2nd International Conference on Social Science, Public Health and Education (SSPHE 2018)

Relationship Research between Chinese-character Continuation and Language Difficulty Level in Chinese Text

Zezhi Zheng*
College of Humanities
Xiamen University
Xiamen, China
zezhizheng@126.com

Dongjie Zhou
College of Humanities
Xiamen University
Xiamen, China
zhoudongjie0396@163.com

Abstract—Text grading is beneficial to compile, recommend and evaluate the textbook, etc. Therefore, the measure of difficulty level of texts has received considerable attentions from researchers. In this paper, the static difficulty of units of language and knowledge, like Chinese-characters and words or phrases, which we named it as "potential", and the dynamic combined difficulty, like the usage of Chinese-character Continuations and word Continuations, which we named it as "state", the two type measurements were researched in Chinese text. With our studies of metrological analysis, evidences and data, the results suggested that (i) in the “potential” measurement, the amount of word type shows statistical reliability; (ii) in the "state" measurement, the Chinese-character Continuation is better than the word Continuation and the two-Chinese-character Continuation is better than the other length Chinese-character Continuation in the comprehensive performance for measuring the language difficulty level of Chinese text; (iii) the Chinese-character Continuation could break the encapsulation of word, and to a considerable extent, shows the difficulty of combination, semantic and logical relations of language expression. As a result, the Chinese-character Continuation has high statistical reliability for measuring the complexity of Chinese text.

Keywords—language difficulty level, two-Chinese-character Continuation, “potential” and "state" measurement

I. INTRODUCTION

The difficulty fitness of teaching materials has an important influence on learners’ learning effectiveness. The textbook compilers have clear reference indexes in knowledge arrangement by using the syllabus as reference. Taking Science teaching material as an example, the knowledge and skills to be introduced in different grades are specified and explained in the 2017 edition *Primary Schools Science Curriculum Standard For Compulsory Education*. However, the requirements of the language usage are hardly mentioned. Textbooks are essentially used to show and explain knowledge system for readers. The using effect of textbooks is directly affected by the style of language expression and the readability of text language. However, the readability of language lacked of quantitative, scientific and reasonable measurements in Chinese texts. The most important feature of Chinese is that Chinese-characters could be directly put together into sentence to express meaning according to the context, and that most words are made up of Chinese-characters. As a result, both Chinese-character and word are the constituents of sentence which is the basic unit of semantics and information transmission, reflecting the content of text. Therefore, the relationship between the Chinese-character, word, Chinese-character Continuation, word Continuation and the language difficulty level of text in the textbook was studied to find out the appropriate measurement index to compile, recommend and evaluate the textbook.

The key points to carry on the researches are finding out what elements cause language difficulty of text and which indexes can be used to measure the language difficulty of text. The language difficulty of text should be measured in the static and dynamic aspects, because the text is a combination of linguistic knowledge and the application of linguistic knowledge. According to the static and dynamic aspects, the text language difficulty could be divided into "potential" and "state" difficulty.

As we all know, "potential energy" is the energy stored in a system, which could be released or converted into other form of energy. Chinese-characters and words in a text are generally recognized as static units or language knowledge units in Chinese language expression. Here, the measurement of Chinese-character quantities and word quantities in the Chinese text was used as the "potential" measurement of text language difficulty, which was regarded as the basic measurement of text language difficulty.

However, the combinatorial state among the elements including characters, words, phrases, etc. in the text, was used as the "state". Text is a system formed by the combination of Chinese-characters and words for expressing the meaning. The “state” difficulty is a comprehensive index for measuring the semantic, syntax, logic and structure of the text. Therefore, the “state” measurement formed by the combination of Chinese-characters and words also should be considered and valued. The "potential" and "state" language difficulty in the text were firstly measured by examining the relationship between Chinese-character, word, Chinese-character Continuation, word Continuation and the language difficulty level of text.

The goal to research the grading of text language difficulty is to provide suitable learning or reading materials for the readers of different ages. Up to now, some scholars have studied it. For example, Sun(2004) and Jiang(2006) studied the selection and classification of Chinese-characters and words in

This research was supported by the center of the national language monitoring and research for the education and teaching material (EJ20131206-05).

the Chinese-character and word list which was used to test Chinese proficiency level[1-2]. Zhu (2012) analyzed the internal grading reasonableness of *Zhongwen Tiantian Du* based on the factors, such as lexical difficulty, text length and semantic chunks length[3]. Zhou(2013) made a systematic study on four sets of grading reading materials of Chinese based on examining the indicators of specific users, the selection of content, language difficulty, and the set of grading, and pointed out the problems existed in grading reading materials of Chinese by comparing grading reading materials of China and foreign country[4]. Wang (2017) established a formula for calculating the reading difficulty of Chinese books by extracting the characteristics of the Chinese-characters difficulty and sentence difficulty (including lexical difficulty level, sentence length, sentence phrase number) of books, so as to realize the automatic calculation of the difficulty of modern vernacular books[5]. The above researches have promoted the study of the grading of text language difficulty, but these researches are basically carried out based on the static aspect of lexical and sentence length, instead of dynamic combinatorial aspect of static units and compared aspect of static units and dynamic combination.

