統計数理（1996）
第44巻 第1号 121-125

公開講演会要旨

読点から現代作家のクセを検証する

札幌学院大学* 金 明哲

（1995年11月1日，統計数理研究所 講堂）

1. はじめに

ある文学作品は何という作家が書いたか？その作品はいつ書かれたものなのか？そうした問題を作品の中の文の長さ，単語の長さ，品詞の使用頻度など文体の特徴について統計分析を行ない，問題の解決を試みる研究が注目されている．文体の特徴の統計分析により，文章の執筆者の推定や執筆時期の推定を行なう問題では，文章に関するどのような要素を用いるかが問題解決の鍵である．

日本文に関しては，源氏物語の宇治十帖の著者の推定に関する研究（安本（1958)），「由良物語」の著者の判別に関する研究（韮沢(1965)），日蓮遺文の真偽に関する研究（村上(1994)），「源氏物語」の言葉の計量分析（村上(1996)）など文章の著者の推定・判別を試みる研究はいくつかあるが，文章に関するどのような要素が文章の著者の推定・判別に有効となるかに関する基礎的な研究はほとんどない状況である．文章における著者の特徴に関する情報は，言語によって異なると考えられる．したがって，外国語での研究成果（Holmes (1994)）が日本語の場合にもあてはまるのか，もしあてはまらないとすれば日本語の文章ではどのような要素に著者の特徴が現れるかというようなことが研究の課題の一つである．本稿では，このような基礎研究の一環である，現代文における作家の特徴に関する研究結果の一部を紹介する．

著者の文体の特徴は，文章を書くとき頻繁に使われている要素に現われやすい．どのような要素が多く使われ，かつ書き手の個性を表しているであろうか？

我々は日本語の文章を書くとき，文の切れ目や文の続きを明確にするため，意味の切れ目と思われるところによく読点を付ける．しかし，読点の付け方には明確な基準がない．どこが意味の切れ目であるかに関しては，人によって異なる場合もある．よく観察してみると，多くの書き手は読点を付けるとき，言葉の選び方と同様に神経を使う．そこで，読点の付け方に注目し分析をすることにした．

2. 計量分析の方法

2.1 読点の付け方に関する計量の方法

読点の付け方をどのように計量するかが問題であるが，本研究では，まず読点をどの品詞の後に付けるか，読点を付ける間隔，読点をどの文字の後に付けるかについて比較分析を行なった．その結果，読点をどの文字の後にどのぐらいの比率で付けるかに関する分析データに著者の特徴が最も明確に現われることがわかった（金(1994b)）．

* 社会情報学部：〒069 北海道江別市文京台11.

読点をどの文字の後にどのぐらいの比率で付けるかに関するデータは，比率が割に高い25項目を選択し，それ以外のものは「その他」の項目でまとめ，計26項目にした（Jin and Murakami (1993)，金 他 (1993)，金 (1994b)）．データを抽出する際には，抽出されたデータに著者の特徴が最大限に含まれるようにするため，会話文と並立した要素に付ける読点は除いた．

2.2 計量分析の方法

文章から抽出したデータに著者の特徴が現われるか否かについては，文章を著者毎に分類する視点から分析・評価を行なった(金(1994b))．

文章の分類は，文章の間の距離を求め，距離のマトリックスを用いて行なった．

いま，文章 i から抽出したデータの第 j 番目の項目の使用頻度を x_{ij} で表すと，n 編の文章における m 項目に分けたデータは

$$P_{n\times m}=[p_{ij}]$$

$$p_{ij}=\frac{x_{ij}}{\sum_{v=1}^{m}x_{iv}},\qquad \sum_{j=1}^{m}p_{ij}=1$$

で表記できる．

さて，文章 i と文章 l との間の距離 d_{il} を次のように定義する（以下 SIR 距離と呼ぶ）．

$$d_{il}=\frac{1}{2}\sum_{j=1}^{m}\left(p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{lj}}+p_{lj}\log\frac{2p_{lj}}{p_{ij}+p_{lj}}\right)$$

ただし

$$p_{ij}=0 \quad \text{なら} \quad p_{ij}\log\frac{2p_{ij}}{p_{ij}+p_{lj}}=0$$

$$p_{lj}=0 \quad \text{なら} \quad p_{lj}\log\frac{2p_{lj}}{p_{ij}+p_{lj}}=0$$

とする．また上式で求められた分布間の距離マトリックスを

$$D_{n\times n}=\begin{bmatrix}0 & d_{ji}\\ d_{ij} & 0\end{bmatrix}$$

で表記する．

3. 読点の付け方のクセ

まず，井上靖，三島由紀夫，中島敦の3人の作品について，読点をどの文字の後にどのぐらいの比率で付けているかを求め，文章の分類を行なった．その結果，文章が著者毎に分類されることがわかった．さらに，この現象が普遍性を持っているか否かを実証するため，いくつかの観点から統計分析を行なった．

