contexts involving languages with limited resources (Mallinson et al., 2020; Feng et al., 2023). We initiated our investigation by conducting zero-shot experiments to establish baseline performances. Subsequently, we implemented the LLMs in a few-shot learning scenario to assess their capacity for acquiring specialized knowledge. Our research encompassed zero- and few-shot experiments using ChatGLM2-6B, GPT−3.5-turbo-0613, GPT-4–0613 and Meta-Llama-3-8B. The prompts, adapted from those used in a text classification study Shi et al. (2023), are shown in the Table 8. For each readability level within the corpus, the few-shot prompt provides one sample sentence. The following few-shot example is taken from OSE dataset:

> *You are a sentence classifier and your task is to assess the readability level of the given sentence into a number from 1 to 3, in which 1 denotes easiest and 3*

denotes hardest. You should directly output the assessed level only. You answer should be a number. Do not output any explanation.
Sentence: In the speech, he asked the country to join together.
Level: 1
Sentence: Both the Republican House Speaker, John Boehner, and the Democratic Leader in the Senate, Harry Reid, spoke about a need to work together to resolve the economic crisis.
Level: 2
Sentence: Up to 50 metres high, these steel-framed supertrees not only have owers and ferns growing up them but their metallic canopies act to absorb and disperse heat too.
Level: 3

5 Experimental settings

5.1 Set-up

We used stratified ten-fold cross validation for the WSJ, OSE, CLC and TCSOL datasets, with a 8:1:1 split for training, development and testing .[16] For the CEFR-SP dataset, we adhered to the trainset, devset, and testset constructed by (Arase et al., 2022). In the case of the CLC and TCSOL datasets, all sentences originating from the same text were grouped within the same fold, so that entities and topics in the test sentences would not be exposed to the model during training.

For both the Hard Label and Soft Labels models, we experimented with employing solely *corpus-invariant features* as well as in combination with *corpus-specific features*. This was done to explore the potential impact of *corpus-specific features* on further enhancing sentence-level automatic readability assessment (ARA). In addition, we conducted cross-domain evaluation on both Neural Models and Hybrid Models to evaluate the contribution of linguistic features in the sentence ARA task. For Large Language Models (LLMs), due to computational resource constraints, evaluation was performed on only one fold, randonly selected from each dataset.

5.2 Evaluation metrics

Our evaluation metrics include accuracy (Acc.), F1-score, precision, recall, and the Quadratic Weighted Kappa (QWK) score. For the CEFR-SP dataset, we report the Macro-F1 score to facilitate a direct comparison with previous results (Arase et al. (2022). McNemar's Test will be used throughout for calculating statistical significance in accuray.

[16] The hyperparameters for learning rate, dropout and batch size are tuned on the dev set.

Table 9 Performance of the sentence-level ARA models

Dataset	Metric	Linguistic	Neural	Hybrid Model		LLM
				Corpus-invariant	With Corpus-specific	
CEFR-SP	Macro-F1	0.252	0.625	**0.752**	0.747	0.318
	Acc.	0.489	0.773	0.810	**0.812**	0.370
	F1	0.451	0.761	0.795	**0.800**	0.371
	Prec.	0.490	0.760	0.812	**0.813**	0.436
	Recall	0.488	0.773	0.810	**0.812**	0.370
	QWK	0.338	0.876	0.875	**0.879**	0.629
WSJ	Acc.	0.709	0.717	0.721	**0.747**	0.540
	F1	0.694	0.691	0.698	**0.729**	0.568
	Prec.	0.697	0.696	0.702	**0.742**	0.624
	Recall	0.709	0.717	0.721	**0.747**	0.540
	QWK	0.766	0.791	0.800	**0.814**	0.773
OSE	Acc.	0.451	0.571	**0.587**	0.581	0.452
	F1	0.428	0.558	**0.576**	0.564	0.453
	Prec.	0.441	0.565	**0.583**	0.574	0.489
	Recall	0.451	0.571	**0.587**	0.581	0.452
	QWK	0.288	0.549	**0.568**	0.560	0.331
CLC	Acc.	0.206	0.470	0.492	**0.493**	0.239
	F1	0.139	0.461	**0.480**	0.478	0.224
	Prec.	0.177	0.468	**0.491**	0.491	0.211
	Recall	0.206	0.470	0.492	**0.493**	0.239
	QWK	0.067	0.556	**0.594**	0.592	0.019
TCSOL	Acc.	0.377	0.472	**0.489**	0.484	0.353
	F1	0.323	0.444	**0.477**	0.464	0.324
	Prec.	0.338	0.452	**0.524**	0.466	0.327
	Recall	0.377	0.472	**0.489**	0.484	0.353
	QWK	0.374	0.618	**0.682**	0.633	0.407

Bold numerical values indicate robust performance

In cross-domain experiments, since the train and test datasets have different difficulty scales, we report pairwise accuracy as follows. Sentence pairs (A, B) were randomly selected from the test dataset. If the predicted levels of A and B preserve the relatively order of their gold levels, the model is considered to be accurate.

6 Results

Table 9 shows the main results of the sentence-level ARA models. Since the four subsets of the Wall Street Journal (WSJ) data produced similar outcomes, only the WSJ14 subset is displayed.

