Evaluate Lexical Richness Measures Using Coefficient of Variation and Relative Value

Wanwan Zheng[1] Mingzhe Jin[2]

[1] Doshisha University, Japan
[2] Doshisha University, Japan

Abstract. Although numerous lexical richness measures have been proposed, a positive evaluation method has not been established to select measures independent of text length to authors' best knowledge. As an existing evaluation method, it is common to view the transition curves of the measure's original data or standardized data. However, this method is mostly judged visually and cannot sufficiently capture the change of measures. In other words, this method cannot compare and evaluate lexical richness measures directly by viewing transition curves of either original data or standardized data. In this article, evaluation statistics CV (coefficient of variation) and RV (relative value) are proposed as two possible methods to evaluate lexical richness. Both statistics make it possible to compare the stability of measures by visual observation. The effectiveness and validity of CV and RV are verified with *TTR*, *M*, *K*, *R*, and *C*. Meanwhile, Japanese, Chinese, and English corpora are used to avoid the possible influence of the languages.

Keywords: Lexical richness measure, Coefficient of variation, Relative value

1 Introduction

In recent years, research in stylistics, neuropathology, language acquisition, and even forensics continue to use lexical richness measures (McCarthy and Jarvis, 2007). Many statistical measures have been proposed to express lexical richness.

The most basic idea in measuring lexical richness is *TTR* (Type Token Ratio, Templin, 1957), which is the ratio of the number of different words V(N) to the number of total words N. However, as the text becomes longer, the rate of increase in V(N) declines gradually, while N increases rapidly. Therefore, the influence of N tends to be noticeable. As a well-known problem, *TTR* is highly text-length dependent. This means the longer a text is, the lower the *TTR* value. To minimize N's impact on the *TTR* value, researchers have proposed many improved measures of *TTR* using the characteristics of square root (Guiraud'*R*, 1954; Carroll'*S*, 1967) or logarithms (Herdan'*C*, 1960 & 1964; Summer'*s*, 1966; Mass'*M*, 1966; Dugast'*Uber*, 1978 & 1979; Tuldava'*LN*, 1977; Dugast'*k*, 1979). In contrast, Vermeer (2004) argued that no measure based on the ratio of V(N) to N would be valid. Further, according to Malvern and Richards (2002), for all of the improved versions of *TTR*, it is impossible to escape the influence of text length. To solve this problem, researchers have proposed that the frequency spectrum of words should be incorporated along with V(N) and N. Examples of these measures are the *K* characteristic value (Yule, 1944); *m* (Michea, 1966 & 1971); and *S* (Sichel, 1975). Orlov (1983) proposed *Z*, according to the generalized Zipf distribution, V(N), N, and the frequency of the most common word divided by the text length (p^*) were concerned.

Lately, software-based lexical richness measures such as *D* and *MTLD* (Measure of Textual Lexical Diversity) have been proposed. *D* can be calculated by the special tool vocd (McKee et al., 2000) and D_Tool (Meara and Miralpeix, 2007). Calculating *D* is complicated because it is the result of a series of random text samplings. On the other hand, McCarthy and Jarvis (2010) showed there was a high correlation between *HD-D* and *D*. As *HD-D* could be directly calculated using hypergeometric distribution, it was cited as an alternative to *D*. *MTLD* can be calculated with a special tool called The Gramulator (McCarthy, 2011). As the text is also divided into segments and *TTR* is calculated for each, the starting point of *MTLD* is similar to that of *MSTTR* (Mean Segmental Type-token Ratio). *MSTTR* is an arithmetic mean of *TTR*. Torruella and Capsada (2013) pointed out that although *MTLD* showed low sensitivity to text length, it was more sensitive than *M*.

