Construction of Quantitative Index System of Vocabulary Difficulty in Chinese Grade Reading

Huiping Wang，Lijiao Yang[1]
Institute of Chinese Information Processing,
Beijing Normal University
Beijing, China
18013752306@163.com yanglijiao@bnu.edu.cn

Huimin Xiao
Chinese Language and Cultural College,
Beijing Normal University
Beijing, China
3087136946@qq.com

***Abstract*—Chinese grade reading for children has a broad application prospect. In this paper, Chinese textbooks for grade 1 to 6 of primary schools published by People's Education Press are taken as data sets, and the texts are divided into 12 difficulty levels successively. The effective lexical indexes to measure the readability of texts are discussed, and a regression model to effectively measure the lexical difficulty of Chinese texts is established. The study firstly collected 30 indexes at the text lexical level from the three dimensions of lexical richness, semantic transparency and contextual dependence, selected the 7 indexes with the highest relevance to the text difficulty through Person correlation coefficient, and finally constructed a Regression to predict the text difficulty based on Lasso Regression, ElasticNet, Ridge Regression and other algorithms. The regression results show that the model fits well, and the predicted value could explain 89.3% of the total variation of text difficulty, which proves that the quantitative index of vocabulary difficulty of Chinese text constructed in this paper is effective, and can be applied to Chinese grade reading and computer automatic grading of Chinese text difficulty.**

Keywords-readability; grade reading; regression models

I. INTRODUCTION

"Graded reading" refers to a reading method and strategy that starts from the characteristics of children's age, thinking and socialization, selects books suitable for children of different ages and guides them to read (Xu Jianhua, Liang Haoguang, 2011). For teachers, parents and children, it is very important to choose the text with appropriate difficulty among many reading texts. The essence of "graded reading" is the grading of text difficulty.

The concept commonly used to describe the ease of writing is text readability. The readability of a text refers to the degree or nature of its readability and comprehension (Li Shaoshan, 2000; Fry, 2002). At present, the relevant researches on the readability of text are mainly divided into the researches on the readability based on the traditional formula of the readability of text features, the researches on the readability inspired by cognitive theory, the researches on the language model method based on word statistics, and the researches on the readability based on the natural language processing foundation and the machine learning model method. Most studies focus on English texts, focusing on the construction of readability formulas and the verification of their validity, such as Flesch,Spache,Powers-Sumner-Kearl and other formulas (Guo Wanghao, 2009). Although there are many kinds of formulas, and the constants and variable coefficients in these formulas are different, the most important predictors are syntactic features (mainly manifested as sentence length) and lexical features (mainly manifested as word length and word frequency).

Vajjala & Meurers(2012) compared the influences of lexical features and syntactic features on the readability of English texts and found that lexical features have stronger influences than syntactic features, especially the word length index (the average number of syllables and characters per word) has the strongest predictive power on the readability of English texts. Different from the importance of word length in English texts, Chinese words are short and of limited types, with monosyllabic words and disyllabic words dominating. The influence of word length on the intelligibility of meaning and the readability of texts may not be as obvious as that of English words. What are the lexical indexes that affect the readability of Chinese text? Previous studies have found that the token, the type, the type-token ratio, the number of unusual words, the number of unique words, the sum of the normalized value of the whole word set, the ratio of difficult words, the number of notional words, the number of function words, the ratio of notional words to function words, and the number of fixed phrases are important lexical indicators to measure the readability of Chinese text. (Sun Hanyin, 1992; Chen Alin, Zhang Su, 1999; Zhang Ningzhi, 2000; Wang Lei, 2005, 2017; Yang Jinyu, 2008; Guo Wanghao, 2009; Chen Yujia, 2012; Zuo Hong, Zhu Yong, 2014; Sun Gang, 2015 et al.). However, these researches usually refer to the readability of English texts rather than directly focusing on the lexical features of Chinese.

