Project Title	**:**	A Text Difficulty Analysis Tool for Developing Extra-Curricular Reading Materials
Grantee	**:**	City University of Hong Kong
Principal Investigator	**:**	LEE Sie-yuen, John City University of Hong Kong
Co-Investigator(s)	**:**	LIU Mei-chun City University of Hong Kong

Final Report
by
Principal Investigator

End-of-Project Report: A Text Difficulty Analysis Tool for Developing Extra-Curricular Reading Materials

PI: John Sie Yuen LEE
Co-I: Meichun LIU

Department of Linguistics and Translation
City University of Hong Kong

Abstract

Since extensive reading is important for language learning, students should engage in extra-curricular reading as much as possible. To facilitate efficient learning, language teachers need to select reading materials at the appropriate difficulty level, and adapt them if necessary. This report describes a text difficulty analysis tool that assists teachers in preparing reading materials in Chinese. The tool performs Automatic Readability Assessment to determine the difficulty level of the text. On the basis of a graded vocabulary list, the tool also estimates the vocabulary complexity of the text with respect to the target school grade. To support text revision, it highlights words that are expected to be new vocabulary, and suggests alternatives that better fit the expected vocabulary proficiency of students at the target grade. Evaluation results have shown that users can efficiently assess the difficulty of a text and revise the text with the tool. It is recommended that schools be informed of the advantages and pitfalls of integrating automatic text difficulty assessment into the process of lesson preparations and examination paper review.

Keywords: natural language processing; computational linguistics; automatic readability assessment; text revision

1 Introduction

Since extensive reading is important for language learning, students should engage in extra-curricular reading as much as possible (Krashen, 1981). To facilitate efficient learning, the difficulty of the reading materials should match the reader's language proficiency: an overly easy text would not stretch the student's linguistic skills, while a very challenging text could be discouraging and may not be even understood by the student. Language teachers therefore need to carefully select materials at an appropriate level of difficulty, and adapt them if necessary. Since this can be a time-consuming process, teachers would benefit from an intelligent text analyzer that provides assistance in editing reading materials.

Automatic Readability Assessment (ARA) is the task of predicting how difficult it is for the reader to understand a text. Recent advances in natural language processing and computational linguistics have led to ARA models models that can accurately predict the difficulty level of a text (Martinc et al., 2021; Lee et al., 2021). This report describes *Text Difficulty Analyzer*, a browser-based tool that performs ARA to estimate the school grade in Hong Kong for which the text is suitable. Further, on the basis of a graded vocabulary list, the tool estimates the proportion of words in a text that are known to students at the target school grade. To support text revision, it highlights words that are expected to be new vocabulary, and suggests alternatives that better fit the expected vocabulary proficiency at the target grade.

The rest of the report is organized as follows. After a literature review (Section 2), we present the conceptual framework of the project (Section 3). We then describe the methodology (Section 4) and data (Section 5) used in our research. Finally, we discuss the results (Section 6), conclude and make recommendations (Section 7).

Our text difficulty analysis tool is accessible at: https://jega.lt.cityu.edu.hk/ARA/ The neural ARA model underlying the tool is also available as open-source Python code at: https://github.com/hhlim333/ALTA2022Readability

2 Review of Research Literature

To provide the background for this project, we summarize relevant findings on automatic readability assessment (Section 2.1), vocabulary profiling (Section 2.2), and text revision (Section 2.3). Table 1 shows a comparison between key features of our tool and several widely known text analyzers.

2.1 Automatic Readability Assessment (ARA)

Text readability is defined as the cognitive load of a reader to comprehend a text (Martinc et al., 2021). Various models of automatic readability assessment (ARA) have been developed to label a text with its difficulty level on a scale, which could be a school grade, a level in the CEFR (Common European Framework of Reference for Languages), or the HSK (*Hanyu Shuiping Kaoshi*) framework, etc.

Readability formulas (Kincaid et al., 1975) are the earliest attempts to automate readability assessment. Since the advent of statistical natural language processing (NLP), statistical classifiers can be trained on a variety of features to determine the readability or grade level of a text. These features may include the language model score (Collins-Thompson, 2008) as well as handcrafted features capturing lexical, syntactic and semantic characteristics of the text (Dell'Orletta et al., 2011; François and Fairon, 2012; Pitler and Nenkova, 2008; Sung et al., 2015) While these classifiers lend themselves to more explainable and linguistically-motivated results, they mostly rely on one-hot linguistic features. Neural models not only remove the requirement for feature engineering but also, similar to other NLP tasks, can offer superior performance in ARA (Tseng et al., 2019; Martinc et al., 2021).

2.1.1 Hybrid models for ARA

'Hybrid' models, which combine linguistic features and neural models, have been found to benefit a variety of NLP tasks (Lei et al., 2018; Strubell et al., 2018). Various methods for combining these approaches have been investigated. For example, a Bi-LSTM can incorporate part-of-speech information (Azpiazu and Pera, 2019). A statistical classifier can directly use sentence embeddings as features (Imperial, 2021). It can also incorporate the decision of the neural model as a single numeric feature (Deutsch et al., 2020), or 'soft' labels expressing the probabilities of each grade as predicted by the neural model (Lee et al., 2021).

Experimental results on the performance of hybrid ARA models have so far been inconclusive. Some studies observed no effect (Deutsch et al., 2020) or only marginal improvement (Filighera et al., 2019) from linguistic features, while others reported significant improvement, e.g. by combining Random Forest and RoBERTa (Lee et al., 2021). One contribution of this project is to directly compare the performance of the linguistic, neural and hybrid approaches on Chinese data.

2.1.2 ARA models for Chinese

Although a number of ARA systems have been developed for Chinese, none of them aims at the language curriculum in Hong Kong or offers support for text revision. ChineseTA is an

Feature	Our tool	CRIE	Chi-editor	AntWordProfiler
Automatic Readability Assessment	Yes	Yes	Yes	Yes
Lexical substitution	Yes	No	No	No
Word segmentation editing	Yes	No	No	n/a
Vocabulary list editing	Yes	No	No	Yes

Table 1: Comparison of key features in text difficulty analyzers

integrated computer software program that estimates vocabulary difficulty on the basis of word frequencies interpolated from various corpora (Chu, 2005). Coh-metrix consists of 108 features, including word information (e.g., lexical frequency and density), syntactical complexity (e.g., noun-phrase density), semantic cohesion (e.g., conjunctions), and semantic relations. Based on Coh-metrix, Chinese Coh-metrix incorporates some of its features designed for Chinese lexical and textual properties, including part-of-speech and frequency, cohesion, word information, connectives, and sentence structures (McNamara et al., 2014). The Chinese Readability Index Explorer (CRIE) makes use of Chinese Coh-Metrix and the Linguistic Inquiry and Word Count (LIWC) dictionary, as well as 70 linguistic features at the word, syntactic, semantic and cohesion levels, to classify documents into Grade 1 to 6. Trained with an SVM, their classification accuracy reached 75% (Sung et al., 2015).

2.1.3 In-domain vs. cross-domain evaluation

The genre and content of the sample texts in ARA training datasets may not exactly match those of the texts on which the trained ARA model is deployed. Domain adaptation techniques can be applied to address differences between native and non-native texts. For example, scores from an ARA ranking model trained on graded texts for native speakers can help estimate the CEFR level of a text for non-native learners (Xia et al., 2016).

