https://doi.org/10.17323/jle.2024.22221

Predictions of Multilevel Linguistic Features to Readability of Hong Kong Primary School Textbooks: A Machine Learning Based Exploration

Zhengye Xu, Yixun Li, Duo Liu

The Education University of Hong Kong, Tai Po, N.T., Hong Kong, China

ABSTRACT

Introduction: Readability formulas are crucial for identifying suitable texts for children's reading development. Traditional formulas, however, are linear models designed for alphabetic languages and struggle with numerous predictors.

Purpose: To develop advanced readability formulas for Chinese texts using machine-learning algorithms that can handle hundreds of predictors. It is also the first readability formula developed in Hong Kong.

Method: The corpus comprised 723 texts from 72 Chinese language arts textbooks used in public primary schools. The study considered 274 linguistic features at the character, word, syntax, and discourse levels as predictor variables. The outcome variables were the publisher-assigned semester scale and the teacher-rated readability level. Fifteen combinations of linguistic features were trained using Support Vector Machine (SVM) and Random Forest (RF) algorithms. Model performance was evaluated by prediction accuracy and the mean absolute error between predicted and actual readability. For both publisher-assigned and teacher-rated readability, the all-level-feature-RF and character-level-feature-RF models performed the best. The top 10 predictive features of the two optimal models were analyzed.

Results: Among the publisher-assigned and subjective readability measures, the all-RF and character-RF models performed the best. The feature importance analyses of these two optimal models highlight the significance of character learning sequences, character frequency, and word frequency in estimating text readability in the Chinese context of Hong Kong. In addition, the findings suggest that publishers might rely on diverse information sources to assign semesters, whereas teachers likely prefer to utilize indices that can be directly derived from the texts themselves to gauge readability levels.

Conclusion: The findings highlight the importance of character-level features, particularly the timing of a character's introduction in the textbook, in predicting text readability in the Hong Kong Chinese context.

KEYWORDS

Chinese, linguistic features, Random Forest, readability models, Support Vector Machine

Citation: Xu Z., Li Y., & Liu D.(2024). Predictions of Multilevel Linguistic Features to Readability of Hong Kong Primary School Textbooks: A Machine Learning Based Exploration. *Journal of Language and Education, 10*(4), 146-158. https://doi.org/10.17323/jle.2024.22221

Correspondence:
Duo Liu,
duoliu@eduhk.hk

Received: August 14, 2024
Accepted: December 16, 2024
Published: December 30, 2024

INTRODUCTION

Text readability refers to the ease with which a text can be read and understood (Crossley et al., 2019). A number of studies across languages have found that reader-level characteristics, such as linguistic knowledge and motivation, can influence text readability (Stutz et al., 2016; Zhang et al., 2014). At the same time, text-level linguistic features, such as word frequency and sentence length, also play pivotal roles in text readability (Crossley et al., 2023; Mesmer, 2005). While reader-level characteristics have been extensively explored in reading research (McBride-Chang et al., 2005; Stutz et al., 2016), text-level features in the context of text readability, particularly in the Chinese language, have received less attention (Crossley et al., 2023; Fitzgerald et al., 2015; Sung et al., 2015). To address this gap, this study examined how text-level features affect text readability

in Hong Kong primary school Chinese textbooks. The findings aim to improve the alignment between text and children's reading abilities, thereby enhancing learning efficiency in Chinese.

Text Readability and Linguistic Features at Different Levels

Text readability can be quantified by constructing a readability formula (Crossley et al., 2019), which provides an overview of text difficulty. It shows promise in benchmarking children's text-difficulty ability levels more accurately, thus allowing them to read texts at target readability levels. These formulas typically result in an absolute score or a grade level that indicates the level of text an average reader in that grade is expected to be able to read and understand successfully (Kincaid et al., 1975; Solnyshkina et al., 2017). For example, one of the most well-known readability formulas, the Flesch-Kincaid grade level formula (Kincaid et al., 1975), (0.39 × the average number of words used per sentence) + (11.8 × the average number of syllables per word) – 15.59, is designed to result in a grade level that indicates a text's readability. For example, a score of 5.3 indicates that the text is appropriate for fifth graders.

These formulas usually consider a few linguistic features, however, research has shown that features relating to word, syntax, and discourse levels significantly affect text comprehension in various languages, such as English and Chinese (Crossley et al., 2019; Liu et al., 2024; Pinney et al., 2024; Solnyshkina et al., 2017). At the word level, word length, i.e., the number of characters per word, is a key indicator of text readability. Longer words typically signify more challenging texts, while shorter words suggest easier comprehension (Crossley et al., 2023; Mesmer & Hiebert, 2015). Word diversity, which reflects the range of different words used in a text, also influences readability (Sung et al., 2015). Word frequency and psycholinguistic-related indexes, particularly reaction time and error rate in lexical decision tasks, have been associated with text readability (Tsang et al., 2018; Tse et al., 2017). In addition to being recognized, the meanings of words are required for successful understanding, resulting in a role for semantic information of words in text reading (Mesmer & Hiebert, 2015). Also, part-of-speech of words (the grammatical category or classification of words in a language based on their functions and roles within a sentence, e.g., nouns and verbs) influence text readability since higher readability levels of texts are generally associated with higher proportions of conjunction words and adverbs, whereas lower readability text levels are linked to higher proportions of adjectives and modal words (Liu et al., 2024).

At the syntax level, sentence length is important. Longer sentences and greater distances between related words in a sentence imply higher syntactic complexity (Crossley et al., 2023). Word dependency, the average distance between two related words in a sentence, has also been found to be related to syntactic complexity (Crossley et al., 2023). A sentence can be easier if the average distance between two related words is shorter (Crossley et al., 2019). Sentence grammar, which encompasses logical relationships within a sentence, can also contribute to the complexity of syntax (Graesser et al., 2011).

Discourse-level factors, primarily the relationships between the sentences of a text, also impact readability (Pinney et al., 2024). These discourse structures, tied to text cohesion, can influence how clearly a noun, pronoun, or noun phrase can be linked to another element (Givón, 1995). Causal cohesion, related to connective indices, can reduce text readability by building relationships between words, concepts, and paragraphs (Graesser et al., 2011).