The existing research results mainly have the following two problems: First, the correlation statistics between lexical indexes and text readability are carried out directly, and the specific analysis and explanation of the Chinese lexical features subordinated to the indexes are not made, as a result it is difficult for readers to know the role of lexical features in text difficulty classification directly from the statistical data. Second, there is a lack of research on the weights of different indicators in the same lexical features. For example, the number of functional words, the number of notional words and the ratio of notional to function words are all part-of-speech characteristics of vocabulary, which are proved to be important indexes affecting the readability of Chinese texts, but which indexes can best reflect the relationship between part-of-speech characteristics and text readability has not been compared and analyzed in previous

[1] Corresponding autor

studies. Even though some studies have clarified the relationship between lexical indexes and lexical features, the study of lexical features is not comprehensive and systematic, and there are also some errors in the attribution of measurement indexes. For example, Bie Xiaolei (2017) only deals with lexical features such as part of speech and frequency in the selection of lexical factors affecting the readability of Chinese texts, but not with other lexical features such as word length and lexical diversity. The number of characters and strokes should belong to the factor of Chinese characters, but Song Yaoting et al. (2013) took the number of characters and strokes with Chinese character attributes as the index of vocabulary number and vocabulary length respectively, and incorporated them into the measurement of vocabulary factors.

From the perspective of lexical features in Chinese, this paper, by referring to the lexical richness analysis framework and combining with previous research results on Chinese readability, selects possible measurement indexes from the three dimensions of lexical richness, semantic transparency and context dependence, and discusses the effective lexical parameters for measuring text difficulty. Through selecting the items, a regression model is established to measure the vocabulary difficulty of Chinese text effectively, so as to provide support for constructing Chinese readability assessing model and Chinese reading ability assessing tool from the perspective of lexicology.

II. Text Vocabulary Difficulty Measurement Index

A. Lexica lRichness

"Lexical richness" mainly includes lexical diversity, lexical complexity, word frequency profile, lexical errors, and lexical density (Read, 2000). By using some of these dimensions, the researchers measured the depth and breadth of the learner's vocabulary knowledge and assessed their vocabulary levels. Vajjala & Meurers (2012) applied lexical diversity and lexical density in the vocabulary richness analysis framework to the English text difficulty classification study, which achieved good results. In view of this, this paper develops from the perspective of lexical features of Chinese, and draws on the framework of lexical richness analysis to explore effective vocabulary parameters for measuring the difficulty of Chinese text.

1) Lexical diversity

"Lexical diversity" refers to the use of different words by learners in language expression to avoid the reuse of certain words, which can also be called "Lexical variability". The traditional method of measuring vocabulary diversity is to calculate the ratio of the number of different words to the total number of words, namely the Type-Token Ratio (TTR). (Malvernetal., 2004) The number of times the same word appears in the text is counted as a type, and every occurrence of a word in the text is counted as a token. However, the traditional TTR is affected by the length of text, that is, if there are too many words in the text, TTR cannot effectively evaluate Lexical diversity (Arnaud, 1984; Richards, 1987).

In order to solve this problem, some researchers have proposed measures such as Guiruud index, Herdan index, Mass index, Uber index, HD-D value, average segment-like-to-shape ratio (MSTTR), and text diversity measurement (MTLD), in which the Guiruud index value is relatively stable and less affected by text size, proved to be one of the most useful tools for studying vocabulary diversity (McCarthy & Jarvis, 2007, 2010; Torruella & Capsada, 2013)). In order to eliminate the influence of text length, this paper also considers the Guiraud index on the basis of examining the type, token and TTR.

2) Lexical Complexity

Read (2000) points out that "complex word", also known as Lexical rareness, refers to the less commonly used word (Unusualword) or advanced word (Advancedword) that appear in the text. Research suggests that the higher the proportion of complex words used, the higher the text quality and learner's language level (Linnarud, 1986; Vermeer, 2000; Liu Donghong, 2003; McNama-raetal., 2010; Wan Lifang, 2010; Fan, 2012, etc.). In this paper, according to the "Compulsory Education Common Vocabulary", the scope of simple vocabulary is defined as first-level and second-level words. The scope of complex vocabulary is defined as three-level and four-level words. The absolute and relative quantities of simple words and complex words in textbooks of different grades are measured respectively.