In summary, the researches of the grading of the language difficulty of the text only on teaching Chinese as a foreign language and the extracurricular reading materials for children, have been studied by some scholars. The reports of the difficulty grading of teaching materials have not been emerged. Textbook is the model for grading the teaching material, which has the nature attribute of difficulty grading. The difficulty measurement unit for effectively sorting the difficulty grade of teaching material could be found with the aid of the natural grade attribute of the textbooks. Therefore, science textbooks for the primary school[a] were used as the research object, and the "potential" (the usage of Chinese-characters and words) and "state" (the usage of Chinese-character Continuations and word Continuations) were used as the indexes to evaluate the text language difficulty. By examining the performance of these two types evaluation indexes in the grade axis, the aim of this study was to find out the effective index for sorting the grading of the text language difficulty, to provide reference for the grading of teaching material.

II. The Continuation

In order to identify the segmentation unit of words, Wang (2001) proposed the construction of a frequently used Continuation corpus which mainly included the high-frequency and multi-Chinese-character groups. The multi-Chinese-character groups were not included in Chinese dictionaries and could be "seen and understood", and often appeared together, such as "zhurou", “zongshangsuoshu”, etc. However, because the high-frequency and multi-Chinese-character groups contained some cross-phrase phrases (such as ”laiyi”, “zaiye”, “baifenzhi”), they were called “frequency Continuation”, instead of a frequency phrase or structure[6]. Wang(2001) indicated that the Continuation used as the segmentation unit of words was based on Chinese-characters, and was similar to high-frequency chunks, rather than words. The fact that it was not enough for Chinese expressions to only focus on words was showed.

The concept of "Continuation" in this article is different from that of Wang (2001), it means the strings of different lengths which appear together in the text, including both Chinese-character Continuations and word Continuations, including both high-frequency Continuations and non-high-frequency Continuations. Moreover, the new Continuations used in different grade sections are also included. Compared with the text of the previous grade, the new Continuations are the unique Continuations in the text of current grade. In this study, the Continuation break the boundaries of words in dictionary, it could contain the new words and new combinations of Chinese-characters which are contained in real text, and embody the linguistic fact that Chinese-characters could be directly put into sentences to describe new things, new relations and new concepts. Therefore, Continuation was used as research object to study its effect on the text language difficulty, so as to illustrate the validity of Chinese-character Continuation for the grading of language difficulty of text.

Since the proportion of Chinese words whose length ranges from one-Chinese-character to four-Chinese-characters make up more than 99% of Chinese words[7], the length of Chinese words longer than 4 characters are seldom seen in Chinese. The fact that the Chinese words whose continuous length exceeds 4 units have no reference value in the study of the text language difficulty could be concluded. Therefore, the distribution of one to four-Chinese-character Continuations and one to four-word Continuations from the 3th to 6th grade in science textbooks for primary school was studied.

According to the definition of Continuation, the shortest Chinese-character Continuation was the two-Chinese-character Continuation, referring to the adjacent, continuous and the length of two-Chinese-characters group. The definition of three and four-Chinese-character Continuation was similar. For ease of narration, one-Chinese-character was classified into Chinese-character Continuations in the data presentation, which was called one-Chinese-character Continuation, but it was still discussed separately in specific studies. The shortest word Continuation was two-word Continuation, referring to the adjacent, continuous and the length of two words group. The definition of three-word and four-word Continuation was similar. Similarly, for ease of narration, one-word was classified into word Continuations in data presentation, which was called one-word Continuation, but was also discussed separately in specific studies. Based on the above definition, one-Chinese-character Continuation and one-word Continuation belonged to the "potential" measurement element of text. The Continuations of two-Chinese-character and longer than two-Chinese-character, and the Continuations of two-word and longer than two-word were the "state" measurement elements of text, in the measuring of the language difficulty level of text.

[a] The science textbooks (the first edition) from the 3th to 6th grade were published by Educational Science Press, in May, 2004.