Among such numerous measures, evaluation researches are being done to find measures that are immune to text length. To find calculation methods that do not depend on text length is also the biggest problem when studying lexical richness of text (Torrulla and Capsada, 2013). In evaluation studies, appropriate data is essential, as inappropriate data may lead to faults in the analysis and, in extension, the results. In previous studies, Tweedie and Baayen (1998) and Kimura and Tanaka (2011) considered the transition curves of measures drawn with the original data. This method is difficult to evaluate lexical richness measures correctly because the stability of transition curves are compared under the condition that measures have different scales. A method of standardizing data as an improved method was proposed by Koizumi and In'nami in 2012 and Jarvis in 2002. This method of standardization can summarize all measures with different scales into one scatter diagram. However, as the overall distribution of the data does not change even if standardization is carried out, it is basically an evaluation by visual method as with the case using the original data.

This literature review concludes that regardless of whether the analysis data is original or standardized data, the stability of measures cannot be sufficiently and correctly compared and evaluated according to their transition curves’ smoothness. To address this issue, this article proposes two statistics: CV and RV. The transition curves drawn with CV and RV allow for direct comparison of the stability of measures using visual observation.

2 Corpus Description

The works of several authors were chosen to obtain versatile conclusions. As the lexical richness may change depending on the time of writing, only long novels were used. In this article, a corpus of Japanese (84,058 words-393,706 words), Chinese

(67,511 words-257,811 words) and English (71,737 words-193,372 words) was constructed. The list of the corpus is shown in Table 1.

Table 1. Corpus

	ID	Works	Tokens
Chinese	C1	Long River (C. W. Shen)	67,511
	C2	Camel Xiangzi (Laos)	88,437
	C3	The Last Quarter of the Moon (Z. J. Chi)	126,167
	C4	Looks Great (S. Wang)	125,959
	C5	The Lotus Lake (L. Sun)	257,811
	C6	Fortress Besieged (Z. S. Qian)	150,785
	C7	White Deer Plain (Z. S. Chen)	148,979
	C8	Spring Fever (D. F. Yu)	167,750
	C9	Soul Mountain (X. J. Gao)	175,285
	C10	Happy accuser (H. S. Zhang)	220,996
English	E1	When All The Woods are Green (S. W. Mitchell)	103,334
	E2	The Sheik (E. M. Hull)	87,825
	E3	The Life of Charlotte Bronte (E. Gaskell)	78,407
	E4	White Nights and Other Stories (F. Dostoevsky)	89,632
	E5	King Coal (U. Sinclair)	71,737
	E6	Wastralls (C. A. D. Scott)	93,266
	E7	The Essays of George Eliot (G. Eliot)	102,280
	E8	Eve (S. B. Gould)	118,014
	E9	Sister Carrie (T. Dreiser)	155,980
	E10	The Financier (T. Dreiser)	193,372
Japanese	J1	The Makioka Sisters (J. Tanizaki)	102,729
	J2	Kozakura Hime Story (A. Asano)	93,059
	J3	Aphrodisiac of play (Y. Mishima)	100,667
	J4	The paradox of Youth (S. Oda)	111,485
	J5	Female genealogy (K. Izumi)	131,230
	J6	Heralds (S. Natsume)	105,917
	J7	A Certain Woman (T. Arishima)	131,993
	J8	Thirst for Love (Y. Mishima)	84,058
	J9	I am a Cat (S. Natsume)	213,319
	J10	September Affair (R. Yokomitu)	393,706

3 Two New Statistics

This section introduces the two statistics, CV and RV, and give an example to describe where the previous problem is and why CV and RV can overcome this problem.

In previous studies, Tweedie and Baayen (1998) and Kimura and Tanaka (2011) divided text into chunks. In this article, to examine changes of measures in detail, the size of a chunk is set to 100 words. In this way, the text is divided into several chunks. Then, the chunks are cumulated separately to calculate each measure’s value.

