Open Access Library Journal
2018, Volume 5, e4660
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Comparison of Word Length Distributions in Spoken and Written Chinese

Heng Chen

Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China
Email: chenheng@gdufs.edu.cn

How to cite this paper: Chen, H. (2018) Comparison of Word Length Distributions in Spoken and Written Chinese. *Open Access Library Journal*, **5**: e4660.
https://doi.org/10.4236/oalib.1104660

Received: May 15, 2018
Accepted: June 11, 2018
Published: June 14, 2018

Abstract

In this study we apply Zipf-Alecseev's function to word length distributions of Chinese prose and dialogue texts. Since there are two potential measurement units of Chinese word length, we applied Zipf-Alecseev's function to both of them. The results show that all the word length distributions fit Zipf-Alecseev's function, no matter the word length is measured in characters or components. The parameters *a* and *b* in Zipf-Alecseev's function $y = cx^{a+b\ln(x)}$ show no difference in different text styles (which are prose and dialogue in our case). However, the parameters are different when word length is measured in different units (character and component respectively). This indicates that the Zipf-Alecseev's function is sensitive to word length measurement units, but not text styles.

Subject Areas

Linguistics

Keywords

Word Length, Chinese, Zipf-Alekseev's Function, Measurement Units, Text Styles

1. Introduction

Word length plays a crucial role in the development of quantitative linguistics, especially in Köhler's lexical control circuit. There has been a wealth research into word length studies in different languages including Chinese [1]-[8], yet some boundary conditions are still not specified clearly [9] [10] [11] [12] [13]. A fundamental problem throughout the investigation of word length is the question if there is a universal model with which word length distributions can generally be theoretically described. To this end, many efforts have been made (see

[9] for more).

Recently a unified model of length distribution of any unit in language was suggested ([9], p. 5) and the authors assumed that "the relative rate of change of the dependent variable (here the frequency) is proportional to the rate of change of the independent variable (here the length)", which yield the Zipf-Alecseev's function $y = cx^{a+b\ln(x)}$. In the unified model there are merely differences in the parameters, and the parameters themselves are part of a dynamic system displaying self-regulation. The most significance lies in that if we succeed in applying the formula to any level of linguistic entities, we arrive at an enormous simplification.

In this book ([13], p. 17), the author stated that the parameter a in Zipf-Alecseev's function increases with the age of a language, and its values may differ in different languages. Based on the analyses of the values of parameter a in many different languages, Popescu *et al.* conclude that "one can see that Indo-European languages have in general a smaller parameter a than the languages of other genetic groups. However, Chinese is an exception." ([13], p. 77)

In this study, we will explore whether the text styles or measurement units of word length influence the value of a in Zipf-Alecseev's function or not. What is more, since the parameters are part of a dynamic system displaying self-regulation, the dependence of the parameter b on parameter a is also tested.

Specifically, the following questions will be explored in this study.

Question 1: Can the word length distributions of Chinese prose and dialogue texts be modeled by Zipf-Alecseev's function $y = cx^{a+b\ln(x)}$?

Question 2: Do the parameters in fitting Zipf-Alecseev's function to Chinese word length distributions display any self-regulation (the dependence of the parameter b on parameter a)?

Question 3: Are the parameters in Zipf-Alecseev's function sensitive to different measurement units of word length (the potential measurement units of Chinese word length are the character and the component)?

Question 4: Are the parameters in Zipf-Alecseev's function sensitive to different text styles (which are prose and dialogue texts in our case)?

This paper contains four sections. Section 2 describes the materials and methods used; Section 3 presents the results of fitting Zipf-Alecseev's function to Chinese word length distributions, as well as the comparisons of the values of parameter a between different text styles and different measurement units of word length; Section 4 concludes this study.

