Understanding Academic Writing Style from the Perspective of Linguistic Closeness of the Speaker's Native Language to English

Chao Lu[1,2], Yi Bu[2], Ying Ding[2], Chengzhi Zhang[1]
[1]School of Economics and Management, Nanjing University of Science and Technology, Nanjing, Jiangsu, China
[2]School of Informatics and Computing, Indiana University, Indiana, U.S.A.

Abstract
This poster examines the relationship between researchers' Linguistic Closeness of the native language to English (LCoE) and their writing styles. The "LCoE of articles" were defined according to the LCoE of their first and corresponding authors which are output by Ethnea, a system with an input of authors' full name and an output of their possible native languages. The mean and standard deviation of the sentence length as well as average proportion of different parts of speech are utilized as two major indicators. Preliminary results showed three features of academic writing on researchers with higher LCoE: (1) longer sentences; (2) the combination between long and short sentences; and (3) more proportion of verbs.
Keywords: Native speaker; Academic writing styles; Text mining; Natural language processing
Citation: Editor will add citation

Contact: luchaonjust@gmail.com
Acknowledge: This poster is supported by Chinese Scholarship Council (CSC).

1 Introduction

More non-native speakers have begun to participate in English-based academic activities. However, language becomes an obstacle to non-native speakers including academic writing. Articles written by non-native speakers are often required extra proof-reading due to unsatisfying writings. To a large extent, academic writing has become an unavoidable issue for many non-native speakers. Therefore, it will be beneficial to understand the differences in academic writing styles between native and non-native speakers, which can help non-native speakers improve their academic writing and reduce the language-related barriers in their scientific careers. This poster mines the relationship between linguistic closeness of the researchers' native language to English and their academic writing styles.

2 Assumptions

A researcher's LCoE indicates the linguistic closeness of his/her native language to English. Here we also employ LCoE into the article level. An article's LCoE is defined as an indicator reflecting its authors' LCoE.

Three assumptions are proposed:

A1: The leading authors, include the first authors (FAs) and the corresponding authors (CAs), have the major contributions to their articles compared with other co-authors (Coats, 2009).

A2: Regarding article *writing*, it is more likely for the FAs to contribute more than other co-authors.

A3: The linguistic closeness of the native language to English (LCoE) of a given article should *increase* if the LCoE of its FA is lower than the CA (if applicable); otherwise it not.

According to A1, an article's LCoE is mainly determined by its FA and CA's LCoE due to their main contributions. In the case of A3, the increase of LCoE of an article is because the potential guidance of the CA to the FA. Please be kindly reminded that LCoE of a given article will *not* decrease if the LCoE of its CA is lower than the FA due to A2.

3 Methodology

3.1 Data

We downloaded all of the full-text articles published on *PLoS ONE* between 2006 and 2015. Our dataset includes 113 thousand articles with 143 thousand authors (including CAs' information if applicable)[1]. Among them 20 thousand articles were randomly selected with an approximate number of 31 thousand authors.

[1] Author name disambiguation is not done because we pursue their mother languages rather than who they actually are. Therefore, the number of authors counted might be a little bit more.

3.2 English Native Speaker Identification

Using all of the authors' full names in the dataset, we employ Ethnea[2] to identify the mother languages of the authors. Founded by Dr. Vetle I. Torvik (2016) from the University of Illinois, Ethnea is a novel system applied to predict ethnicity – first language in this poster – based on the geo-temporal distribution of names of authors in PubMed, DBLP, MAG, ADS, NIH, NSF, and USPTO with a good performance. The output of Ethnea is a list of languages the author's mother language might be as well as their probabilities. The most-likely language is chosen as the author's first language. After this, 27 languages are detected based on the selected 31,009 authors in our dataset.

All of the authors are then divided into three types according to their first languages output by Ethnea: (1) English native speakers, (2) non-English native speakers but Indo-European language family native speakers (e.g., German, French), and (3) non-Indo-European-language-family native speakers (e.g., Chinese, Korean). Commonly the LCoE of (1) is higher than (2), and (2) higher than (3). We mark the first types of authors as 1.0 score, the second types 2.0, and the third types 3.0. Obviously the fewer score an author is marked, the higher LCoE he/she will have. Note that for bilingual native speakers, we still determine their mother languages according to their names by Ethnea.

3.3 LCoE of an Article Identification

LCoE of an article is decided by the scores of its leading authors. In this poster, the score of an article reflecting its LCoE will decrease by 0.5 if the score of its FA is larger than its CA according to A3. That is to say, LCoE will be improved due to better language ability of the CA. The relationship between the score of an article and their leading authors is shown as Table 1, in which **the less score an article is marked, the higher LCoE it will get**. Table 2 displays the number of articles among different types.

FA score	CA score	Article score	FA score	CA score	Article score
1.0	1.0	1.0	2.0	3.0	2.0
1.0	2.0	1.0	3.0	1.0	2.5 =(3.0-0.5)
1.0	3.0	1.0	3.0	2.0	2.5 =(3.0-0.5)
2.0	1.0	1.5 (=2.0-0.5)	3.0	3.0	3.0
2.0	2.0	2.0			

Table 1. The score of an article and their leading authors (Smaller score refers to higher LCoE).

Article Score	Number (%)	Article Score	Number (%)	Article Score	Number (%)
1.0	3728 (18.6%)	2.0	7917 (39.6%)	3.0	5911 (29.6%)
1.5	1003 (5.0%)	2.5	1441 (7.2%)		

Table 2. Descriptive statistics among different types of articles.