6.1 Linguistic model

Across all evaluated datasets, the XGBoost (XGB) model consistently surpassed the performance of the Random Forest (RF) model. Specifically, on the CEFR-SP and OSE datasets, XGB attained accuracies of 0.252 and 0.451, respectively, outperforming RF's accuracies of 0.220 and 0.412. For the four WSJ datasets, XGB demonstrated a progressive increase in accuracy: 0.709, 0.726, 0.780, and 0.792, respectively. This performance was notably higher than that of RF, which scored 0.675, 0.690, 0.757, and 0.774, respectively. The performance difference between XGB and RF was narrower for the CLC and TCSOL datasets, with XGB achieving a marginal lead with accuracies of 0.206 and 0.377, as opposed to RF's 0.188 and 0.353. Given XGB's superior performance across both the English and Chinese datasets, we will focus on results from the XGB model in the subsequent discussion.

6.2 Neural model

The performance metrics detailed in Table 9 for the CEFR-SP and OSE datasets are derived from the XLNet and BART-large models, the respective best-performing transformers in the two datasets (see Table 15 in Appendix A). These models emerged as the top performers on the mentioned English datasets, as can be seen in Table 15 within Appendix A. XLNet delivered an impressive Macro-F1 score of 0.524 on the CEFR-SP dataset. Meanwhile, BART-large excelled in the WSJ dataset, achieving an accuracy of 0.791. Turning to the Chinese datasets, CLC and TCSOL, the LERT-large and RoBERTa-base models were the frontrunners, with accuracies of 0.470 and 0.472, respectively, as detailed in Table 14 in Appendix A.

Overall, the Neural Models outshone the Linguistic Models in terms of performance on the CEFR-SP, OSE, CLC, and TCSOL datasets, consistent with the trends observed in passage-level ARA studies. However, it's worth noting that for the WSJ dataset, the performance gap narrowed as the amount of training data decreases.

6.3 Hybrid model

Detailed results of the Hybrid Model can be found in Table 10.

6.3.1 Hard label vs soft labels vs concatenated model

For the CEFR-SP dataset, the Hard Label model achieved the highest Macro-F1 score, at 0.752. The Soft Labels Model produced the second best performance, with the Concatenated Model also showing commendable results. For the WSJ datasets, the Hard Label Model scored the highest accuracy (0.747), also followed by the Soft Labels Model and the Concatenated Model.

On the OSE dataset, the Soft Labels Model emerged as the top performer in terms of accuracy, albeit with a lower accuracy rate (0.587) compared to its performance on the WSJ dataset. This discrepancy in accuracy may be attributable to the less distinct boundaries between the difficulty levels of the sentences in the OSE

Table 10 Performance of the Hybrid ARA models

Dataset	Metric	Concatenated	Hard Label		Soft Labels	
			Corpus-invariant	With Corpus-specific	Corpus-invariant	With Corpus-specific
CEFR-SP	Macro-F1	0.391	**0.752**	0.747	0.611	0.664
	Acc	0.583	0.810	**0.812**	0.782	0.777
	F1	0.564	0.795	**0.800**	0.768	0.767
	Prec	0.584	0.812	**0.813**	0.764	0.778
	Recall	0.583	0.810	**0.812**	0.782	0.777
	QWK	0.804	0.875	**0.879**	0.859	0.867
WSJ	Acc	0.654	0.719	**0.747**	0.721	0.731
	F1	0.626	0.691	**0.729**	0.698	0.717
	Prec	0.634	0.697	**0.742**	0.702	0.732
	Recall	0.654	0.719	**0.747**	0.721	0.731
	QWK	0.764	0.786	**0.814**	0.800	0.799
OSE	Acc	0.568	0.579	0.578	**0.587**	0.581
	F1	0.559	0.566	0.565	**0.576**	0.564
	Prec	0.584	0.592	**0.593**	0.583	0.574
	Recall	0.568	0.579	0.578	**0.587**	0.581
	QWK	0.540	0.539	0.537	**0.568**	0.560
CLC	Acc	0.460	0.491	**0.493**	0.492	0.491
	F1	0.450	0.476	0.478	**0.480**	0.476
	Prec	0.469	**0.494**	0.491	0.491	**0.494**
	Recall	0.460	0.491	**0.493**	0.492	0.491
	QWK	0.560	0.579	0.592	**0.594**	0.579
TCSOL	Acc	0.462	**0.489**	0.484	0.483	0.481
	F1	0.450	0.477	0.464	**0.531**	0.460
	Prec	0.456	**0.524**	0.466	0.506	0.468
	Recall	0.462	**0.489**	0.484	0.483	0.481
	QWK	0.634	**0.682**	0.633	0.653	0.630

Bold numerical values indicate robust performance

data, which were produced by text revision levels. The Hard Label Model secured the second-best results on the OSE dataset, while the Concatenated Model also gave competitive performance.

Although it was marginally outperformed by the Soft Labels Model on the OSE dataset, the Hard Label Model consistently showed the most impressive results for the CEFR-SP, WSJ, CLC, and TCSOL datasets. It achieved a significantly important improvement over the Soft Labels Model on the CEFR-SP dataset, at $p < 2.4 \cdot e^{-2}$ according to McNemar's Test. In contrast, on the OSE dataset, the Soft Labels Model demonstrated a significant improvement over the Hard Label Model, at

$p < 2.0 \cdot e^{-4}$. Nevertheless, no significant differences were observed between these two models on the WSJ, CLC, and TCSOL datasets, with $p > .05$.

