Quantitative Research on Readability Formula of Texts for Chinese Heritage Language

Ruo Lin and Juan Xu*

School of Information Science, Beijing Language and Culture University, Beijing, China;
Email: 202021296069@stu.blcu.edu.cn (R.L.)
*Correspondence: xujuan@blcu.edu.cn (J.X.)

***Abstract*—The compilation of Chinese heritage textbooks has always adopted the standards of teaching Chinese as mother language or as second language. The difficulty level of the texts is mostly measured separately from the aspects of Chinese characters, vocabulary, and grammar. There is no comprehensive quantitative evaluation standard. Readability formula is one of the methods to comprehensively measure the difficulty of text. The present study made use of Chinese heritage textbooks as the source of data and constructed readability formula by the method of multiple regression analysis. The formula includes four language features, i.e., proportion of difficult words, number of different characters, average sentence length, and proportion of function words, which could explain 72.9% of the variation of text difficulty level. Limitations and future works of the formula are discussed.**

***Keywords*—readability formula, text feature analysis, Chinses heritage language texts**

I. Introduction

Readability refers to the degree that a text can be read and understood. (Dale & Chall, 1949) For second language learners, readability has a direct impact on readers' understanding, reading speed and degree of interests. Therefore, the evaluation of readability is of great significance to both educators and learners.

Many methods have been proposed to measure the readability of text, which includes subjective evaluation, answering questions according to the text, cloze and readability formula [1]. However, some researchers pointed out that it is still common to use subjective evaluation in most studies of Chinese readability [2].

Regarding the development of natural language processing technology, the construction of readability formula is increasing, and has recently been applied in the field of teaching Chinse Language as mother language and as second language. Previous research has shown that Chinese L2 learners with different levels and language backgrounds have different reading processing patterns. In other words, readability research should consider both the text type and the reader's background, for example, learning Chinese as a heritage language. However, in some Chinese heritage language curriculum, it is common that teachers and students are using Chinese L2 textbooks, which points to a gap in readability for Chinese as a heritage language.

The research paradigm of readability formula is: adopting the knowledge and methods of reading psychology, linguistics and statistics, the readability level of the text is obtained by measuring the syntactic complexity and semantic difficulty of the text and using the method of mathematical statistics. The construction of readability formula can be divided into two steps: data construction and formula verification. In the construction stage, text corpus, language features, and the relationship between text difficulty level and language features are to be constructed. In the verification stage, it is necessary to verify whether the constructed formula can predict the difficulty of new text, and evaluate the performance of the readability formula by comparing the predicted results with the real level of the text.

The study of readability formula originated in the United States, and it is mainly used to calculate the reading difficulty of English texts. Among which, the most classical formulas include Dale&Chall formula [3]; Flesch Kincaid formula which is embedded in Microsoft Word [4]; Smog formula [5] and Lexile Scores [6]. The study of Chinese text readability began in the 1970s, which includes mother language texts [7–11] and second language texts [2, 12–14]. However, there are few studies on the readability for Chinese as a heritage language, and the text of which is different from Chinese as mother language textbooks and as second language textbooks in the aspect of vocabulary, sentence, and text structure. For now, it remains to be shown whether the existing formula will perform well in text of Chinese as a heritage language.

Researchers have reached a common view on the characteristics of students learning Chinese as a heritage language. Chinese heritage language textbooks are compiled to help students acquire knowledge of Chinese language as well as culture. The instruction of Chinese heritage language is "different from the Chinese language education in China as mother language and the teaching of Chinese as second language" Currently, the research on Chinese heritage language textbooks mainly focuses on the compilation and application of textbooks, but the overall study on text readability is rather rare. Based on previous studies, this study aims to re-fit the existing Chinese readability formulas, explore the current

Manuscript received January 22, 2022; revised April 15, 2022; accepted December 26, 2022.

readability formulas for Chinese heritage language textbooks, and try to build a more scientific readability formula for students learning Chinese as a heritage language.

II. Methods

A. Corpora

Two Chinese heritage language textbooks *Zhongwen* and *Hanyu* were combined as our corpora in this research. These textbooks are the major textbooks used in overseas Chinese heritage language classroom, and are classified into 12 volumes for primary and 6 volumes for secondary (Table I).

