A Statistical Analysis of Text for Inferring Authenticity

Kun-Moo Rhee

Kyungju university ,Departmet of computer science
Hyohyen Dong 1
Kyungju city, Korea
rheekm@thrunet.com

Eunil Kim
Pukyung Natinal University,Department of English Literature & Language
599-1 Daeyeun Dong Nam Gu
Pusan , Korea
eikim@pknu.ac.kr

1. introduction

To find out real authors for literary works, the following methods have traditionally been used: (i) the comparison of the recorded texts with historical facts; (ii) the examination of hand-written texts if any; or (iii) the inference of the recorded year through chemical analyses of paper and ink. Neither of these is available in some cases such as old hand-written texts or literary works produced by computers. Therefore, we need new scientific methods of telling the authenticity of literary works. It has been suggested that one can use statistical analyses of grammatical structures such as the length of a sentence, the number of alphabets within a word, the frequency of parts of speech, and the frequency of idiomatic expressions. These statistical methods are based on the assumption that writers have their own peculiar stylistic characteristics just as their finger-prints. These statistical methods are seldom used in Korea, while they began to be used in Europe 100 years ago. This pilot study is to show how these statistical methods can be used in Korean literature by reviewing the previous studies.

2. A Sample Study

Mikhail Sholokhov(1905-1984)'s "And Quiet Flows The Don", a great epic that describes the Russian Revolution, began to be written in 1925. Its first volume was published in 1928 but the last volume was published in 1940. It took 15 years for him to finish the work. This work became sensational and he won the nobel prize in 1965. There began a suspicion that Sholokhov might not be the real author of "And Quiet Flows The Don", because of the following reasons: (i) He is not a native Cossack; (ii) he was too young(21) to describe very detailed lives of Cossack people; and (iii) he has no other work except this.

A book that argues that "And Quiet Flows The Don" was plagiarized was published 9 years after he won the nobel prize in Paris. The book argues that 95% of Volumes 1 & 2, and 68-70% of Volumes 3 & 4 were written by Fyodor Kryukov, a native Cossack who died in 1920', and that Sholokhov added some touches on them. The book was written by an anonymous author D* but prefaced by Russian Nobel Prize Winner Alexander Solzhenitsyn, who also stands up for D*. To verify D*'s argument, G. Kjetsaa and his Norwegian and Swiss coworkers analyzed 150,000 word length of Sholokhov's and Kryukov's sentences and compare them with the sentences in "And Quiet Flows The Don": They compared the length of a sentence, the length of a word, and the frequency of a word.

The results show that "And Quiet Flows The Don" is closer to Sholokhov's than Kryukov's. The statistical results are as follows: The average number of words in a sentence is 12.9 in Sholokhov's, 13.9 in Kryukov's and 12.4 in "And Quiet Flows The Don". The difference between Sholokhov's and "And Quiet Flows The Don" is 0.5 while the difference between Kryukov's and "And Quiet Flows The Don" is 1.5. The average number of alphabets in a word also shows a similar difference: The difference between Sholokhov's and "And Quiet Flows The Don" is 0.2 while the difference between Kryukov's and "And Quiet Flows The Don" is 0.9. The ratio of the number of different words per the number of the whole words, which is shown in Table 1, can be used as an index that shows the richness of the author's vocabulary.

<Table 1> Ratio of Number of Different Words per Number of Whole Words in "And Quiet Flows The Don", Sholokhov's and Kryukov's.

	Ratio (%)
Sholokhov's Work #1	48.3
Sholokhov's Work #2	48.8
Average	48.6
Kryukov's Work #1	43.2
Kryukov's Work #2	46.9
Average	45.1
Quiet Don Vol. 1	49.3
Quiet Don Vol. 2	51.5
Quiet Don Vol. 3	50.7
Average	50.5

Table 1 shows that "And Quiet Flows The Don" has rich vocabulary and it is closer to Sholokhov's than Kryukov's in that the difference in ratio between "And Quiet Flows The Don" and Sholokhov's is 1.5 while the difference between "And Quiet Flows The Don" and Kryukov's is 5.4.

Kjetsaa and his coworkers' statistical analysis such as t-verification and chi-square verification leads to the conclusion that Sholokhov is the real author of "And Quiet Flows The Don".

3. Conclusions

To suggest a quantitative method of judging the authenticity of the literary work, this study has reviewed the previous studies and reanalyzed the quantitative data on who's the real author of "And Quiet Flows The Don". This study will provide us with various types of empirical authority on the gigantic discourse such as plagiarism, ancient Korean history and the identity of the Korean people, which have continually been debated among Korean scholars of humanities.

Reference

R. D. Lord, Studies in the history of probability and statistics VIII. De Morgan and the statistical study of literature style. ***Biometrika,45,*** 1958

S. E De Morgan, Memoir of Augustus de Morgan by his wife Sophie Elizabeth De Morgan with selection from his letters. Longman,Green and Co.,1882.

G. U. Yule, The statistical study of Literacy Vocabulary . Cambridge University press, 1944

L. L. Adams and A. C. Rencher, The popular critical view of the Isaiah problem in light of statistical style analysis. ***Computer Studies in the Humanities and Verbal Behavior*** . 7(3-4),1973.

Stremja Tichogo Dona, Zakadki romana. YMCA-prews, Paris,1974.

RESUME

First Author

KunMoo Rhee

1998-Present	Kyungju University (Professor of computer Science Department)
1999	Keimyung University (Phd in Computer Education)
1991	Keimyung University (MA in Computer Science)
1984	University of KeiMyung (MA in Education)
1982	Keimyung University (BA in Education)

Second Author

Eunil Kim

1993-Present	Pukyong National University (Associate Professor)
1992	University of Colorado (Ph.D. in Linguistics)
1988	University of Oregon (MA in Linguistics)
1985	Hankuk University of Foreign Studies (MA in English Linguistics)
1983	Keimyung University (BA in English Language and Literature)