6.3.2 Corpus-invariant vs corpus-specific

Table 9 compares the performance of two sets of linguistic features, based on the best Hybrid Model. For the CEFR-SP dataset, the corpus-invariant features produced a higher Macro-F1 score (0.752). Notably, adding corpus-specific features to the model brought about only a minor increase in accuracy (from 0.810 to 0.812). Similarly, the inclusion of corpus-specific features slightly boosted the accuracy (from 0.492 to 0.493).

On the WSJ dataset, the corpus-specific features led to an increase in accuracy (from 0.721 to 0.747) for the Hybrid Model, a statistically significant improvement ($p < 3.1 \cdot e^{-2}$). As for CEFR-SP, CLC and TCSOL, the addition of corpus-specific features did not yield a statistically significant improvement.($p > .05$). Interestingly, for the OSE dataset, the model that incorporated only corpus-invariant features outperformed the one with corpus-specific features (accuracy 0.581 vs. 0.587), a statistically significant improvement at $p < 1.4 \cdot e^{-3}$. Overall, our findings suggest that the exclusive use of corpus-invariant linguistic features can already achieve excellent model performance. Additional, corpus-specific linguistic features, on the other hand, may lead to overfitting. Nonetheless, they may prove beneficial for a genre-specific dataset of sufficient size.

6.3.3 Hybrid model vs baselines

As shown in Table 9, Hybrid Model has demonstrated superior performance over the Neural Models and traditional classifiers across all examined datasets. Consistent with Liu and Lee (2023), the Concatenated Model performed worse than the Neural Model across datasets, likely because long complex sequences more easily overfit on the transformer-based models. The superiority of the best Hybrid Model over the Neural Model was found to be statistically significant for all five datasets ($p < .05$).

The Hybrid Model achieved 0.752 macro-F1 on the two publicly available subcorpora of the CEFR-SP dataset. The proposed model described in Arase et al. (2022) obtained the best macro-F1 on CEFR-SP, at 0.845, based on all three subcorpora. As we were not able to obtain the third subcorpus, a direct comparison with the Arase et al. (2022) model is unfortunately not possible.

Notably, the Hybrid Model posted 0.747 accuracy on the WSJ14 dataset, exceeding the formerly top accuracy of 0.600 attained by an SVM model (Brunato et al., 2018). The Hybrid Model also outperformed that model in the WSJ15 and WSJ16 datasets by a margin of 3% and 2% respectively, and achieved comparable results on the WSJ17 dataset. This slight difference in performance might be attributable to the small size and sampling size of datasets. For example, WSJ17 comprises only 234 sentences, including fewer than 5 sentences classified as level 4 and 5.

6.3.4 Cross-domain

As shown in Table 12, the Hybrid Model consistently exhibited superior performance compared to the Neural Model in the cross-domain setting. This suggests that some of the textual characteristics learned by the Neural Model may be peculiar to the domain of the training corpus. Since the corpus-invariant linguistic features likely have broader relevance across domains, the Hybrid Model was more effectively in transferring the learned patterns to new domains.

The models trained on the CEFR-SP and WSJ datasets demonstrated a high degree of generalizability across domains. For instance, the Hybrid Model trained on the WSJ dataset achieved a pairwise accuracy of 74.4% in predicting the readability of sentences in the CEFR-SP dataset, which utilizes a similar annotation schemes. On the other hand, the model trained on the OSE dataset appeared less generalizable to the CEFR-SP and WSJ datasets. This discrepancy can be attributed to the fact that the OSE dataset was derived from a parallel corpus for text simplification.

For Chinese, the models trained on the CLC dataset achieving a pairwise accuracy of 0.526 when predicting relative readability levels on the TCSOL dataset. In the opposite direction, the pairwise accuracy was 0.597. The relatively low accuracy could be due to differences in the intended audience: while the CLC dataset caters to native Chinese speakers, while the TCSOL dataset reflects text difficulty from the perspective of non-native learners.

6.4 Large language models (LLMs)

Table 9 outlines the performance of various LLMs on the sentence-level ARA task. For details, please refer to Table 16 in Appendix B.

For the WSJ dataset, the few-shot approach with GPT-4–0613 emerged as the top performer, achieving an accuracy of 0.540, with its zero-shot version following closely behind. Although the few-shot approach tends to yield improved performance over zero-shot setting in most of the cases, its efficacy can be compromised by the choice of samples. For example, on the OSE dataset, the zero-shot GPT-4–0613 outperformed its few-shot counterpart, reaching an accuracy of 0.452. Effective strategies for constructing few-shot prompts for Large Language Models (LLMs) in sentence-level ARA is an area that requires additional research and analysis.

In general, GPT-4–0613 achieved the best performance, with Meta-Llama-3-8B and GPT−3.5-turbo-0613 trailing behind, which conform with our expectation. Yet, on the whole, the Large Language Models (LLMs) did not achieve better results than the Hybrid Models and Neural Models. In some cases, their performance was even inferior to that of traditional linguistic classifiers (Table 9). The limited number of samples provided to the LLMs may not have enabled them to grasp the difficulty scale, which is essential to the ARA task. Fine-tuning the LLMs on the entire dataset could mitigate this issue. This was not attempted in our study because of constraints in computational costs, but would be well worth considering in future work.