3) Lexical frequency profile

"Lexical frequency profile" refers to the use of words with different degrees of commonality in a text, which reflects the proportion of words used in each word frequency level (Laufer & Nation, 1995). In the study of English word recognition, researchers have found that word frequency is an important variable affecting word cognition. The higher the frequency of occurrence and use of a word, the easier it is to extract and activate from the psychological dictionary when reading. Conversely, the use of low- frequency words has a relatively long activation time and a relatively low level of activation. In general, if a word is often used, it becomes easier to understand because of repeated cognition. Therefore, the role of word frequency factors in distinguishing text legibility cannot be ignored. This paper refers to the frequency of use of modern Chinese words published by the National Language Committee, and divides the words into the most common words, common words, sub-common words and very words, and counts the number and proportion of different levels of words used in textbook texts. In addition, this article also considers the number and proportion of high-frequency words in the text.

4) Lexical density

Lexical density refers to the ratio of the number of words with lexical attributes (ie, words with practical meaning, including nouns, verbs, adjectives, etc.) to the total number of words (Ure, 1971). Gilliland (1972), based on a series of experimental studies, proposes that the ratio of notinoal words to functional words will affect the legibility of the article. Zhang Biyin also found in the book Reading Psychology (1992) that readers mainly pay attention to the notional words in the reading process, focusing on the meaning of the notional words. And people who read quickly tend to work on notional words. The more substantive words in the text, the greater the vocabulary density, the more information is passed, and the difficulty of the text increases accordingly. Vajjala & Meurers (2012) also considers the notional word ratio to be an effective indicator of the difficulty of measuring English text. Chen

Yujia (2012) used six sets of commonly used Chinese second language textbooks (496 texts) as the analysis object, and used the chi-square statistical method to select the most relevant measurement indicators. It was found that the number of notional words and the ratio of notional words were important vocabulary indicators for measuring text readability.Wang Lei (2005, 2017), Zuo Hong and Zhu Yong (2014) found that the number of function words is an important lexical factor affecting readability in the study of Chinese two-language text readability.

In view of the classification criteria for Chinese notional words and function words, there is still controversy in the academic circles. This paper adopts two ways to calculate the notional word density: one is to take the usual practice in foreign countries, dividing the notional words into nouns, substantive verbs, adjectives and adverbs, and the rest are classified as function words; the other is to draw on the word classification criteria of Zhu Dexi (1982), and to use nouns (including time words and position words, place words).), verbs (including auxiliary verbs and directional verbs), adjectives (including attribute words and status words, distinguishing words), numerals, quantifiers, pronouns (including personal pronouns, demonstrative pronouns, and interrogative pronouns) are classified as notional words, adverbs, prepositions, conjunctions, auxiliary words (including modal particles), interjections, onomatopoeia, etc. are classified as function words.

B. Semantic transparency

In 1962, S. UIImann, a British functionalist semantician, put forward the idea of "semantic transparency/obscurity words" in his book "Semantics: An Introduction to the Science of Meaning". Semantic transparency refers to the degree to which the semantics of a compound word can be inferred from the semantics of each morpheme that constitutes a compound word. Its operation is defined as the degree of semantic correlation between the whole word and its morpheme (Wang Chunmao, Peng Yuling, 1999). Semantic transparency is usually divided into four types: total transparent, transparent, obscurity and total obscurity. The study of semantic transparency has application value in children's reading and Chinese as a foreign language. Xu Caihua (2001) studied the influence of semantic transparency on children's reading, and believed that transparent words can promote learning, while opaque words will cause certain obstacles to learning. Based on this, this paper assumes that the transparency of word meaning is also one of the characteristics that affect the readability of text.

In this paper, we use the word vector to represent the semantics of the whole word, the char vector represents the semantics of the morpheme, and calculate the semantic transparency of the compound word by calculating the semantic similarity between the word and the char. We add the morpheme char vectors and then take the mean value. Then we calculate the similarity between the morpheme char vectors and the whole word vectors. Then we can get the semantic transparency of the whole word. Our formula for calculating semantic transparency is as follows:

(1) Adding and averaging the morphological prime word vector:

$$C_m = \frac{\sum_{i=1}^{n} c_i}{n}$$

Cm is the average of the morpheme vectors, ci is the morpheme vector, and n is the number of morphemes that make up the compound word.