TABLE I. Information of Chinese Heritage Language Textbooks

Textbook 1	*Zhonngwen* (Primary) (2006)	*Zhonngwen* (Secondary) (2010)
Editor	College of Chinese Language and Literature, Jinan University	
Criterion	*Chinese Proficiency Standard and Grammar Outline* (1996) *Graded Vocabulary and Characters for Chinese Proficiency* (2001) *Basic Vocabulary Table of Modern Chinese Characters* (1988)	
Textbook 2	*Hangyu* (Primary) (2007)	*Hanyu* (Secondary) (2010)
Editor	Beijing Chinese Language and Culture College	
Criterion	Higher than the standard requirements of Youth Chinese Test (YCT) Chinese Proficiency Test Program Chinese language syllabus in China mainland's primary and secondary school Chinese Proficiency Test Program (HSK bank 6)	

In terms of text selection, this study has eliminated poetry, riddles, dialogues, and ancient Chinese texts. As Chinese heritage education has the obligation to inherit cultural knowledge [15], both sets of textbooks contain contents of Chinese culture, including idioms, myths, legends, and historical stories. Some of the same content appears in both of two textbooks, but is written at different volumes, such as "Lan Yu Chong Shu" (滥竽充数) in volume 12 of *Hanyu* and volume 9 of *Zhongwen*; "San Ge He Shang" (三个和尚) is in volume 10 of *Hanyu* and volume 8 of *Zhongwen*. In this case, considering that although the topic of the text is roughly the same, the language features such as vocabulary and sentences cannot be completely replaced equally, so both of the texts are retained and labeled according to the volume in textbook.

In total, the number of texts selected in this paper contains 730 texts, including 510 texts in 12 primary textbooks and 220 texts in 6 secondary textbooks, totally 304,633 words.

B. Language Features

The difference between the current Chinese readability formulas is mainly in the selection of text features, in addition, there are differences in the way the features are calculated and the reference standards. Reference [10] compares the features used in different formulas and divides the selected features into three categories, namely, the difficulty, length, and category features.

TABLE II. Language Feature of Chinese Readability Formulas

Formula user	Formula constructors	Language features		
		Difficulty	**Length**	**Category**
Learning Chinese as mother language	Wang (2020) [16]	Familiar words	Average number of sentences, Average number of words	
	Liu (2021) [11]	Average difficulty of words	Number of different Characters	Percentage of function words
Learning Chinses as second language	Hong (2014) [13]	Number of easy words, Percentage of hard words		Number of function words
	Wang (2008) [2]	Percentage of easy words	Number of sub-sentences, Number of different words	Percentage of function words

As shown in Table II, different researchers have a certain preference for the choice of linguistic features, considering the different user of the formulas: almost all Chinese readability formula constructors believe that length characteristics have a greater impact on readability; Most researchers believe that the function words in the lexical features is an important indicator to judge the Chinese readability.

Based on the previous research results, this paper combines the reading characteristics of Chinese heritage students and extracts text features from three levels: word difficulty, length, and category.

1) Word difficulty

Different from English, Chinese characters and words are two different reading units. From the reading point of view, the difficulty of words reading is mainly related to the complexity of Chinese characters, character frequency and word frequency information.

The complexity of Chinese characters is affected by the number of strokes, the number and the arrangement of parts. Relevant studies showed that the familiarity of Chinese characters has effect on the comprehension process. Research has shown that Chinese heritage students and non-Chinese heritage students have different feature of Chinese character acquisition, in which the former tends to process whole characters because of their high familiarity with Chinese characters, while the latter tends to do component analysis.

Character frequency and word frequency are closely related to the grade of Chinese characters and words. In this paper, we used "International Chinese Education Standard for Chinese Language" (refer to as "New Standard") [17] in extracting the frequency of words and characters. The new Standard divides levels into three stages and nine levels. Moreover, for the first time, the recognition and handwriting of Chinese characters are clearly quantified, so that the Chinese character instruction mode "recognize more and write less" has a

specific indicator and includes more recognition vocabulary that is helpful for reading. In the study of readability formulas, the proportion of easy words and hard words has a strong correlation with text difficulty. In this paper we take the number of elementary words, the proportion of elementary words and the proportion of difficult words (the percentage of advanced and super-outline words in the total number of words) as the index of word difficulty.

2) Length

With the development of natural language processing technology, feature extraction gradually tends to mine deep linguistic features such as semantics and cohesion.

Relevant studies have shown that deep language features can effectively improve the predictive performance of readability formulas, but shallow features on vocabulary and sentence level also have inherent advantages, which are simple, intuitive and easy to quantify.

As can be seen from the table above, length is the most used factor for researchers, and it plays an important role in practical applications.

Referring to previous studies, this paper chooses the length features including number of characters, number of different characters, number of words, number of different words, number of whole sentences, number of sub-sentences, and average sentence length.

3) Word category

As we know, words of a language can be divided into content words and function words. When reading, the semantics of the content words are relatively fixed and thus easy to understand. While function words are difficult due to their flexible semantics, and it makes differences of the structure and meaning of the sentence. In this paper, the definition of function words are prepositions, conjunctions, auxiliaries, mood words, adverbs and positional words.

Due to the characteristics of the learners and the task of cultural inheritance, Chinese heritage education contains many idioms with rich cultural connotations, which requires a rich grasp of cultural knowledge when students encounter the idioms.