Text Readability Research in Chinese

Text readability research in Chinese incorporates word, syntax, and discourse-level features as in alphabetic languages, but also considers character-level features due to that the character is the basic writing unit in the Chinese language (Cheng et al., 2020; Sung et al., 2015). Specifically, a character can stand alone to form a one-character word (e.g., 筆/bat1/ *pen*) or can be combined with others to form two-character words (e.g., 筆記/bat1-gei3/ *note*), or three- or more-character words (e.g., 筆記本/bat1-gei3-bun2/ *notebook,*). Each Chinese character has its own form, sound, and meaning(s); therefore, related linguistic features attached to characters can influence text readability (Sung et al., 2015). Traditional Chinese text readability formulas include character-level features, like the average number of characters, but often overlook other influential factors, particularly at the discourse level, such as text cohesion (Cheng et al., 2020; Jing, 1995). They also assume a linear relationship between readability level and linguistic features, limiting the accuracy of the model (Rodriguez-Galiano et al., 2015).

To address these issues, machine learning techniques have been employed to improve readability estimation. Unlike traditional formulas, machine learning can handle a large number of linguistic features and identify complex relationships among them (Rodriguez-Galiano et al., 2015). This approach presents the predicted readability level as a category (e.g., Grade 5), indicating the appropriate reading level for readers in that grade. In addition to aiding reader-text matching, the machine learning approach can enhance our understanding of text readability by identifying key linguistic features (Rodriguez-Galiano et al., 2015). For instance, Fitzgerald et al. (2015) analyzed 238 features and determined that nine features related to word structure, semantics, and cohesion were crucial for understanding English text complexity. Therefore, the current study employed machine learning approaches to provide insights into text readability.

Machine Learning Based Text Readability Formulas in Chinese

Machine learning techniques have been utilized to explore text readability in Chinese, with studies primarily focusing

on the support vector machine (SVM) algorithm (e.g., Chen et al., 2011; Sung et al., 2015; Wu et al., 2020). These studies, mostly conducted in Taiwan, used traditional Chinese writing systems and multilevel linguistic features to train SVM models for classifying text readability, achieving high accuracy rates for lower (first and second, 95%) and middle-grade levels (third and fourth grades, 84%; Chen et al., 2013). For a more nuanced classification, using the grade level (i.e., Grades 1–6) as the indicator for text readability, Sung et al. (2015) combined SVM with 31 linguistic features from lexical, semantic, syntax, and discourse levels. Sung et al. (2015) found that models incorporating features from multilevel offered higher accuracy in predicting text readability (71.75%) than models using features from a single level (43.97%–65.13%).

Research on text readability in simplified Chinese, predominantly used in Mainland China, has also been conducted. Wu et al. (2020) utilized SVM models to examine the impact on text readability of 104 linguistic features from character, word (comprising two or more characters), syntax, and discourse. The findings of Wu et al. (2020) indicated that among the models with single-level features, the word-level features (accuracy: 62.1%) performed the best. Moreover, the inclusion of character (accuracy: 63.8%) and syntax (accuracy: 63.1%) level features improved prediction accuracy more than the word-level model did.

A recent study (Liu et al., 2024) examined linguistic features on simplified Chinese text readability using a detailed semester-level scale (i.e., 1-12). They used the random forest (RF) and SVM algorithms along with numerous lexical and discourse features, confirming that models using features from multiple levels outperformed those using features from a single level, with higher accuracies (RF: 27%; SVM: 28%) and lower mean absolute error, the average absolute difference between the true and predicted readability levels (MAE, RF: 1.24; SVM: 1.25). Furthermore, Liu et al. (2024) identified that character and word frequency, semantic features, lexical diversity, syntactic categories, and referential cohesion were the most important features.

However, compared to the situations in Mainland China and Taiwan, less attention has been paid to text readability in Hong Kong, where the Chinese community has unique text-related features that differ from those in Mainland China and Taiwan (McBride-Chang et al., 2005). To address this gap, the current study aimed to develop an appropriate model for approximating text readability in Hong Kong.

The Present Study

The present study focuses on text readability in Hong Kong, where the traditional writing system is used, and texts are processed in Cantonese, differing from Mainland China and Taiwan. Cantonese possesses some unique features, such as additional tones and vocabulary, specific spoken language terms, and regional variations. For example, the character 是/si6/ is used in written language, while 係/hai6/ is more commonly used in spoken language to express the meanings of *yes*. Moreover, in the spoken language, Cantonese has some words to indicate the ends of utterances, such as 啊/aa3/, 㗎/gaa3/, and 囉/lo1/, which are not commonly used in formal books. Also, some terms used in Hong Kong differ from those used in Mainland China and Taiwan. For instance, the concepts of *bus* and *taxi* are often represented by 巴士/baa1-si6/ and 的士/dik1-si6/, respectively, in Hong Kong. However, they are expressed as 公交/gong1-jiao1/ and 出租车/chu1-zu1-che1/, respectively, in Mainland China, and公車/gong1-che1/ and 計程車/ji4-cheng2-che1/, respectively, in Taiwan. These features of Cantonese make it necessary to develop readability formulas using a corpus developed with locally used texts (McBride-Chang et al., 2005).

This study uses a corpus of articles from Chinese language arts textbooks commonly used in Hong Kong, particularly for primary school students. Following previous studies in the Chinese language (e.g., Liu et al., 2024; Sung et al., 2015), the study incorporates linguistic features from character, word, syntax, and discourse levels to estimate text readability. It employs a more nuanced scale based on semesters, with a readability level scale of 1-24. The study also uses a subjective indicator, teacher-rated semesters, for each selected text. SVM and RF were adopted, and the importance of linguistic features was analyzed to comprehensively understand text readability in the Chinese language. The current study sought to answer two research questions: 1) Whether and to what extent do the levels of features affect text readability models' performance? 2) What are the features that are most important to the current best model(s)?

METHOD

Study Design

This study utilized a text corpus from Chinese language arts textbooks for primary school students published by three major Hong Kong publishers. Due to copyright issues, the text materials can not be publicly shared. Each publisher contributed four textbooks per grade level, divided into two textbooks per semester, yielding a total of 72 textbooks. Two research assistants meticulously digitalized and proofread the texts three times to ensure accuracy. The study considered 723 texts after excluding non-passage elements such as ancient Chinese prose, illustrations, tables of contents, bibliographies, and indexes. Then, linguistic features to represent character-, word-, syntax-, and discourse-level characteristics of each text were extracted and calculated using the CKIP Chinese word segmentation system (Ma & Chen, 2005). This study was approved by the Human Research Ethics Committee of The Education University of Hong Kong and conformed to the Declaration of Helsinki.