Another type of mismatch is caused by cross-domain or cross-corpus data. Real-word applications of ARA models are often targeted at cross-domain or cross-corpus data. Consider the task of retrieving extra-curricular reading materials for language learning from web texts, which likely diverge in style and content from the training data. In-domain evaluation therefore may not accurately reflect the actual performance on such tasks. Some past ARA studies on ranking in ARA have addressed this issue. For example, when ranking models are trained on Newsela, they suffered a performance degradation when tested on OneStopEnglish and Vikidia (Lee and Vajjala, 2020). For the grade prediction task, however, experiments have mostly been conducted in-domain, with the training and test data drawn from the same source. Cross-domain evaluation has been reported mainly in terms of correlation (Chen and Meurers, 2016). This may be due to the fact that different grade scales are adopted in the major benchmarks, such as Newsela, OneStopEnglish and WeeBit. In this project, we leverage two comparable datasets in Chinese (Section 5) to conduct cross-domain evaluation on hybrid models to assess the contribution of linguistic features in the grade prediction task.

2.2 Vocabulary profiling

As most existing ARA models are not designed to support revision, they operate as a black box and do not explain its prediction. Vocabulary profiling can complement them by providing an explainable analysis, and serving as an initial step towards automatic text recommendation (Lee et al., 2022).

Most vocabulary profilers work on the basis of a graded vocabulary list. They match words in the input text against the list, display the grade for each word, and compute the vocabulary coverage — the proportion of words in the text covered by the list — for a target grade. Beyond these basic functions, they vary in their approach to customization, word segmentation, and text revision support (Table 1).

2.2.1 User customization

For English, vocabulary profilers such as the Online Graded Text Editor (OGTE) (https://www.er-central.com/ogte) and AntWordProfiler (AWP) (https://www.laurenceanthony.net/software/antwordprofiler/) (Anthony, 2023) support a wide variety of vocabulary lists, such as the New General Service List (https://www.newgeneralservicelist.com), CEFR (Capel, 2015) and BNC/COCA lists. Both of them allow users to specify words to be ignored; AWP, in addition, can work with user-supplied graded vocabulary lists.

2.2.2 Word segmentation

Word segmentation is critical for vocabulary profiling of Chinese text, since inaccurate segmentation can drastically alter the profiling result (Section 4.2.2). To our knowledge, existing vocabulary profilers and ARA systems for Chinese use standard parsers for word segmentation (Sung et al., 2016; Lim et al., 2022). Chi-editor, a widely used Chinese vocabulary profiler, perform segmentation according to the *Graded Chinese Syllables, Characters and Words for the Application of Teaching Chinese to Speakers of Other Languages* (Bo et al., 2019). However, it does not allow users to correct word segmentation mistakes or revise vocabulary lists.

2.3 Text revision

Most text simplification models aim to reduce the difficulty of the input text, without specifying a target grade or difficulty level (Siddharthan, 2002; Belder and Moens, 2010; Kajiwara et al., 2013; Paetzold and Specia, 2017). Recent work has increasingly recognized the need for text simplification to an absolute target grade (Štajner et al., 2017; Agrawal and Carpuat, 2023), as the teacher may wish to revise the difficult words in the text in order to suit the target reader.

However, most editors for text simplification aim to simplify all words, and to suggest the simplest replacement, regardless of the language proficiency of the target reader (Jill Burstein and Ventura, 2007). The editor described in (Lee et al., 2016) considers only two levels of difficulty, and supports English only. A closely related task is Complex Word Identification (CWI) (Yimam et al., 2018), or Lexical Complexity Prediction (LCP) (Shardlow et al., 2021), which aims to identify the words in a passage that may pose difficulties for the reader. To our knowledge, there has not been any CWI or LCP tools aimed at the Chinese curriculum in Hong Kong.

2.4 Mandarin VerbNet

Mandarin VerbNet (MVN) is a verbal semantic database with annotation of frame-based constructional features (Liu and Chiang, 2008). In addition to frame elements, its frames make use of a schema-based meaning representation and constructional patterns. Adopting a hybrid approach to the semantic analysis of the lexical-constructional behavior of Chinese verbs, it incorporates tenets of Frame Semantics (Fillmore and Atkins, 1992) and Construction Grammar (Goldberg, 1995).

Archi-frames and Basic frames. The semantic hierarchy in MVN includes Archi-frames and Basic frames. At the top of the hierarchy, the Archi-frame represents the broadest possible scope of an eventive background, providing an overarching conceptual schema that serves as the semantic foundation for the individual frames within a relatively large and self-contained domain.

Basic frames correspond to the cognitively salient categories. They are more informative in terms of semantics, more prevalent and common in terms of distribution, and easier to acquire and acquire early in terms of acquisition.

Core and non-core frame elements. Similar to FrameNet (https://framenet.icsi.berkeley.edu), Mandarin VerbNet distinguishes between "core" or "non-core" frame elements. Core frame elements are fundamental; they commonly appear as a necessary argument in a sentence and plays an essential role in the event frame. Non-core frame elements are optional; they are "potentially relevant" and can be added to a sentence as an adjunct (Liu and Chiang, 2008).

Some non-core frames describe a process based on the main verb and provide more information about this verb. They are often considered to be more advanced structures, and some are found only at higher grades (Lee et al., 2020).

3 Conceptual framework of project

The *Text Difficulty Analyzer* is designed to help teachers analyze Chinese texts for pedagogical use with primary and secondary school students in Hong Kong. The user works with the tool in a cycle that involves analyzing text difficulty (Section 3.1), adjusting the parameters (Section 3.2), revising the text (Section 3.3), possibly with reiterations of the cycle until the text is ready.

3.1 Analyze text difficulty

In this step, the tool analyzes the difficulty of the text by running the Automatic Readability Assessment (ARA) model (Section 4.1), which classifies the text at one of the school grades in Hong Kong.

Further, the tool estimates the vocabulary coverage of the text at any target grade (Section 4.2.1), i.e., the percentage of words that are known to the typical student in that grade. Figure 1 shows an example vocabulary profile of a text with the target grade set to Grade 1. As indicated in the leftmost tab on top, 79.2% of the words in the text are predicted to be known to the typical Grade 1 student. Vocabulary that is new to Grade 1 students are color-coded, with deeper colors indicating a higher level of difficulty.

3.2 Adjust parameters

In this step, the user adjusts the word boundaries and word grades (Section 4.2.2). If the vocabulary profiling results contain word segmentation errors, the user can immediately revise the word boundaries in the text input field by inserting and deleting white space. Further, if the user disagrees with the grade of a word, he or she can dynamically revise the vocabulary list or add new words. After these adjustments, the user can refresh the vocabulary coverage estimation.

3.3 Revise

After any necessary adjustments in the parameters, the user may start revising the text to the difficulty level of the target grade. To assist in this task, the tool provides suggested lexical substitutions (Section 4.3). For each difficult word, a drop-down list offers word substitutions that conform to the target grade. Consider the sentence '... the bed suddenly became unusually bright' in Figure 1. The word 異常 *yichang* 'unusually' is highlighted as a Grade 5 word since it exceeds the target grade (Grade 1). The user can click on the word to open a drop-down list, which contains simpler words with related meaning. In this case, the drop-down list proposes the words 非常 *feichang* 'extremely' (Grade 1), 十分 *shifeng* 'very' (Grade 1), etc. After clicking on a suitable word in the list to apply the substitution, the user may re-analyze the difficulty of the text (Section 3.1) and possibly repeat the cycle until a satisfactory version of the text is obtained.

4 Methodology

According to the conceptual framework presented above, we describe our research methodology for each of its three components: automatic readability assessment (Section 4.1), vocabulary profiling (Section 4.2), and text revision (Section 4.3).