(2) Calculate semantic similarity using cosine distance:

$$s = \frac{w \times c_m}{\sqrt{w^2 + {c_m}^2}}$$

In this paper, the cosine distance is chosen to calculate the semantic similarity. As shown in Equation 2, s is the similarity of the average of the morpheme vectors and the compound word vector, and w is the word vector.

(3) Normalized semantic transparency：

$$w_t = s + 0.5 \times s$$

In order to facilitate analysis, as in equation (3), we normalize the semantic similarity obtained at (2). W2 is the semantic transparency of compound words.

Based on the above method for calculating the semantic transparency of words, we use the cleaned 6.2G Wikipedia unlabeled corpus as the data set, and choose Word2vec as the word vector training tool to train the word vector. This paper calculates the semantic transparency of all the words in the textbook. We divide words with transparency higher than 0.4 into high-transparency words, such as “花园” and “吃饭”, and divide words with a transparency lower than or equal to 0.4 into low-transparency words, such as “马上” and “须眉”. We investigate the distribution of high transparency words and low transparency words in different difficulty levels of text.

C. Contextual Dependence

Context dependent refers to whether the understanding of lexical meaning depends on the context. For example, “刘爷爷今天早上走了”. This is an ambiguous sentence, the ambiguity caused by the polysemy “走”. “走” means “离开” (leave) and “死亡” (die). It depends on the context to know which one to take. In practice, there are cases where some polysemous words can have two or more interpretations in a sentence, and the specific meaning needs to be determined by context. We believe that the meaning of words is related to the difficulty of text comprehension. But is it that polysemous words will increase the difficulty of text reading? The answer is not always. The common words such as "打" and "大" have more meanings, but because they usually have specific meanings in sentences, they will not cause reading difficulties for readers. We believe that polysemy with high context dependence is difficult to read. Based on the Modern Chinese Polysyllabic Dictionary, we manually screens 2000 ambiguous words with high degree of context dependence, and investigates the distribution of these polysemous words with high context dependence in textbooks of different grades.

III. REGRESSION MODEL OF TEXT VOCABULARY DIFFICULTY

The regression model constructed in this paper can be divided into five stages: (1) pre-processing stages such as automatic word segmentation, part-of-speech tagging and punctuation removal; (2) obtain each characteristic value; (3) feature selection through Pearson correlation coefficient; (4) Lasso and Ridge models are used to carry out regression analysis of text vocabulary difficulty; (5) verify the model.

A. Corpus processing

1) Selection of text

The readability of the text depends on the difficulty of the text. It is very important to select a series of Chinese texts with distinct difficulty levels, because the main indexes affecting readability and their weights should be analyzed and refined on the basis of comparing the linguistic features of Chinese texts. Under the present condition, the primary school Chinese textbooks compiled and approved by many Chinese experts have a systematic gradient from easy to difficult, which is an ideal language material for the study of the readability of Chinese texts. This paper takes the Chinese textbooks for grade one to grade six published by People's Education Press as the corpus, selects the modern narrative, expository, argumentative and applied texts, and excludes the poems, lyrics, ballads, fu and other rhymes to build the text library. There are 576 articles in total, and the grade level of the text is its difficulty value, which is divided into 12 levels. The distribution at all levels is as follows: 36 at level 1, 41 at level 2, 42 at level 3, 38 at level 4, 42 at level 5, 52 at level 6, 51 at level 7, 55 at level 8, 54 at level 9, 56 at level 10, 54 at level 11, and 55 at level 12.

Considering that the number of texts and words at the lower level are smaller than those at other levels, in order to avoid data skew, we added 50 articles and invited 10 educational experts to grade them. The final total of each level is: 47 for level 1, 45 for level 2, 48 for level 3, 49 for level 4, 51 for level 5, 52 for level 6, 51 for level 7, 55 for level 8, 54 for level 9, 60 for level 10, 58 for level 10, and 55 for level 12.