The study used machine learning models built with scikit-learn version 1.1.2 in Python 3.10 to explore the predictive roles of multiple levels of linguistic features in text readability (Pedregosa et al., 2011). Text readability was represented by two indicators: publisher-assigned semester (Y1) and teacher-rated semester (Y2). The teacher-rated semester was the average ratings of text readability levels of 11 experienced primary school teachers (whose written informed consent to participate in the study was obtained), using a 1-9 scale tailored for an average reader in the corresponding grade. for these 11 teachers. We then assigned a teacher-rated semester to the texts in each publisher-assigned semester based on the rearranged average ratings within each grade. Both Y1 and Y2 ranged from 1-24, with higher values indicating greater text readability. A total of 15 combinations of linguistic features (referred to as Xs) at different levels were developed: character (C), word (W), syntax (S), and discourse (D). These included single-level Xs (C, W, S, D), two-level Xs (C_W, C_S, C_D, W_S, W_D, S_D), three-level Xs (C_W_S, C_W_D, C_S_D, W_S_D), and a four-level X (Xall).

According to Liu et al. (2024) and Sung et al. (2015), two machine learning algorithms, SVM and RF, were employed. A five-fold cross-validation approach was used to evaluate the performance of the machine learning models. The 723 texts were randomly divided into five subsets, with each subset containing an equal percentage of texts from each semester. Four subsets were used for training, and one subset was used for testing in each iteration. The predicted Y values from the models were compared to the actual Y values to assess accuracy and MAE. Linear mixed models (LMMs) were constructed with the the *lmer* package in R 4.0.3 to compare the prediction performance (Baayen et al., 2008). The LMMs used *z*-score transformations to address collinearity and included X, the machine learning algorithm, and their interaction as fixed factors. RF and Xall were used as reference levels. Random intercepts and slopes were included, and a more complex model was accepted if it improved the fit.

After comparing the readability models, the best model(s) for predicting publisher-assigned and teacher-rated semesters were chosen. Then, the importance of each feature was ascertained using *permutation importance* in the Python *ELI5* package, in which feature importance is estimated by measuring how predictor power decreases when a feature is not available (Korobov & Lopuhin, 2019). The loss of predictive power was evaluated by both accuracy and MAE. The data and analysis code are openly available in Open Science Framework.[1]

Linguistic Features

Character-Level Features

In total, 110 character-level features, relating to four aspects: 1) character diversity (N = 4), 2) character structural complexity (N = 8), 3) character frequency (N = 40), and 4) psycholinguistic information for characters (N = 58), were calculated for each text.

Four indicators were considered for character diversity: the raw number of the token (for all characters) count, the raw number of type (for different characters) count, the ratio of type count to token count, and the proportion of characters that only occur once. For example, in one of the texts, «我和妈妈玩捉迷藏, » there are eight token characters. Since the third and fourth characters of the sentence are the same, i.e., 妈, there are seven types of characters.

Character structural complexity was measured with four indicators: the average number of strokes, the proportion of characters with less than ten strokes, between 10 and 20 strokes, and more than 20 strokes. Token count and type count were calculated for each indicator, resulting in eight lexical features for character structural complexity.

Character frequency was measured using five corpora: The Balanced Corpus of Modern Chinese, CNCORPUS (Jin et al., 2005), The SUBTLEX-CH corpus (Cai & Brysbaert, 2010), Sinica Corpus (Huang, 2006), Chinese text computing (Da, 2004), and Hong Kong, Mainland China & Taiwan: Chinese character frequency (this corpus will be identified as HK-MCT hereafter[2]). Characters that were not found in a given corpus were considered difficult characters. Four indicators were considered: the average frequency scores for frequent characters, the standard deviation of frequency scores for frequent characters, the raw number of difficult characters, and the proportion of difficult characters. Token count and type count were used to calculate these indicators.

Psycholinguistic information for characters was assessed using 26 indicators from previous studies (i.e., Liu et al., 2007; Su et al., 2023). These indicators were: the age at which it is expected a particular character can be learned, character familiarity, the ease of describing the meaning of a character, the ease of creating an image of a character, the grade and semester in which a character is first introduced in the textbook, the number of meanings of a character, the number of homophones of a character, the summed frequency of

[1] https://osf.io/adqw7/?view_only=c4343a96bd86419b88a8e11d1e0c4426

[2] https://humanum.arts.cuhk.edu.hk/Lexis/chifreq/

all characters that share the same pronunciation (calculated based on the five aforementioned corpora), the number of words that a character can form, the summed frequency of all words that contain a character (calculated based on the five aforementioned corpora), the availability of pronunciation cues in a character (1 for a reliable cue, 0 for the absence of cues, and -1 for unreliable/misleading cues), and the reaction times and error rates for character naming by Chinese adults.

Two indicators concerning pronunciation cues in characters were included: the proportion of characters with reliable pronunciation cues (the pronunciation cue has the same pronunciation as the character) and those with unreliable/misleading pronunciation cues (the sounds of the pronunciation cue and its corresponding character are different; Su et al., 2023). Semantic radical transparency of characters, which refers to the degree of meaning correspondence between the semantic radical and the whole character, was also involved. For instance, while both 海/hoi2/ (sea) and 测/cak1/ (measure) contain the semantic radical氵 (a variant of 水/seoi2/ water), the former is semantically transparent and the latter opaque. Token count and type count were calculated for each indicator, resulting in 58 lexical features for this category.

Word-Level Features

A total of 105 word-level features was calculated for each text, covering six aspects: word length (N =12), word diversity (N = 4), word frequency (N = 32), psycholinguistic information for words (N = 22), set structure (N = 1), and part-of-speech syntactic categories (N = 34).

For word length, six facets were considered: the average word length and the percentages of one-character, two-character, three-character, four-character, and five-or-more-character words in a text. Token count and type count were calculated for each facet, resulting in 12 linguistic features.

Word diversity refers to the richness of words in a text. Four indicators were considered: the raw number of words (both token count and type count), the ratio of type count to token count, and the proportion of words that only occurred once.

Word frequency was measured based on the frequency of a word in five corpora, except for HKMCT, as it does not have statistics for word frequency. Four indicators were considered: average frequency scores for frequent words, standard deviation of frequency scores for frequent words, raw number of difficult words, and proportion of difficult words. Token count and type count were used to calculate these indicators, resulting in 32 lexical features.

Psycholinguistic information for words was calculated based on the corpus MELD-SCH (Tsang et al., 2018), which provided reaction times and error rates for Chinese adults. Twelve features were calculated based on the mean and standard deviation of reaction times and error rates. Words that are not included in MELD-SCH were considered low-frequent words, which were identified by their proportions and the raw numbers. The semantic radical transparency of two-character words was extracted based on the work of Su et al. (2023). Three indicators were considered for each character and for the word as a whole. For all psycholinguistic features, both token count and type count were calculated for each facet.