4.1 Automatic Readability Assessment (ARA)

We compare the performance of hybrid, neural and linguistic models on ARA in the cross-domain setting. Our hypothesis is that the hybrid model outperforms the neural model both in-domain and cross-domain in Chinese datasets, providing further evidence on the contribution of linguistic features. Further, the hybrid model is expected to exhibited smaller performance degradation on cross-domain data, suggesting their robustness and ability to capture more salient indicators of text difficulty.

In terms of metrics for ARA performance, we use accuracy, F1, adjacent accuracy and quadratic weighted kappa (QWK). For adjacent accuracy, the system is considered correct if the predicted label is within one grade higher or lower than the gold grade. QWK also helps capture the distance between gold and predicted grades. These metrics give a comprehensive evaluation of model performance from different perspectives.

4.1.1 Neural Model

We fine-tuned MacBERT (Cui et al., 2020), RoBERTa (Cui et al., 2020), BERT (Devlin et al., 2019) and BERT-wwm (Cui et al., 2020) on the Mainland dataset for grade prediction. We used the code by Lee et al. Lee et al. (2021) in default parameters for fine-tuning, accessed from https://github.com/yjang43/pushingonreadability_transformers. We used the versions `macbert-large`, `chinese-roberta-wwm-ext`, `bert-base-chinese`, and `chinese-bert-wwm`, respectively.

4.1.2 Linguistic Model

We built *ChiLingFeat*, a toolkit that extracts most features used in previous Chinese ARA studies (Sung et al., 2015; Lu et al., 2020). This toolkit is open-source and publicly available at https://github.com/ffliu6/ChiLingFeat. Using *ChiLingFeat*, we trained a statistical classifier on the 221 linguistic features provided. We evaluated SVM, Random Forest (RF), and XGBoost (XGB) using the implementation in scikit-learn (Pedregosa et al., 2011). We applied Variance Threshold algorithm in scikit-learn for feature selection, but obtained the best result with the full feature set.

4.1.3 Hybrid Model

Our hybrid model follows Lee et al. (2021) in wrapping linguistic features and neural model output in a non-neural, statistical classifier. We evaluated three types of hybrid models:

Hard labels (Deutsch et al., 2020): The grade of the input text, as predicted by the neural model (Section 4.1.1) serves as an additional feature in the classifier.

Soft labels (Lee et al., 2021): The probabilities of each grade, as predicted by the neural model (Section 4.1.1), serve as additional features.

Sentence Embeddings (Imperial, 2021): The sentence vectors, produced by SBERT (Reimers and Gurevych, 2019) from the sentences in the input text, serve as additional features.

4.1.4 Syntactic and Semantic Features

To study the contribution of syntactic and semantic features to ARA performance, we selected the following syntactic features for training a classification model for ARA:

Sentence Length The Sentence Length feature of a passage is defined to be the average sentence length of its sentences. Length is measured in terms of the number of Chinese characters in the sentence, excluding punctuation. Easier passages can be expected to have shorter sentences than more difficult passages.

Character Difficulty The Character Difficulty feature of a passage is defined as the proportion of character types that are listed at Grade N or below, according to the graded character lists published by the Ministry of Education in China. We used the *Chinese Curriculum Standards of Compulsory Education* 《义务教育语文课程标准》 (downloaded from http://www.gov.cn/zhengce/zhengceku/2022-04/21/content_5686535.htm). Graded character lists designed for learners of Chinese as a foreign language, such as the European Benchmarking Chinese Language (ECBL) List or those from the *Hanyu Shuiping Kaoshi* (Hanban, 2014), can also potentially be helpful.

Word Difficulty The Word Difficulty feature of a passage is defined as the proportion of word types that are among the most frequently used N words. Word frequency is based on the 10,000 most frequent words in the Sinica Corpus.

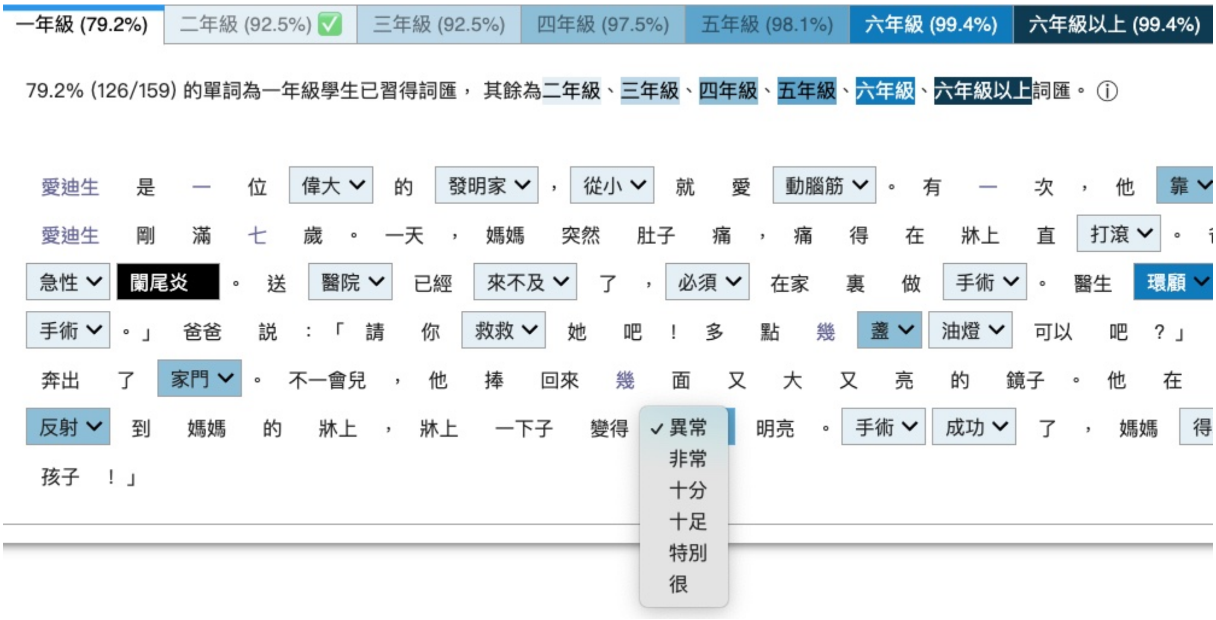
Further, we selected the following semantic features:

Non-core Verb Frame Elements The Non-Core Frame Element feature is defined as the percentage of verb frames that have non-core frame elements in Mandarin VerbNet. Non-core frame elements tend to appear more rarely in easy passages than in difficult passages (Lee et al., 2020).

Subject Omission To reduce repetition, a writer may omit a verb argument from a sentence, expecting the reader to infer the information from the context, This phenomenon is frequent in Chinese even for some core arguments, such as pro-dropped subjects (Kim, 2000). The number of zero pronouns is likely correlated with the effort needed for resolution.

We focus on core frame elements that normally occupy the subject position before the verb. These include the frame elements `Placer`, `Agent`, `Mover`, `Cognizer`, `Speaker`, `Exp`, `Affector`, `Intls`, `Intl_1`, `Coactors`, `Co-actor_1`, `Coordinator`, `Perceiver`, `Perceiver_Exp`, and `Perceiver_Agentive`. We define the Subject Omission feature as the proportion of verb frames in which these core frame elements are missing. Overall, sentences are more likely to lack subjects in passages at higher grades.

