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

小野　洋平

博士(学術)

総合研究大学院大学
複合科学研究科
統計科学専攻

平成 28 年度

目次

第 1 章 研究目的

1.1 研究動機

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(ウォード法、最長距離法、最短距離法など)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、言語学や文学等言語に関わる分野における従来の計量的研究では、この点について、十分な注意が払われてきたとは言い難いようである。これに鑑みて、本研究においては、より望ましい解析方針を模索することとした。本論で言語データと言っているものは典型的には 2 相 2 元の多変量データを指す。具体的には、3 章では源氏物語の 54 帖における 21 の助動詞の頻度データ、4 章では言語類型論における 129 言語の 19 の語順に関する 87 変数のデータを指す。ただし、5 章ではアイヌ語の 19 方言の語彙の同根性に関する 2 相 3 元のデータのことである。

1.2 研究史

1.2.1 国外における言語データに対する統計的アプローチ

まず、言語データに対する統計的アプローチの歴史を Oakes(1998)と Moisl(2015)を参考として振り返る。最初に、比較的歴史の長い計量文献学周辺の研究についてやや詳しく述べ、続いてコンピュータの発展により可能になった分野としてコーパス言語学の分野に言及する。最後に、それ以外についても言及するが、これらは言語データという語で括ることができるすべての分野を網羅したものでは勿論ない。

計量文献学とは、書簡、文学作品や様々な歴史的分析資料など、様々な文書資料に対して計量的な手法によってその特徴を明らかにし、何らかの文献学的な知見を明らかにしようとする試みである。Morton(1978)によれば、1851 年に De Morgan が記した手紙の中に、計量文献学の基本的な原理が示されているとされる。この 1851 年の De Morgan の手紙を初めとすれば、計量文献学は 150 年以上の歴史を持っている。

今日に至る計量文献学の研究のなかで、利用される変数の特徴と、分析手法は密接な関係が見られる。Oakes(1998)によれば、コンピュータによる計量文献学的研究においては次の 4 つの特徴量がよく利用される。

1. 語や文章の長さ
2. 文章における語の位置
3. どのような語彙がよく利用されるか

4. どのような文法的な特徴があるか

たとえば、Mendenhall(1887)の研究では、「語の長さの分布」を考えることにより、Shakespeare や Bacon といった著名な作家や思想家の文体を比較している。これは Shakespeare の素性が十分に知られておらず、この両者が同一人物であるという俗説があることから、一種の著者推定の問題を扱ったものと言える。また、Williams(1970)の報告によれば、Yule(1939)が「キリストにならいて(De Imitatione Christi)」という書物の著者推定問題を扱った際にも「語の長さ」を統計量として利用している。さらに Yule(1944)は「語の長さ」だけでなく、ある文章の中で扱われている「語彙の豊富さ」を表す K 統計量(K-characteristics)を発案し、「キリストにならいて」の著者推定問題について詳細な検討を行っている。

このような、文章の特徴を一つの分布や複数の特徴量に集約することで比較を容易にする方法は、Morton(1965)や Ellegård(1962)にも見られる。しかし、Yule(1944)などの初期の研究においては注目する特徴量は少数の場合が多いが、時代を経るにつれて複数のテキスト(作品)に関する多数の特徴量に注目することなっていく。これに伴いデータの形式は、個体(作品)×特徴量(品詞)もしくは特徴量(品詞)×個体(作品)の頻度表になり、「頻度表のデータを如何に分析していくか」ということが計量文献学における分析上の焦点の一つとなっていく。

こうした文脈の中で、頻度表のデータには判別分析や因子分析、主成分分析、クラスター分析など様々な多変量解析の手法が用いられていくことになる。

「複数のテキストに関する複数の特徴量についてコンピュータによって検索をかけ、得られた特徴量について統計的な処理を行う」という現在の計量文献学で一般的に用いられる手法の適用は、Austin(1966)にその萌芽がみられる。いわゆる多変量解析と呼ばれる手法の適用としては、複数の特徴量の組み合わせによって判別分析を行い、Shakespeare と Fletcher の作品を分類した Baillie(1974)の研究などが典型例の一つである。また、Mosteller and Wallace(1963)においては著者推定にベイズ統計の手法が使われていることも注目に値する。さらに、Baayen, van Halteren and Tweedie (1996)においては、単語のレベルの特徴量を離れ、文法的な観点から注目した特徴量について主成分分析を適用することで作品の分類を行い、軸の意味を解釈している点で計量文献学では先駆的である。