III. The Effect of Token and Type Number of Continuation on the Language Difficulty Grading of Textbooks

The hypothesis that the language difficulty level of text increased with the increase of grade was used as research premise. The distribution of token and type number of Continuations of text in the grade axis was investigated to verify the validity of the above statistical indexes in measuring the language difficulty level of text.

A large number of character Continuations existed in textbooks, and needed to be processed for the research. Therefore, a software product (CENASS) was independently developed based on VB for extracting Continuations and automatically counting the number of Continuations.

A. The Token Number Distribution of Continuations in the Grade Axis

CENASS was used to analyze the token number of one to four-Chinese-character Continuations and one to four-word Continuations in textbooks from the 3th to 6th grade respectively. Fig. 1 and Fig. 2 show the statistical results.

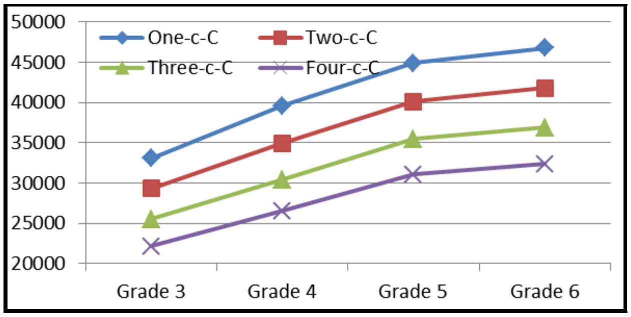

Fig. 1. Statistics on the token number of Chinese-character Continuations in the different grades.

As shown in the Fig. 1, it could be found that the token number of Chinese-character Continuations of different lengths increases synchronously with the increase of grade, indicating that the token number of Chinese-character Continuations is positively correlated with grade and keeps consistency with the trend that the language difficulty level of text increases with the increase of grade. Thus, the token number of one-Chinese-characters and Chinese-character Continuations as the measure unit of language difficulty level of Chinese text has reliability of quantity statistics. Moreover, in terms of different grades, the token number distribution from one-Chinese-character to four-Chinese-character Continuations is the same, which is monotonically increasing.

As shown in Fig. 2, the token number of word Continuations from the 3th to 5th grades increases monotonously, but the token number of word Continuations of each length in the 6th grade is slightly lower than that in the 5th grade. With the trend that the quantity of Chinese-character Continuations number increases with the increase of grade, the token number of word Continuations in the 6th Grade decreases. Although the decreased amount is not many, it is enough to clarify the fact that it is inconvenient to use the word Continuation to measure the text language difficulty.

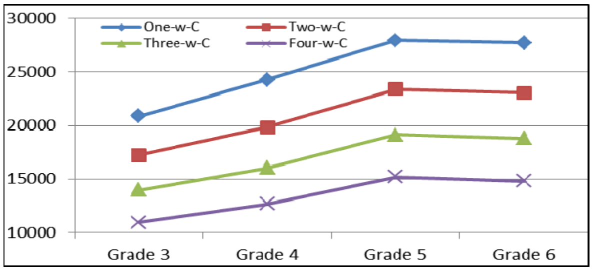

Fig. 2. Statistics on the token number of word Continuations in the different grades.

To some extent, the reason for explaining the different distribution between Chinese-character and word, Chinese-character Continuation and word Continuation in the grade axis, may be that most of the Chinese-characters are independent and meaningful morphemes which have form, sound and meaning[8-9]. In Chinese expression, Chinese-characters have the ability to express new phenomena and things by direct and temporary combination, while temporary combination could directly produce object-oriented words or directly be put into sentence to express the intended meaning. This intrinsic characteristic of Chinese-characters, theoretically and logically, would lead to that the measure of text language difficulty based on the combination of Chinese-characters may be better than the word-based measurement. In addition, to a certain extent, the encapsulation of words conceals the difference of reading difficulty caused by the stable and real-time or temporary combination of Chinese-characters. For example, in terms of the perspective of Continuation, the "rencai" as a word belongs to stable combination in the sentence of "Ta shi ge rencai.", the "rencai" as a Chinese-characters combination belongs to the real-time combination needed for expressing meaning in the sentence of "Zheyang de ren cai neng chengdan daren.". Moreover, the long word could be divided into multiple Chinese-character Continuations which could be used for measuring the language difficulty inside word by breaking the static encapsulation of word. Therefore, the Chinese-character Continuation could be used to measure the language difficulty existed in word, inside-word, and between-words. However, the validity of Chinese-character Continuation still needs to be further verified by data.