Metaphoric Usage Since metaphoric usage involves cognitive transfer from one domain to another, it tends to make a sentence harder to understand, even when the vocabulary and syntactic structures are simple. The Metaphoric Usage feature is defined to be the percentage of verbs in the passage with metaphoric usage frame elements. These verbs should be physical action verbs, since abstract ideas get their meaning and structure from the metaphorical mapping of concrete, embodied schema, such as Motion (Lakoff and Johnson, 2008). Physical actions verbs are defined by real and specific actions (Gao, 2001) and have an extensive involvement in the way we conceive and understand abstract meanings (Panunzi and Vernillo, 2019). Metaphoric usage can be expected to appear more often at higher-grade texts (Lee et al., 2020).



79.2% (126/159) 的單詞為一年級學生已習得詞匯， 其餘為二年級、三年級、四年級、五年級、六年級、六年級以上詞匯。ⓘ

愛迪生 是 一 位 偉大 的 發明家 ， 從小 就 愛 動腦筋 。 有 一 次 ， 他 靠
愛迪生 剛 滿 七 歲 。 一天 ， 媽媽 突然 肚子 痛 ， 痛 得 在 牀上 直 打滾 。
急性 闌尾炎 。 送 醫院 已經 來不及 了 ， 必須 在家 裏 做 手術 。 醫生 環顧
手術 。」 爸爸 說 ：「 請 你 救救 她 吧 ！ 多 點 幾 盞 油燈 可以 吧 ？」
奔出 了 家門 。 不一會兒 ， 他 捧 回來 幾 面 又 大 又 亮 的 鏡子 。 他 在
反射 到 媽媽 的 牀上 ， 牀上 一下子 變得 ✓異常 明亮 。 手術 成功 了 ， 媽媽 得
孩子 ！」

Figure 1: Interface of *Text Difficulty Analyzer* providing feedback on a Chinese passage: (1) **Vocabulary profiling**: Words expected to be new vocabulary at the target grade (Grade 1) are highlighted; words at different grades are color-coded from light blue to dark blue. (2) **Text revision**: For each highlighted word, a drop-down list suggests simpler substitutions.

Clause as Verb Argument Some verbs can take either a noun phrase (NP) or a clause as argument. The distinction is reflected by the frame element. A sentence such as 我担心[你会生病]$_{Target\_Possible\_Situation}$, which contains the clause 'you might get sick' as object, has the frame element `Target_Possible_Situation`. In contrast, a sentence such as 我担心[你的健康]$_{Target\_Entity}$, with the NP 'your health' as object, has the `Target_Entity` element. Similar distinctions are made in other frame categories, for example with `Phenomenon` (clause) vs. `Topic` (NP), and `Topic_Proposition` (clause) vs. `Topic_Description` (NP). Given a choice between NP and clause for a complement, clause is more often used in easy texts.

The Clause as Verb Argument feature is defined as the number of instances of verbs that take an clause as argument, out of the total instances of verbs that take either an NP or clause as argument.

4.2 Vocabulary Profiling

Among the existing vocabulary profilers, only a few are designed to support Chinese-language teachers in the text revision process (Jin et al., 2018; Chu, 2005). The *Text Difficulty Analyzer*, designed for teaching and learning Chinese in Hong Kong, is distinguished from its peers in several significant aspects. First, it calculates the vocabulary coverage with respect to the target grade and highlights the words that are above the target grade (Section 4.2.1). Second, recognizing that word segmentation is critical to vocabulary profiling accuracy, the tool tailors the segmentation to the graded vocabulary list through Forward Maximal Matching, and supports dynamic editing of word boundaries (Section 4.2.2).

4.2.1 Vocabulary coverage

Figure 1 shows the profile of a text with the target grade set to Grade 1. As indicated in the leftmost tab on top, 79.2% of the words in the text are predicted to be known to the typical

Seg.	Left context	Word with Segmentation output	Right context	Word grade
(1a) without FMM	我 *wo* ‘I’	受[Gr 1] | 不[Gr 1] | 了[Gr 1] *shou bu liao* ‘receive’ NEG ASP	委屈 *weiqu* ‘humiliation’	Grade 1 Unnecessary segmentation
(1b) with FMM	我 *wo* ‘I’	受不了[Grade 2] *shoubuliao* ‘cannot stand’	委屈 *weiqu* ‘humiliation’	Grade 2 (Gold)
(2a) without FMM	這是一個 *zhe shi yi ge* ‘This is a’	爭吵聲[Grade 6+] *zengchaosheng* ‘argument’	不絕的城市 *bu jue de chengshi* ‘constant’ DE ‘city’	Grade 6+ Missed segmentation
(2b) with FMM	這是一個 *zhe shi yi ge* ‘This is a’	爭吵[Grade 3] | 聲[Grade 1] *zengchao sheng* ‘argue’ ‘voice’	不絕的城市 *bu jue de chengshi* ‘constant’ DE ‘city’	Grade 3 Gold

Table 2: Word segmentation output with and without Forward Maximal Matching (FMM)

Grade 1 student. Vocabulary that is new to Grade 1 students are color-coded, from light blue for Grade 2 words (e.g., 偉大 *weida* ‘great’), to deep blue for Grade 6 words (e.g., 反射 *fanshe* ‘reflect’), to black for words above Grade 6 (e.g., 闌尾炎 *lanweiyan* ‘appendicitis’), i.e., words that are not in the vocabulary lists. Proper nouns, temporal nouns, numbers, and non-Chinese words are identified as those given the NR, NT, OD and CD part-of-speech tags, respectively, by the HanLP Chinese Parser (https://github.com/hankcs/pyhanlp). They are colored in grey (e.g., 愛迪生 *aidisheng* ‘Edison’) and are excluded from the calculation of vocabulary coverage.

4.2.2 Word segmentation

As suggested by the variety of segmentation guidelines adopted in different corpora, there are often multiple possible word segmentation for Chinese text (Xue et al., 2005). In the context of vocabulary profiling, the ‘correct’ segmentation should in principle conform to the content in the vocabulary list. We will refer to the examples in Table 2 in the rest of this section. Consider the term 受不了 *shoubuliao* ‘cannot stand’, a Grade 2 word according to the EDB List. As shown in segmentation (1a), the HanLP Chinese parser segments the term into three independent words *shou*, *bu*, and *liao*. This segmentation interprets the characters to be a sequence of three words at Grade 1, hence underestimating the difficulty. Analogous to the treatment of multiword expressions in English vocabulary lists (Lee and Uvaliyev, 2023), the three characters should be considered a single word, as shown in segmentation (1b).

Suboptimal word boundaries could also inflate the difficulty. In segmentation (2a), the example 爭吵聲 *zhengchaosheng* ‘argument’ is taken as one word by the parser. This word is considered to be above Grade 6, as it is not on the EDB List. However, the term is in fact simpler since it can be understood by semantic compositionality through its constituent words *zhengchao* ‘argue’ and *sheng* ‘voice’, which are listed in Grades 3 and 1, respectively.

To alleviate these issues, one could train a word segmentation model on texts segmented according to the vocabulary list, but it would be difficult to obtain sufficient training data. Instead, after obtaining an initial word segmentation with the parser, we perform Forward Maximal Matching (FMM) to identify multi-character words that match the EDB List. The word lexicon for FMM includes not only the expanded EDB list, but also the HSK (Hanban, 2014), TOCFL (Tseng, 2014), and the *Yiwu Jiaoyu Changrong Cibiao* (So, 2019), the standard vocabulary list used in Mainland China, resulting in a lexicon with a total of 37,984 words. Following

FMM, the unmatched text spans follow the original word boundaries from HanLP parser. Using this algorithm, the word *shoubuliao* ‘cannot stand’ is now correctly recognized as a Grade 2 word in segmentation (1b). Likewise, the word *zhengchaosheng* is analyzed as two words in segmentation (2b), with *zhengchao* ‘argue’ in Grade 3 and *sheng* ‘voice’ in Grade 1.