The lexical richness measure is likely to show a completely different change for different authors and works. In this article, to obtain versatile conclusions, for measure *I*, the average value of ten texts per chunk *i* is used.

$$\bar{v}_{Ii} = \frac{1}{10}\sum_{j=1}^{10} v_{Ii,j} \tag{1}$$

3.1 Coefficient of Variation

3.1.1 Relationship between *R* and *CTTR*

The standard deviation represents the degree of variation from the average value. It cannot be compared simply when the average value and the unit of data are different. When comparing the dispersion of data with different average values and units, CV should be used. The formula is shown below.

$$CV = \frac{\sigma}{\bar{x}} \tag{2}$$

For example, *CTTR* is the result of *R* over $\sqrt{2}$. Here, a problem occurs when dividing a number greater than 1, whether the measure can be improved or not. The standard deviation of *R* is $\sqrt{2}$ times *CTTR*; However, as CV of *R* and *CTTR* are the same, *R* and *CTTR* should have an equal effect in dispersion. This means *CTTR* fails to improve *R*. Alternatively, to consider transition curves drawn with the original data of *R* and *CTTR*, chunks were cumulated one by one to calculate values of each lexical richness measure every time. Transition curves drawn with the original data of *R* and *CTTR* are shown in Fig. 1. By considering Fig. 1, it is likely to incorrectly recognize that *CTTR* is more stable than *R*.

$$R = \frac{V(N)}{\sqrt{N}} \quad (3) \qquad\qquad CTTR = \frac{V(N)}{\sqrt{2N}} \quad (4)$$

3.1.2 Comparison between Standardized Data and CV

In this section, the transition of standardized data and CV of *TTR*, *M*, *K*, *R*, and *C* are compared. Previous studies have pointed out that *M*, *K*, *R*, and *C* are unlikely to be influenced by text length. On the other side, it is well known that *TTR* is highly dependent on text length.

Standardization does not change the overall distribution of measures. In this article, data is standardized by the following formula to put all of the measures into one chart diagram.

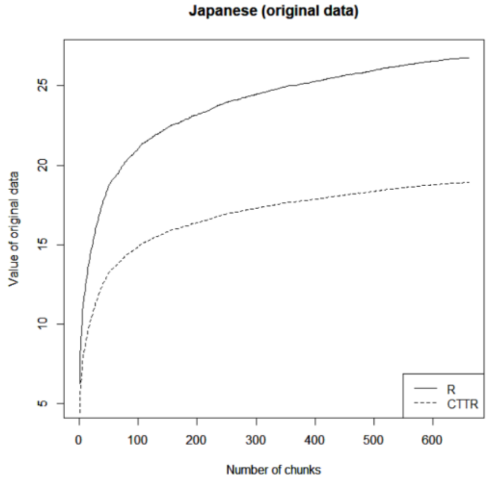


Fig. 1. Transition curves of the original data of *R* and *CTTR*

$$x' = \frac{x - mean}{sd} \times 10 + a \qquad (5)$$

x is the original data, x' is the standardized data, and a is a constant to prevent the negative value.

Subsequently, for each five chunks, a moving CV is computed while shifting from the first chunk. Fig. 2 shows the transition of standardized data of *TTR*, *M*, *K*, *R* and *C* in Japanese corpus, and the CV transition is shown in Fig. 3. The vertical axis is standardized data or CV, and the horizontal axis is the number of chunks. In Fig. 2, *TTR* is obviously more stable than *C*, *M*, and *R* in the latter part, and *K* is the most stable. However, considering the CV transition shown in Fig. 3, CV of *TTR* is the largest. As CV is a statistic for comparing data dispersion with different average values and units, it can be said that *TTR* is more dependent on text length than *M*, *K*, *R*, and *C*. Further, *C* is more stable than *K*. When compared, the standardized data and the CV are showing inconsistent results.