本論文の 3 章では源氏物語の各帖に関する多変量データを扱う。これは上記の計量文献学の歴史に連なるものと言える。

次に、コーパス言語学における適用について簡単に紹介しよう。ここでいうコーパス言語学とは「特定の研究目的のために集められたテキストの集合に関する研究」とする。Moisl(2015)によれば、コーパス言語学においても頻度表(個体[作品]×特徴量[品詞])の統計的分析がなされている。先駆的な研究としては Miller(1971)、Kiss(1973)などがあり、これらの研究では頻度表のデータにク

ラスター分析を適用している。しかし、当時の伝統的な記述言語学や、生成文法などにおいては、依然として話者の内省によって文が文法的に適正かどうかを判断することが行われていた。そのため、当初はコーパス言語学やコーパスデータにクラスター分析を適用する研究は、言語学の主流とはならなかった。しかし、コーパスデータを基に量的分析を行うという方法論は、近年言語学のなかで重要性が増し、特に心理言語学(Gilquin & Gries, 2009; McEnery & Hardie, 2012)や認知言語学(Gries & Stefanowitsch, 2007; Gries 2012; Stefanowitsch, 2010)などで用いられている。

また、Moisl(2015)によれば、量的な研究手法は、広く言語の歴史的変化や空間的変化、社会的変化を研究する有効な手段としても用いられてきた。言語の歴史的な変化に量的手法を適用した先駆的研究として、Kroeber and Chrétien(1937)や Kroeber and Chrétien(1939)があげられる。空間的な変化を扱ったものとしては Reed and Spicer(1952)が、社会的変化を扱ったものとしては Labov(1963)や Labov(1966)が、それぞれ先駆的な研究である。本論文の 4 章で扱う言語類型論のデータ、5 章で扱うアイヌ語方言のデータも、広い意味で、このような言語変化に関わるものと言える。

1.2.2 国内における言語データに対する統計的アプローチ

日本においては、文章の著者を推定する問題に関しては古くから歴史があったが、計量文献学的な研究はこの半世紀で進展し、安本(1958)の源氏物語の宇治十帖の著者推定に関する研究を嚆矢とした一連の研究、村上(1994)の日蓮遺文の真偽に関する研究などが代表的な研究としてある。

また、作品(日本文学)×特徴量の頻度表のデータに因子分析を適用し、日本文学の作品の類型化を試みた研究としては、安本·本多(1981)があげられる。同様に頻度表のデータにクラスター分析を適用した研究としては、Jin and Murakami(1993)や金·樺島·村上(1993)がある。彼らは、日本語の現代文の著者推定の問題において、読点の前にくる文字の比率が重要であることを指摘し、それらの比率から距離行列を作成し、クラスター分析を適用することによって高い精度で著者を推定することを可能にした。

さらに、作品(文章)×品詞(助詞, 助動詞)の頻度表のデータに判別分析を適用し、文章のジャンルの分類に成功した先行研究として村田(2000)が挙げられる。

このように、国内外の研究を俯瞰すると大きく「単変量」から「多変量」へと分析対象となるデータの形式が拡大していることがわかる。

1.2.3 数量化Ⅲ類の言語データへの適用に関する研究史

前述の通り、頻度表を様々なタイプのデータには各種の多変量解析の手法が用いられてきた。ここでは多変量解析の一つである数量化Ⅲ類やそれに関連する手法についての研究史を述べておく。

言語データを分析する際には、様々な種類の頻度データが存在し、その種類と分析目的に応じて分析手法も変わる。本研究では、主に個体(作品)×特徴量(品詞)の頻度表のデータと、アイテムカテゴリー型データ(岩坪, 1987)を扱う。