Table 11 ELI5 feature weights of linguistic features led to better performance of Hybrid Models

Rank	Feature	Weight	Category	Description
1	ra_AvVeT_C	0.0162	Synta	Ratio of Adverb POS count to Verb POS count
2	ra_NoVeT_C	0.0163	Synta	Ratio of Noun POS count to Verb POS count
3	MTLDTTR_S	0.0142	LxSem	Measure of Textual Lexical Diversity (default TTR = 0.72)
4	to_UEnti_C	0.0129	Disco	Total number of unique Entities
5	to_EntiM_C	0.0101	Disco	Total number of Entities Mentions counts
6	ra_VeSuT_C	0.0070	Synta	Ratio of Verb POS count to Subordinating Conjunction count
7	ra_NoSuT_C	0.0057	Synta	Ratio of Noun POS count to Subordinating Conjunction count
8	ra_VeCoT_C	0.0019	Synta	Ratio of Verb POS count to Coordinating Conjunction count
9	ra_AvCoT_C	0.0006	Synta	Ratio of Adverb POS count to Coordinating Conjunction count

Table 12 Pairwise accuracy of cross-domain readability

Train Dataset	Test Dataset	Neural Model	Hybrid Model
CEFR-SP	WSJ	0.680	0.685
	OSE	0.618	0.619
WSJ	CEFR-SP	0.728	0.744
	OSE	0.596	0.603
OSE	CEFR-SP	0.702	0.702
	WSJ	0.638	0.638
CLC	TCSOL	0.519	0.526
TCSOL	CLC	0.589	0.597

6.5 Analysis on linguistic features

To understand why the hybrid model achieved the best result, it could be insightful to examine the linguistic features that enabled it to outperform the other models. This section investigates the features learned by the hybrid model that contributed most to its performance gains over the neural model. Using the hybrid model trained on the corpus-invariant linguistic features, we compiled the set of sentences for which the hybrid model performed better than the neural model, i.e., predicted a level that is closer to the gold.

The ELI5 tool, which stands for "explain like I'm five," is a Python package that helps explain the inner workings of complex machine learning models. We applied the ELI5 tool to analyze the influence of each feature on the predictions of the hybrid model. Table 11 presents the most important features, sorted according to their respective weights.

Table 13 Sample sentences that contain corpus-invariant features

Sentence	Neural	Hybrid	Gold
to_EntiM_C, total number of Entities Mentions Counts			
(a) 如果上来一个10元_MONEY 的起步价、大约10分钟_TIME 路程的乘客，加上每次载客之间的平均空驾时间7分钟_TIME，等于我花了17分钟_TIME 赚了10元_MONEY，而17分钟_TIME 的成本价格是9.8元_MONEY，不划算，20元_MONEY 到50元_MONEY 之间的生意性价比最高。	3	5	5
MTLDTTR_S, Measure of Textual Lexical Diversity (default TTR = 0.72)			
(b) A few months ago your paper reported the results of a study to determine why Asians who arrive in this country without any money, and unable to speak English, become overnight successes.	2	3	3
ra_VeSuT_C, ratio of Verb POS count to Subordinating Conjunction count			
(c) 只要_SCONJ 两_NUM 只_NUM 小_ADJ 狼_NOUN 一_ADV 碰到_VERB ，_PUNCT 它们_PRON 就_ADV 会_VERB 用_ADP 刚长_VERB 出来_VERB 的_PART 乳牙_NOUN 和_CCONJ 爪子_NOUN 来_PART 厮打_VERB ，_PUNCT 胜者_NOUN常用_VERB 脚踏_NOUN 在_VERB 败者_NOUN 的_PART 身上_NOUN ，_PUNCT 并_ADV 翘起_VERB 小_ADJ 尾巴_NOUN ，_PUNCT 非常_ADV 得意_VERB ；_PUNCT 而_ADV 败者_NOUN 则_ADV 躺_VERB 在_ADP 地上_NOUN ，_PUNCT 一面_ADV 摇尾_VERB 一面_ADV 喘气_VERB 地_PART 向_ADP 胜利者_NOUN 乞求_VERB “_PUNCT 宽恕_NOUN ”_PUNCT ，_PUNCT 显出_VERB 一_NUM 副_NUM 可_VERB 怜相_VERB 。_PUNCT	1	5	6

Reaction conditions: **1** (0.5 mmol), CuPT (5 mol%), KOH (25 mol%), DMSO (1 mL), 25 °C, air; isolated yield

[a]CuPT (10 mol%).

Feature	Weights
Textual Lexical Diversity (to_EntiM_C)	0.3342
Verb-to-subordinating-conjunction Ratio (ra_VeSuT_C)	0.2384
Noun-to-subordinating-conjunction Ratio (ra_NoSuT_C)	0.1911
Neural Pred.	0.1311

Fig. 2 Sentence explained weights sample

Our analysis revealed that 9 of the 10 corpus-invariant linguistic features contributed to the superior performance of the Hybrid Models. The only exception was the SquaAvV_S feature, which is concerned with adverb variation (see Table 7). Although this feature shows significant correlation with the sentence readability level, the neural model may have effectively learned this feature Table 12.

The sample sentences in Table 13 illustrate the contribution of these features. The readability of sentence (a), a Grade 5 sentence, is underestimated by the Neural Model to be at Grade 3. One reason is the large number of time and money expressions in the sentence, which could be hard for the reader to follow. This difficult aspect is captured by the linguistic feature to_EntiM_C, which tracks the number of entity mentions in the sentence. This feature thus helped nudge the Hybrid Model towards a higher estimation of readability, in this case successfully predicting Grade 5. Figure 2 illustrates the weights assigned to sample sentence (a) by the hybrid

model. In this particular sentence, the primary influential factor is identified as the measure of lexical diversity (to_EntiM_C), followed by the ratio of Verb POS count to Subordinating Conjunction POS count (ra_VeSuT_C).

