TYPE： Posters

To Be or Not to Be: Will Scientific Writing Affect Scientific Impact?

Chao Lu[1,2] Yi Bu[2] Xianlei Dong[3] Jie Wang[1] Ying Ding[2] Chengzhi Zhang[1,*]

luchaonjust@gmail.com;
[1] Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

buyi@iu.edu;
Indiana University, 107 S Indiana Ave, Bloomington, IN, 47405(U.S.A).

sddongxianlei@163.com;
Shandong Normal University, 88 Wenhua E Rd, Lixia Qu, Jinan, 250000 (China)

1342234559@qq.com;
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

dingying@indiana.edu;
Indiana University, 107 S Indiana Ave, Bloomington, IN, 47405(U.S.A).

zhangcz@njust.edu.cn
Nanjing University of Science and Technology, 200 Xiaolingwei Street, Nanjing, 210094 (China)

Introduction

When it comes to possible benefit of linguistic complexity to scientific impact, opinions are split in communities (e.g., Gopen & Swan, 1990; O'Conner, 2010). To be or not to be complex? This is a worthy question for us to investigate—whether linguistic complexity is crucial to scientific writing so that impact can be expanded. Can scholars succeed when they write pleasant English, usually accompanied with complexity to some extent? Or they just employ writing as a tool to report their findings and ignore the beauty of language. After all, publishing papers of high impact as many as possible is one of scholars' genuine concerns. This preliminary study is to dig this controversial questions based on large-scale fulltext and citation data of academic articles using regression analysis.

Methodology

Data

We collected 170,000 fulltext journal articles with publishing history (the dates when the paper is received, revised, accepted, and published) detailed in dates from 2006-2015 published in *PLoS*[1] and their corresponding citation data harvested from Scopus between 2016 February 3-6, a very short time period, so that we can neglect the potential error in citation data caused by different harvest timelines. We only kept all the articles in Biology by retrieving the assigned disciplinary information to reduce disciplinary differences. Our final dataset contained 49,350 fulltext articles in Biology with their publishing history and citation data with date.

Independent variables

The independent variable, Linguistic Complexity comprises Syntactic Complexity, the sentence-level complexity of language performance, and Lexical Complexity, the vocabulary-level language performance. Variables for syntactic complexity can be divided into two sub-groups: Sentence Length and Sentence Complexity (Vajjala & Meurers, 2012). And Lexical Complexity includes three sub-groups: Lexical Diversity, Lexical Sophistication, and Lexical Density (Ellis & Yuan, 2004; Kormos, 2011) (Table 1). Pearson's Correlation Analysis showed no correlation between these variables.

Table 1. variables of linguistic complexity.

Aspects	Indicators	Descriptions	Formulas
Syntactic Complexity	Sentence Length	Calculating average number of words in sentences and corresponding standard deviation of each article	$MSL = \sum_{i=1}^{N} SLi/N$ $\mathrm{SSTD} = \sqrt{\frac{\sum_{i=1}^{N}(SL_i - MSL)^2}{N}}$
	Sentence Complexity	Counting the ratio of complex sentences that contain "that" or "which" in each article	$CR = \frac{(N_{that} + N_{which})}{N_{all}}$
Lexical Complexity	Lexical Diversity	Type-Token Ratio per 1000 words in each article	$TTR = \frac{\#\ of\ Distinct\ words}{\#\ of\ tokens} \times 1000$
	Lexical Density	Counting the ratio of lexical items in tokens in each paper based on their part of speech	$NR = \frac{\#\ of\ Nouns}{\#\ of\ Tokens}$ $VR = \frac{\#\ of\ Verbs}{\#\ of\ Tokens}$ $JR = \frac{\#\ of\ Adjectives}{\#\ of\ Tokens}$ $DR = \frac{\#\ of\ adverbs}{\#\ of\ Tokens}$
	Lexical Sophisticati on (4)	Counting the length of nous, verbs, adjectives, and adverbs	$MNL = \frac{\sum_{i=1}^{N} Noun\ Len_i}{N}$ $MVL = \frac{\sum_{i=1}^{N} Verb\ Len_i}{N}$ $MJL = \frac{\sum_{i=1}^{N} Adj\ Len_i}{N}$ $MDL = \frac{\sum_{i=1}^{N} Adv\ Len_i}{N}$