数量化Ⅲ類は様々な形で言語データに対して応用されてきた。海外において、代表的な雑誌の一つである「Literary & Linguistic Computing」において、創刊号から現在までの内容について数量化Ⅲ類と同等の手法である「Correspondence Analysis」で検索をかけると、1986 年 1 月から 1989 年 12 月までが 2 件、1990 年 1 月から 1999 年 12 月までが 11 件、2000 年 1 月から 2009 年 12 月までが 6 件、2010 年 1 月から現在までが 3 件が該当した。その中で特に論じられているものは共観福音書(The Synoptic Gospels)に関するものであり、Mealand(1995)や Linmans(1997)などの研究は典型的な個体×特徴量の頻度表データに対して数量化Ⅲ類(Correspondence Analysis)を適用している。これらは、コーパス言語学において数量化Ⅲ類を適用する研究の系統に属しているが、一方方言学において数量化Ⅲ類を適用した研究は Cichocki(1989)が嚆矢である。Cichocki はフランス語の方言について地域と方言の特徴量とデータに対して数量化Ⅲ類(双対尺度法)を適用している。Google scholar を検索した限りでは Cichocki(1989)を引用している論文は現在まで 5 件しか存在しないが、Google scholar において「"Correspondence Analysis" "dialectology"」と検索すると現在までに 96 件の研究があり、その中には Cichocki(2006)もある。地域×方言の特徴量というデータ形式は数量化Ⅲ類の適用に向いており、今後の研究の発展が期待される。

一方、日本国内においては、1990 年代以降コーパス言語学の中で数量化Ⅲ類が利用されてきた。例えば、統計数理研究所の共同利用研究制度の中でもコーパス言語学関係の課題が 2005 年度以来毎年度複数件実施されており、その成果の一部は統計数理研究所の共同研究リポート等に詳述されている[1]。一方、方言学においては方言の地理的変化について数量化Ⅲ類を用いて探った井上史雄の一連の研究が有名である(例えば、Inoue, 1988; 井上, 1990; Inoue, 1992; Inoue, 2009)。さらに、その流れを引き継いだ上田博人の研究も挙げられる（例えば、Ueda, 1993; 上田, 2013; Ueda & Perea, 2014)。

1.3「数量化Ⅲ類クラスタリング」をめぐる研究史

本研究では、頻度表(もしくはアイテムカテゴリー型データ)のデータに林の数量化Ⅲ類 (Hayashi's Quantification Method Ⅲ [Hayashi, 1952]; Correspondence Analysis または Multiple Correspondence Analysis[CA または MCA][Benzécri, 1973])を適用し、個体もしくは変数(特徴量など)に関して得られた多次元の座標に基づきに階層的クラスター分析を実行する方法（ここでは「数量化Ⅲ類クラスタリング」と呼ぶ）を言語学データに対して試行する。この方法の意味づけについては、2.4 節で改めて述べる。

データに対して MCA もしくは主成分分析(PCA)などを適用し、次元圧縮を行い、少数の次元の座標値に基づいて(非)階層クラスター分析を実行する手法

[1] 該当する研究課題は多数存在するので列挙することは避けるが、広い意味で言語における分野の統計分析に関わる課題については、統計数理研究所のホームページの「共同研究データベース」http://www.ism.ac.jp/kyodo/index_i.html で参照することができる。

は「タンデムクラスタリング」と称され(「数量化Ⅲ類クラスタリング」もその一つである)、有用性が主張される一方で、先行研究においてさまざまな問題点が指摘されている (Kiers & Vichi, 2010; Timmerman, Ceulemans, Yamamoto, 2012; Vichi & Kiers, 2001; Yamamoto & Hwang, 2014)。一つ目の批判点は理論的な観点から、「タンデムクラスタリング」においては、MCA もしくは PCA のステップと、(非)階層クラスタリングの部分が独立して実行されるため、二つの違う基準を別々に最適化していることになり、第一の手続きの基準で得られた解を所与とした第二の手法での最適解が、全体として見たときに最適の解を与えているとは限らないことである。そのため、近年においては、それらを同時に最適化する手法が提案されている(De Soete & Carroll, 1994; Gattone & Rocci, 2012; Vichi & Kiers, Yamamoto, 2012)。しかし、これらの同時最適化の手法は、MCA もしくは PCA と非階層クラスタリングを組み合わせる場合に提案された手法であり、MCA もしくは PCA と階層クラスタリングを同時最適化する手法は筆者の知る限り開発されていない。本研究のように、言語データにクラスター分析を適用しようとする時、個体もしくは特徴量のクラスター間の階層性を明らかにしたいという目的によることが多い。そのような目的のためには、前述のように提案されている同時最適化の手法は適さない。