2. Materials and Methods

In order to measure the word length in spoken Chinese and written Chinese, we built a dialogue text collection (spoken language) and a prose text collection (written language), with 20 texts respectively. The number of words in each text ranges from 726 to 3792. The spoken language texts come from a TV talk show named "QiangQiang San Ren Xing" (in English *Three People*) on Phoenix TV from 2013.06 to 2013.09, 5 texts each month and 20 texts in total, in the form of daily conversation. This TV program mainly discusses the current social hot issues. The written language texts come from a well-known Chinese prose journal *Selective Prose*[1], from 2013.06 to 2013.09, 5 texts each month and 20 texts in total.

We need to explain in detail here that, the word “汉语” (means Chinese) consists of two characters “汉” “语”, and five components: “氵 ” “又” “讠 ” “五” “口”. Since there are no natural boundaries between words, word segmentation is needed before measuring word length. Word segmentation involves the definition of the word, which is a difficult problem especially in Chinese. But it is not the issue we will discuss here, in the present investigation we segment words with unified standard. Firstly, we use the ICTCLAS, one of the best Chinese word segmentation software, to segment words automatically. Then we did the manual checking and corrected the errors. Table 1 and Table 2 show the number of characters and words tokens in each text.

After word segmentation, we developed a java program to measure word length. To measure the number of components of a word, we used a list consisting of 20902 characters (CJK Unified Ideographs) with numbers of strokes and components of each character.[1]

We used Matlab 2012b to do the fitting work, and the goodness of fitting can be seen from the determination coefficients R^2. As for the statistical comparisons, we used t-test through SPSS 19, and we set the significance level to 0.05 in this study.

3. Results and Discussions

Results of fitting Zipf-Alecseev’s function to Chinese word length distributions. In this part we show the results of fitting Zipf-Alecseev’s function to word length distributions of Chinese prose and dialogue texts, including the parameters and the determination coefficients R^2. What is more, the dependence of the parameter b on parameter a is tested to see if Chinese word length distributions display any self-regulation.

Table 3 presents the results of prose texts, the word length of which is measured in characters.

Using the data from Table 3, the relation between the parameters a and b in Table 3 is visualized in Figure 1. The existence of this link is a sign of self-regulation.

Table 4 also presents the results of prose texts as in Table 3, but the word length is measured in components.

The relationship between a and b in Table 4 is visualized in Figure 2. The existence of this link is a sign of self-regulation.

Table 5 displays the results of dialogue texts, and the word length is measured in components.

The relation between the a and b in Table 5 is visualized in Figure 3. The existence of this link is a sign of self-regulation.

[1] *Selected Prose* Website: http://swsk.qikan.com.

Table 1. Number of characters and words in spoken Chinese texts.

Text	Character tokens	Word tokens	Text	Character tokens	Word tokens
1	2168	1589	11	5441	3792
2	1561	1068	12	5419	3783
3	2520	1763	13	5216	3592
4	2245	1526	14	5021	3444
5	1373	941	15	4959	3498
6	1002	726	16	5251	3609
7	2287	1567	17	5093	3571
8	1306	883	18	5127	3437
9	2047	1445	19	4848	3329
10	1822	1278	20	4668	3197

Table 2. Number of characters and words in written Chinese texts.

Text	Characters tokens	Word tokens	Text	Characters tokens	Word tokens
1	1920	1366	11	1928	1368
2	1309	952	12	2655	1861
3	2055	1490	13	1423	948
4	2394	1657	14	2318	1779
5	2014	1502	15	1471	962
6	1550	1119	16	4128	2876
7	1786	1269	17	5143	3654
8	1466	993	18	5012	3512
9	1830	1366	19	4423	3057
10	2693	1928	20	4403	2953

Table 6 also presents the results of prose texts as in Table 5, but the word length is measured in components.

The relation between the a and b in Table 6 is visualized in Figure 4. The existence of this link is a sign of self-regulation.

It can be concluded from the above results that Chinese word length distributions can be modeled by the Zipf-Alecseev's function, and the dependence of the parameter b on parameter a is testified.

3.1. Parameters with Regard to Different Measurement Units and Text Styles

3.1.1. Comparisons between Different Text Styles

1) Character as the measurement unit

Table 7 presents the comparison results between Prose and Dialogue texts for parameter a.