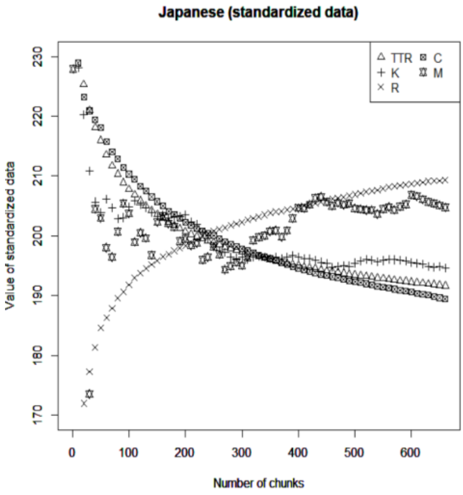
Similarly, the results of Chinese corpus are shown in Fig. 4 and Fig. 5, and the results of English corpus are shown in Fig. 6 and Fig. 7. For Chinese, according to Fig. 4 using the standardized data, *M* and *C* are more dependent on text length than *R*, *K*, but it is reversed when considering Fig. 5 using CV. Similar results were observed in English.

Meanwhile, by considering the transition curves of the original data or standardized data, Tweedie and Baayen (1998) did not judge *C*, Koizumi and In'nami (2012) did not judge *M* as good measure.

Although CV is good for comparing the dispersion of measures, when computing CV, the section must be set first. Moreover, the transition of CV should be different depending on the size of the section and it will affect the conclusion to some extent.

In this article, to examine variations of each measure in detail, CV is obtained for every five chunks. However, this may not be the best setting. Further, it is extremely difficult to decide how much the section should be sized.

To solve the problem of the section setting for CV, we improve the formula of CV and propose the RV as a new statistic.



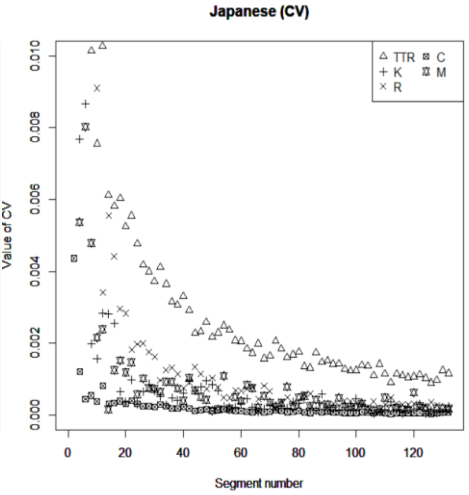
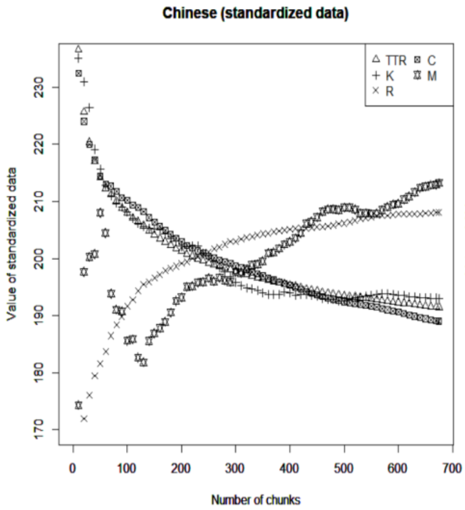