B. The Type Number Distribution of Continuations in the Grade Axis

Table I and Table II show the type number of one to four-Chinese-character Continuations and one to four-word Continuations in the textbooks from the 3th to 6th grade, which is counted by CENASS.

TABLE I. THE TYPE NUMBER DISTRIBUTION OF CHINESE-CHARACTER CONTINUATIONS IN DIFFERENT GRADES

Type / Grade	Chinese-character Continuations			
	One-C	***Two-C***	***Three-C***	***Four-C***
Grade 3	1492	10865	16812	17982
Grade 4	1628	12826	20168	21592
Grade 5	1617	14193	22867	24692
Grade 6	1815	15775	25188	27018

TABLE II. THE TYPE NUMBER DISTRIBUTION OF WORD CONTINUATIONS IN DIFFERENT GRADES

Type Grade	**Word Continuations**			
	One-W	***Two-W***	***Three-W***	***Four-W***
Grade 3	3280	10595	11951	10272
Grade 4	3738	12332	13722	11782
Grade 5	4127	14431	16216	14047
Grade 6	4667	15141	16661	14113

TABLE III. THE INCREMENT AND GROWTH RATE OF CHINESE-CHARACTER CONTINUATIONS TYPE NUMBER BETWEEN ADJACENT GRADES

Increase Grade	**Chinese-character Continuations increment[a]/growth rate[b]**			
	One-C	***Two-C***	***Three-C***	***Four-C***
Grade 4-3	136/9%	1961/18%	3356/20%	3610/20%
Grade 5-4	-11/-1%	1367/10.7%	2699/13.4%	3100/14.4%
Grade 6-5	198/12%	1582/11.1%	2321/10.1%	2326/9.4%

Table I and Table III show the trend that the type number of one-Chinese-character Continuations increases and decreases, then increases with the increase of grade in the textbook from the 3th to 6th grade, indicating that the type number distribution of one-Chinese-character Continuations in the grade axis is disordered, which is inconsistent with the trend that the language difficulty level of text increases with the increase of grade. The type number of one-Chinese-character Continuations could not be used to measure the language difficulty level of text.

Although the type number of one-Chinese-character Continuations could not distinguish the language difficulty level of different grades, more than 9% monotonous increment of Chinese-character Continuations in Table III shows that the increment of new knowledge or new expression pattern is obvious, indicating that the Chinese-character combinations are more diverse and the expression is more novel, namely the difficulty and novelty of text increase significantly. It is further illustrated that Chinese-character Continuation or the Chinese-character combination as a sensitive measure element of text language difficulty has obvious hierarchical distinction in text language difficulty measuring.

TABLE IV. THE INCREMENT AND GROWTH RATE OF WORD CONTINUATIONS TYPE NUMBER BETWEEN ADJACENT GRADES

Increase Grade	**Word Continuations increment/growth rate**			
	One-W	***Two-W***	***Three-W***	***Four-W***
Grade 4-3	458/14.0%	1737/16.4%	1771/14.8%	1510/14.7%
Grade 5-4	389/10.4%	2099/17.0%	2494/18.2%	2265/19.2%
Grade 6-5	540/13.1%	710/4.9%	445/2.7%	66/0.5%

Table II and Table IV show that the type number of one-word, two-word, three-word and four-word Continuations from the 3th to 6th grade also increases monotonously with the increase of grade. Especially, the increasing rate of the type number of one-word Continuations in grade axis is higher than 10%, which is consistent with the trend that language difficulty level of text increases with the increase of grade, indicating that the one-word type number as "potential" measurement could be used as a sensitive measure element for measuring the language difficult level. Compared with the lowest growth rate of the type number of Chinese-character Continuations and the one-word Continuations higher than 9%, although the type number of word Continuations also increases monotonically with the increase of grade, the lowest growth rate of them is less than 5%. The little obvious increment indicates that the type number of word Continuations is not suitable for distinguishing the text language difficulty of different grades.

The conclusion that the encapsulation effect of word on the Chinese-characters may cause the above phenomenon, needs to be further studied.

IV. THE DISTRIBUTION OF NEW CONTINUATION IN GRADE AXIS

According to the analysis in the previous section, the Chinese-character Continuation is better than word Continuation in distinguishing the text language difficulty of different grades. But how do the new Chinese-character Continuation in different grades perform? What is the relationship between new Chinese-character Continuation and the language difficulty level of text? Thus, the new Chinese-character Continuation used in different grades which represents the new content and new expression of the text, would be investigated to study.