Table 3. Results of fitting Zipf-Alecseev's function to word length distributions of Chinese prose texts (word length measured in characters).

Prose texts	*a*	*b*	*c*	R^2
1	4.829	−6.46	239	0.9988
2	3.674	−5.507	243	0.999
3	4.377	−5.984	272	0.9979
4	5.924	−7.737	320	0.9978
5	5.769	−7.967	273	0.9993
6	4.841	−6.905	257	0.9985
7	5.317	−6.823	211	0.9998
8	5.601	−7.539	205	0.9952
9	4.77	−6.735	261	0.9992
10	5.543	−7.226	272	0.9992
11	4.519	−5.919	224	0.9978
12	5.241	−6.558	260	0.9988
13	5.31	−6.827	199	0.9974
14	3.626	−5.61	409	0.9991
15	6.602	−8.21	177	0.9984
16	5.239	−6.592	411	0.994
17	5.332	−6.777	465	0.9967
18	5.985	−7.578	470	0.9973
19	6.034	−7.439	412	0.9913
20	5.611	−6.799	420	0.998

Table 4. Results of fitting Zipf-Alecseev's function to static word length distributions of Chinese prose texts (word length measured in components).

Prose texts	*a*	*b*	*c*	R^2
1	2.918	−1.479	30.78	0.9785
2	2.362	−1.277	36.1	0.9456
3	2.709	−1.394	37.42	0.9607
4	2.983	−1.41	35.3	0.9605
5	3.31	−1.657	27.97	0.9796
6	2.777	−1.48	35.14	0.9552
7	3.025	−1.468	25.26	0.9442
8	3.2	−1.525	19.96	0.9548
9	3.02	−1.531	29.34	0.9608
10	3.533	−1.685	25.07	0.9564
11	3.45	−1.621	19.67	0.9701
12	3.787	−1.727	20.1	0.967
13	3.042	−1.448	22.32	0.9504
14	3.084	−1.608	43.07	0.9939
15	3.177	−1.436	18.4	0.943
16	3.407	−1.572	38.57	0.9684
17	3.495	−1.597	39.33	0.9747
18	3.798	−1.703	34.34	0.9753
19	4.169	−1.782	23.02	0.9496
20	3.617	−1.61	34.96	0.9686

Table 5. Results of fitting Zipf-Alecseev's function to static word length distributions of Chinese dialogue texts (word length measured in characters).

Dialogue texts	*a*	*b*	*c*	R^2
1	4.706	−6.446	211	0.9992
2	4.724	−5.981	148	0.9995
3	5.618	−7.159	219	0.9991
4	4.345	−5.546	195	0.9997
5	5.425	−6.959	116	0.9999
6	5.922	−8.256	128	1
7	5.461	−6.748	176	0.9991
8	4.241	−5.569	139	0.9989
9	5.138	−6.485	180	0.9998
10	5.083	−6.666	177	1
11	4.597	−5.633	323	0.9996
12	5.964	−7.485	305	0.9996
13	5.292	−6.288	268	0.999
14	4.932	−5.903	292	0.9996
15	5.243	−6.452	248	0.9996
16	5.781	−6.997	289	0.9997
17	4.708	−5.771	303	0.9979
18	5.685	−6.672	258	0.9989
19	5.627	−6.812	293	0.999
20	5.07	−6.3	283	0.9994

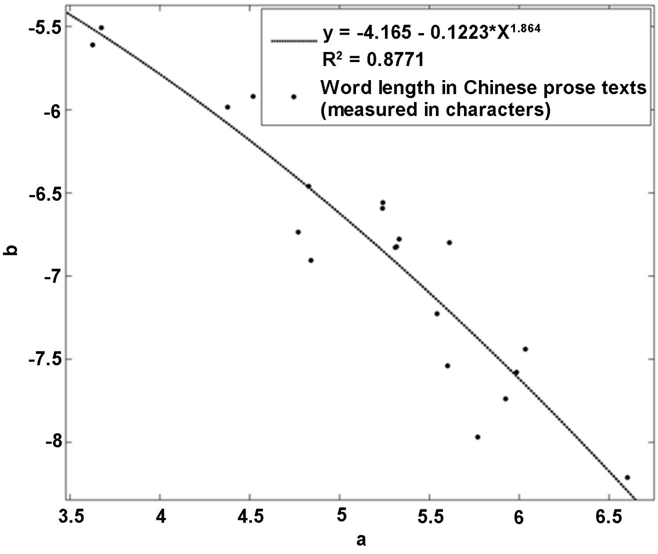

Figure 1. Word length (measured in characters) in Chinese prose texts.

