読点の情報に基づく文献の分類

4B−7

金　明哲
総合研究大学院大学

村上　征勝
総合研究大学院大学
及び統計数理研究所

概要

執筆者の文章の特徴がどの部分に現れるかを知ることは、文献の真偽の判定や執筆者の推定を行う際の鍵となる。Jin and Murakami(1993)では、四人の作家の文章について読点がどの文字の後に付けられているかを分析し、読点の位置の情報だけで、四人の作家の作品が分類できることを報告した。本研究では、他のジャンルの文章においても、このようなことが言えるかを検証するため、三人の研究者の33編の科学論文について分析を試みた。その結果、科学技術論文においても、読点の位置は執筆者によって異なり、したがって執筆者の文章の一つの特徴と見なせ、執筆者の推定に有効な情報を与えることが分かった。

1．はじめに

文献の執筆者の推定や真偽判定などの執筆者推定問題(authorship problem)や、執筆年の同定、執筆順序の推定などの執筆時期推定問題(chronology)などの研究方法には二種類ある。一つは文献の内容や、成立に関する歴史的事実の考察を主とした方法で、従来の研究はこの方法が中心であった。しかし、近年コンピュータの進歩や統計分析法の発展にともない、文献の数量的な性質（特徴）に注目した、統計分析を主とした研究法が用いられるようになった(伊藤、村上、1992)。文献の数量的な性質としては、単語の出現頻度、延べ語数に対する異なり語数の比率、平均単語長、平均文長、品詞の出現頻度などの文の構造や文法に関する情報や、ある特殊な言葉の出現頻度などの言葉に関する情報が用いられてきた。このような統計的手法を中心とした研究が行われるようになったのは、欧米では今世紀に入ってからであり、日本語の文献については20世紀の中頃からである。ところで、日本語の文献は文章が英文のように「分かち書き」されていないために、単語の認識が困難であり、これが日本語の文章を扱う際の最大の難点となっている。また、漢字、仮名など、文字の種類が多いため、コンピュータでの処理が複雑になるうえに、漢字をどのように読むかというような問題もある。このようなことが原因で日本語の文献の計量分析に関する研究は、欧米に比べかなり遅れている。本研究では、これまで研究されていなかった日本語の文章における「読点」の位置に注目し、執筆者によって「読点」の付け方に違いがあるかどうか、つまり読点の付け方が執筆者の文章の特徴と見なせるか否かを、科学技術論文について計量分析した。

2．分析に用いた資料

今回分析したのは、次の三人の研究者の合わせて33編の論文である。この中でY11だけは『行動計量学会論文誌』に掲載されたもので、これ以外は雑誌『数理科学』に掲載されたものである。

	論文の記号	発表の年
今井　功		
「電磁気学を考える」の連載	I1〜I12	1985〜1986
佐藤　文隆		
「相対論」の連載	S1〜S10	1978〜1979
安本　美典		
「日本語の誕生」の連載	Y1〜Y8	1972
「語彙の量の構造」	Y9	1977
「計量文体小史」	Y10	1981
「比較言語学における計量研究」	Y11	1974

本研究では一回の連載を一編の論文として扱った。例えば、今井の「電磁気学を考える」では12回連載した論文を12編の論文として扱った。

3　データの解析

本研究では読点の前の文字の出現頻度を調べ、出現頻度の多い文字(と、て、は、が、で、に、ら、も、し、を、り、の、く、時、か、ば、た、い、後、ず、れ、

Characteristic styles of writing as seen through the use of commas

Mingzhe Jin: The Graduate University for Advanced Studies
Masakatsu Murakami: The Graduate University for Advanced Studies and The Institute of Statistical Mathematics

き、る、え、う、)とそれ以外の文字の出現頻度を1組にした計26の変数を用いた。分析では、まず主成分分析によって、26変数から少数個の主成分を抽出し、それを用いて判別分析を試みた。分析の結果、第四主成分までを用いると、見かけの的中率が100%で、33編の論文が全て正しく判別されている。第4主成分までの累積寄与率は83.08%である。図1は第1判別得点と第2判別得点をプロットしたものである。図の中の「＋」記号は判別空間における各群の重心である。このように科学技術文章の場合でも、執筆者によって、読点の付け方に癖が見られることが分かった。