The Neural Model also underestimates the difficulty of sentence (b), a Grade 3 sentence, to be Grade 2. This sentence utilizes varied vocabulary, with few repeated words, which could be challenging for a Grade 2 reader. The Hybrid Model, which considers lexical diversity as a feature (MTLDTTR_S), is able to predict the gold grade.

As for example sentence (c), a Grade 6 sentence, the Neural Model erroneously assessed it as Grade 1. This sentence stands out in its large number of verbs — it contains 16 verbs but only one subordinating conjunction. The ratio of verbs to subordinating conjunction enabled the Hybrid Model to make a close prediction (Grade 5), while the Neural Model does not consider this factor.

7 Conclusion

We have presented the first study that compares traditional classifiers, neural models, hybrid models and Large Language Models (LLMs) on the task of sentence-level automatic readability assessment (ARA). Building on our preliminary study on this topic (Liu and Lee, 2023), we have extended our experiments in terms of language, method and evaluation settings. First, we assessed the efficacy of various representative LLMs in both zero-shot and few-shot settings. The performance of LLMs on sentence-level ARA tasks was found to be lacking, which suggests a need for further exploration into the construction of prompts and refinement of fine-tuning approaches. Second, we have included an additional dataset in English as well as two newly annotated Chinese datasets, which facilitated the first sentence-level ARA evaluation on Chinese. Finally, for both the in-domain and cross-domain settings, and on both English and Chinese datasets, we demonstrate the superior performance of hybrid models in comparison to neural models, traditional classifiers and LLMs. This result suggests that linguistic features are crucial in capturing key elements that reflect sentence complexity, and are more generalizable to new domains. We analyzed the linguistic features that most significantly contributed to the performance of the hybrid model, and identified 10 linguistic features that are correlated to readability across all datasets. The best hybrid model obtained state-of-the-art results on the Wall Street Journal dataset, surpassing the previous best result by almost 15% absolute.

Our findings are expected to inform and guide future research in sentence-level ARA. However, caution must be exercised when applying these pretrained models and linguistic features to different contexts. The labeling approaches for assessing sentence readability vary across corpora, as they are influenced by annotators' subjective understanding of readability and their prior teaching experience. Consequently, further investigation and discussion are required to explore the theoretical framework of sentence-level readability and its distinction from text-level readability. It is worth noting that researchers have been investigating how humans process sentences using eye tracking and constructing psycholinguistic models (Howcroft

and Demberg, 2017; Mertzen et al., 2023; Prystauka et al., 2024). Factors such as source text genres, target audiences, and the relationship between sentences and their source texts should be taken into account. Nonetheless, this study offers a comprehensive analysis of bilingual sentence readability corpora and demonstrates superior model performance compared to previous baselines. Future research should delve deeper into the connection between psycholinguistic models and diverse human labeling approaches in different contexts. Additionally, it should explore the theoretical interactions and experimental applications between sentence-level and text-level readability.

Appendix A Appendix: Performance of Neural Models

Table 14 reports the performance of various neural models on the CLC and TCSOL datasets. LERT-large achieved the highest level of performance on the CLC dataset, while RoBERTa-base stood out as the superior model on the TCSOL dataset. Table 15 reports the performance of various neural models on the English dataset. XLNet outperforms the others on the CEFR-SP dataset, demonstrating the most effective performance. BART-large achieved the best results on the WSJ and the OSE datasets.

Table 14 Performance of neural models on CLC and TCSOL. Bold numerical values indicate robust performance.

Dataset	Metric	BERT Base	RoBERTa Base	ELECTRA Base	MacBERT Base	LERT Base	RoBERTa Large	MacBERT Large	LERT Large
CLC	Acc	0.459	0.455	0.414	0.454	0.462	0.464	0.462	**0.470**
	F1	0.456	0.450	0.408	0.449	0.450	0.453	0.454	**0.461**
	Prec	0.476	0.468	0.445	0.463	0.468	**0.472**	0.468	0.468
	Recall	0.459	0.455	0.414	0.454	0.462	0.464	0.462	**0.470**
	QWK	**0.573**	0.543	0.542	0.535	0.536	0.541	0.562	0.556
TCSOL	Acc	0.466	**0.472**	0.464	0.468	0.470	0.465	0.464	0.446
	F1	0.444	0.444	0.442	**0.445**	0.444	0.444	0.444	0.427
	Prec	0.447	**0.452**	0.440	0.446	0.450	0.450	0.447	0.435
	Recall	0.466	**0.472**	0.464	0.468	0.470	0.465	0.464	0.446
	QWK	0.621	0.618	0.627	0.625	**0.631**	0.612	0.620	0.572

Table 15 Performance of neural models on CEFR-SP, WSJ and OSE. Bold numerical values indicate robust performance.