2) Preprocessing

Word is the basic unit of sentence composition. The first step of Chinese processing is word segmentation and part-of-speech tagging. In this paper, the NPIR system for modern Chinese is used for word segmentation and part-of-speech tagging. Because there are some errors in the system, we need to proofread the results manually. The proofreading work is mainly based on the Modern Chinese Dictionary (7th edition), and exclusion of proper words such as names and place names, Western letters and alphabetic words (such as "father") in statistics. After word segmentation and part-of-speech tagging are completed, punctuation is removed.

B. Feature acquisition

On the basis of examining the correlation between the difficulty of Chinese textbooks for primary school students and various factors of lexical richness framework, semantic transparency and context dependence, this paper puts forward 30 lexical parameters to measure the difficulty of texts. The results are shown in table I.

TABLE I. QUANTITATIVE INDEX SYSTEM OF TEXT VOCABULARY DIFFICULTY

b	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**
Number of type	Number of token	TTR	IOG	Number of most common words	Number of common words	Number of sub-common words	Number of non-common words	Ratio of most common words	Ratio of common words
m	**n**	**o**	**p**	**q**	**r**	**s**	**t**	**u**	**v**
Ratio of sub-common words	Number of non-common words	Number of complex words	Number of simple words	Number of outline words	Ratio of complex words	Ratio of simple words	Ratio of outline words	Number of highly transparent words	Number of lowly transparent words
w	**x**	**y**	**z**	**aa**	**ab**	**ac**	**ad**	**ae**	**af**
Ratio of highly transparent words	Ratio of lowly transparent words	Number of notional words 1	Number of notional words 2	Substantive density 1	Substantive density 2	Number of polysemous words	Ratio of polysemous words	Number of textbook high-frequency words	Ratio of textbook high-frequency words

C. Feature selection

The common feature selection methods can be roughly divided into three categories: filtering, enveloping and embedded. In this paper, the filtering method is selected. By evaluating the degree of correlation between individual features and difficulty level, the ranking leaves the features with high degree of correlation. We draw the correlation thermal diagram of each index, and measure the correlation of index based on Pearson correlation coefficient. The correlation thermal diagram is shown in figure 1.

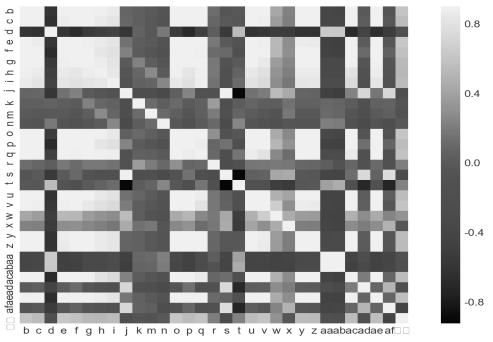

Figure1. Thermal Diagram of Index Correlation

Usually, the correlation intensity of variables is judged as follows: correlation coefficient 0.8-1.0 is extremely

strong correlation; correlation coefficient 0.6-0.8 is strong correlation; correlation coefficient 0.4-0.6 is medium; correlation coefficient 0.2-0.4 is weak correlation; correlation coefficient 0.0-0.2 is weak correlation or no correlation. As shown in the figure, the lighter the color, the higher the correlation, we can draw the following conclusions:

(1) IOG, number of complex words, number of most common words, number of notional words, number of lowly transparent words, and number of polysemous words are all significantly positively correlated with the text difficulty.

(2) the proportion of non-common words, common words and highly transparent words is not related to the text difficulty.

(3) the correlation coefficient of variables such as the number of most common words, the number of common words, the number of frequently used words, and the number of non-common words is close to 1, and there may be collinearity. These variables can be used as features to predict sentence difficulty, but because of collinear problems, only one can be selected as features.

