動詞の長さの分布に基づいた文章の分類と和語及び合成語の比率

金　　明　哲[†]

文章 (テキスト) の執筆者の推定問題などに対して，文章の内容や成立に関する歴史的事実の考証とは別に，文章から著者の文体の計量的な特徴を抽出し，その統計分析によって問題解決を試みる研究が多くの人々の注目をあつめつつある．文章に関するどのような要素に著者の特徴が現れるかについて，欧米文に関してはいくつかの研究の例があるが，それは言語によって異なるとも考えられるため，欧米文に関する研究成果が日本文の場合にもあてはまるかについて実証的な研究が必要である．また，各言語はその言語における著者の特徴を表す独特な要素があることも考えられる．本論文では，今まで明らかにされていない，日本文における動詞の長さの分布に著者の特徴が現れることと，その結果が動詞中の漢語・和語，合成語・非合成語の使用率の影響ではないことを著者 3 人の計 21 の文章を用いて明らかにした．計量分析の手法としては，同一著者の文章における動詞の長さの分布間の距離の平均値と，異なる著者の文章における動詞の長さの分布の距離の平均値との差，および距離マトリックスを用いて主成分分析を行うという方法を用いて数量・視覚的文章の分類を試みた．

キーワード: 文章の分類, 動詞の長さの分布, 和語・漢語, 合成語・非合成語, 相関係数, 主成分分析

Classification of Sentences by Word-length Distribution of Verbs with Proportions of Japanese Words and those of Compound Words

Jin Ming-Zhe

For the identification of a unknown author of a certain literature, it is significant to find the main elements of writing styles which characterize authors or distinguish them. It was previously known that the authors' writing styles are characterized in distributions fo lengths of words are categorized by parts of speech. This paper presents a quantitative analysis of the classification of sentences by word-length distribution of verbs, and the relationships between writing styles in word-length distribution of verbs and the proportions of Japanese words and Chinese words or those of compound words and non-compound words in verbs. This analysis shows that word-length distribution of verbs characterizes writing styles even where there are no differences in the proportions of Japanese words and Chinese words or those of compound words and non-compound words.

KeyWords: *classification of sentences, verb-length distribution, Japanese words and Chinese words, compound words and non-compound words, correlation coefficient, principal component analysis*

† 国立国語研究所 特別研究員, National Language Institute, The JSPS Postdoctoral Fellowship