The set structure was measured by calculating the raw number of named entities (e.g., the names of people, organizations, and locations) in a text using HanLP[3]. Part-of-speech syntactic categories were determined by assigning one of 16 categories to each word using the Natural Language Processing & Information Retrieval Sharing Platform (Liu et al., 2004). The 16 categories include nine types of content words (nouns, verbs, adjectives, numerals, quantifiers, pronouns, time words, place words, and position words) and seven types of function words (adverbs, prepositions, conjunctions, particles, interjections, differentiators, and state words). The raw number and proportion of words for each category were calculated, resulting in 32 features. Additionally, the number and proportion of all content words were calculated, resulting in 34 discourse features.

Syntax-Level Features

At the syntax level, 15 features were considered, focusing on sentence length (N = 4), word dependency (N = 3), and grammar (N = 8). Sentence length was represented by four features: the average numbers of characters and words were calculated separately for each sentence and each clause. Individual sentences and clauses were identified based on the punctuation. A sentence ends with a full stop, exclamation mark, question mark, ellipsis, or dash, whereas a clause ends with a comma, colon, or semicolon (Wang & Wu, 2020).

Word dependency was analyzed on a per-sentence basis. Three indicators were considered within each sentence: the numbers of characters and words before the main verb and the average word distance between any pairs of related words. Related words refer to words that are syntactically governed or dependent on another word.

Four grammar-related indicators reflecting the presence of complex Chinese grammar were considered: negative, metaphorical, passive, and contrastive sentences, based on the Baidu Open Platform (https://cloud.baidu.com). We calculated the number and the proportion of each of these four sentence patterns in a text and, therefore, identified eight discourse features concerning grammar.

[3] https://hanlp.hankcs.com/

Discourse-Level Features

For the discourse-level features, there were 24 features of referential cohesion and 20 features of causal cohesion. Referential cohesion features were included following the work of Graesser et al. (2011). We tracked four types of words: the overlap of all words, content words, nouns, and verbs, with six indicators for each. The six indicators were the proportion of adjacent sentence/paragraph pairs that shared the same words, the proportion of all possible sentence/paragraph pairs that shared the same words, and the weighted proportion of all possible sentence/paragraph pairs that shared the same words (i.e., the distance of two sentences/paragraphs was quantified into the number of sentences between them, and a score of 1/[L + 1] was granted when the distance was L sentence). The same calculations were carried out separately for sentences and paragraphs.

Causal cohesion features were adopted from Graesser et al. (2011). These were the raw numbers of precedents (e.g., at first), causes (e.g., because), adversatives (e.g., however), coordinations (e.g., and), additives (e.g., furthermore), successors (e.g., then), inferences (e.g., only if), conditions (e.g., unless), suppositions (e.g., if), concessions (e.g., even though), purposes (e.g., in order to), frequencies (e.g., always), parentheses (e.g., as everyone knows), abandonments (e.g., would rather not), results (e.g., so), comparatives (e.g., rather than), preferences (e.g., instead of), summaries (e.g., in sum), recounts (e.g., for example) and temporal (e.g., when) connectives.

RESULTS

The Roles of Linguistic Features in Predicting Text Readability in Chinese

The means and standard deviations for accuracy and MAE of the five-fold cross-validation are shown in Table 1. The results of the Y1 and Y2 models were similar (see Figure 1). Across all four LMMs, a significant effect of the machine learning algorithm was observed. RF outperformed SVM in predicting text readability with higher accuracy (Y1: Estimate = -1.27, *SE* = 0.25, $t(145) = -5.08$, $p < .001$; Y2: Estimate = -1.38, *SE* = 0.24, $t(145) = -5.87$, $p < .001$) and lower MAE (Y1: Estimate = 0.76, *SE* = 0.19, $t(139.99) = 3.90$, $p < .001$; Y2: Estimate = 0.56, *SE* = 0.18, $t(145) = 3.08$, $p = .003$). In terms of X, Xall demonstrated superior performance compared to the Xs without character-level features, except for the W_S_D in the Y2 models (*p*s > .05). This was evident in terms of accuracy (Estimates = -1.25 – -3.03, *SE*s = 0.24 – 0.25, *t*s = -12.86 – -4.99, *p*s < .001) and MAE (Estimates = 0.50 – 2.72, *SE*s = 0.18, *t*s = 2.76 – 15.21, *p*s < .01). However, there were no significant differences between Xall and Xs that include the character-level features (*p*s > .05).

Table 1

Means and Standard Deviations of Accuracy (ACC) and Mean Absolute Error (MAE) of All Machine Learning Models

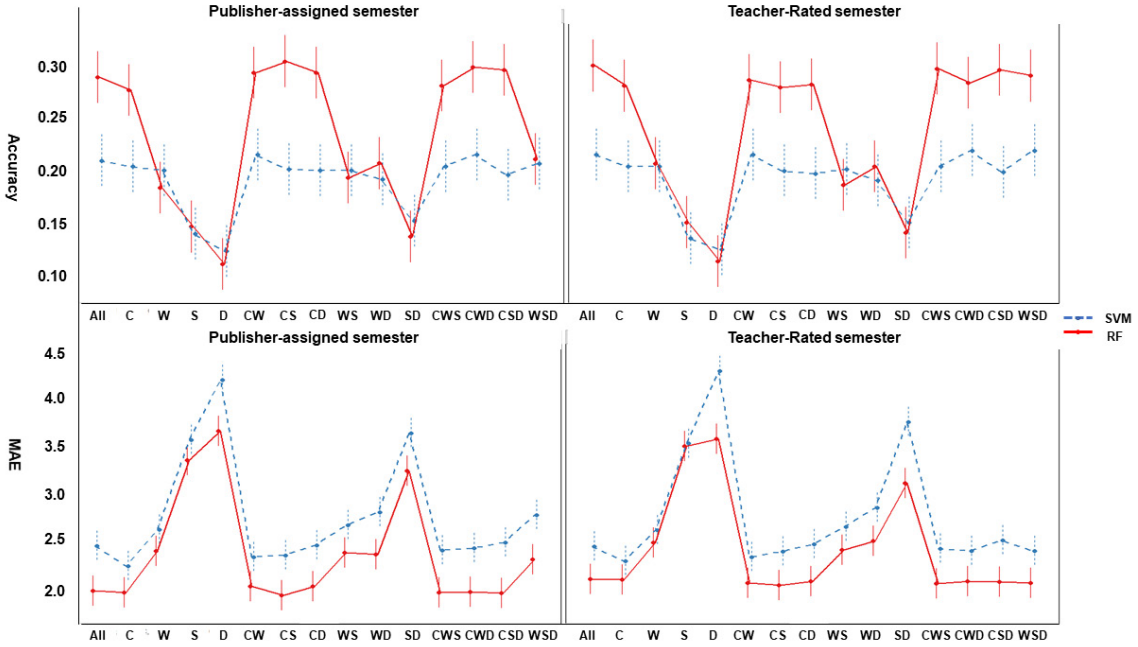
		Y1 (Publisher-assigned semester)		Y2 (Teacher-rated semester)	
	X	**ACC**	**MAE**	**ACC**	**MAE**
SVM	All	0.21 (0.02)	2.45 (0.18)	0.21 (0.02)	2.45 (0.14)
	C	0.20 (0.02)	2.24 (0.09)	0.20 (0.02)	2.29 (0.14)
	W	0.20 (0.03)	2.63 (0.23)	0.20 (0.02)	2.62 (0.20)
	S	0.14 (0.03)	3.59 (0.33)	0.13 (0.03)	3.55 (0.27)
	D	0.12 (0.02)	4.23 (0.53)	0.12 (0.02)	4.32 (0.54)
	C_W	0.21 (0.04)	2.34 (0.27)	0.21 (0.03)	2.33 (0.22)
	C_S	0.20 (0.02)	2.35 (0.08)	0.20 (0.02)	2.40 (0.11)
	C_D	0.20 (0.03)	2.46 (0.21)	0.20 (0.03)	2.47 (0.20)
	W_S	0.20 (0.03)	2.68 (0.22)	0.20 (0.03)	2.66 (0.22)
	W_D	0.19 (0.04)	2.81 (0.18)	0.19 (0.03)	2.86 (0.12)
	S_D	0.15 (0.02)	3.66 (0.37)	0.15 (0.01)	3.78 (0.36)
	C_W_S	0.20 (0.02)	2.41 (0.18)	0.20 (0.02)	2.43 (0.15)
	C_W_D	0.21 (0.02)	2.43 (0.20)	0.22 (0.02)	2.40 (0.12)
	C_S_D	0.20 (0.03)	2.49 (0.10)	0.20 (0.03)	2.52 (0.07)
	W_S_D	0.21 (0.03)	2.78 (0.18)	0.22 (0.02)	2.40 (0.12)