4.3 Text revision

Our tool supports text revision by suggesting lexical substitutions that reduce vocabulary complexity, hence pushing the text towards the target school grade. We used the Chinese-LS model (Qiang et al., 2021) (https://github.com/luxinyu1/Chinese-LS) and the pretrained language model `bert-base-chinese` to generate possible word substitutions for each highlighted word. Using Fig. 1 as an example, in order to generate the drop-down list for *yichang* ‘unusually’, we input the sentence with *yichang* replaced with `<MASK>` and asked the model to generate the top 50 word candidates for the masked word. We retained only those candidates that appear at a lower grade in the EDB list than the original word. To ensure preservation of meaning, we compute the SBert (version `paraphrase-MiniLM-L3-v2`) sentence embeddings (Reimers and Gurevych, 2019) of the sentence produced with each candidate substitution. The five substitutions producing sentence embeddings with maximum cosine similarity to the original ones are displayed in the drop-down list.

5 Data collection and analysis

5.1 Corpus of Hong Kong Textbooks

We constructed a corpus of Hong Kong Chinese-language pedagogical texts, containing passages from the following 12 textbooks: 二十一世紀中國語文, 快樂學語文, 啓思語文新天地, 啓思中國語文, 我愛學語文, 現代普通話, 現代中國語文, 新語文, 學好中國語文, 中國語文, 生活中國語文, 啓思新高中中國語文. Published by major publishers including Oxford University Press (OUP) and Modern Education, all of these 12 textbooks are used in primary and secondary schools in Hong Kong and thus reflect the local Chinese usage. We digitized all passages in these textbooks and compiled them in our dataset for training the text difficulty assessment model. The dataset contains a total of 1,051,977 Chinese characters, belonging to 2,953 passages spanning 12 levels, from Primary 1 to Secondary 6.

5.2 Corpus of Mainland Textbooks

While Chinese-language textbooks in Mainland China follow similar language proficiency standards, they are compiled from different sources and use simplified rather than traditional characters. They can thus provide a realistic cross-domain scenario for ARA evaluation. Drawn from textbooks for Chinese language used in Mainland China (Lee et al., 2020), this dataset consists of 7.15M characters distributed in 4,831 passages in 12 grades (Cheng et al., 2019).

To study the contribution of surface and semantic features, we selected a total of 35 verbs from the six Archi-frames in Mandarin Verbnet that have the highest frequency across all grades in this corpus. We then retrieved passages that contains at least five occurrences of these 35 verbs, which yielded a dataset containing 1,939 passages. We manually and exhaustively annotated the verb frame usage in these sentences. The final dataset contains a total of 41,718 frame elements in 23,328 frames. This represents an average of 12.0 frames per passage and 1.8 frame elements per frame. The passages are divided into two categories: the *Easy Passages* are the textbook passages used in primary school; the *Difficult Passages* are those used in middle school and high school.

Transformer	Hybrid model type	In-domain	Cross-domain
BERT	Hard labels	0.312	0.288
	Soft labels	**0.342**	**0.290**
	Sentence Embeddings	0.322	0.269
BERT-wwm	Hard labels	0.295	**0.283**
	Soft labels	**0.341**	0.278
	Sentence Embeddings	0.318	**0.283**
RoBERTa	Hard labels	0.301	0.285
	Soft labels	**0.341**	**0.301**
	Sentence Embeddings	0.318	0.287
MacBERT	Hard labels	0.305	0.283
	Soft labels	**0.353**	**0.309**
	Sentence Embeddings	0.329	0.269

Table 3: Accuracy of the three hybrid model types (Section 4.1.3)

5.3 Graded vocabulary lists

Designed for Chinese teachers in Hong Kong, *Text Difficulty Analyzer* follows the vocabulary guidelines from the Hong Kong Education Bureau (EDB). Published by EDB, the *Lexical Lists for Chinese Learning in Hong Kong* (https://www.edbchinese.hk/lexlist_ch) consist of 5,037 words for Key Stage 1 (KS1), which corresponds to Grades 1 to 3 in primary school; and 4,870 words for Key Stage 2 (KS2), which corresponds to Grades 4 to 6. We will henceforth refer to this as the 'EDB List'.

The words within each key stage were split evenly to the three grades in the stage as follows. We computed the frequency of each word in the training set of the Hong Kong textbook corpus. Among the KS1 words, those with the highest frequency in the Grade 1 texts were assigned to Grade 1; then, those most frequent in the Grade 2 texts were assigned to Grade 2; with the remainder assigned to Grade 3. The KS2 words were processed in an analogous manner.

The EDB List is not exhaustive; in the Corpus of Hong Kong Textbooks (Section 5.1), for example, there are 6,799 unique words that are not found in the list. To incorporate these words into the list, we estimated their grades as follows:

Constituent-based grade The grade of a word is taken to be the grade of its most difficult constituent character. The grades of the constituent characters are based on the EDB Graded Character List (https://www.edbchinese.hk/lexlist_ch), supplemented with a list (https://ephpth.ephhk.com/resource/tools/lcprichi) published in 2003 and another list (http://code.web.idv.hk/misc/hkpri90.php) published in 1990. The character with the highest grade determines the constituent-based grade of the word.

Empirical grade The lowest grade at which the word appears in a passage in the corpus.

Finally, the grade of the word is assigned to be either the constituent-based grade or empirical grade, whichever is higher. Following this procedure, the expanded EDB List contains 16,706 words.

6 Results and Discussion

We present experimental results on readability assessment (Section 6.1), vocabulary complexity analysis (Section 6.2), and text revision (Section 6.3), respectively.

Metric	Linguistic Model		Neural Model			Hybrid Model	
	In-domain	Cross-domain	Transformer	In-domain	Cross-domain	In-domain	Cross-domain
Acc.	0.276	0.263 (-0.013)	BERT	0.303	0.197 (-0.106)	0.342	0.290 (-0.052)
			BERT-wwm	0.308	0.196 (-0.112)	0.341	0.278 (-0.063)
			RoBERTa	0.293	0.196 (-0.097)	0.341	0.301 (-0.040)
			MacBERT	0.333	0.239 (-0.094)	**0.353**	**0.309** (-0.044)
Adj. Acc.	0.596	0.561 (-0.035)	BERT	0.618	0.504 (-0.114)	0.690	0.656 (-0.034)
			BERT-wwm	0.627	0.485 (-0.142)	0.688	0.639 (-0.049)
			RoBERTa	0.599	0.488 (-0.111)	**0.699**	**0.683** (-0.016)
			MacBERT	0.644	0.563 (-0.081)	0.685	0.677 (-0.008)
F1	0.259	0.221 (-0.038)	BERT	0.273	0.154 (-0.119)	0.338	0.262 (0.076)
			BERT-wwm	0.280	0.154 (-0.126)	0.337	0.249 (-0.088)
			RoBERTa	0.256	0.147 (-0.109)	0.335	0.273 (-0.062)
			MacBERT	0.307	0.198 (-0.109)	**0.348**	**0.276** (-0.072)
QWK	0.739	0.475 (-0.264)	BERT	0.759	0.633 (-0.126)	**0.841**	0.817 (-0.024)
			BERT-wwm	0.755	0.612 (-0.143)	0.833	0.782 (-0.051)
			RoBERTa	0.731	0.597 (-0.134)	0.841	0.822 (-0.019)
			MacBERT	0.768	0.712 (-0.056)	0.829	**0.832** (+0.003)

Table 4: Performance of the Hybrid Model and the two baselines. The gap between in-domain and cross-domain performance is shown in brackets

6.1 Automatic Readability Assessment (ARA)

In the cross-domain evaluation, we used the entire Corpus of Mainland Textbooks (Section 5.2) as training data, and the entire Corpus of Hong Kong Textbooks (Section 5.1) as test data. In the in-domain evaluation, we used stratified 5-fold cross-validation on the Corpus of Mainland Textbooks. Among the three classifiers, Random Forest (RF) outperformed SVM and XGBoost (XGB) in most settings and metrics. The rest of the report will focus on results based on RF.