Dataset	Metric	BERT base	BART base	RoBERTa base	XLNet base	ELECTRA base	BART large	RoBERTa large	ELECTRA large
CEFR-SP	Macro-F1	0.419	0.507	0.498	**0.524**	0.413	0.506	0.521	0.518
	Acc	0.577	0.617	0.637	0.638	0.636	0.625	**0.650**	0.646
	F1	0.567	0.613	0.626	0.630	0.609	0.621	**0.640**	0.635
	Prec	0.621	0.616	0.622	0.632	0.597	**0.664**	0.636	0.627
	Recall	0.577	0.617	0.637	0.638	0.636	0.625	**0.650**	0.646
	QWK	0.792	0.818	0.808	**0.826**	0.797	0.820	0.822	0.817
WSJ	Acc	0.687	0.704	0.703	0.689	0.675	**0.717**	0.716	0.687
	F1	0.667	0.687	0.674	0.657	0.636	**0.691**	0.690	0.669
	Prec	0.689	**0.700**	0.688	0.663	0.637	0.696	0.689	0.684
	Recall	0.687	0.704	0.703	0.689	0.675	**0.717**	0.716	0.687
	QWK	0.786	0.786	0.766	0.767	0.761	0.791	**0.794**	0.792
OSE	Acc	0.547	0.571	0.569	0.562	0.555	**0.571**	0.570	0.566
	F1	0.532	0.555	0.554	0.543	0.533	**0.558**	0.555	0.549
	Prec	0.549	0.570	0.566	0.554	0.552	0.565	**0.567**	0.566
	Recall	0.547	0.571	0.569	0.562	0.555	**0.571**	0.570	0.566
	QWK	0.500	0.537	0.537	0.535	0.512	**0.549**	0.541	0.532

Appendix B Appendix: Performance of Large language models

Table 16 presents the performance of several Large Language Models (LLMs) on sentence-level ARA. In general, GPT-4-0613 yielded the best performance. The few-shot approach did not consistently surpass zero-shot results, indicating room for improvement in sample selection and prompt engineering.

Table 16 Performance of LLMs in sentence-level ARA. Bold numerical values indicate robust performance.

Dataset	Metric	ChatGLM2-6B		GPT−3.5-turbo-0613		GPT-4-0613		Meta-Llama-3-8B	
		Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
CEFR-SP	Macro-F1	0.119	0.129	0.231	0.062	**0.318**	0.214	0.282	0.238
	Acc	0.268	0.145	**0.384**	0.114	0.370	0.251	0.382	0.342
	F1	0.191	0.173	0.328	0.063	0.371	0.263	**0.383**	0.341
	Prec	0.374	0.275	0.428	0.085	0.436	0.411	**0.447**	0.366
	Recall	0.268	0.145	**0.384**	0.114	0.370	0.251	0.382	0.342
	QWK	0.043	0.178	0.376	0.151	0.629	0.456	0.636	**0.641**
WSJ	Acc	0.241	0.209	0.302	0.413	0.476	**0.540**	0.270	0.397
	F1	0.260	0.184	0.192	0.274	0.446	**0.568**	0.277	0.415
	Prec	0.372	0.168	0.145	0.244	0.503	**0.624**	0.341	0.441
	Recall	0.241	0.209	0.302	0.413	0.476	**0.540**	0.270	0.397
	QWK	−0.102	0.050	0.187	0.162	0.451	**0.773**	0.512	0.488
OSE	Acc	0.345	0.370	0.354	0.390	**0.452**	0.441	0.417	0.440
	F1	0.204	0.312	0.240	0.341	**0.453**	0.441	0.456	0.443
	Prec	0.267	0.423	**0.497**	0.467	0.489	0.522	0.575	0.462
	Recall	0.345	0.370	0.354	0.390	**0.452**	0.441	0.417	0.440
	QWK	0.020	0.058	0.087	0.173	**0.331**	286	0.469	0.324
CLC	Acc	0.098	0.069	0.084	0.122	0.106	0.108	0.236	**0.239**
	F1	0.050	0.036	0.034	0.085	0.058	0.088	0.223	**0.224**
	Prec	0.045	0.076	0.024	0.110	0.043	0.111	0.211	**0.211**
	Recall	0.098	0.069	0.084	0.122	0.106	0.108	0.236	**0.239**
	QWK	0.006	0.013	0.037	0.186	0.122	**0.230**	−0.021	0.019
TCSOL	Acc	0.082	0.145	0.246	0.200	**0.353**	0.307	0.278	0.333
	F1	0.050	0.096	0.214	0.177	**0.324**	0.305	0.312	0.300
	Prec	0.320	0.319	**0.417**	0.194	0.327	0.371	0.357	0.278
	Recall	0.082	0.145	0.246	0.200	**0.353**	0.307	0.278	0.333
	QWK	0.009	0.142	0.202	0.194	0.407	0.499	0.468	**0.556**

Acknowledgements This work was partly supported by the Language Fund from the Standing Committee on Language Education and Research (project EDB(LE)/P&R/EL/203/14); the General Research Fund (Project 11207320); and the CityU Strategic Research Grant (project 7005803).

Funding Open access publishing enabled by City University of Hong Kong Library's agreement with Springer Nature.

Data availibility The datasets generated during the current study are available from the corresponding author on request.

Declarations

Conflict of interest All authors declare that they have no conficts of interest to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ambati, B.R., Reddy, S., & Steedman, M. (2016). Assessing relative sentence complexity using an incremental ccg parser. *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 1051–1057).

Arase, Y., Uchida, S., & Kajiwara, T. (2022). CEFR-based sentence difficulty annotation and assessment. *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 6206–6219). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.416

Azpiazu, I. M., & Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics, 7*, 421–436.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Bo, W., Chen, J., Guo, K., & Jin, T. (2019). Data-driven adapting for fine-tuning chinese teaching materials: Using corpora as benchmarks. *Computational and Corpus Approaches to Chinese Language Learning*, 99–118,

Bosco, G. L., Pilato, G., & Schicchi, D. (2021). Deepeva: A deep neural network architecture for assessing sentence complexity in italian and english languages. *Array, 12*, 100097.