3.1 読点の付け方に変化が大きい谷崎潤一郎

谷崎潤一郎は文体が大きく変化していった作家と言われている．谷崎潤一郎は文章を書くとき，句読点の付け方で文体を変えようと試みていた．例えば，谷崎の「痴人の愛」,「吉野葛」,「青春物

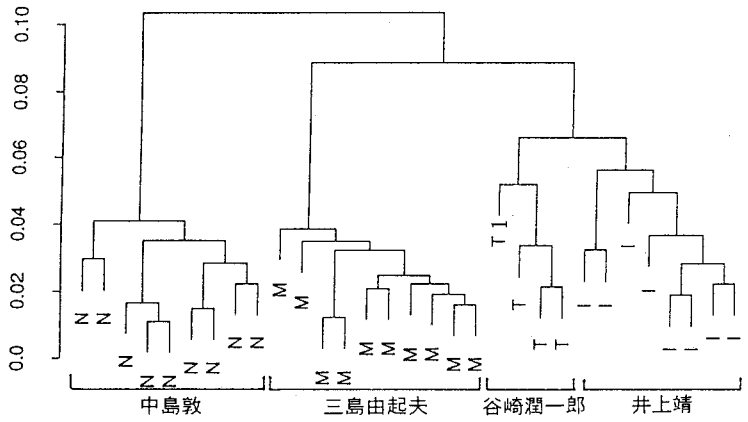

図1．井上，谷崎，中島，三島の作品の分類の樹形図.

語」では平均約24～27文字置きに一つの読点を付けているが，「春琴抄」の場合は平均約195文字置きに一つの読点を付けている．読点を付ける間隔から見ると，「春琴抄」の文体は谷崎潤一郎の他の作品とは異なると考えられる．しかし，読点をどの文字の後にどのぐらいの比率で付けているかに関するデータを用いて分析すると，疑問なしに谷崎潤一郎の作品であると判定される．視覚的に考察を行なうため，文章が著者毎に分類されるか否かを見ることにする．分類を視覚化するクラスタ分析の方法としては，階層的クラスターリングの方法，2または3次元での散布図がよく用いられている．本稿では，紙面上の都合により，階層的クラスターリングの樹形図だけを示す．樹形図は，まず*ISR*距離のマトリックスを二重中心化し，そのマトリックスの分散共分散の行列の主成分得点を用いて，群平均法により生成した．

図1に井上靖（8編，Iで表記する.），谷崎潤一郎（4編，Tで表記する.），中島敦（9編，Nで表記する.），三島由紀夫（10編，Mで表記する.）の計31編の文章の階層的クラスターリングの樹形図を示す．図では，谷崎の作品「春琴抄」（T1）は谷崎の独自の枝・グループに帰属する．つまり，谷崎潤一郎が読点の付け方で文体を変えるために読点を付ける間隔を変えていても，読点をどの文字の後に付けるかに関しては，依然谷崎のクセが残されていると言えるであろう．

3.2　一人三人作家

昭和初期の流行作家，一人三人作家と言われている長谷川海太郎は一人で林不忘・谷譲次・牧逸馬の三役をこなし，三つのペンネームで多くの作品を書いた．長谷川海太郎は明治三十年生まれ，大正十三年アメリカから帰国し，作家活動を始めた．仕事をしすぎたせいか35才の若さで亡くなった．その代表的な作品として林不忘のペンネームでは「丹下左膳」，谷譲次のペンネームでは「テキサス無宿」，牧逸馬のペンネームでは「運命のＳＯＳ」が挙げられる．三つのペンネームで書いた文体はペンネーム毎に異なりを見せていると言われている．しかし，読点の付け方から見た場合でも，ペンネーム毎に文体が変わっていると言えるであろうか．前述の林（2編，HAで表記する.），谷（3編，TAで表記する.），牧（2編，MAで表記する.）の作品と前節の井上（8編），中島（9編），三島（10編）の計34編の文章について，読点をどのような文字の後にどのぐらいの比率で付けているかを求め，文章の分類を行なう．図2に上記の34編の作品の階層的クラスターリングの樹形図を示す．図でわかるように，文章がそれぞれの著者毎に四つの枝・グループにきれいに分れる．長谷川海太郎の作品は一つ独立の枝・グループを形成するが，ペンネーム

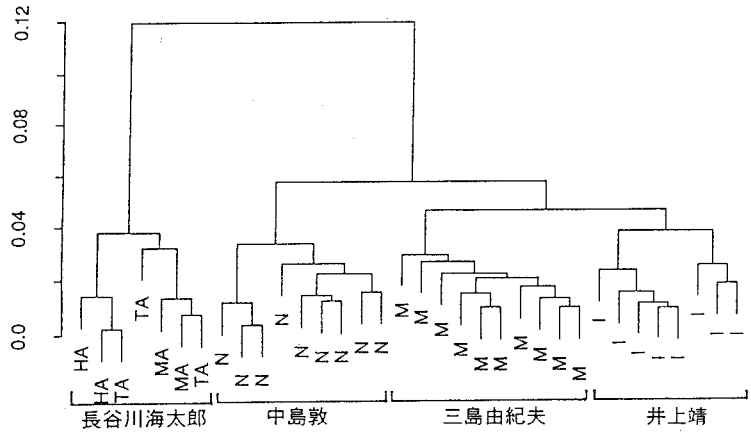

図2．井上，中島，長谷川，三島の作品の分類の樹形図.