		Y1 (Publisher-assigned semester)		Y2 (Teacher-rated semester)	
	X	**ACC**	**MAE**	**ACC**	**MAE**
RF	All	0.29 (0.04)	1.97 (0.13)	0.30 (0.02)	2.10 (0.19)
	C	0.28 (0.05)	1.96 (0.22)	0.28 (0.04)	2.10 (0.11)
	W	0.18 (0.04)	2.40 (0.04)	0.21 (0.02)	2.49 (0.08)
	S	0.15 (0.03)	3.37 (0.18)	0.15 (0.03)	3.52 (0.25)
	D	0.11 (0.02)	3.69 (0.20)	0.11 (0.02)	3.60 (0.11)
	C_W	0.29 (0.04)	2.02 (0.20)	0.29 (0.03)	2.06 (0.18)
	C_S	0.30 (0.04)	1.93 (0.15)	0.28 (0.02)	2.03 (0.10)
	C_D	0.29 (0.04)	2.02 (0.19)	0.28 (0.03)	2.08 (0.21)
	W_S	0.19 (0.03)	2.38 (0.13)	0.19 (0.04)	2.41 (0.19)
	W_D	0.21 (0.04)	2.37 (0.07)	0.20 (0.05)	2.51 (0.17)
	S_D	0.14 (0.02)	3.26 (0.09)	0.14 (0.03)	3.12 (0.28)
	C_W_S	0.28 (0.04)	1.96 (0.17)	0.30 (0.03)	2.05 (0.10)
	C_W_D	0.30 (0.04)	1.96 (0.15)	0.28 (0.04)	2.08 (0.14)
	C_S_D	0.30 (0.05)	1.95 (0.18)	0.30 (0.05)	2.07 (0.22)
	W_S_D	0.21 (0.04)	2.31 (0.14)	0.29 (0.03)	2.06 (0.13)

Note. SVM = support vector machine; RF = random forest; C = Character features; W = Word features; S = Syntax features; D = Discourse features; All = features at all levels; Letter combinations represent the combination of features at different levels, e.g., C_W = features at Character and Word levels.

Figure 1

Results of Prediction Accuracy and Mean Absolute Error (MAE) in Readability Models Using Different Linguistic Features



Note. SVM = support vector machine; RF = random forest; C = Character; W = Word; S = Syntax; D = Discourse; All = Character_Word_Syntax_Discourse; Letter combinations represent the combination of features at different levels, e.g., C_W = features at Character and Word levels.

The interactions between the machine learning algorithm and the contrasts between Xall and Xs without character-level features, except for W_S_D, were significant in the accuracy models (Estimates = 1.03 – 1.63, *SE*s = 0.33 – 0.35, *t*s = 2.91 – 4.88, *p*s < .01). Regarding MAE, the interactions between the machine learning algorithm and the contrasts of Xall with S (Estimate = -0.52, *SE* = 0.26, *t* = -2.01, *p* = .047) and D (Estimate = 0.61, *SE* = 0.26, *t* = 2.39, *p* = .019) were significant, while the other interactions were not significant (*p*s > .05). Post-hoc analyses indicated that among the RF models, the differences between C and Xall in terms of both accuracy and MAE, were not significant (*p*s > .05). Moreover, C exhibited higher accuracy than the Xs that did not include the character features (Y1: Estimates = 1.12 – 2.66, *SE*s = 0.28, *t*s = 3.99 – 9.48, *p*s < .05; Y2: Estimates = 1.20 – 2.72, *SE*s = 0.26, *t*s = 4.56 – 10.32, *p*s < .01). Additionally, C had lower MAE than those with Xs without the character and word features (i.e., S, D, and S_D, Y1: Estimates = -2.74 – -2.06, *SE* = 0.20, *p*s < .001; Y2: Estimates = -2.43 – -1.66, *SE* = 0.20, *t*s = -8.20 – -11.98, *p*s < .01).

Feature Importance Analyses in the Best-Fitting Models

The LMMs showed that the RF models with all linguistic features (all-RF) and character-level (character-RF) features were optimal for predicting text readability. Feature importance analyses showed that both Y1 and Y2 features were similar, indicating that the character-level features are superior to other features, especially those from the syntax and discourse levels. More importantly, all models highlighted the importance of psycholinguistic information for characters. Specifically, the semester and grade when a character is first introduced in the textbook, measured either in token or type counts, played the most important roles in all optimal models. In Y1, the character-RF models revealed that the reaction times and error rates of character naming by Chinese adults (Liu et al., 2007) were highly important. Age of acquisition was important in three of the four accuracy models, but not in the all-RF model of Y2. The availability of pronunciation cues in a character, the ease of describing the meaning of a character, and the semantic radical transparency of characters were also ranked in the top 10 features in all models, except for the two character-RF models for Y2.