Finally, we selected the following features based on the two principles of removing indexes with high collinear rate and preferential selection of features with high contribution to the result value:

(1) vocabulary diversity -- IOG

(2) vocabulary complexity -- number of complex words

(3) word frequency overview -- number of unusual words and number of textbook high-frequency word

(4) notional word density -- notional word density 1

(5) semantic transparency -- number of lowly transparent words

(6) context dependence -- number of polysemous words

D. Experimental results

Taking 500 texts as training samples, this paper constructs a regression algorithm based on Lasso Regression, Elastic Net, Ridge Regression, Gradient Boosting Regression and XGBoost. In order to ensure the objectivity and accuracy of model evaluation and avoid over-fitting caused by uneven data distribution, we adopt the 10 fold cross-validation method to record the mean square error as the evaluation index, and finally take the average of 10 tests as the experimental results. Finally, it is found that the effects of the five algorithms are similar. Lasso model has the best effect, the mean square error is 0.1040 and the standard deviation is 0.0069.

Different models have their own advantages and differences, and model fusion can give full play to the advantages of each model. We choose the above five models as the basic model, and use the average value of the predicted values of the five models as the final predicted value. The final experimental results are shown in Table II.

TABLE II. EXPERIMENT RESULT OF TEXT DIFFICULTY REGRESSION

Model	Neg-mean-squared-error	
	Mean	***Std***
Lasso	0.1040	0.0069
ElasticNet	0.1109	0.0074
Ridge	0.1137	0.0075
Gradient Boosting	0.1068	0.0080
XGBoost	0.1151	0.0069
Averaged-Model	0.1022	0.0052

Model evaluation

To evaluate the model, we selected 100 texts for validation. The regression model trained above was used for prediction. The input values were IOG, number of complex word, number of high-frequency word, number of most common word , number of notional word, number of low transparent word and number of polysemous word in each text, and the output value was predicted difficulty value of each text. Regression analysis was conducted with predicted value as independent variable and actual difficulty value as dependent variable, and the results were shown in table 3. The regression results showed that the model fitted well and the predicted value could explain 89.3% of the total variation of text difficulty.

TABLE III. REGRESSION ANALYSIS OF MODEL PREDICTION VALUE ON TEXT DIFFICULTY

	R2	**Adjusted R2**	**F Variation**	**Standardization Beta**	**Value of t**
Prdictive **Value**	0.895	0.893	376.818***	0.946	19.412***

Note :* means p<0.05, ** means p<0.01, *** means p<0.001

IV. RESULTS ANALYSIS

Through experiments, we found that IOG index, number of most common words, number of high frequency words in textbooks, number of complex words, notional word density, polysemous words and lowly transparent words are the most effective indexes affecting the text difficulty. We can find that:

(1) lexical diversity affects the readability of texts. With the increase of text difficulty level, the more diverse the lexical types are, the more diverse the use of word meanings is.

It is proved that IOG index and number of polysemy are effective indexes to predict lexical diversity of text readability. According to the IOG index, the higher the difficulty level of the text, the richer the words in the text and the more diverse the expressions. From the use of different meanings of polysemous words, with the increase of text difficulty level, the use of meaning is more. For example, the locational noun "上" has two main meanings in the Modern Chinese Dictionary. One is to express the specific position, and the other is to extend to time, scope and rank, which is more obscure than the meaning of the first one. In the primary difficulty level, it generally refers to "1 position word, used after a noun, means on the surface of an object", while in the higher difficulty level, it is more inclined to use "2 position word, used after a noun, means within the scope of something", which is a more abstract meaning, and the use of sense 2 obviously increases.

(2) word frequency profile affects the readability of text, and the number of most common words increases with the increase of text difficulty level.

The topic of low difficulty text is limited, the vocabulary in the text is also limited, the coverage of most common words is low. In the middle and high stage, the text content is more abundant, including narrative, exposition, argumentation and lyric prose and other genres. These contents make the vocabulary to organize the text structure and express the feelings of the text more diverse, the total text vocabulary has increased significantly, and the coverage of the most common words also increases.

(3) lexical density affects the readability of a text. As the difficulty level of a text increases, the number of notional words increases, while the content word ratio decreases.