二つ目の批判点としては、タンデムクラスタリングが、次元圧縮を伴わない(つまり全変数をそのまま利用した)クラスター分析の結果より常に解釈の容易な解を与えるとは限らないことである。(Arabie & Hubert, 1994; DeSarbo, Jedidi, Cool & Schendel, 1991)

「数量化Ⅲ類クラスタリング」を含むタンデムクラスタリングの応用は、例えば、マーケティング分野で、ブランドのポジショニングや消費者セグメントの発見のために多用される(例えば、Arabie & Hubert [1994]や菅[2001])。しかし、筆者が調べた限り言語データに対して、「数量化Ⅲ類クラスタリング」を適用した例はほとんど見られない。よって、本研究では、「数量化Ⅲ類クラスタリング」が「タンデムクラスタリング」が抱える問題と同様の批判点を持つことを念頭に置きながら、実際に言語データに「数量化Ⅲ類クラスタリング」を試行することにより、今後の言語データ解析への指針を探る。

1.4 本研究の目的

以上の研究史を見て気づく点は、言語学や文献学における個体×特徴量の頻度表データやカテゴリカルデータ(質的変数についての多変量データ)の分析にあたって、我々は様々な段階で手法の選択に迫られるが、言語データにおいてはこの問題に関する議論が不足していることである。ここで、迫られる選択とは、例えば、データの予備分析段階での検討、そもそもクラスター分析の適用自体が必要かどうか、クラスタリングが必要な場合次元圧縮を伴う手法(タンデムクラスタリング)を採るべきか否か、クラスタリングにおける具体的な分類手法や距離の選択、次元圧縮を伴う手法を用いる際の次元数の決定、クラスタリ

ング結果を如何に評価するか、などである。

本研究では、具体的な言語データでの試行と考察の繰り返しを通じてこれらの問題について何らかの指針を得ることを目的とする。

本研究で「数量化Ⅲ類クラスタリング」等の統計手法を言語データに試行することには、以下のような意義もある。

通常の言語学の研究においては、得られた質的データに対して、文献学や歴史学、考古学など様々な知識を総合し、より妥当性のある結論を導くが、質的データに対して統計学の手法(数量化Ⅲ類やクラスター分析など)を適用することによって、量的な視点を提供し、これまでの言語学の知見に「統計学から見た」結果を提示することができる。さらに、「統計学から見た」知見に対して、言語学の側において既存の知見に対する再検討が行われ、今度は、言語学の側から、統計学に対して問題が提起される。このように、言語学と統計学の対話がなされることによって、言語学の営みがより生産的になり、統計学にとっても、言語学の要請から新しい分析手法を見いだすきっかけを得る。

このような言語学と統計学との対話のきっかけの一つとなる議論を提示することも本研究の目的である。

1.5 本論文の構成

本論文の構成は以下のようにまとめられる。

表1. 本論文の構成

		3章	4章	5章
分野		文学	言語学	方言学
サブテーマ		計量文体論	言語類型論	アイヌ文化・言語論
データの内容と形式		源氏物語54帖の助動詞の頻度表	世界の129言語の言語規則に関するカテゴリカルデータ	アイヌ語19方言の語彙の同根性に関する2相3元データ
分析の目的		54帖の分類	語順規則と129言語の同時分類	19方言の分類
主な分析手法	階層的クラスター分析	◎	( )	( )
	MCA	( )	◎	○
	「数量化Ⅲ類クラスタリング」	△	△	○
	Neighbor-Net	○	△	×
	MCA _Neighbor-Net	○	△	◎
分析上着眼したポイント		変数変換 誤分類個体 タンデム法における次元数選択 クラスタリングにおける距離と手法の選択	尺度水準 変数個体のプロット	尺度水準 欠測データの扱い 2番目以降に強い構造への着目
個別分野への貢献		成立順序論 作者論	語順言語類型論	アイヌ語方言地理学 周圏論