Fig. 2. Transition of Japanese standardized data

Fig. 3. Transition of Japanese CV



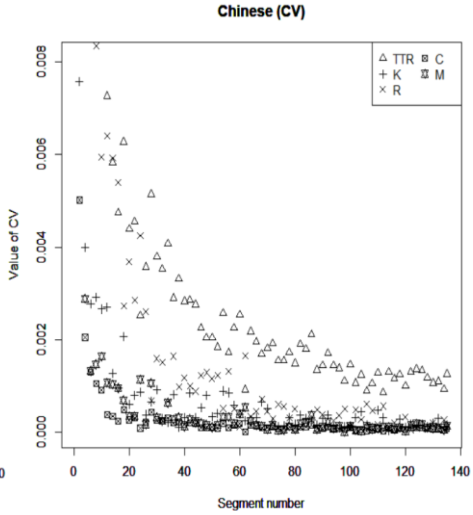


Fig. 4. Transition of Chinese standardized data

Fig. 5. Transition of Chinese CV

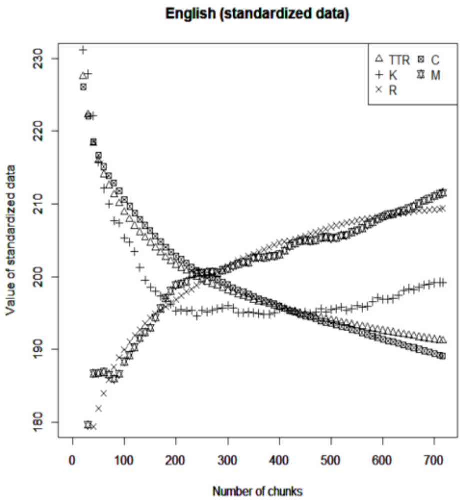


Fig. 6. Transition of English standardized data

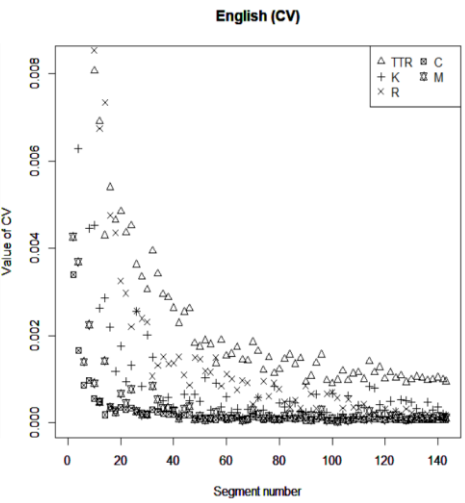


Fig. 7. Transition of English CV

3.2 Relative Value

The original data x divided by the average value $\bar{x}$. In this article it is called RV. RV can avoid setting section size and can also remove the influence of unit.

$$RV = \frac{x}{\bar{x}} \tag{6}$$

3.2.1 The Validity of RV

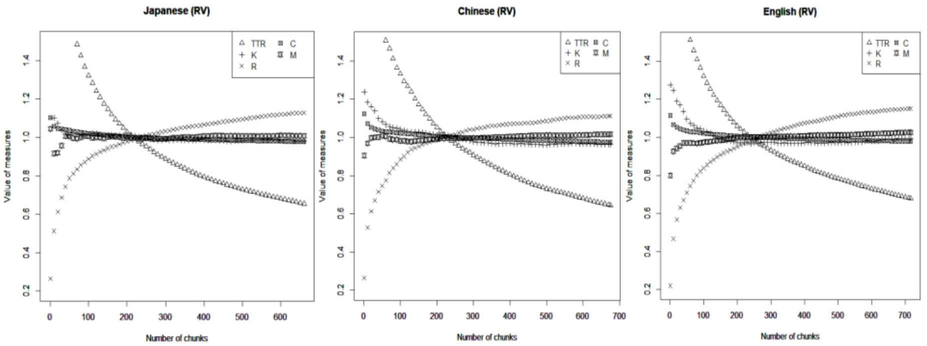
First, the RV's validity is verified from the formula. As shown below, for measure *I*, the standard deviation of RV is the same number as CV.

$$CV_I = \sigma(RV_I) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)\bar{x}^2}} \tag{7}$$

Therefore, the dispersion of measures can be directly compared with the standard deviation of RV. RV relieves the drawback that the transition curves of original data and standardized data cannot sufficiently capture the change of measures in previous studies.

When applied to lexical richness measures, it can be estimated that RV curves of *R* and *CTTR* always overlap, so the same result as CV can be obtained.

Then, the transition of RV and CV is compared to verify the RV's validity in detail. The RV of *TTR*, *M*, *K*, *R*, and *C* are computed and used to draw the transition curves (Fig. 8). As shown in Fig. 8, the distribution of RV of these five measures shows almost the same result as the distribution of CV shown in Fig. 3, 5, and 7. *TTR* is the most unstable, followed by *R*, and the transition curves of *M*, *K*, *C* overlap.