It can be seen from Table 7 that the mean values of *a* (word length measured in characters) between prose and dialogue texts make no difference, and the T-test also verified that there is no significant difference.

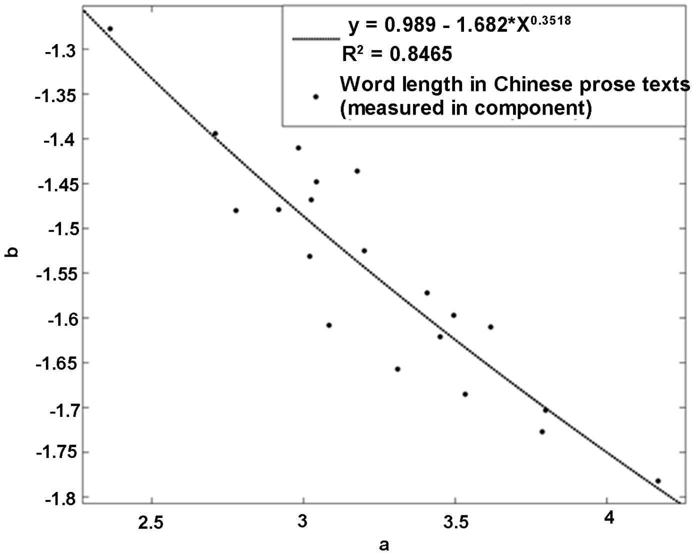

Figure 2. Word length (measured in components) in Chinese prose texts.

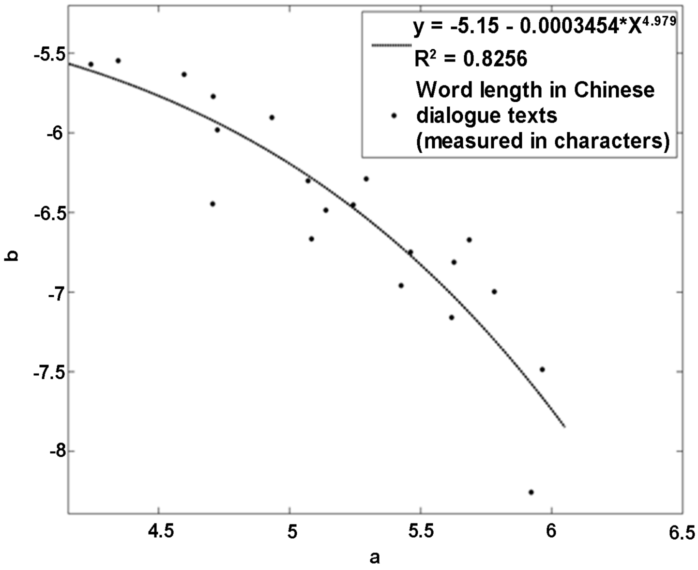

Figure 3. Word length (measured in characters) in Chinese dialogue texts.

2) Component as the measurement unit

When using component as Chinese word length measurement unit, the comparison results are given in Table 8.

Table 8 displays the comparisons of parameter *a* (word length measured in components) in Chinese prose and dialogue texts, and the T-test result also shows no significant difference as in the case of Table 7.

3.1.2. Comparisons between Different Measurement Units

1) Prose texts

As for prose texts, *i.e.* Written Chinese, when word length is measure in different units, the comparison of values of parameter *a* is displayed in Table 9.