TABLE III. CORRELATION COEFFICIENT BETWEEN LANGUAGE FEATURES AND TEXT DIFFICULTY

	Language features	Correlation Coefficient
1	Number of strokes	−0.211*
2	Number of elementary words	0.484**
3	Proportion of elementary words	0.132**
4	Proportion of difficult words	0.471**
5	Number of characters	0.813**
6	Number of different characters	0.921**
7	Number of words	0.785**
8	Number of different words	0.826**
9	Number of whole sentences	0.663**
10	Number of sub-sentences	0.689**
11	Average sentence length	0.617**
12	Number of function words	−0.140*
13	Proportion of function words	0.037*
14	Number of idioms	0.120*

* $p < 0.05$, ** $p < 0.01$

Regarding the previous studies, this paper identifies 14 language features that affect text difficulty from three levels which have mentioned above. The correlation coefficient between language features and text difficulty is shown in Table III.

III. RESULT

A. Constructing Readability Formula

Before the multiple regression analysis, there may have significant correlations among selected factors. These highly correlated factors should be filtered out because they yield similar effects on text difficulty and will affect the accuracy of parameter estimation.

Filter method is as follows: 14 text features are added to the multiple regression analysis as predictive variables in turn. If the Variance Inflation Factors (VIF) is over 5 after adding a feature, which determined the collinearity problem, then further compared the adjust R^2 and retain the larger one into the readability formula. Based on the results of filtering, the selected language features were proportion of difficult words, number of different characters, average sentence length, and proportion of function words and the results are shown in Table IV.

TABLE IV. MULTIPLE REGRESSION ANALYSES

	Unstandardized coefficient		Standardized coefficient	*t*	*p*	VIF	R^2	Adjusted R^2
	B	**Standard error**	**Beta**					
constant	−54.079	24.095	−	−2.244	0.026*	−		
Average sentence length	3.797	0.844	0.170	4.500	0.000**	1.267		
Number of different characters	1.943	0.121	0.693	16.033	0.000**	1.659	0.733	0.729
Proportion of difficult words	6.041	1.899	0.121	3.180	0.002**	1.277		
Proportion of function words	52.006	84.854	0.023	0.613	0.002**	1.222		

* $p < 0.05$, ** $p < 0.01$

Based on these results, the following formula was constructed:

$$Y = -54.079 + 3.797X_4 + 1.943X_1 + 6.041X_2 + 52.006X_3$$

where Y is the readability score, X_1 is the number of different characters, X_2 is the proportion of difficult words, X_3 is the proportion of function words, and X_4 is the average sentence length.

The results of our multiple regression analysis indicate that the combination of proportion of difficult words, number of different characters, average sentence length, and proportion of function words produces a multiple correlation of 0.733 and a corresponding R^2 of 0.729. That is to say, this result signifies that the combination of the four variables alone accounts for 72.9% of the variance based on Chinese heritage language text. In other words, using these four variables, the formula can predict 72.9% of the difficulty for these textbooks.

B. Validation and Evaluation

To further evaluate the validation of which the formulas predict readability, correlation coefficient between the predicted and actual text values and the prediction accuracies were generated for the texts.

There is a significant positive correlation between the predicted and actual text difficulties ($r = 0.863$, $p < 0.01$), indicating that the indices calculated from this formula are most consistent with the original volume levels of the texts.

TABLE V. THE ACCURACY OF OUR FORMULA

	Textbook Volume	Absolute Accuracy	Relative Accuracy
Primary	1	0.34	0.56
	2	0.34	0.67
	3	0.23	0.76
	4	0.43	0.89
	5	0.36	0.78
	6	0.35	0.79
	7	0.39	0.72
	8	0.32	0.65
	9	0.28	0.56
	10	0.21	0.45
	11	0.15	0.21
	12	0.12	0.25
Secondary	1	0.15	0.2
	2	0.12	0.18
	3	0.04	0.12
	4	0.08	0.13
	5	0.05	0.10
	6	0.20	0.15
Average		0.23	0.45

As can be seen from Table V, the indices of absolute accuracy are more consistent in primary level text than secondary level. This result is consistent with previous study [12], which found that elementary and intermediate texts are strongly influenced by lexical factors, while intermediate and higher texts are influenced by sematic factors.

C. Comparison with Previous Work

Table VI provides a comparison of a multiple correlation result produced by our formula and previous readability formula. We used the corpus of this study as dataset, reconstructed 4 formulas with original feature sets as independent variables and the results are shown in the table. It is apparent that the correlation made by our formula is stronger than those made by the previous formulas.