(a) Japanese (b) Chinese (c) English

Fig. 8. Transition of RV of *TTR*, *M*, *K*, *R*, *C*

In Fig. 8 the CV and the RV of *M*, *K*, *C* overlap and therefore to differentiate between these measures Fig. 9 and Fig. 10 were constructed. For Japanese, the distribution of CV of *K*, *M* is mixed, and the dispersion of RV of *K*, *M* are also approximated correspondingly. For Chinese, *M* and *C* take approximate CV. On the other hand, the RV curves of *M* and *C* are similar. For English, similar results are also obtained.

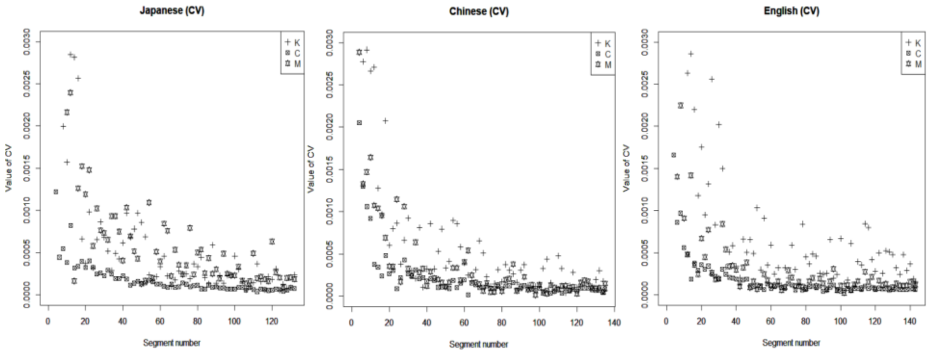
According to the results of comparing Fig. 9 and Fig. 10, RV shows almost the same result as CV, and the transition curve of RV is easier to differentiate.

4 Conclusion

When viewing the transition curves drawn with the original data of lexical richness measures, as the scales of measures are different, it is not appropriate to directly compare the stability. Further, as the average values of measures are different, when comparing the dispersion of the latter part of the transition curve (standard deviation), the average value must also be considered simultaneously. Even if the data has been standardized, as overall distribution does not change, it is the same as the case of using the original data. Therefore, there is a drawback that the stability of measures cannot be sufficiently and correctly compared and evaluated by the smoothness of curves drawn with either the original data or the standardized data.

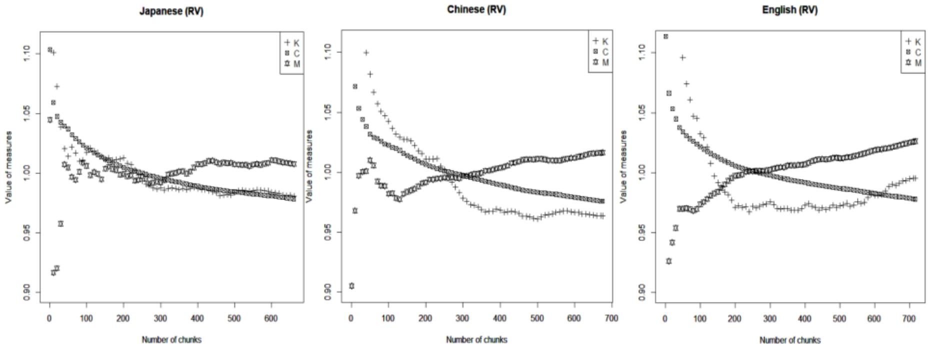
To directly compare the stability of measures using visual observation, this article proposes a method using the coefficient of variation. The standardized data and CV transitions were compared using Japanese, Chinese, and English corpora. As a result, *TTR* is highly dependent on text length and the improvement effect of *M*, *R*, and *C*, which were not seen by the standardized data that can be clearly viewed by CV. Although the standardized data of *K* and the CV of *K* both indicate that *K* is a good measure, the measure of *C* is more appropriate than *K* as reflected through its stability and ease of calculation.