3.4 Writing Style Indicators

To quantify writing styles of the authors, two indicators are applied, (1) the mean/standard deviation of sentence length (measured by the average number of words per sentence) and (2) the proportions of different parts of speech. In (2), only content words are included and four parts of speech, i.e., nouns, verbs, adjectives, and adverbs, are categorized. Generally speaking, higher (1) refers higher fluency in English academic writing while higher (2) reflects the preference of using English words.

4 Results and Discussion

Figure 1 shows the average sentence length of the articles in different scores (LCoE), in which we can see that the articles in higher LCoE groups tend to contain longer sentences. Specifically, those with 1.0 scores use average two words more in one sentence compared with the articles marked with 3.0. Since LCoE of an article is determined by LCoE of its leading authors, it indicates that the authors with higher LCoE have a tendency to use longer sentences in their academic writing. Generally speaking, longer sentences indicate higher mastery of English usage.

On the other hand, Figure 2 shows the standard deviation of sentence length among different LCoE of articles. We can find that overall higher LCoE articles tend to use diverse lengths of sentences. Therefore, higher LCoE authors prefer to mix with long and short sentences in their papers.

[2] Ethnea is available here: http://abel.lis.illinois.edu/cgi-bin/ethnea/search.py

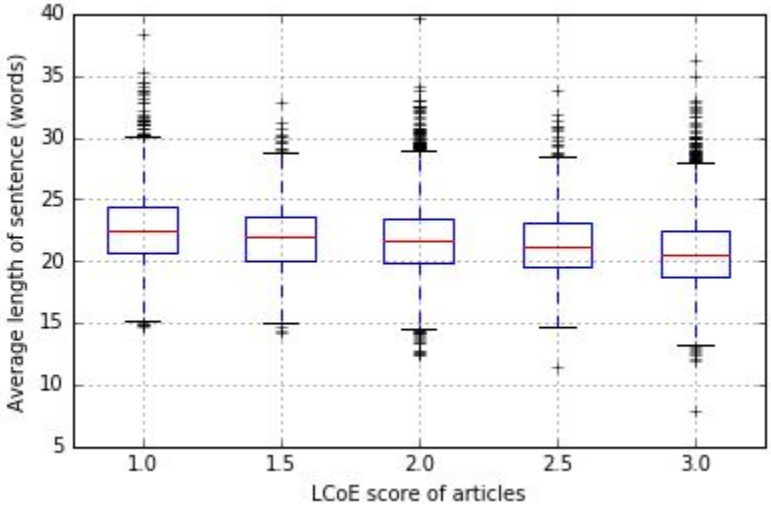

Figure 1. Average sentence length among different LCoE of articles.

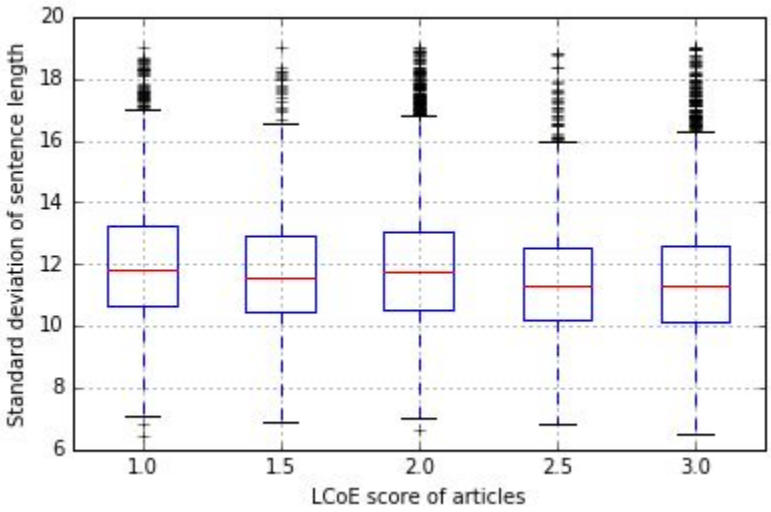

Figure 2. Standard Deviation of sentence length among different LCoE of articles.

Table 3 shows average proportion of different parts of speech among different LCoE of articles. We can find that higher LCoE articles tend to include more proportion of verbs than lower LCoE articles. In another word, higher LCoE authors prefer more fractions of verbs in their academic writing. As for adverb, there is a relatively unobvious trend that higher LCoE articles use more adverbs.

Score of articles	adjective	adverb	noun	verb
1.0	13.86%	5.08%	56.44%	24.62%
1.5	12.78%	4.82%	57.80%	24.60%
2.0	13.74%	4.95%	56.77%	24.54%
2.5	12.57%	4.69%	58.23%	24.51%
3.0	13.14%	4.62%	58.06%	24.18%

Table 3. Average proportion of different parts of speech among different LCoE of articles.

5 Conclusions

This study examined the relationship between linguistic closeness of the researchers' native language to English (LCoE) and their academic writing style. The "LCoE of articles" were defined as the LCoE of their leading authors originated from Ethnea. The mean and standard deviation of the sentence length as well as average proportion of different parts of speech are utilized as two indicators. Preliminary results showed three features of academic writing on researchers with higher LCoE: (1) longer sentences; (2) the combination between long and short sentences; and (3) more proportion of verbs.

In the future, we will explore researchers' writing style in a more detailed way. From the perspective of words, for example, function words (empty words) and the length of the words can be examined. Paragraph-level and semantic-level information can also be added. Moreover, the identification of articles' LCoE might be improved by specifying the relationship between distinct authors and their contributed articles.

6 References

Coats, A. J. S. (2009). Ethical authorship and publishing. *International Journal of Cardiology*, 131(2): 149-150.

Torvik, V. I., & Agarwal, S. (2016). Ethnea: An instance-based ethnicity classifier based on geo-coded author names in a large-scale bibliographic database. *International Symposium on Science of Science*, March 22-23, 2016, Washington D.C., U.S.A.