TABLE VI. THE COMPARISON OF A MULTIPLE CORRELATION RESULT

Formula constructor	R^2	F test
Wang (2020)	0.485	120.412
Liu (2021)	0.706	190.532
Hong (2014)	0.684	98.773
Wang (2008)	0.586	83.735
Our Formular	0.733	162.839

IV. CONCLUSION AND DISCUSSION

This study constructs a readability formula for Chinse heritage language text using texts from primary and secondary Chinese heritage language textbooks as its source. This formula is formed from four language features, namely, proportion of difficult words, number of different characters, average sentence length, and proportion of function words. The results of analyses show that the formula can explain as much as 72.9% of the total variance, which is most predictive comparing to previous formula.

As in all studies, this one has limitations. First, there is the question of the text set, which are all texts taken from editors of China mainland and examined a relatively small text set. Future studies would do well to use a larger text set, together with a separate and comparable training set. Second, there is the question of the method, although multiple regression analysis and readability formula have been an enduring method of most readability studies, their validity has been debated [18]. More and more readability tasks rely on machine learning models, artificial neural networks and SVM to integrate multilevel linguistic features. Future studies would benefit from readability models that combine new methods. Finally, there is the question of the language features, which are all shallow text features. Future works will pay attention on study the further level of the features, such as semantics and cohesion, as well as the learner factors.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTION

Ruo Lin is mainly responsible for text data statistics and analysis, and wrote the main content of the article. Juan Xu supervised and helped with finalizing the article, she acted as corresponding author. All authors had approved the final version.

FUNDING

This research was funded by the 2022 Key Project of International Chinese Language Research of the Center

for Language Education and Cooperation under grant number 22YH50B, and the BLCU supported project for young researchers program (supported by the Fundamental Research Funds for the Central Universities) under grant number 22YCX048.

REFERENCES

[1] J. Gilliland, *Readability*, London: Hodder and Stoughton, 1972.
[2] L. Wang, "Research on Chinese readability formula of texts for elementary and inntermediate South Korean and Japanese learners," *Lang. Teach. Linguist. Stud.*, vol. 6, pp. 46–53, 2008.
[3] D. Edgar and S. C. Jeanne, "A formula for predicting readability," *Educ. Res. Bull.*, vol. 27, no. 1, pp. 11–20, 28, 1948.
[4] R. Flesch, "A new readability yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, pp. 221–233, 1948.
[5] G. H. McLaughlin, "SMOG grading: A new readability formula," *J. Read.*, vol. 12, no. 8, pp. 639–646, 1969.
[6] D. R. Smith, "The lexile scale in theory and practice. Final report," *Anal. Var.*, 1989.
[7] H. Y. Sun, "Chinese reading ease formula," Beijing Normal University, Beijing, China, 1992.
[8] Y. T. Sung, J. L. Chen, J. H. Cha, *et al.*, "Constructing and validating readability models: The method of integrating multilevel linguistic features with machine learning," *Behav. Res. Methods*, vol. 47, no. 2, pp. 340–354, Jun. 2015.
[9] H. C. Tseng, H. T. Hung, Y. T. Sung, *et al.*, "Classification of text readability based on deep neural network and representation learning techniques," in *Proc. the 28th Conference on Computational Linguistics and Speech Processing*, 2016, pp. 255–270.
[10] Y. Cheng, D. Xu, and J. Dong, "On key factors of text reading difficulty grading and readability formula based on Chinese textbook corpus," *Applied Linguist.*, vol. 1, pp. 132–143, 2020.
[11] M. M. Liu, Y. Li, X. M. Wang, *et al.*, "Leveled reading for primary students: Construction and evaluation of Chinese readability formulas based on textbooks," *Appl. Linguist.*, vol. 2, pp. 116–126, 2021.
[12] W. H. Guo, "Research on readability formula of Chinese text for foreign students," Shanghai Jiao Tong University, Shanghai, 2010.
[13] Z. Hong, "Research on Chinese readability formula of texts for internediate level European and American students," *Chinese Teach. World*, vol. 28, no. 2, pp. 263–276, 2014.
[14] J. Y. Cai, "Research on L2 Chinese text readability," Beijing Language and Culture University, Beijing, 2020.
[15] X. Zhou, N. Chen, and J. Guo, "Overview of Chinese teaching materials based on the global Chinese teaching material e-library," *Overseas Chinese Educ.*, vol. 2, no. 75, pp. 225–234, 2015.
[16] H. B. Wang, "A review of research on the difficulty assessment of reading materials based on L2 graded reading education," *J. Int. Chinese Teach.*, vol. 2, pp. 75–88, 2020.
[17] Y. Liu and P. LI, "A path to the globalization of Chinese proficiency standards in international Chinese education," *Chinese Teach. World*, vol. 2, pp. 147–157, 2020.
[18] S. A. Crossley, J. Greenfield, and D. S. McNamara, "Assessing text readability using cognitively based indices," *TESOL Q.*, vol. 42, no. 3, pp. 475–493, 2008.