毎に枝・グループを作らない．彼が作品を書くとき，ペンネーム別に意識的に文体を変えることがあったとしても，読点の付け方までは変えることができず，個人のクセがそのまま現われたものであろう．

3.3 研究論文

前節までは，文学作品についての統計分析であった．本節では，文章のジャンルを変え，学術論文においても同様な現象が見られるかについての統計分析を行なう．統計分析の対象としては，「数理科学」に連載された今井功（12回，Iで表記する.），佐藤文隆（10回，Sで表記する.），安本美典（8回，Yで表記する.）の文章を主とし，一回毎の連載をそれぞれ一編の論文として扱う．安本の場合は「数理科学」の連載が8回しかなかったため，他の論文誌に掲載された論文3編を加え11編を用いる．3人の計33編の文章について，読点をどの文字の後にどのぐらいの比率で付けているかを求め，文章の分類を行なう．図3に33編の論文の階層的クラスターリングの樹形図を示す．図でわかるように，論文が著者毎にきれいに三つの枝・グループに分れる．つまり，研究論

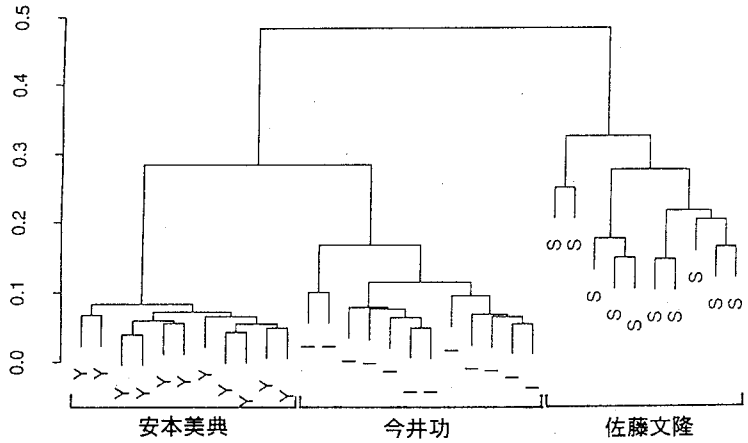

図3．3人の論文の分類の樹形図.

文の場合でも，読点をどの文字の後にどのぐらいの比率で付けるかに関するデータに著者のクセが明確に現われている．

4. おわりに

本研究では，今まで研究されていない読点の付け方と文章の著者（書き手）との関係について統計分析を行なった．この統計分析では文体の変化が大きいと言われている現代作家の文学作品だけではなく，研究論文の文章についても分析を行なった．読点をどの文字の後にどのぐらいの比率で付けるかに関するデータには，文学作品はもちろんのこと，研究論文でも著者の特徴が明確に現われることが分かった．

また，一方で文体の計量分析で，著者の特徴を表す情報としてよく用いられている文の長さ，単語の長さ（金(1995)），品詞の使用率，漢字・カナの使用率などとの比較分析も行なうと読点の付け方に著者の特徴が最も明確に現われるという結果が得られた（金(1994a)）．

読点をどの文字の後に付けるかに関するデータの抽出は，読点の前の文字をカウントするだけであるため，簡単に統計分析を試みることができるという利点があり，今後も多様なジャンルの文章の分析に活用できるであろう．

謝　辞

本研究に用いたデータベースの一部は，文部省統計数理研究所の村上征勝教授の研究費で作成したものである．本研究をご支援及びご指導くださった村上先生，公開講演および本稿の執筆の機会を与えてくださった方々に，心より感謝致します．

参　考　文　献

Holmes, D.I. (1994). Authorship attribution, *Computers and Humanities*, **28**, 87-106.
金　明哲（1994a）．自然言語におけるパターンに関する計量的研究，総合研究大学院大学統計科学専攻博士学位論文．
金　明哲（1994b）．読点の打ち方と文章の分類，計量国語学，**19**, 318-329.
金　明哲（1995）．単語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，自然言語処理，**2**, 58-75.
Jin, M.Z. and Murakami, M. (1993). *Authors' characteristic writing styles as seen through their use of commas, Behaviormetrika*, **20**, 63-76.
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，**18**, 382-391.
村上征勝（1994）．計量的文体研究の威力と成果，言語，**23**(2), 30-37.
村上征勝（1996）．「源氏物語」の言葉を分析する，統計数理，**44**, 127-131.
韮沢　正（1965）．由良物語の著者の統計的判別，計量国語学，No. 33, 21-28.
安本美典（1958）．著者の推定　――源氏物語，宇治十帖の著者について――，心理学評論，2(1), 147-156.