In addition, character frequency was also highlighted across the optimal models. The summed and averaged character frequency was highlighted in all models, although the results that were calculated according to different corpora were selected in different models. The ratio of type count to token count was only selected in the character-RF model of Y2 in terms of MAE. Differing from the models for Y1, three of the four models for Y2 showed that the character structural complexity, i.e., the number of strokes, was important. At the same time, some word-level features demonstrated high importance. The summed frequency of one-character words and the numbers of different kinds of words (including multiple-character words, one-character words, and low-frequent words) played critical roles in the all-RF models. Two indicators of part-of-speech syntactic categories, i.e., the raw numbers of adverbs and quantifiers, were only selected in the all-RF model of Y1 in terms of MAE. Only one model—the all-RF model of Y2 in terms of ACC—highlighted a discourse-level feature about referential cohesion, i.e., the proportion of all possible paragraph pairs that share the same content words.

DISCUSSION

Using machine learning techniques, research has demonstrated the importance of linguistic features from diverse levels, such as word, syntax, and discourse levels, on text readability (Fitzgerald et al., 2015). In Chinese, studies in Mainland China and Taiwan consistently found that using features from multiple levels outperformed those using features from a single level (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020). This study was one of the first to investigate text readability in Hong Kong. It extracted 274 linguistic features from 723 digitized Chinese language-arts textbooks commonly used in Hong Kong primary schools, representing character, word, syntax, and discourse levels. Two machine learning algorithms, namely SVM and RF, were utilized to examine the predictive capacity of these features in assessing text readability. The present study extended previous studies in Chinese (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020) by focusing on a finer, semester-level scale for text readability and introducing a subjective index: the teacher-rated semester level, along with the publisher-assigned semester level. Meanwhile, the important linguistic features for predicting text readability in the context of Hong Kong were identified. The current findings showed that the models with single character-level features and those with multilevel features incorporating character-level features performed similarly to the models with all 274 features. The findings demonstrate the central role of character features in predicting text readability in Chinese. The results of feature importance indicated similarities between the perspectives of publishers and teachers. Results from both perspectives showed that the character-level features, i.e., the semester and grade when a character is first introduced in the textbook, were crucial. Meanwhile, findings from these two perspectives had differences. Models with teacher-rated semesters underscored the importance of the number of strokes, while those with publisher-assigned semesters highlighted the influence from research results, i.e., adults' response times and error rates in lexical decision tasks (Liu et al., 2007).

The Central Role of Character Features in Predicting Text Readability in Chinese

Consistent with previous studies conducted in Mainland China (e.g., Wu et al., 2020) and Taiwan (e.g., Sung et al., 2015), the current findings illustrated that lexical features (i.e., character and word levels) were more advantageous than syntax-level and discourse-level features in determining text readability. More specifically, the models for both publisher-assigned semester level and teacher-rated semester level were similar and demonstrated that models with single character-level features and models with all 274 features performed best in terms of accuracy and MAE in predicting text readability. The RF models further demonstrated that models with character-level features outperformed those without, showing higher accuracy and lower MAE. These similar results suggest that the lexical features had greater effects than the syntax-level and discourse-level features on text readability across Chinese communities using different written and spoken languages. Although it has been found that character-level and word-level features are more predictive of text readability, this doesn't mean syntax-level and discourse-level features have no influence. Prior research highlights the influence of syntactic and discourse skills on reading comprehension (Chik et al., 2012). For instance, a study involving Hong Kong fourth graders (Yeung et al., 2013) revealed that after controlling for word reading, syntactic skills (word-order knowledge, morphosyntactic knowledge) and discourse skills (sentence-order knowledge) uniquely contributed to reading comprehension. Thus, even though character- and word-level features may significantly impact text readability, syntax- and discourse-level features also play a vital role.

On the other hand, the current finding was inconsistent with previous studies in Taiwan (i.e., Chen et al., 2013; Sung et al., 2015), which did not find that models with single-level features could perform as well as those with multiple-level features in predicting text readability. Such a difference might be due to differences in the models' design in the current study compared to previous studies. The present study incorporated both RF and SVM algorithms and used a more granular indicator (semester level) than the grade level used in previous studies. Also, we used more linguistic features (274) compared to previous studies and distinguished between character-level and word-level features, which were not separated in the previous studies.

Differing from the current study, where character-level features outperformed other features in predicting text readability, a study conducted in Mainland China (i.e., Wu et al., 2020), where simplified Chinese is used, found word-level features had an advantage over character-level features. The strong performance of character-level features in our study might be attributable to the use of traditional Chinese in Hong Kong (McBride-Chang et al., 2005). Traditional Chinese characters, known for their ideographic origins, have a close connection between form and meaning. The simplification process used in Mainland China often weakens this connection, which could make simplified characters more difficult to recognize and read, especially for beginning readers. For example, the simplified character 爱 (love) was developed by removing an element associated with the whole character's meaning, i.e., 心/sam1/ (heart), from its traditional counterpart 愛/oi3/. Studies have shown that children learning simplified characters perform better in visual skill tasks compared to those learning traditional characters (e.g., McBride-Chang et al., 2005). This suggests that some character-level features make traditional characters relatively easy to recognize and read, thereby influencing text comprehension. These aforementioned differences between our study and those conducted in Taiwan and Mainland China may be due to language-related differences between Hong Kong and Taiwan, underscoring the need for specific text readability formulas for different Chinese communities.

Similar and Dissimilar Roles of Features Across Perspectives of Publishers and Teachers The current study advanced previous research (e.g., Liu et al., 2024; Sung et al., 2015; Wu et al., 2020) by incorporating both the publisher-assigned semester level and the teacher-rated semester level as indicators of text readability. The present study revealed consistent findings from both publishers and teachers in terms of feature importance. Specifically, the semester and grade at which a character is first introduced in the textbook significantly impacted text readability. This influence remained notable even when all features across the four aspects were considered. Additionally, the age of acquisition, which correlates with the semester and grade of a character's introduction, was also found to significantly influence text readability. These features indicate the learning sequence of characters, which is not arbitrary. In Hong Kong, publishers are required to refer to *Lexical Lists for Chinese Learning in Hong Kong* (Education Bureau, 2007) when editing textbooks. According to this document, characters that appear at the early stages of learning commonly have lower visual complexity and higher frequencies, e.g., 一/jat1/ (one), 我/ngo5/ (me), and 你/nei5/ (you), than those that are usually taught later, e.g., 勢/sai3/ (power), 滲/sam3/ (seep), and 癒/jyu6/ (heal). This suggests that characters taught at initial stages are designed to be simpler than those introduced later. Moreover, characters taught earlier may also have more exposure, enabling children to better understand them through contextual reading (Brent & Siskind, 2001). Consequently, children might master early-introduced characters, which could enhance text readability.