By comparing the distribution of parts of speech in different levels of corpus, we find that nouns, verbs, adjectives, auxiliary words and adverbs are the most commonly used five parts of speech in all levels of texts, among which nouns, verbs, adjectives and other notional words have the largest number and the largest proportion. Compared with low-difficulty texts, the proportion of notional words in high-difficulty texts is decreasing. This may be related to the theme setting of the text. The content of primary difficulty text is mainly fairy tales, fables and stories. The language of these subjects is popular, simple and vivid. Words are often presented by notional words with rich meanings. Therefore, notional words expressing real meanings such as things, actions, behaviors, places and time occupy an absolute advantage in primary difficulty text. There are fewer functional words with legal meaning or function, mainly "的", "了", "着", "都" and other commonly used functional words. Highly difficult texts are richer in content, including narrative, expository, argumentative and Lyric prose. These texts increase the proportion of functional words which bear the functions of organizing text structure, expressing words and logical relations between words and sentences, while the proportion of notional words decreases accordingly.

(4) the transparency of words affects the readability of text. With the increase of text difficulty level, the frequency of words with low transparency becomes more prominent.

Semantic transparency plays an important role in compound word processing. Previous studies have found the semantic transparency effect, that is, transparent words are processed faster than opaque words. In the single-word priming condition, compared with opaque words, words with high transparency have shorter vocabulary judgment time. Morpheme information of words with high transparency promotes the processing of whole words, while morpheme information with low transparency hinders the processing of whole words(Wang Chunmao et al. ,1999). Semantic transparency is an important factor in the process of Chinese vocabulary cognition. By examining the transparency of compound words in textbooks, we find that, compared with low-difficulty texts, the proportion of words with low transparency in high-difficulty texts increases gradually and the comprehension difficulty increases.

V. Conclusion

In this paper, Chinese textbooks for grade 1 to 6 of primary schools published by People's Education Press are taken as data sets, and the texts are divided into 12 difficulty levels successively. The effective lexical indexes to measure the readability of texts are discussed, and a regression model to effectively measure the lexical difficulty of Chinese texts is established. The study firstly collected 30 indexes at the text lexical level from the two dimensions of lexical richness and semantic transparency, selected the 7 indexes with the highest relevance to the text difficulty through Person correlation coefficient, and finally constructed a Regression to predict the text difficulty based on Lasso Regression, ElasticNet, Ridge Regression and other algorithms. The regression results show that the model fits well, and the predicted value could explain 89.3% of the total variation of text difficulty, which proves that the quantitative index of vocabulary difficulty of Chinese text constructed in this paper is effective, and can be applied to Chinese grade reading and computer automatic grading of Chinese text difficulty.

Based on the analysis and summary of the experimental results, we believe that the text vocabulary difficulty regression in this paper still needs to be improved in the following aspects :(1) the number of samples is small, so it is necessary to increase the number of samples to alleviate the problem of data sparsity and improve the prediction effect of text vocabulary difficulty; (2) in this paper, the lexical level features that affect the text difficulty are selected, which are completely based on statistical methods. The lexical difficulty of the text is the result of the interaction of multiple factors, and some interventions will be carried out in combination with rules in later work.

References

[1] Chall J S, Dale E.Readability revisited:the new Dale-Chall readability formula[J].Brookline Books, 1995:149.

[2] Colins-Thompson K, Calan J P.A language modeling approach to predicting reading dificulty[C]//Hu- man Language Technologies:the 2004 Conference of the North American Chapter of the Asociation for Computational Linguistics. Asociation for Computational Linguistics, 2004: 193-200.

[3] East, Martin. The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. Assessing Writing, 2006, 11(3) .

[4] Engber, Cheryl A. The relationship of lexical proficiency to the quality of ESL compositions. Journal of Second Language Writing, 1995, 4(2) .

[5] Fan, Fengxiang. A quantitative study on the lexical change of American English. Journal of Quantitative Linguistics, 2012, 19(3) .

