氏　　名	小野 洋平
学位(専攻分野)	博士(学術)
学 位 記 番 号	総研大甲第 1923 号
学位授与の日付	平成29年3月24日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	言語データにおける統計的分類手法の適用に関する探索的研究
論 文 審 査 委 員	主　査　准教授 前田 忠彦 教授 吉野 諒三 准教授 足立 淳 名誉教授 村上 征勝 統計数理研究所 名誉教授 林 文 東洋英和女学院大学

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

言語データの統計解析においては，品詞や語順などに関する計量データに基づく各文学作品や各言語の分類を主要な目的とし，研究が行われる。この目的のために,「数量化Ⅲ類」などの多次元データ解析法や各種のクラスタリング手法が用いられることが多い。そのような解析を遂行する中で，研究者はしばしば，作品や言語のクラスタリング，あるいは品詞や語順などの変数のクラスタリングにおいて，対象間の距離の定義，多様なクラスタリング手法の中からより適切な手法の適用などについて,様々な選択に迫られるはずである。しかしながら，言語学における従来の計量的研究では，これらの点について，十分な注意が払われてきたとは言い難いようである。この点に鑑みて，本論文においては，より望ましい解析方針を模索することとした。

まず，1 章では上記の研究動機を説明し，言語データに関する統計的アプローチの研究史を述べ本論文の目的を位置付けた。2 章では，本論文の 3 章以降で用いる統計手法について概観した。具体的には，クラスター分析，数量化Ⅲ類（多重対応分析),「数量化Ⅲ類クラスタリング」, MCANeighbor-Net などである。数量化Ⅲ類クラスタリングとは，多変量質的データの次元圧縮の方法である数量化Ⅲ類によりカテゴリカルな変数に適切な尺度値を与えた上で対象間の距離行列を算出し，階層的なクラスタリングの手法を適用する方法一般を指す。また MCANeighbor-Net とは，数量化Ⅲ類クラスタリングの一種であるが，系統樹推定の方法として利用される Neighbor-Net と呼ばれるアルゴリズムを用いたクラスリングを利用する手法のことを指す。

つづいて，3 章や 4 章，5 章では実データに基づいて，先述した様々な分析プロセスにおける選択の実例を紹介し，本論文での工夫を述べた。

3 章では源氏物語の文体研究に関する記念碑的研究である村上· 今西(1999)論文を統計科学的に再検討した。解析対象のデータは 54 帖における 21 助動詞の相対出現頻度である。この先行研究が抱える潜在的な課題を確認した上で，本論文では主に 2 つの工夫を行なった。第一に，次元圧縮を伴うクラスター分析と通常のクラスター分析を含む様々な分析手法を比較すること，第二に，データの事前処理として適切な変数変換を施すことである。様々な方法や条件の組み合わせの中で，最も妥当な結果を与えたのは平方根変換を施したデータにクラスター分析を適用する条件であった。この結果得られた樹形図は，村上· 今西(1999)で推測された源氏物語の成立論とは矛盾しないが，それより少ない高々2 グループの成立順序を示唆するものであった。本章ではさらに，Neighbor-Net Analysis を用い，従来の文献学的知見と矛盾する分類となった個体（帖）について考察した。

4 章では，言語類型論のデータに対して統計的な手法を適用した先駆的な研究である Tsunoda, Ueda and Itoh (1995a)論文を統計科学的に再検討した。彼らが分析した 129 の言語の 19 の語順に関する 87 の変数のデータについて尺度水準の観点から考察を行い，彼らの研究では明らかにならなかった構造を発見することを期待し，より適切と思われる方法である多重対応分析を適用した。分析結果は，Tsunoda et al. (1995a)の知見の再考が必要であることを示唆するとともに，言語類型論の観点からも新たに興味深い知見を得た。

さらに,「数量化Ⅲ類クラスタリング」を適用することによって，これらの知見を明瞭なものとした。

5 章では，アイヌ語方言に関する基礎語彙統計学的研究の記念碑といえる服部・知里(1960)論文を統計科学的に再検討した。服部・知里(1960)が扱ったのは 19 地域の方言間の基礎語彙の同根性を判断した 2 相 3 元のデータであるが，本章では，彼らの研究やそれを再分析した後続の研究は，いずれもそのデータの「方言間の単語の同根性がわからない」という情報を無視している点に注目した。すなわち，その情報を無視できない 1 つのカテゴリーとして多重対応分析を適用し，採用した次元の座標から距離行列を求めクラスター分析を適用することで，樺太の 6 つの方言について言語地理学的な観点からより妥当な結果を得た。同様に，MCANeighbor-Net の適用によって，アイヌ語学において今まで文献学的な言及はあったものの，統計学的には示されていなかった様々な言語地理学的パターンを考察できる可能性を示唆した。

最後に 6 章において，本論文の成果を方法面と実質科学的貢献という 2 つの観点からまとめた。はじめに，1 章で述べた目的に関する各種統計分析における手法の選択に関する検討という目的に合わせて本論文の提案についてフローチャートの形で述べた。この提案の中には，(1)分析目的の確認(2)データの予備的検討(3)多重対応分析や主成分分析による変数と個体の同時分類の考察(4)クラスター分析の適用にあたっての様々な注意点（手法の選択や結果の評価等の内容）を含んでいる。分析の開始前に現実のデータ生成過程の性格を充分に考慮し，その過程に合わせた手法選択を行うことの重要性を強調した。次に，本論文を通じて明らかになった計量文献学，言語類型論及びアイヌ語学における成果をまとめた。さらに，これらの分析を通じて示唆される統計的手法の実質科学分野への貢献のあるべき姿，いわば分析上の哲学について「データの科学」の観点から私見を述べた。

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

本博士論文は，言語データの統計解析において利用される統計的分類手法について扱ったものである。解析プロセスの中で分析者はさまざまな場面で適切な手法の選択に迫られるが，本論文においては，実データの解析を通じた探索的検討を通じて有力な選択指針を提言し，同時に実質科学的な発見を得ることを目指した。

1，2 章が導入的部分で，3 章～5 章で実データ解析に基づくさまざまな工夫を述べている。3 章は源氏物語の 54 帖を分類する問題，4 章では言語類型論分野での言語の側と語順に関する変数の側の双方を同時に分類する問題，5 章ではアイヌ語方言に関する基礎語彙統計学的な先行研究データを扱い，19 地域の方言を分類する問題，をそれぞれ考察した。これらの章を通じて主に活用されている方法は数量化Ⅲ類クラスタリングと呼ばれるものである。それらの結果に基づき 6 章で，本論文の成果を方法面と実質科学的貢献という 2 つの観点からまとめている。

本論文は，言語関連分野で適用される統計的分類手法におけるさまざま場面での手法の選択について，実例の分析に基づき推奨される工夫を取りまとめたもので，これらの工夫は今後の同分野での計量分析を考える際の指針として有用といえる点が本論文の大きな成果である。また 3 章～5 章の実データ解析の結果は，実質科学的な観点からも新しい発見に相当する。

既発表の成果についても 3 章と 5 章がそれぞれ査読付き学術誌に研究ノートと論文として採択された他，第 4 章に密接に関連し，より広範囲のデータを扱った英文論文が，査読を経て書籍の 1 章として採択されており，専攻内基準を満たしている。学力確認の結果についても，博士（学術）を授与するに十分な学識を有するものと判断される。

以上のことから，5 名よりなる審査委員会は全会一致で，本論文が統計科学の関連分野への十分な貢献をなし，博士（学術）に値するものと判断した。