In line with Liu et al. (2024), the feature-importance analyses highlighted the significance of character frequency and

word frequency in text readability. This aligns with previous research showing a strong frequency effect where high-frequency words are read more accurately and faster across multiple languages (e.g., Cai & Brysbaert, 2010). It was suggested that higher frequencies could facilitate character and word comprehension.

Meanwhile, there were a few differences between the present findings regarding the feature importance for the publisher-assigned semesters and teacher-rated semesters models. Specifically, the features about the number of strokes, which correlate with the visual complexity of characters, featured prominently in the top 10 features of the optimal models for teacher-rated semesters, but not in the models for publisher-assigned semesters. Our feature-importance analysis reported fewer complexity-related features compared to frequency-related ones, suggesting a relatively minor influence of complexity on readability. Consistently, a study on Chinese children (Su & Samuels, 2010) found a diminishing effect of visual complexity on word processing as children's reading skills matured.

On the other hand, the features linked with adults' response times and error rates in lexical decision tasks (Liu et al., 2007) were only observed in the publisher-assigned semesters' analysis. This discrepancy could be attributed to the relative readability for teachers in directly grasping information about adults' response times and error rates in lexical decision tasks, compared to the number of strokes. Consequently, while publishers might rely on diverse information sources to assign semesters, teachers likely prefer to utilize indexes that can be directly derived from the texts themselves to gauge readability levels.

Limitations and Future Directions

As one of the first studies to investigate text readability in Hong Kong, our corpus only covered textbooks from three publishers. Future research could include a wider variety of texts, such as storybooks. Although we engaged experienced teachers to rate the texts considering an average reader at a certain grade, it remains challenging to directly reflect children's readability. Future studies could involve children's ratings and their reading comprehension performances, which are closely related to text readability (Mesmer & Hiebert, 2015). Furthermore, future research could employ additional machine learning algorithms suitable for classification, such as the K-nearest neighbor and decision-tree classifier (Rodriguez-Galiano et al., 2015).

Initially, the accuracy rates of our models were not particularly high, but they improved when the grade level (Grades 1-6) was used as the readability level. Specifically, the SVM and RF models performed similarly. Both models, whether using the single level of character features (SVM: mean accuracy = 66.08%, SD = 0.04; RF: mean accuracy = 70.94%, SD = 0.04) or all features (SVM: mean accuracy = 65.57%, SD = 0.04; RF: mean accuracy = 68.34%, SD = 0.04), performed equally well and outperformed other models that did not include character-level features. Our models achieved accuracy rates for grade levels comparable to previous studies conducted in Mainland China (e.g., Wu et al., 2020) and Taiwan (e.g., Sung et al., 2015), which contributes to the existing research on readability across Chinese communities. However, this also emphasizes the need for further exploration of models with finer scales that can achieve higher accuracy in predicting readability. Future studies should focus on investigating such models.

CONCLUSION

The primary aim of this study was to investigate the predictive power of linguistic features at the character, word, syntax, and discourse levels in assigning texts to primary school semester levels. By employing robust machine learning techniques, the study demonstrated the significant predictive power of linguistic features, particularly at the character level. In addition, as a secondary objective, the study analyzed two optimal RF models based on all features and character-level features, which achieved high accuracy and low MAE in predicting semester levels. The feature importance analyses specifically revealed that character learning sequences, character frequency, and word frequency are crucial in predicting text readability. These findings directly address our research questions by identifying the key linguistic features that influence readability assessments from the perspectives of both publishers and teachers.

Practically, these findings offer valuable insights for teaching. Teachers can concentrate on lexical-level features, especially when teaching new characters Furthermore, future studies could develop an automated text readability analyzer centered on character-level features using the two optimal RF models identified in the current study. Such an analyzer could streamline the semester assignment of textbooks and identify readability levels of texts from other resources, like storybooks. Consequently, children, parents, and teachers could more easily select formal and informal reading materials that align with children's reading abilities.

ACKNOWLEDGEMENTS

This work was partially supported by the Research Seed Fund of the Department of Special Education and Counselling, The Education University of Hong Kong (Ref. No. 04670) to Dr. Duo Liu. This work was also partially supported by a fellowship award to Dr. Zhengye Xu, from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. EdUHK PDFS2122-8H09).

DATA AVAILABILITY STATEMENTS

The data supporting this study's findings are available from https://osf.io/h9ew4/

DECLARATION OF COMPETITING INTEREST

None declared.

AUTHOR CONTRIBUTIONS

Zhengye Xu: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Software; Supervision; Visualization; Writing – original draft; Writing – review & editing.

Yixun Li: Data curation; Investigation; Methodology; Resources; Writing – original draft.

Duo Liu: Conceptualization; Data curation; Funding acquisition; Investigation; Methodology; Project administration; Software; Writing – original draft; Writing – review & editing.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*(2), B33-B44. https://doi.org/10.1016/S0010-0277(01)00122-6

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles [Data set]. *PloS One, 5*(6), Article e10729. https://doi.org/10.1371/journal.pone.0010729

Chen, Y., Chen, Y., & Cheng, Y. (2013). Assessing Chinese readability using term frequency and lexical chain. *Journal of Computational Linguistics & Chinese Language Processing, 18*(2), 1-18.

Chen, Y., Tsai, Y., & Chen, Y. (2011). Chinese readability assessment using TF-IDF and SVM.

Cheng, Y., Xu, D., & Dong, J. (2020). 基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究 [A study on the analysis of key factors of text reading difficulty grading and the readability formula based on a corpus of language teaching materials]. 语言文字应用*, 1*, 132-143. https://doi.org/10.16499/j.cnki.1003-5397.2020.01.014

Chik, P. P., Ho, C. S., Yeung, P., Chan, D. W., Chung, K. K., Luan, H., Lo, L., & Lau, W. S. (2012). Syntactic skills in sentence reading comprehension among Chinese elementary school children. *Reading and Writing, 25*, 679-699. https://doi.org/10.1007/s11145-010-9293-4

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3-4), 541-561. https://doi.org/10.1111/1467-9817.12283

Crossley, S., Heintz, A., Choi, J. S., Batchelor, J., Karimi, M., & Malatinszky, A. (2023). A large-scaled corpus for assessing text readability. *Behavior Research Methods, 55*(2), 491-507. https://doi.org/10.3758/s13428-022-01802-x

Da, J. (2004). A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction [Data set]. *Proceedings of the Fourth International Conference on New Technologies in Teaching and Learning Chinese,* 501-511.