Compared with the adjacent grades, the amount of new Chinese-character Continuations used in the higher grade reflects the novelty of the text content, so the higher the proportion of the type number and token number of new Chinese-character Continuations in the Chinese-character Continuations, the more obvious the change of text content between the adjacent grades, the more suitable the new Chinese-character Continuation used as an observation point to distinguish the language difficulty level in different grades.

As shown in Table V, it is clearly that except for the new one-Chinese-character Continuations type number, the type number of new Chinese-character Continuations in different grades increases steadily with the increase of grade, which is consistent with the trend that the language difficulty level of text increases with the increase of grade, implying that the new Chinese-character Continuation has reliability for distinguishing the language difficulty level of text in different grades. Moreover, the coverage rate of token number of new Chinese-character Continuations in different grades is close (Table VI), which shows the reliability for the hypothesis of that the grade is used as the criteria for distinguishing language difficulty level of text.

The fact that the type number of new one-Chinese-character Continuations do not increase steadily with the increase of grade proves again that the one-Chinese-character has no clear correspondence relation with the text language difficulty.

[a] The increment of type number of Continuations refers to the amount of type number of Continuations in the current grade minus the number of type number of Continuations in the previous grade.

[b] The growth rate of type number of Continuations refers to the value which is 100% multiplies by the number of that the type number of Continuations used in last grade is divided by increment of type number of Continuations.

TABLE V. THE TYPE NUMBER AND PROPORTION OF NEW CHINESE-CHARACTER CONTINUATIONS BETWEEN ADJACENT GRADES

Type Grade	New one-C type number[a]/percent[b]	New two-C type number/percent	New three-C type number/percent	New four-C type number/percent
4-3	436/26.8%	9387/73.2%	17849/88.5%	20544/95.1%
5-4	413/25.5%	10419/73.4%	20405/89.2%	23616/95.6%
6-5	477/26.3%	11153/70.7%	22084/87.7%	25650/94.9%

TABLE VI. THE TOKEN NUMBER AND COVERAGE RATE (C_R) OF NEW CHINESE-CHARACTER CONTINUATIONS BETWEEN ADJACENT GRADES

Token Grade	New one-C number[c]/C_R[d]	New two-C number/ C_R	New three-C number/ C_R	New four-C number/ C_R
4-3	1669/4%	16061/46%	23951/78.7%	24430/92.2%
5-4	1482/3.3%	18068/45%	27738/78.2%	28634/92.1%
6-5	1599/3.4%	17911/43%	28404/77%	29608/91.6%

Table V and VI observed in the transverse direction show that the type number and token number of new Chinese-character Continuations increases monotonously with the increase of the length of new Chinese-character Continuations, indicating that the longer the length of Continuation is, the higher the novelty of the text is. However, the longer the length of Continuation means the lower repetition of Continuation in the text. For instance, the proportion of the new three-Chinese-character and the new four-Chinese-character Continuations have reached a very high value. The token number of the new three-Chinese-character and the new four-Chinese-character Continuations in current grade have approached or exceeded the token number of the three-Chinese-character and the four-Chinese-character Continuations of previous grade. Moreover, the proportion of the new three-Chinese-character and new four-Chinese-character Continuations whose frequency is one, has reached more than 73% and 84% respectively, indicating the poor repeatability and stability of three-Chinese-character and four-Chinese-character Continuation. The three-Chinese-character and four-Chinese-character Continuation are not suitable for measuring the text language difficulty. However, compared with the three-Chinese-character and four-Chinese-character Continuation, although the proportion of the type number of new two-Chinese-character Continuations reaches more than 70%, the token number of the new two-Chinese-character Continuations reaches less than 46%, indicating that the repetition rate of the two-Chinese-character Continuations of the previous grade in the current grade has reached more than 54%, illustrating that the two-Chinese-character Continuation has better statistical significance, reproducibility, and stability.

V. THE ANALYSIS OF TWO-CHINESE-CHARACTER CONTINUATION

This section the two-Chinese-character Continuation would be studied in detail by taking all the two-Chinese-character Continuations of the 4th grade in the science textbook as an example, because the two-Chinese-character Continuation could break the encapsulation of word and is better than the Continuations of other length in the performance of language difficulty grading of text. The type and type usage situation of two-Chinese-character Continuation, and the function of each type of two-Chinese-character Continuation in the language expression are investigated to find out the sensitive element for measuring the language difficulty of text.