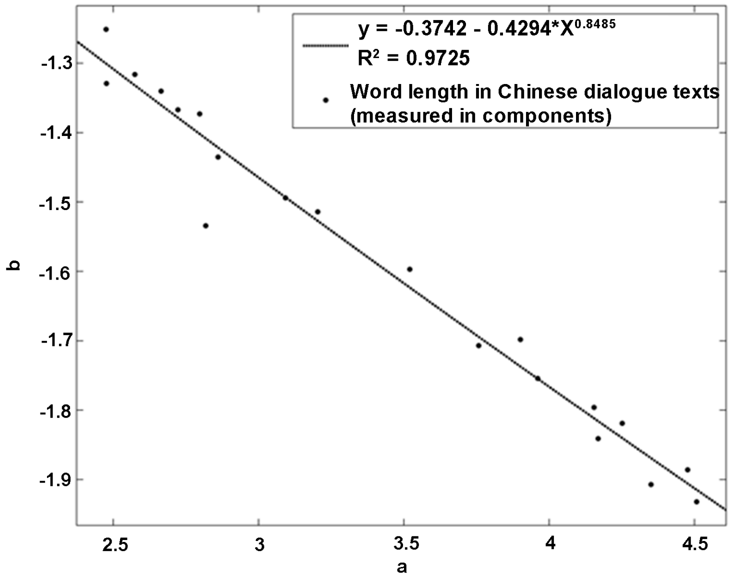

Figure 4. Word length (measured in components) in Chinese dialogue texts.

Table 6. Results of fitting Zipf-Alecseev's function to static word length distributions of Chinese dialogue texts (word length measured in components).

Dialogue texts	*a*	*b*	*c*	R^2
1	2.476	−1.329	34.03	0.976
2	3.092	−1.494	17.25	0.9603
3	2.664	−1.34	33.72	0.9404
4	2.86	−1.435	26.95	0.9523
5	2.475	−1.251	18.79	0.9053
6	2.818	−1.534	19.07	0.9809
7	3.203	−1.514	20.16	0.9405
8	2.797	−1.373	17.46	0.9273
9	2.722	−1.367	26.99	0.9467
10	2.574	−1.316	26.62	0.9621
11	3.757	−1.707	25.6	0.9656
12	4.168	−1.841	18.25	0.9584
13	4.476	−1.886	12.63	0.9432
14	4.154	−1.796	17.18	0.9377
15	3.96	−1.754	16.69	0.9387
16	4.507	−1.932	14.12	0.9581
17	3.52	−1.597	26.34	0.9703
18	4.251	−1.819	15.29	0.9326
19	3.901	−1.698	20.1	0.9396
20	4.35	−1.907	14.9	0.9384

Table 7. Comparisons of parameter *a* between prose and dialogue texts (word length measured in characters).

	Style	N	Mean value	StDev	SE Mean
a	Prose	20	5.2072	0.76146	0.17027
	Dialogue	20	5.1781	0.51235	0.11456

Table 8. Comparisons of parameter *a* between prose and dialogue texts (word length measured in components).

	Style	N	Mean value	StDev	SE Mean
a	Prose	20	3.2432	0.42575	0.09520
	Dialogue	20	3.4363	0.73874	0.16519

Table 9. Comparisons of parameter *a* between different measurement units of word length (prose texts).

	Measurement units	N	Mean value	StDev	SE Mean
a	character	20	5.2072	0.76146	0.17027
	component	20	3.2432	0.42575	0.09520

Table 10. Comparisons of parameter *a* between different measurement units of word length (dialogue texts).

	Measurement unit	N	Mean value	StDev	SE Mean
a	character	20	5.1781	0.51235	0.11456
	component	20	3.4363	0.73874	0.16519

It can be seen from Table 9 that parameter *a* has quite different values when word length is measured by different measurement units, and the T-test results show that there is significant difference between them.

2) Dialogue texts

Then is the dialogue texts, *i.e.* Spoken Chinese, the comparison results are illustrated in Table 10.

Table 10 shows the results of comparisons between different word length measurement units, and it can be seen that the values of *a* are quite different. The T-test result corroborates our observations.