Brunato, D., De Mattei, L., Dell'Orletta, F., Iavarone, B., & Venturi, G. (2018). Is this sentence difficult? do you agree? *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2690–2699).

Cheng, Y., Xu, D., & Lv, X. (2019). Automatically grading text difficulty with multiple features. *Data Analysis and Knowledge Discovery, 3*(7), 103–112.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (p.1724).

Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial efl learners. *Multiple affordances of language corpora for data-driven learning*, 109–128,

Clark, K., Luong, M.-T., Le, Q.V., & Manning, C.D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint*[SPACE]arXiv:2003.10555, ,

Collins-Thompson, K. (2008). Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics, 165*(2), 97–135.

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. *arXiv preprint*[SPACE]arXiv:2004.13922, ,

Cui, Y., Che, W., Wang, S., & Liu, T. (2022). Lert: A linguistically-motivated pre-trained language model. *arXiv preprint*[SPACE]arXiv:2211.05344, ,

Dell'Orletta, F., Wieling, M., Venturi, G., Cimino, A., & Montemagni, S. (2014). Assessing the readability of sentences: which corpora and features? *Proceedings of the ninth workshop on innovative use of nlp for building educational applications* (pp. 163–173).

Deng, J., Cheng, L., & Wang, Z. (2021). Attention-based bilstm fused cnn with gating mechanism model for Chinese long text classification. *Computer Speech and Language, 68*, 101182.

Deutsch, T., Jasbi, M., & Shieber, S. (2020b). Linguistic Features for Readability Assessment. *Proceedings of the fifteenth workshop on innovative use of nlp for building educational applications.*

Deutsch, T., Jasbi, M., & Shieber, S.M. (2020a). Linguistic features for readability assessment. *Proceedings of the fifteenth workshop on innovative use of nlp for building educational applications* (pp. 1–17).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. *Proc. North American chapter of the association for computational linguistics - human language technologies (naacl-hlt).*

Feng, Y., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. (2023). Sentence simplification via large language models. *arXiv preprint*[SPACE]arXiv:2302.11957, ,

Filighera, A., Steuer, T., & Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. *European conference on technology enhanced learning* (p.335–348). Springer.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 513–578.

Fry, E. (1990). A readability formula for short passages. *Journal of Reading, 33*(8), 594–597.

Garbacea, C., Guo, M., Carton, S., & Mei, Q. (2021). Explainable prediction of text complexity: The missing preliminaries for text simplification. *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1086–1097).

Grießhaber, D., Vu, N. T., & Maucher, J. (2020). Low-resource text classification using domain-adversarial learning. *Computer Speech and Language, 62*, 101056.

Hanban/Confucius Institute Headquarters. (2008). *International curriculum for Chinese language education*. Beijing: Foreign Languages Teaching and Research Press.

Howcroft, D.M., & Demberg, V. (2017). Psycholinguistic models of sentence processing improve sentence readability ranking. *Proceedings of the 15th conference of the european chapter of the association for computational linguistics: Volume 1, long papers* (pp. 958–968).

Imperial, J.M. (2021). Bert embeddings for automatic readability assessment. *Proceedings of the international conference on recent advances in natural language processing (ranlp 2021)* (pp. 611–618).

Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7943–7960). Online: Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.709

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlỳ, P. (2008). Gdex: Automatically finding good dictionary examples in a corpus. *Proceedings of the xiii euralex international congress* (Vol. 1, pp. 425–432).

Kincaid, P.J., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas for Navy enlisted personnel. *Research branch report 8-75.* Chief of Naval Technical Training: Naval Air Station Memphis.

Kingma, D.P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proc. 3rd international conference for learning representations.* San Diego.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lee, B.W., Jang, Y.S., & Lee, J.H.-J. (2021). Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. *Proceedings of the 2021 conference on empirical methods in natural language processing.*

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint*[SPACE]arXiv:1910.13461, ,

Liu, F., & Lee, J.S. (2023). Hybrid models for sentence readability assessment. *Proceedings of the 18th workshop on innovative use of nlp for building educational applications (bea 2023)* (pp. 448–454).

Liu, X., You, X., Zhang, X., Wu, J., & Lv, P. (2020). Tensor graph convolutional networks for text classification. *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 8409–8416).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*[SPACE]arXiv:1907.11692, ,

Logar, N., Grčar, M., Brakus, M., Erjavec, T., Holdt, Š.A., & Krek, S. (2020). *Korpusi slovenskega jezika gigafida, kres, ccgigafida in cckres: gradnja, vsebina, uporaba*. Znanstvena založba Filozofske fakultete.

Lu, D., Qiu, X., & Cai, Y. (2020a). Sentence-level readability assessment for l2 chinese learning. *Chinese lexical semantics: 20th workshop, clsw 2019, Beijing, China, june 28–30, 2019, revised selected papers 20* (pp. 381–392).

Lu, D., Qiu, X., & Cai, Y. (2020). Sentence-level readability assessment for l2 Chinese learning. CLSW 2019. *LNAI, 11831*, 381–392.

Mallinson, J., Sennrich, R., & Lapata, M. (2020). Zero-shot crosslingual sentence simplification. *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 5109–5126).