6.1.1 Hybrid model types

Table 3 reports the performance of the three hybrid model types (Section 4.1.3). For in-domain evaluation, hybrid models with soft labels outperformed those with hard labels and sentence embeddings, regardless of the transformer. For cross-domain evaluation, that was also the case for BERT, RoBERTa and MacBERT. The only exception was BERT-wwm, for which hard labels and embeddings performed slightly better (0.283), but still less accurate than the other transformers. The results presented below will be based on soft labels.

6.1.2 In-domain evaluation

Baselines. As shown in Table 4, the Linguistic Model yielded 0.276 accuracy in the in-domain setting. It was outperformed by the Neural Model regardless of the transformer used. MacBERT achieved the best performance for the Neural Model on accuracy (0.333) and all other metrics.

Hybrid Model. The Hybrid Model trained on MacBERT attained the highest accuracy (0.353) and F1, while RoBERTa led to the best adjacency accuracy and QWK (tied with BERT). Regardless of the choice of transformer or metric, the Hybrid Model outperformed both baselines. The absolute accuracy gains over the Neural Model ranged from 2.0% (MacBERT) to 4.8% (RoBERTa). The improvement is statistically significant for all four models at $p < 0.01$ according to McNemar's Test with continuity correction. Consistent with previous results on English datasets (Lee et al., 2021), linguistic features enhance the performance of neural models on the Chinese datasets.

Training dataset size	Linguistic Model		Neural Model		Hybrid Model	
	In-domain	Cross-domain	In-domain	Cross-domain	In-domain	Cross-domain
20%	0.281	0.247 (-0.034)	0.267	0.231 (-0.036)	0.325	0.294 (-0.031)
60%	0.286	0.259 (-0.027)	0.307	0.236 (-0.071)	0.337	0.299 (-0.036)
100%	0.276	0.263 (-0.013)	0.333	0.239 (-0.106)	0.353	0.309 (-0.044)

Table 5: Model accuracy at different training dataset size, expressed in percentage of the full dataset. The gap between in-domain and cross-domain performance is shown in brackets

6.1.3 Cross-domain evaluation

Our two datasets of Chinese-language textbook materials, graded under comparable scales but drawn from different sources, facilitate a cross-domain evaluation on ARA performance.

Baselines. As expected, model performance degraded in the cross-domain setting. MacBERT produced the best-performing Neural Model in terms of all four metrics. Unlike the in-domain evaluation, the Linguistic Model outperformed the Neural Model in terms of accuracy (0.263 vs. 0.239) and F1, though worse in terms of adjacent accuracy and QWK. Its competitive performance can be attributed to the robustness of linguistic features in the face of dissimilar materials. While the Linguistic Model degraded only slightly (-0.013) in accuracy on cross-domain data, the Neural Model suffered a much more substantial drop (-0.094).

Hybrid Model. The Hybrid Model outperformed both baselines in all metrics and all transformers. The improvement of the hybrid model over the neural model is statistically significant for BERT, BERT-wwm and RoBERTa at $p < 0.00001$ according to McNemar's Test. MacBERT again led to the best performance in terms of accuracy (0.309), F1 and QWK, but was slightly worse than RoBERTa in adjacent accuracy.

The superior performance of the Hybrid Model resulted from its smaller degradation on cross-domain data. This can be seen by the gap between in-domain and cross-domain performance, shown in brackets in the "Cross-domain" column in Table 4. For all transformers and all metrics, the gap was substantially smaller with the Hybrid Model. For example, the gap was only 0.044 cross-domain but more than doubled (0.094) in-domain for MacBERT. This suggests that some textual characteristics learned by the Neural Model may be only accidentally correlated with readability in the training corpus, while the Hybrid Model benefits from linguistic features that are more generally relevant to readability and therefore transferable to new domain.

Our hypothesis can be corroborated with the analysis on various dataset sizes in Table 5. When trained on only 20% of the dataset, all three models exhibited a similar gap between in-domain and cross-domain performance. With additional training data, the Neural Model became more accurate in-domain (0.267 to 0.333). However, the improvement hardly carried over cross-domain, leading to a growing performance gap (-0.036 to -0.106), possibly indicating overfit to corpus-specific textual characteristics. In contrast, the gap shrank for the Linguistic Model, and remained relatively stable for the Hybrid Model, even as it improved steadily in accuracy.

6.1.4 Surface vs. semantic features

All experimental results are based on stratified 10-fold cross-validation. Table 6 compares the performance of the Surface Model, which is trained on surface features only, and the Frame

Model	RF classification accuracy			SVM classification accuracy		
	Overall	Easy	Difficult	Overall	Easy	Difficult
Surface Model	83.09	82.58	83.60	80.90	81.69	80.11
Frame Model	**87.30**	**87.19**	**87.42**	**82.70**	**84.83**	**80.56**

Table 6: Syntactic vs. semantic features: classification accuracy in percentage (Section 6.1.4)

Model, which is trained on both surface and semantic features based on verb frames (Section 4.1.4).

The Surface Model derived from the RF classifier provided a strong baseline with 83.09% accuracy, based only on sentence length, graded character lists and word frequency lists. By incorporating the verb frame features, the Frame Model outperformed the Surface Model at 87.30% accuracy. The improvement is statistically significant at $p < 0.05$ by McNemar's Test. The Frame Model achieved improved performance among both easy passages and difficult passages. On the one hand, it was able to recognize some difficult passages that appear easy on the surface. On the other hand, it can also recognize some easy passages despite seeming difficulty on the surface.

The Frame Model also outperformed the Surface Model when trained with the SVM classifier. The classification accuracy is similar on the easy and difficult passages for both models, likely attributable to the balanced dataset.

6.2 Vocabulary Profiling

We evaluate the quality of the word segmentation on their ability to predict the difficulty of a text. One way to predict the difficulty grade is according to the vocabulary coverage at the grade. While there is no consensus on the minimum vocabulary coverage for text comprehension, it is generally agreed that over 90% is required (Laufer, 1989). In our experiment, we set 95% as the threshold; in other words, the lowest grade at which vocabulary coverage exceeds 95% would be predicted as the grade of the input text. We compare the following settings for word segmentation:

Without FMM Use the word segmentation output directly from the HanLP parser.

With FMM Enhance the word segmentation output with FMM, following the procedure outlined in Section 4.2.2.

With FMM and dynamic revision Use the revised segmentation after user editing of word boundaries on the vocabulary profiler.

6.2.1 ARA accuracy

We performed an automatic evaluation to compare 'Without FMM' and 'With FMM'. As shown in Table 7, direct use of the segmentation output by the parser yielded an accuracy of 28.75% and adjacent accuracy of 67.50% in grade prediction. The use of FMM led to an improvement of over 7% absolute, raising the accuracy and adjacent accuracy to 36.25% and 75%, respectively. Segmentation errors could be due to both categorial (whether the string AB is a single word or should be split into A and B) and overlapping ambiguities (whether AB or BC forms a word in the string ABC) (Maosong and Tsou, 1995). Anecdotal analysis suggested that the HanLP output contained mostly overlapping errors, and the FMM made accuracy gains by resolving these errors.