Education Bureau (2007). *Lexical Lists for Chinese Learning in Hong Kong*.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology, 107*(1), 4–29. https://doi.org/10.1037/a0037289

Givón, T. (1995). Coherence in text vs. coherence in mind. *Coherence in Spontaneous Text, 1995*, 59-116.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234. https://doi.org/10.3102/0013189X11413260

Huang, C. (2006, March, 6-7). *Automatic acquisition of linguistic knowledge: From sinica corpus to gigaword corpus* [Conference presentation] [Data set]. The 13th National Institute of Japanese Language International Symposium Language Corpora: Their Compilation and Application, Tokyo.

Jin, G. J., Xiao, H., Fu, L., & Zhang, Y. F. (2005). 现代汉语语料库建设及深加工[Construction and further processing of Chinese National Corpus] [Data set]. 语言文字应用*, 2*, 111–120. https://doi.org/10.16499/j.cnki.1003-5397.2005.02.017

Jing, X. (1995). 中文国文教材的适读性研究：适读年级值的推估 [A study on the readability of Chinese national language teaching materials: Estimation of readability values of grade levels]. 教育研究资讯, *5*, 113-127.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Flesch-kincaid grade level*. United States Navy.

Korobov, M., & Lopuhin, K. (2019). *Permutation importance.*

Liu, M., Li, Y., Su, Y., & Li, H. (2024). Text complexity of Chinese elementary school textbooks: Analysis of text linguistic features using machine learning algorithms. *Scientific Studies of Reading, 28*(3), 235-255. https://doi.org/10.1080/10888438.2023.2244620

Liu, M., Li, Y., Wang, X., Gan, L., & Li, H. (2021). 分级阅读初探: 基于小学教材的汉语可读性公式研究 [Leveled reading for primary students: Construction and evaluation of Chinese reeadability formulas based on textbooks]. 语言文字应用, *2*, 116-126. https://doi.org/10.16499/j.cnki.1003-5397.2021.02.010

Liu, Q., Zhang, H. P., Yu, H. K., & Cheng, X. Q. (2004). 基于层叠隐马模型的汉语词法分析 [Chinese lexical analysis using cascaded hidden Markov model]. 计算机研究与发展, *41*(8), 1421–1429.

Liu, Y., Shu, H., & Li, P. (2007). Word naming and psycholinguistic norms: Chinese [Data set]. *Behavior Research Methods, 39*(2), 192–198. https://doi.org/10.3758/BF03193147

Ma, W., & Chen, K. (2005). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society, 14*(3), 235-249.

McBride-Chang, C., Chow, B. W., Zhong, Y., Burgess, S., & Hayward, W. G. (2005). Chinese character acquisition and visual skills in two Chinese scripts. *Reading and Writing, 18*, 99-128. https://doi.org/10.1007/s11145-004-7343-5

Mesmer, H. A. E. (2005). Text decodability and the first-grade reader. *Reading & Writing Quarterly, 21*(1), 61-86. https://doi.org/10.1080/10573560590523667

Mesmer, H. A., & Hiebert, E. H. (2015). Third graders' reading proficiency reading texts varying in complexity and length: Responses of students in an urban, high-needs school. *Journal of Literacy Research, 47*(4), 473-504. https://doi.org/10.1177/1086296X16631923

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel. M., Prettenhofer. P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Pinney, C., Kennington, C., Pera, M. S., Wright, K. L., & Fails, J. A. (2024). Incorporating word-level phonemic decoding into readability assessment. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, Italia,* 8998-9009.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. J. O. G. R. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews, 71*, 804-818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Solnyshkina, M., Zamaletdinov, R., Gorodetskaya, L., & Gabitov, A. (2017). Evaluating text complexity and Flesch-Kincaid grade level. *Journal of Social Studies Education Research, 8*(3), 238-248.

Stutz, F., Schaffner, E., & Schiefele, U. (2016). Relations among reading motivation, reading amount, and reading comprehension in the early elementary grades. *Learning and Individual Differences, 45*, 101-113. https://doi.org/10.1016/j.lindif.2015.11.022

Su, I., Yum, Y. N., & Lau, D. K. (2023). Hong Kong Chinese character psycholinguistic norms: Ratings of 4376 single Chinese characters on semantic radical transparency, age-of-acquisition, familiarity, imageability, and concreteness [Data set]. *Behavior Research Methods, 55*(6), 2989-3008. https://doi.org/10.3758/s13428-022-01928-y

Su, Y., & Samuels, S. J. (2010). Developmental changes in character-complexity and word-length effects when reading Chinese script. Reading and Writing, 23, 1085-1108. https://doi.org/10.1007/s11145-009-9197-3

Sung, Y., Chen, J., Cha, J., Tseng, H., Chang, T., & Chang, K. (2015). Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods, 47*, 340-354. https://doi.org/10.3758/s13428-014-0459-x

Tsang, Y., Huang, J., Lui, M., Xue, M., Chan, Y. F., Wang, S., & Chen, H. (2018). MELD-SCH: A megastudy of lexical decision in simplified Chinese [Data set] . *Behavior Research Methods, 50*, 1763-1777. https://doi.org/10.3758/s13428-017-0944-0

Tse, C., Yap, M. J., Chan, Y., Sze, W. P., Shaoul, C., & Lin, D. (2017). The Chinese lexicon project: A megastudy of lexical decision performance for 25,000 traditional Chinese two-character compound words. *Behavior Research Methods, 49*, 1503-1519. https://doi.org/10.3758/s13428-016-0810-5

Wang, F., & Wu, F. (2020). Postnominal relative clauses in Chinese. *Linguistics, 58*(6), 1501-1542. https://doi.org/10.1515/ling-2020-0226

Wu S., Yu D., & Jiang X. (2020). 汉语文本可读性特征体系构建和效度验证 [Development of linguistic features system for Chinese text readability assessment and its validity verification]. 世界汉语教学, *34*(1), 81-97.

Yeung, S. S., Siegel, L. S., & Chan, C. K. (2013). Effects of a phonological awareness program on English reading and spelling among Hong Kong Chinese ESL children. *Reading and Writing, 26*, 681-704. https://doi.org/10.1007/s11145-012-9383-6

Zhang, J., McBride-Chang, C., Wong, A. M., Tardif, T., Shu, H., & Zhang, Y. (2014). Longitudinal correlates of reading comprehension difficulties in Chinese children. *Reading and Writing, 27*, 481-501. https://doi.org/10.1007/s11145-013-9453-4