Dependent Variable

Since we had only obtained the total citation number for each article, we used the number of citations per month (CPM) as an alternative to eliminate the

possible effect caused by different periods of citation history. The variable was calculated as follows:

$$CPM = \frac{total\ number\ of\ citations}{Month(harvest\ date - published\ date)}$$

Regression Analysis

To investigate the relationship between Linguistic Complexity in scientific writing with scientific impact, we conducted regression analysis, using CPM as dependent variable and the 12 explanatory variables shown in Table 1. Considering that the variables had shown a strong trend of decrease after increase (two samples in Figure 1). Multinomial model $Y = a\mathbf{X}^2 + b\mathbf{X} + c$ was to fit the data. articles with top 1% CPM ranking were selected as the most highly cited articles to compare the regression models.

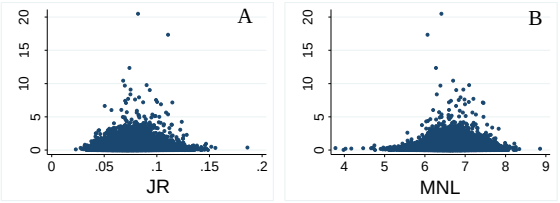

Figure 1. Scatter plots between CPM and JR (A) and MNL (B).

Results

Table 2 shows descriptions of the two models: one for all the full text articles and another for the top 1% articles based on CPM ranking.

Table 2. Descriptions of two regression models.

Variables	Model_All	Model_Top
MNL	0.504***(0.111)	
MNL^2	-0.0364***(0.008)	
MVL	0.698***(0.152)	
MVL^2	-0.0487***(0.011)	
MJL	0.219**(0.069)	
MJL^2	-0.015**(0.005)	
TTR	-0.00265***(0)	
TTR^2	0.00000397***(0)	
CR	0.330***(0.075)	
CR^2	-0.487***(0.118)	
MSL	0.00775***(0.002)	
MSL^2	-0.0000788**(0)	0.00135**(0)
SSTD	0.00314***(0.001)	-0.232***(0.047)
$SSTD^2$	-0.00000861***(0)	0.0045***(0.001)
NR	0.970***(0.085)	-119.5*(50.16)
NR^2		164.5*(69.48)
VR	11.5***(2.161)	
VR^2	-36.24***(7.482)	
JR	6.184***(0.86)	
JR^2	-31.82***(5.365)	
VR^2	53.91***(5.388)	
cons	-6.154***(0.659)	25.72**(8.994)
N	49350	493
R^2	0.017	0.134

Notes: (1) Standard errors in parentheses; (2) $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

In the first model, most features show quadric relationships with CPM ($p<0.01$), except for adverb length and noun ratio (positive linear relationship). The parameters suggest that moderate level of complexity in paper can promote the CPM and that the abundance of vocabulary provides a positive support for CPM of an article. In other words, moderate linguistic complexity is helpful to improving average citation of an article; however, too much complexity in syntactic level or word length may affect gain of citation since the difficulty in reading might dramatically increase according to studies (e.g., Juhasz, 2008).

However, the low level of R^2s in the preliminary models might suggest that major efforts should be made in other areas to improve the content of an article, e.g., the novelty or contribution of the study articled, which is supported by top CPM articles.

Conclusion

This preliminary study is to find out the relationship between linguistic complexity in scientific writing and scientific impact using regression analysis. The results might suggest that complex scientific writing helps improve scientific impact of an article to some extent but the major effort should be made in the content of it, e.g., the novelty and contribution to academic community. Given that the mutual effects of these variables to the impact has not been considered, more types of models should be uses to clarify the relationships between them in the future.

Acknowledgments

This work is supported in part by Major Projects of National Social Science Fund of China (No. 16ZAD224).

References

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), 59-84.

Juhasz, B. J. (2008). The processing of compound words in English: Effects of word length on eye movements during reading. *Language and Cognitive Processes, 23*(7-8), 1057-1088.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20*(2), 148-161.

Gopen, G. D., & Swan, J. A. (1990). The science of scientific writing. *American Scientist, 78*(6), 550-558.

O'Conner, P. T. (2010). *Woe is I: The grammarphobe's guide to better English in plain English*. New York City, NY: Penguin Books.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pp. 163-173, July 8-14, 2012, Jelu Island, South Korea.

http://www.plos.org/

Corresponding author: Chenghzi Zhang