Approach	Accuracy	Adjacent Accuracy
With FMM	**36.25%**	**75.00%**
Without FMM	28.75%	67.50%

Table 7: Evaluation of Forward Maximal Matching word segmentation on ARA accuracy (Section 6.2)

Statistics	Passage 1	Passage 2	Passage 3
# difficult words (Grade 3 or above)	6	9	11
# difficult words with revision options	5	8	7
# revision options	7	16	9
# usable revision options	6	12	7

Table 8: Results in the user study on text revision (Section 6.3)

6.2.2 Tool usability

The usability of the tool depends on whether the user can successfully correct word segmentation to gain more accurate profiling results. To answer this question, we performed a user study to compare 'With FMM' and 'With FMM and dynamic revision'.

We recruited 7 students at a university in Hong Kong to participate in the study. All native speakers of Chinese, they were asked to predict the grade level of 13 passages with the tool. The average length of these passages was 260 characters. The subjects were instructed to examine the vocabulary profile of each passage, and revise the word segmentation if necessary. As in the automatic evaluation, 95% of vocabulary coverage was taken as the minimum threshold for the text to be suitable for students at that grade.

In the 'With FMM' approach, where the grades of the passages were predicted without intervention from the subjects, the Mean Average Error (MAE) in grade prediction was 2.23.

In the 'With FMM and dynamic revision' approach, the grades were predicted using the word boundaries after revision by the subjects. This approach reduced the MAE to 2.15. On the one hand, for 29.7% (27/91) of the passages, the subjects predicted a grade that was closer to the correct grade than the original. On the other hand, the predicted grade was further away from the correct one in only 16.5% (15/91) of the passages. These results suggest that our profiler can support users in revising the segmentation to achieve more accurate text difficulty assessment.

6.3 Text revision

We conducted a user study on text revision using *Text Difficulty Analyzer*. The subjects of the study were drawn from students enrolled in the Master of Language Studies Program at City University of Hong Kong. We recruited only those specializing in the Language Pedagogy Stream. Since these students are preparing for a career in language teaching, they constitute our target users and their feedback would be especially relevant for this project.

6.3.1 User study set-up

A total of 92 subjects participated in the study. Before the study, they were instructed to watch the demonstration video on the website of the *Text Difficulty Analyzer*. The video showed them how to estimate the grade level of a passage, to analyze vocabulary coverage and to use the drop-down lists to replace difficult vocabulary with simpler alternatives. The subjects were then given three Chinese passages, and were asked to revise one or more of them for Grade 2 students to read. At the end of the study, they anonymously submitted their revised version(s), and filled out an anonymous questionnaire with the three questions shown in Table 9.

Question	Response				
	Strongly agree	Agree	Neutral	Dis-agree	Strongly disagree
(1) Did you find it easy to use this website for predicting text difficulty?	41.2%	**50.6%**	7.1%	1.2%	0.0%
(2) Did you find the website useful for identifying difficult words in a passage?	25.9%	**58.8%**	10.6%	4.7%	0.0%
(3) Did you find the website helpful for simplifying a passage?	10.6%	**43.5%**	29.4%	16.5%	0.0%

Table 9: Questionnaire result from user study of text revision (Section 6.3.3)

6.3.2 Revision results

All three passages in the study were at the Grade 3 level. There were a total of 107 revision attempts, with 37 revised versions for Passage 1, 27 for Passage 2 and 43 for Passage 3. Of the 107 attempts, 86% (92) were successful in simplifying the passage so that the assessment of the *Text Difficulty Analyzer* was lowered from Grade 3 to Grade 2.

Revision support. Since the subjects' task was to revise the text for Grade 2 students to read, all words listed at Grade 3 or above in the EDB Graded Vocabulary List were considered 'difficult'. As shown in Table 8, the number of difficult words in the three passages ranged from 6 to 11. With the relatively low target grade (Grade 2), it could be challenging to find suitable alternatives to preserve the meaning of the text. Nonetheless, *Text Difficulty Analyzer* was able to suggest synonyms for most of the difficult words in all three passages, ranging from 7 to 16 options. On average, 1.6 alternative words were offered for each difficult word.

Revision outcomes. The subjects selected options from all of the provided drop-down lists. We call an option *usable* if it is chosen by at least one subject in the study. As shown in Table 8, evaluation results suggest that most options are deemed useful to some of the subjects, with 85.7% (6/7) of the options being usable in Passage 1; 75.0% (12/16) in Passage 2; and 77.7% (7/9) in Passage 3.

6.3.3 Questionnaire results

As shown in Table 9, the subjects were polled on their opinion on the user-friendliness of the *Text Difficulty Analyzer* (Question 1), its effectiveness in identifying difficult vocabulary (Question 2) and in text revision (Question 3). In response to Question 1, over 91% of the subjects agreed or strongly agreed with that the web interface was easy to use. For Question 2, a majority of the subjects (84.7%) agreed or strongly agreed that the website was useful for identifying difficult words in a text. Finally, when asked about their experience with the drop-down lists in Question 3, a slight majority (54.1%) of the users agreed or strongly agreed that the tool was helpful for text simplification.

7 Conclusions and Recommendation

The curation of reading materials for students can be a time-consuming task for language teachers. We have presented a *Text Difficulty Analyzer*, a tool that can perform automatic readability assessment (ARA) on a Chinese text to determine its grade. To help the user adapt the text as reading material for students at the target grade, the tool analyzes its vocabulary complexity, and estimates the percentage of words that are known to the typical student. Further, it suggests word substitutions that better match their expected vocabulary proficiency.

Experimental results on a corpus of Chinese textbook passages show that a hybrid ARA model outperforms the neural model and linguistic model, in both the in-domain and cross-domain settings. Analyses suggest that this is due to the robustness of linguistic features, which enable more stable performance even with fewer training samples and on texts in other domains. In evaluations, users were able to assess the difficulty of a text and to revise it efficiently with the tool. In particular, Forward Maximal Matching and dynamic revision of word segmentation were shown to lead to more accurate assessment, and revision suggestions were mostly deemed helpful by some users.

In conclusion, we make the following recommendations. First, researchers in computational linguistics should pursue further research on linguistic features, to investigate if and how they could help capture more generalizable characteristics for text difficulty assessment based on Large Language Models. Second, language teachers and other frontline practitioners should make use of automatic text difficulty assessment, which can offer an objective estimate as a complement to one's professional judgment. Third, during preparation of pedagogical or examination materials, it is well worth considering the use of a text analyzer tool to identify words that are above the target grade, and to obtain editing suggestions for more efficient revision. Finally, schools should be informed of the potential advantages and pitfalls of integrating automatic text difficulty assessment into the process of lesson preparations and examination paper review.

References

Agrawal, S. and Carpuat, M. (2023). Controlling pre-trained language models for grade-specific text simplification. In et al., W. M., editor, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 12807–12819.

Anthony, L. (2023). *AntWordProfiler (Version 2.1.0)*. http://www.antlab.sci.waseda.ac.jp/, Waseda University, Tokyo, Japan.

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

Belder, J. D. and Moens, M. F. (2010). Text Simplification for Children. In *Proc. SIGIR Workshop on Accessible Search Systems*.

Bo, W., Chen, J., Guo, K., and Jin, T. (2019). Data-Driven Adapting for Fine-Tuning Chinese Teaching Materials: Using Corpora as Benchmarks. In Lu, X. and Chen, B., editors, *Computational and Corpus Approaches to Chinese Language Learning. Chinese Language Learning Sciences*, Singapore. Springer.