A. The Type Distribution of two-Chinese-character Continuation in Different Frequency Segments

According to whether the Chinese-characters in Continuation are word or not, two-Chinese-character Continuations are divided into three kinds as follow. One-word Continuation is the Continuation whose two Chinese-characters constituted one word. Between-words Continuation is the Continuation whose two Chinese-characters belong to two different words respectively. Inside-word Continuation is the Continuation whose two Chinese-characters just constituted part of one word, the length of the word must be longer than two Chinese-characters. One-word Continuation and inside-word Continuation are static unit or encapsulated by static unit, while between-words Continuation shows the combined relationship between word and word, reflecting the connection between concept and concept, representing the narrative nature of language ("state" difficulty).

According to the emerging frequency of the two-Chinese-character Continuations in the textbook of the 4th grade, the two-Chinese-character Continuations are divided into accidental occurrence (A-occurrence, frequency=1) and non-accidental occurrence Continuation (N-A-occurrence). According to the frequency range, non-accidental Continuations are divided into high-frequency Continuation (frequency≧10), medium-high-frequency Continuation (9≧frequency≧5) and medium-low-frequency Continuation (4≧frequency≧2).

According to the application of Continuations in text, the domain features of Continuations could be divided into discipline(DI) and universality(UN). The Continuations with discipline feature are used to express subject knowledge, belonging to knowledge system (knowledge system is the direct teaching object, which is also called object language) [10]. The Continuations with universality feature are used to translate and explain the knowledge system, belonging to narrative system[11]. The knowledge system and the narrative system constitute the whole content of the textbook. Here, the domain features of the one-word Continuation and inside-word Continuation are labeled, according to the actual usage in the textbook. However, due to that the between-words Continuation is the intersection of knowledge system and narrative system, and has both discipline and universality features, it would be discussed in detail in the next section.

[a] The type number of new Chinese-character Continuations refers to the number of new Chinese-character Continuations that emerges in the current grade but do not emerge in the last grade.

[b] The proportion of the type number of new Chinese-character Continuations(percent) refers to the value which is 100% multiplies by the number that the type number of Chinese-character Continuations used in current grade is divided by the type number of new Chinese-character Continuations.

[c] The token number of new Chinese-character Continuations refers to the sum of the frequency of each new Chinese-character Continuations used in the current grade.

[d] The coverage rate of the token number of new Chinese-character Continuations is the value which is 100% multiplies by the number that the token number of Chinese-character Continuations used in the current grade is divided by the token number of new Chinese-character Continuations.

Table VII shows the type distribution of two-Chinese-character Continuations in different frequency segments in the 4th grade.

TABLE VII. THE TYPE DISTRIBUTION OF TWO-CHINESE-CHARACTER CONTINUATIONS IN DIFFERENT FREQUENCY SEGMENTS IN THE 4TH GRADE

Frequency type	N-A-occurrence						A-occurrence		Total/rate	
	f≧10		9≧f≧5		4≧f≧2		f=1			
	DI	*UN*	*DI*	*UN*	*DI*	*UN*	*DI*	*UN*	*DI*	*UN*
One-w-C	214	116	209	99	508	264	519	490	1450	969
Inside-w-C	17	4	64	5	181	27	279	152	541	188
Between-w-C	207/37.1%		449/54.3%		2159/68.8%		5863/82.6%		9678/75.4%	
Total	558		826		3139		8303		12826	

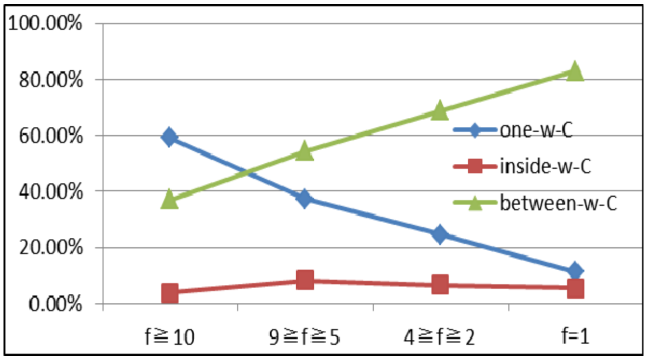

Fig. 3. Type distribution of the two-Chinses-character Continuations in the different frequency segments in the 4th grade.

As shown in Table VII and Fig. 3, several results could be obtained as follow.