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

Mertzen, D., Paape, D., Dillon, B., Engbert, R., Vasishth, S., Mertzen, D., ... & Vasishth, S. (2023). Syntactic and semantic interference in sentence comprehension: Support from english and German eye-tracking data. *Glossa Psycholinguistics*, *2*(1), ,

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*[SPACE]arXiv:1301.3781, ,

Mohammadi, H., & Khasteh, S.H. (2019). Text as environment: A deep reinforcement learning text readability assessment model. *arXiv preprint*[SPACE]arXiv:1912.05957, , ,

Nadeem, F., & Ostendorf, M. (2018). Estimating linguistic complexity for science texts. *Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications* (pp. 45–55).

Nahatame, S. (2021). Text readability and processing effort in second language reading: A computational and eye-tracking investigation. *Language Learning, 71*(4), 1004–1043.

Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., & Yuret, D. (2007). The conll 2007 shared task on dependency parsing. *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (emnlp-conll)* (pp. 915–932).

OpenAI (2021). *Chatgpt.*[SPACE]https://openai.com/.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., & Grisel, O. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825-2830,

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Applied Linguistics, 7*(1), 143–159.

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. *Proceedings of the ninth workshop on innovative use of nlp for building educational applications* (pp. 174–184).

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and Machine Learning Approaches for Second Language Sentence-level Readability. *Proc. 9th workshop on innovative use of nlp for building educational applications.*

Prystauka, Y., Altmann, G. T., & Rothman, J. (2024). Online eye tracking and real-time sentence processing: On opportunities and efficacy for capturing psycholinguistic effects of different magnitudes and diversity. *Behavior Research Methods, 56*(4), 3504–3522.

Qiu, X., Chen, Y., Chen, H., Nie, J.-Y., Shen, Y., & Lu, D. (2021). Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment. *Proc. 59th annual meeting of the association for computational linguistics (acl)* (p.3013–3025).

Schicchi, D., Pilato, G., & Bosco, G.L. (2020). Deep neural attention-based model for the evaluation of italian sentences complexity. *2020 ieee 14th international conference on semantic computing (icsc)* (pp. 253–256).

Schumacher, E., Eskenazi, M., Frishkoff, G., & Collins-Thompson, K. (2016). Predicting the relative difficulty of single sentences with and without surrounding context. *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1871–1881).

Segler, T.M. (2007). Investigating the selection of example sentences for unknown target words in icall reading texts for l2 German.

Shi, Y., Ma, H., Zhong, W., Mai, G., Li, X., Liu, T., & Huang, J. (2023). Chatgraph: Interpretable text classification by converting chatgpt knowledge to graphs. *arXiv preprint*[SPACE]arXiv:2305.03513, ,

Shiffler, R. E. (1988). Maximum z scores and outliers. *The American Statistician, 42*(1), 79–80.

Song, D., Ma, S., Sun, Z., Yang, S., & Liao, L. (2021). Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems, 230*, 107408.

Štajner, S., Ponzetto, S.P., & Stuckenschmidt, H. (2017). Automatic assessment of absolute sentence complexity. *Proceedings of the 26th international joint conference on artificial intelligence, ijcai* (Vol. 17, pp. 4096–4102).

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., & Wang, G. (2023). Text classification via large language models. *arXiv preprint*[SPACE]arXiv:2305.08377, ,

Sung, Y.-T., Chen, J.-L., Cha, J.-H., Tseng, H.-C., Chang, T.-H., & Chang, K.-E. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*, 340–354.

Todirascu, A., François, T., Bernhard, D., Gala, N., & Ligozat, A.-L. (2016). Are cohesive features relevant for text readability evaluation? *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 987–997).

Vajjala, S., & Lučić, I. (2018). Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. *Proceedings of the thirteenth workshop on innovative use of nlp for building educational applications* (pp. 297–304).

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 163–173).

Vajjala, S., & Meurers, D. (2014). Assessing the relative reading level of sentence pairs for text simplification. *Proceedings of the 14th conference of the european chapter of the association for computational linguistics* (pp. 288–297).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, ,

Volodina, E., Pilán, I., Borin, L., & Tiedemann, T.L. (2014). A flexible language learning platform based on language resources and web services. *Lrec* (pp. 3973–3978).

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., & De Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. *Proceedings of the 33rd international conference on international conference on machine learning-volume 48* (pp. 1995–2003).

Weiss, Z., & Meurers, D. (2022). Assessing sentence readability for german language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference? *Proceedings of the 17th workshop on innovative use of nlp for building educational applications (bea 2022)* (pp. 141–153).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A.M. (2019). Huggingface's transformers: State-of-the-art natural language processing. ArXiv:abs/1910.03771

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics, 3*, 283–297.

Yang, S., Sun, R., & Wan, X. (2023). A new dataset and empirical study for sentence simplification in chinese. *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (p.8306-8321).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., & Le, Q.V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*, ,

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489).

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 7370–7377).

Zeng, J., Xie, Y., Yu, X., Lee, J.S.Y., & Zhou, D.-X. (2022). Enhancing Automatic Readability Assessment with Pre-training and Soft Labels for Ordinal Regression. *Findings of the association for computational linguistics: Emnlp 2022* (p.4586-4597).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Fengkai Liu[1] · Tan Jin[2] · John S. Y. Lee[1]

✉ Fengkai Liu
fengkaliu3-c@my.cityu.edu.hk

Tan Jin
tjin@scnu.edu.cn

John S. Y. Lee
jsylee@cityu.edu.hk

[1] Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR, China

[2] School of International Culture, South China Normal University, Guangzhou, Guangdong, China