Capel, A. (2015). The English Vocabulary Profile. In Harrison, J. and Barker, F., editors, *English profile in practice*, page 9–27, Cambridge, UK. Cambridge University Press.

Chen, X. and Meurers, D. (2016). Characterizing Text Difficulty with Word Frequencies. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, page 84–94.

Cheng, Y., Xu, D., and Lv, X. (2019). Automatically Grading Text Difficulty with Multiple Features. *Data Analysis and Knowledge Discovery*, 3(7):103–112.

Chu, C. (2005). ChineseTA 1.1. Beijing, China. Beijing Language and Culture University Press.

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., and Hu, G. (2020). Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP*. Association for Computational Linguistics.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proc. 2nd Workshop on Speech and Language Processing for Assistive Technologies*.

Deutsch, T., Jasbi, M., and Shieber, S. (2020). Linguistic Features for Readability Assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*.

Filighera, A., Steuer, T., and Rensing, C. (2019). Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, page 335–348. Springer.

Fillmore, C. J. and Atkins, B. T. (1992). Towards a Frame-based organization of the lexicon: the semantics of RISK and its neighbors. In Lehrer, A. and Kittay, E., editors, *Frames, Fields, and Contrasts: New Essays in Semantics*, pages 75–102, Hillsdale. Lawrence Erlbuan.

François, T. and Fairon, C. (2012). An "AI Readability" Formula for French as a Foreign Language. In *Proc. EMNLP-CONLL*.

Gao, H. (2001). A specification system for measuring relationship among near-synonyms of physical action verbs. In *Proceedings of the 2nd Chinese Lexical Semantics Workshop*, pages 45–51. Citeseer.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago.

Hanban (2014). *International Curriculum for Chinese Language and Education*. Beijing Language and Culture University Press, Beijing, China.

Imperial, J. M. (2021). BERT Embeddings for Automatic Readability Assessment. In *Proc. Recent Advances in Natural Language Processing*, page 611–618.

Jill Burstein, Jane Shore, J. S. Y.-W. L. and Ventura, M. (2007). The Automated Text Adaptation Tool. In *Proc. NAACL-HLT Demonstration Program*.

Jin, T., Lu, X., Lin, Y., and Li, B. (2018). *Chi-Editor: An online Chinese text evaluation and adaptation system*. LanguageData (languagedata.net/editor), Guangzhou, China.

Kajiwara, T., Matsumoto, H., and Yamamoto, K. (2013). Selecting Proper Lexical Paraphrase for Children. In *Proc. 25th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 59–73.

Kim, Y.-J. (2000). Subject/object drop in the acquisition of Korean: A cross-linguistic comparison. *Journal of East Asian Linguistics*, 9:325–351.

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

Krashen, S. D. (1981). The fundamental pedagogical principle in second language teaching. *Studia Linguistica*, 35(1-2):50–70.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Laufer, B. (1989). What percentage of text is essential for comprehension? In Lauren, C. and Nordman, M., editors, *Special Language; from Humans Thinking to Thinking Machines*, pages 316–323, Clevedon. Multilingual Matters Ltd.

Lee, B. W., Jang, Y. S., and Lee, J. H.-J. (2021). Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Lee, J., Liu, M., and Cai, T. (2020). Using Verb Frames for Text Difficulty Assessment. In *Proc. International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*.

Lee, J. and Vajjala, S. (2020). A Neural Pairwise Ranking Model for Readability Assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813.

Lee, J., Zhao, W., and Xie, W. (2016). A Customizable Editor for Text Simplification. In *Proc. Proc. 26th International Conference on Computational Linguistics (COLING): System Demonstrations*.

Lee, J. S. Y. and Uvaliyev, A. (2023). Automatic Generation of Vocabulary Lists with Multiword Expressions. In *Proc. 19th Workshop on Multiword Expressions (MWE 2023)*.

Lee, J. S. Y., Yeung, C. Y., and Yang, Z. (2022). Personalized and adaptive text recommendation for learners of Chinese. *Interactive Learning Environments*.

Lei, W., Xiang, Y., YuweiWang, Zhong, Q., Liu, M., and Kan, M.-Y. (2018). Linguistic Properties Matter for Implicit Discourse Relation Recognition: Combining Semantic Interaction, Topic Continuity and Attribution. In *Proc. AAAI*, pages 4849–4855.

Lim, H. H., Cai, T., Lee, J. S. Y., and Liu, M. (2022). Robustness of Hybrid Models in Cross-domain Readability Assessment. In *Proc. 20th Workshop of the Australasian Language Technology Association (ALTA)*.

Liu, M. and Chiang, T. Y. (2008). The construction of mandarin verbnet: A frame-based study of statement verbs. *Language and Linguistics*, 9(2):239–270.

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

Maosong, S. and Tsou, B. K. (1995). Ambiguity resolution in chinese word segmentation. In *Proc. 10th Pacific Asia Conference on Language, Information and Computation*.

Martinc, M., Pollak, S., Marko, and Robnik-Šikonja (2021). Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., and Cai, Z. (2014). *Automated evaluation of text and discourse with Coh Metrix*. Cambridge University Press.

Paetzold, G. H. and Specia, L. (2017). Lexical Simplification with Neural Ranking. In *Proc. EACL*.

Panunzi, A. and Vernillo, P. (2019). Metaphor in action. action verbs and abstract meaning. *Perspectives on abstract concepts. Cognition, language and communication*, pages 215–237.

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

Pitler, E. and Nenkova, A. (2008). Revisiting Readability: a Unified Framework for Predicting Text Quality. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiang, J., Lu, X., Li, Y., Yuan, Y.-H., and Wu, X. (2021). Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. EMNLP-IJCNLP*.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proc. 15th International Workshop on Semantic Evaluation (SemEval)*.

Siddharthan, A. (2002). An Architecture for a Text Simplification System. In *Proc. Language Engineering Conference (LEC)*.

So, X. (2019). Yiwu Jiaoyu Changyong CibiaoCaoan). *Huayu Xuekan* .

Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. (2018). Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Sung, Y.-T., Chang, T.-H., Lin, W.-C., Hsieh, K.-S., and Chang, K.-E. (2016). Crie: An automated analyzer for chinese texts. *Behavior Research Methods*, 48:1238–1251.

Sung, Y.-T., Lin, W.-C., Dyson, S. B., Chang, K.-E., and Chen, Y.-C. (2015). Leveling L2 Texts Through Readability: Combining Multilevel Linguistic Features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Tseng, H.-C., Chen, H.-C., Chang, K.-E., Sung, Y.-T., and Chen, B. (2019). An Innovative BERT-Based Readability Model. In Rønningsbakk, L., Wu, T. T., Sandnes, F., and Huang, Y. M., editors, *Lecture Notes in Computer Science, vol 11937*.

Tseng, W.-H. (2014). Huayu baqianci ciliang fenji yanjiu (Classification on Chinese 8000 Vocabulary). *Huayu Xuekan* , 6:22–33.

Štajner, S., Ponzetto, S. P., and Stuckenschmidt, H. (2017). Automatic Assessment of Absolute Sentence Complexity. In *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI)*.

Xia, M., Kochmar, E., and Briscoe, T. (2016). Text readability assessment for second language learners. In *Proc. 11th Workshop on Innovative Use of NLP for Building Educational Applications*, page 12–22.

Xue, N., Xia, F., Chiou, F. D., and Palmer, M. (2005). The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11:207–238.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A Report on the Complex Word Identification Shared Task 2018. In *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*.