- Most of Continuations in high frequency segment are the one-word Continuations which mainly are words in discipline domain and mainly consist of the core knowledge terms of the grade. The universality words in one-word Continuations in high frequency segment are common teaching language, such as “keyi, women”. The inside-word Continuations also mainly consist of knowledge terminology, such as “shengsu, gangyan”. The above analysis shows that except the between-words Continuations, the two-Chinese-character Continuations in the high-frequency segment mainly exhibit the core knowledge system of textbooks.
- The one-word Continuations and inside-word Continuations in the medium-high frequency segment also mainly consist of discipline words. The universality words include a large number of content words which have close relation with the cultivation of pupils' scientific literacy, such as "changshi, zongjie". In this frequency segment, the sum of the one-word Continuations and the inside-word Continuations number is basically equivalent to the number of between-words Continuations, that is, the number of Continuations for expressing the object of knowledge is basically equivalent to the number of Continuations as narrative language.
- In the medium-low and low frequency segments, the number of the between-words Continuations is far more than the sum of the number of one-word and inside-word Continuations. Moreover, the number of universality words in the one-word and inside-word Continuation increases, indicating that there are large amount of narrative language and the high proportion of language for describing and explaining knowledge.

B. The Analysis on the Type of between-words Continuation

The proportion of one-word and inside-word Continuations with discipline property reflects the knowledge difficulty in textbook, while the proportion of between-words Continuations mainly reflects the richness of the narrative language in the textbook, which embodies the "state" measure of text.

The types of the between-words Continuations can be subdivided into four categories as follow by induction.

- The Continuation consists of the morphemes[a] of the three main content words (noun, verb and adjective) [12].
- The Continuation consists of the morpheme of the three main content words and the morpheme of other parts of speech (except the three main content words). One morpheme in the Continuation is core morpheme, the other one is non-core morpheme.
- The Continuation consists of the combination of the morphemes of the content words beyond the three main content words, or the combination between the morpheme of the content words beyond the three main content words and the morpheme of function words[13].
- The Continuation consists of the morphemes of function words.

The statistics data is shown in Table VIII.

TABLE VIII. THE TYPE DISTRIBUTION OF THE BETWEEN-WORDS CONTINUATIONS IN DIFFERENT FREQUENCY SEGMENTS

Frequency Type	f≧10		9≧f≧5		4≧f≧2		f=1		Total	
	type	*num*	*type*	*num*	*type*	*num*	*type*	*num*	*type*	*num*
1th type	4	291	7	556	24	1729	46	2672	81	5248
2th type	29	2908	47	1842	91	3092	146	3631	313	11473
3th type	13	617	23	356	43	434	70	384	149	1791
4th type	2	228	4	59	13	143	23	176	42	606
Total	48	4044	81	2813	171	5398	285	6863	585	19118

Table VIII shows that the number distribution of the four types of between-words Continuation in different frequency segments is roughly the same. In each frequency segment, the quantity and combination type of the second type of between-words Continuation are the most. The second type of between-words Continuation which consists of one core morpheme and

[a] Morphemes not only refer to the linguistic units which are the constituents of words, but also refer to the linguistic units which can be used as words independently (referring to Zhang Bin's *Modern Chinese Descriptive Grammar*, Beijing: Commercial Press, 2010, pp. 13).

one non-core morpheme, is mainly used in the between-words Continuations for describing, restricting things and actions in the nature, shape, relationship and other aspect, such as "yanhe, shizui", which has the ability of connecting the knowledge system and narrative system. Therefore, the second type of between-words Continuation plays a key role for the reflection of language expression complexity for the between-words Continuation. Moreover, the phenomenon which is that Chinese-characters are directly put into sentence to expressing meaning, often occurs in the second type of between-words Continuation, such as “beikou, guanbi”, indicating the flexibility of Chinese characters in Chinese expression.

Although the quantity of the third type of between-words Continuation is not large, the combination type of it is abundant. The third type of between-words Continuation is widely used to describe the entity objects in the knowledge system based on number, tools, location, linguistic relation, etc., which has the ability of supplementary explanation for knowledge system, playing a supplementary role in presenting the novelty of language expression, such as “liangjie, bamei”.

Different from the third type of between-words Continuation, although the combination type of the first type of between-words Continuation is not many, the quantity of it is large. This kind of Continuation whose semantic is substantive, is mainly used to expound the core knowledge, but the expression pattern of which is not complicated, such as “dongfu, chichuan”. Therefore, the first type of between-words Continuation is helpful for presenting the richness of knowledge, but it has little obvious effect on the richness of language expression.

The quantity and combination type of the fourth type of between-words Continuation are least, as a result of that functional morphemes need to be attached to substantive morphemes, and two functional morphemes occurring together are mainly used in the complex expression of semantic relations. With the decrease of frequency, the quantity and combination type of the fourth type of between-words Continuation also increase slowly, indicating that they can present the richness and complexity of linguistic expression. However, the quantity of the fourth type of between-words Continuation is too small to play the very important role for presenting the richness and complexity of linguistic expression.