4. Conclusions

Base on the analyses above, we conclude that:

1) The word length distributions of Chinese prose and dialogue texts can be modeled by Zipf-Alecseev's function $y = \mathrm{cx}a + b\ln(\mathrm{x})$.

2) The dependence of the parameter *b* on parameter *a* is testified, which means that the parameters in fitting Zipf-Alecseev's function to Chinese word length distributions display some self-regulation.

3) Different measurement units of Chinese word length lead to different values of parameter *a* in Zipf-Alecseev's function.

The parameters in Zipf-Alecseev's function are not sensitive to different text styles (which are prose and dialogue texts in our case), which means that it may be only sensitive to different language types.

Acknowledgements

This work is supported by the Education Department of Guangdong Province "Innovative Strong School Project" Youth Innovation Talents Project (Humanities and Social Sciences) (Project Number: 2017WQNCX046).

References

[1] Wimmer, G., Köhler, R., Grotjahn, R. and Altmann, G. (1994) Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics,* **1**, 98-106. https://doi.org/10.1080/09296179408590003

[2] Wimmer, G., Witkovský, V. and Altmann, G. (1999) Modification of Probability Distributions Applied to Word Length Research. *Journal of Quantitative Linguistics,* **6**, 257-268. https://doi.org/10.1076/jqul.6.3.257.6163

[3] Wimmer, G. and Altmann, G. (2005) Unified Derivation of Some Linguistic Laws. In: Köhler, R., Altmann, G. and Piotrowski, R.G., Eds., *Quantitative Linguistics. An International Handbook*, de Gruyter, Berlin, 791-807.

[4] Köhler, R. (2005) Synergetic Linguistics. In: Köhler, R., Altmann, G. and Piotrowski, R.G., Eds., *Quantitative Linguistics. An International Handbook*, de Gruyter, Berlin, 760-774.

[5] Chen, H. and Liu, H. (2018) Quantifying Evolution of Short and Long-Range Correlations in Chinese Narrative Texts across 2000 Years. *Complexity*, **2018**, Article ID: 9362468. https://doi.org/10.1155/2018/9362468

[6] Chen, H. and Liu, H. (2016) How to Measure Word Length in Spoken and Written Chinese. *Journal of Quantitative Linguistics*, **23**, 5-29. https://doi.org/10.1080/09296174.2015.1071147

[7] Chen, H., Chen, X. and Liu, H.T. (2018) How Does Language Change as a lexical network? An Investigation Based on Written Chinese Word Co-Occurrence Networks. *Plos One*, **13**, e0192545. https://doi.org/10.1371/journal.pone.0192545

[8] Chen, H., Liang, J. and Liu, H. (2015) How Does Word Length Evolve in Written Chinese? *Plos One*, **10**, e0138567. https://doi.org/10.1371/journal.pone.0138567

[9] Grzybek, P. (2006) History and Methodology of Word Length Studies. In: Grzybek, P., Ed., *Contributions to the Science of Text and Language*: *Word Length Studies and Related Issues*, Springer, Dordrecht, 15-90.

[10] Grzybek, P. (2013) Homogeneity and Heterogeneity within Language(s) and Text(s): Theory and Practice of Word Length Modeling. In: Köhler, R. and Altmann, G., Eds., *Issues in Quantitative Linguistics* 3, RAM-Verlag, Lüdenscheid, 66-99.

[11] Altmann, G. (2013) Aspects of Word Length. In: Köhler, R. and Altmann, G., Eds., *Issues in Quantitative*; *Linguistics* 3, RAM-Verlag, Lüdenscheid, 23-38.

[12] Popescu, I.I., *et al.* (2013) Word Length: Aspects and Languages. In: Köhler, R. and Altmann, G., Eds., *Issues in Quantitative Linguistics* 3. *Dedicated to Karl-Heinz*

Best on the Occasion of His 70*th Birthday*, RAM, Lüdenscheid, 224-281.

[13] Popescu, I.I., Best, K.H. and Altmann, G. (2014) Unified Modeling of Length in Language. RAM-Verlag, Lüdenscheid.