4．終わりに

本論文では、文章の中の読点がどの文字の後に付けられているかの情報を、主成分分析、判別分析で分析し、科学技術論文が著者別に分類できるかを試みた。その結果、科学技術論文においても文章の中の読点に関する情報は執筆者により異なり、しかも同じ執筆者のものは比較的安定しているということが分かった。この方法は非常に簡単で誰でもが計量分析を繰り返し試みることができるという長所がある。

研究の次の段階として、読点の前にある語が文法的に見てどのような品詞の語であるかを調べ、執筆者の読点の打ちの特徴を文の構造に関係づけて分析する必要がある。

5．謝辞

執筆者の一人金に研究助成金を交付して下さった足銀国際交流財団に深謝する。

参考文献

伊藤　瑞叡、村上征勝(1992). 三大秘法棄承事の計量文献学的新研究、大崎学報、第148号、pp.1-52.

Mingzhe Jin and Masakatsu Murakami(1993). Authors' characteristic writing styles as seen through their use of commas, Behaviormetrika, Vol.20, No.1 に発表予定。

金　明哲、樺島　忠夫、村上　征勝(1993). 読点と書き手の個性、計量国語学、投稿中。

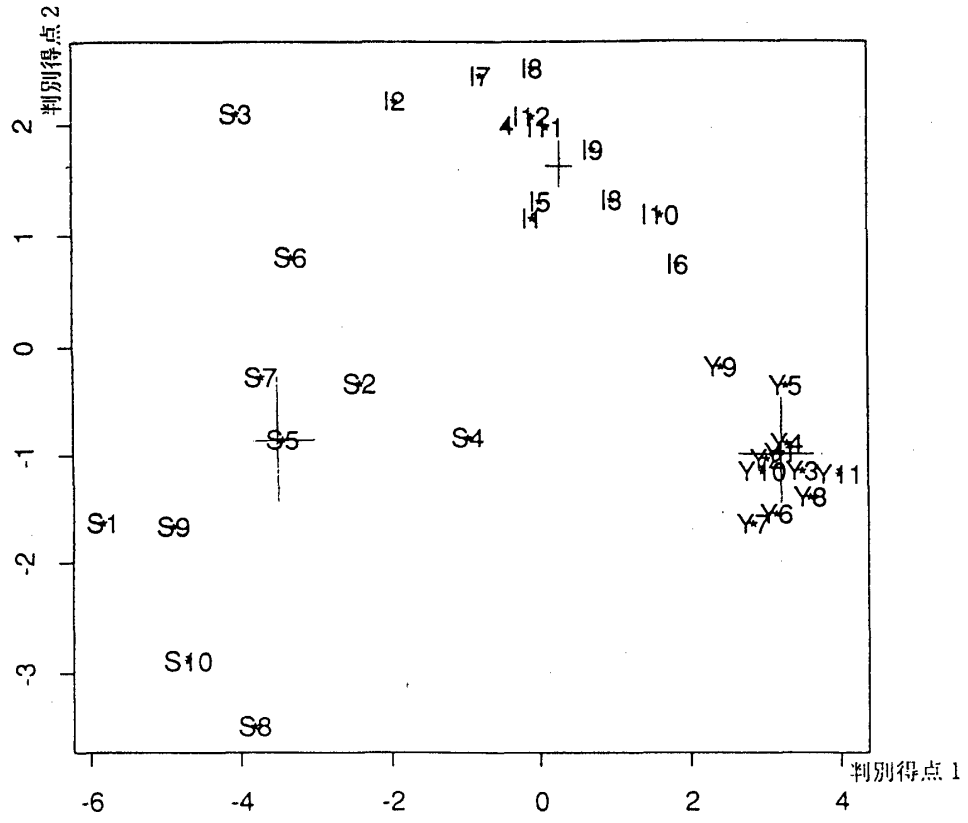

図1　正準判別分析の判別得点のプロット