Furthermore, as it is difficult to decide the section size for CV, RV is proposed as an improved statistic. RV avoids setting section size and cancels the influence of units. As the standard deviation of RV is the same as the CV, the dispersion of transition curves of RV can be compared directly. Moreover, according to the consideration when RV applies to real lexical richness measures, the transition of RV shows the same results as CV, and allows for greater visual differentiation.



(a) Japanese (b) Chinese (c) English

Fig. 9. Transition of CV of *M*, *K*, *C*



(a) Japanese (b) Chinese (c) English

Fig. 10. Transition of RV of *M*, *K*, *C*

References

1. McCarthy, P. M., Jarvis, S.: vocd: a theoretical and empirical evaluation. Language Testing 24, 459–488 (2007).
2. Templin, M.: Certain language skills in children: Their development and interrelationships. Minneapolis: The University of Minnesota Press (1957).
3. Guiraud, H.: Les Caracteres Statistiques du Vocabulaire. Universitaires de Fance Press, Pairs (1954).
4. Carroll, J. B.: On sampling from a lognormal model of word-frequency distribution. Computational analysis of present-day American English, 406–424 (1967).

5. Herdan, G.: Type-Token Mathematics: A Textbook of Mathematical Linguistics. Mouton & Co, The Hague， The Netherlands (1960).
6. Herdan, G.: Quantatative Linguistics，Butterworth，London (1964).
7. Dugast, D.: Sur Quoi se Fonde la Notion D'etendue Theoratique du Vocabulaire?，Le Francais Modern，46(1)， 25-32 (1978).
8. Dugast, D.: Vocabulaire et Stylistique. *I: Theatre et Dialogue*, Travaux de Linguistique Quatitative, Slatkine-Champion，Geneva (1979).
9. Vermeer, A.: The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. Vocabulary in a Second Language, 173–189 (2004).
10. Malvern. D., Richards. B.: Investing accommodation in language proficiency interviews using a new measure of lexical diversity. Language Testing 2002 19 (1):85–104 (2002).
11. Yule, G. U.: The Statistical Study of Literary Vocabulary. Cambridge University Press (1944).
12. Shichel, H. S.: On a Distribution Law for Word Frequencies, Journal of the American Statistical Association, 70, 542-547 (1975).
13. Zipf, G. K.: Human Behaviors and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press (1949).
14. Orlov, Y. K.: Ein Modell der Haufigkeit Struktur des Vokabulars，In Studies on Zipf's Law, Brockmeyer， Bochum, 154-233, (1983).
15. Mckee, G., Malvern, D., Richards, B.: Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing, 15(3), 323–337 (2000).
16. Meara, P. M., Miralpeix, I.: D_Tools (version 2.0; _lognostics: Tools for Vocabulary Researchers: Free software From _lognostics) [Computer Software]. University of Wales Swansea (2007).
17. McCarthy, P. M., Jarvis, S.: MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42, 381–392 (2010).
18. Torrulla, J., Capsada, R.: Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. Procedia-Social and Behavioral Sciences 95, 447–454 (2013).
19. Tweedie, F. J., Baayen, R. H.: How Variable May a Constant be? Measures of Lexical Richness in Perspective. Computers and the Humanities, 32:323–352 (1998).
20. Daisuke Kimura, Kumio Tanaka: A Study on Constants of Natural Language Texts. Journal of Natural Language Processing, Vol. 18, No. 2, 119–137 (2011).
21. Koizumi, R., In'nami, Y.: Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. System 40, 554–564 (2012).
22. Jarvis, S.: Short text, best-fitting curves and new measures of lexical diversity. Language Testing 19 (1) 57-84 (2002).
23. Van Hout, R., Vermeer, A.: Comparing measures of lexical richness. Modeling and assessing vocabulary knowledge, 93–115 (2007).