In conclusion, different types of between-words Continuation play different roles in language expression, and the second type of between-words Continuation plays a key role in presenting the complexity of language expression, because of its diversity of combination type and richness of quantity.

VI. Conclusions

The "potential" measure (the usage of one Chinese-characters and one-words) and the "state" measure (the usage of Chinese-character Continuations and word Continuations) were used as the indexes for measuring the language difficulty level of the textbooks in different grades. The relevant conclusions are as follows.

(I) The "state" measurement is effective to measure the language difficulty level of text, which complements and improves the defect of "potential" measure in the measure of text language difficulty level. In the "state" measurement, Chinese-character Continuation is better than word Continuation in the measure of text language difficulty level, because Chinese-character Continuation has the ability to describe new phenomena and things by the temporary combinations which could break the encapsulation of word. Thus, the Chinese-character Continuation not only could measure the whole content measured by words in the measure of text language difficulty level, but also could measure the language difficulty caused by inside-word Continuation and between-words Continuation.

(II) In Chinese-character Continuations, two-Chinese-character Continuation has the best comprehensive performance in the measure of the text language difficulty level, because most of two-Chinese-character Continuations are the between-words Continuations. The between-words Continuation could reflect the complexity of syntax and semantic logic, and indicate the complexity of language expression in the text, embodying the difficulty of "state" in language.

(III) The second type of between-words Continuation consists of one core morpheme and one non-core morpheme, connecting the knowledge system and narrative system, is mainly used in the between-words Continuation, playing a key role in the between-words Continuation to reflect the language expression difficulty level of text.

The statistical validity of Chinese-character Continuation for distinguishing text difficulty level was verified by researching the distribution of Chinese-character Continuation number and word Continuation number in the grade axis. However, the Chinese-character Continuation used as the evaluation criteria for measuring the language difficulty level of text still needs to be further studied.

References

[1] M. J. Sun, “Scale and statistics of Chinese-characters proficiency testing,” in Applied Linguistics, no. 1, pp. 63-70, 2004.

[2] X. Jiang, G. Zhao, H. Y. Huang, Y. M. Liu, Y. M. Wang, “The effects of frequency, productivity and complexity on learning Chinese-characters and words by foreign students,” in Language Teaching and Linguistic Studies, no. 2, pp. 14-22, 2006.

[3] Y. Zhu and P. C. Zou, “A study of readability of reading Chinese,” in Yunnan Normal University(Teaching and Research on Chinese As A Foreign Language), vol. 10, pp.41-46, 2012.

[4] X. B. Zhou and B. Qian, “An investigation of graded Chinese reading materials – also on the comparison with reading,” in Applied Linguistics, no.2, pp.107-116, 2013.

[5] J. Wang, H. Zhou, G. F. Luo, X. Gu, “Grading the reading difficulty degree based on natural language processing,” in Computer Era, no.8, pp.1-5, 2017.

[6] H. J. Wang, “The internal structure of Wordlist of contemporary Chinese for information processing and the structural characteristics of Chinese,” in Applied Linguistics, no. 4, pp. 90-97, 2001.

[7] J. Zhou, “Combination of Two Character and Dictionary Receipt,” in Studies of the Chinese Language, no.4, pp.304-309, 1999.

[8] R. L. Li, “On the Interactivity and Harmonious Developments of Chinese and the Chinese-characters,” in Jilin University Journal Social Sciences Edition, vol. 49, no. 2, pp. 108-116, 2009.

[9] T. Q. Xu, “Reanalysis of Chinese-characters and the study of Chinese semantic and syntax,” in Linguistic Research, no. 3, pp. 1-9, 2005.

[10] X. C. Su, J. J. Du, J. H. Guan, S. H. Zheng, “The Nature, Characteristics and Significance of Research of Textbook Language,” in Applied Linguistics, no. 4, pp. 86-91, 2007.

[11] Z. Z. Zheng, “Exploration on the Language Research of Subject Textbooks–A Case Study of Math Textbook language Research, “ in Journal of Beihua University (Social Sciences), vol. 18, pp. 1-6, 2017.

[12] Y. S. Zhang, “The Nature, Scope and Classification of Adverbs in Modern Chinese,” in Studies in Language and Linguistics, no. 2, pp.51-63, 2002.

[13] X. Fan, “Reflections on Chinese Empty Words,” in Journal of Shanghai Normal University (Philosophy & Social Sciences Edition), vol. 45, no. 6, pp. 105-115, 2016.