[6] Fiona.J Tweed & Baayen, R Harald. How variables may a constant be Measures in lexical richness in perspective. Computer and the Humanities, 1998, 32(4) .

[7] Flor M, Klebanov B B, Shehan K M.Lexical tight- nes and text complexity[C]//The Workshop on Nat- ural Language Procesing for Improving Textual Accesibility, 2013:29-38.

[8] Gilliland,J. Readability[M]. University of London Press for the United Kingdom Reading Association. 1972.

[9] Heilman M, Colins-Thompson K, Eskenazi M.An analysis of statistical models and features for reading dificulty prediction[C]//Procedings of the Work- shop on Innovative Use of Nlp for Building Educational Applications, 2018:71-79.

[10] Petersen S E, Ostendorf M.A machine learning approach to reading level asesment[J].Computer Spech & Language, 2009, 23(1):89-106.

[11] Pierre, J. A. Rare words, complex lexical units and the advanced learner[A], In Coady and Huckin, ed.,Second Language Acquisition[C].Shanghai,2001.

[12] Vogel M，Washburne C. An objective method of determining grade placement of children's reading material[J].Elementary School Journal，1928，28(5):373- 381.

[13] Chen alin, Zhangsu. On the Difficulty Model of Chinese Text Reading and Readability Formula[J]. Computer Science,1999(11):42-44+27.

[14] Guo Wanghao. Rearch on Readability Formula of Chinese Text for Foreign Students[D]. Shanghai Jiao Tong University, 2010.

[15] Jing Xiyu. Study on Readability of Chinese Textbooks: Estimation of Grade Value[J]. Educational Research Information.1995, (05): pp113-127.

[16] Li Yongkang. Researches in Definition of Words-Difficulty in Second Language[J]. Journal of Anhui University of Technology(Social Sciences) . 2003,20(05):pp122.

[17] Liu Xiao. A Review of Text Readability Studies[J]. Journal of Hubei University(Philosophy and Social Science) ,2015,42(03):141-146.

[18] Liang Maocheng. What is Corpus Linguistics[M]. Shanghai: Shanghai Foreign Language Education Press, 2016.

[19] Sun Hanyin. Chinese intelligibility formula[D]. Beijing: Beijing Normal University.1992.

[20] Sun Gang. Research on Prediction Method of Chinese Text Readability Based on Linear Regression[D]. Nanjing University, 2015.

[21] Wang Yixuan. The Correlation between Lexical Richness and Writing Score of CSL Learner－－the Multivariable Linear Regression Model and Equation of Writing Quality[J]. Applied Linguistics, 2017(02): 93-101.

[22] Wang Lei. Some Concepts of Readability Formula and Relevant Research Paradigm as well as the Research Tasks of Formula in TCFL[J]. Language Teaching and Linguistic Studies.2008,(06):pp50-57.

[23] Wang Chunmao, Peng Danling, The Roles of Surface Frequencies, Cumulative Morpheme Frequencies, and Semantic Transparencies in the Processing of Compound Words[J]. Acta Psychologica Sinica, 1999(03):266-273.

[24] Wu Siyuan, Cai Jianyong, Yu Dong, Jiang Xin. A Survey on the Automatic Text Readability Measures[J]. Journal of Chinese Information Processing, 2018,32(12):1-10.

[25] Xu Caihua, Litang. The Role of Semantic Transparency on Word Recognition and Reading Comprehension:An Experimental Study on Children[J]. Applied Linguistics,2001(01):53-59.

[26] Xu Jianhua, Liang Haoguang. Study on The Current Conditions and Countermeasures of Grade Reading Guide in Children's Library[J]. Library Tribune,2011, 31(06):247-252.

[27] Zhang Biyin. Reading Psychology[M]. Beijing: Beijing Normal University Press.1992.

[28] Zhang Yan, Chen Jiliang. Quantitative measuring approach of lexical richness in speech production[J]. Foreign Language Testing and Teaching, 2012(03): 34-40.

[29] Zhao Shaohui. A Multi-dimension Perspective of Assessment of Lexical Competence[J]. TCSOL Studies, 2008(2).