◎は主たる分析に使用した手法.○は分析プロセスの一部として使用した手法.( )は予備的分析として使用した手法. △は二次的分析に使用した手法. ×は使用しなかった手法.

まず、2 章では、本研究の 3 章以降で用いる統計手法について概観する。言語データに対しては、さまざまな分類手法が存在することを確認したが、実際

には当該データへの適用の試行錯誤や、既存の知見などとの総合的検討を行うことが重要である。

つづいて、3 章、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、本研究での工夫を述べる。表 1 はこの 3 章分についてデータの内容面と手法面の検討事項をまとめてある。

3 章では源氏物語の文体研究に関する記念碑である村上·今西(1999)論文を統計科学的に再検討した。この先行研究が抱える潜在的な課題を確認した上で、本研究では主に 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の比較をすること、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 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についての考察を論じる。

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。Tsunoda et al. (1995a)は、129 の言語の 19 の語順に関する 87 の変数についてのデータ(角田 1991; Tsunoda, Ueda & Itoh, 1995b)に対してクラスター分析を適用した。1990 年代にこのようなデータを独力で作成したこと自体画期的なことであるが、さらに、通常の言語類型論の研究が言語の分類に注目するのに対して、彼らの研究の革新的なのは言語の分類とともに変数の分類をも試みた点である。さらに、上田・伊藤(1995)は、同データについて主成分分析を使うことによって変数間の潜在構造を明らかにしている。しかし、彼等のデータは尺度水準の観点からは、高々順序尺度であり、仮に非線形の効果が考えられるのならば、名義尺度にすぎない。すなわち、上田・伊藤(1995)が用いた間隔尺度を前提とした主成分分析ではデータの構造を十分に捉えているとは言えない可能性がある。そこで本研究では Tsunoda et al. (1995b)のデータに、Tsunoda et al. (1995a)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、多重対応分析を適用した。分析の結果は、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)は、昭和 30 年から 31 年にかけて、北海道 13 地点、樺太 6 地点のアイヌ語の方言について、大規模な調査を行った結果である。日本における、基礎語彙統計学的研究の嚆矢であり、また彼らの調査の後、多くのアイヌ語の話者が亡くなっていることから、アイヌ語の資料の保存という観点からも記念碑的な論文である。

しかし、服部･知里(1960)や、それに基づいた Asai(1974)や Lee and Hasegawa(2013)の研究は、いずれもデータの尺度水準の観点から問題があった。また、いずれの研究も服部･知里(1960)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。本章では、服部･知里(1960)のデータにおける、その情報を無視できない一つのカテゴリーとして数量化Ⅲ類を適用し、採用した次元の座標から距離行列を求めることにした。その距離行列に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、上記の距離行列に Neighbor-Net (Huson & Bryant, 2006)を適用することによって、アイヌ語における方言周圏論的構造を統計科学的に再検証した。

最後に 6 章において、本研究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る。6.1 節では 1.4 節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研究の提案について述べ、6.2 節では本研究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る。さらに、6.3 節では、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る。

(参考)

本論文の 3 章及び 5 章は、それぞれ、小野(2015a)、小野(2015b)を本論文の各章として加筆修正したものである。また、4 章は Whitman and Ono (2017, to appear)と密接に関連するが、その論文とは別のデータ・内容に関して投稿中の論文に依拠している。

第 2 章 統計的「分類」のための各手法

本章では、本論文の 3 章以降で用いる各統計的分類手法について概観する[1]。

2.1 はじめに

物事を分類するのは科学の基礎であり、その意味で当該の対象を統計的に分類するクラスタリング技法の有用性は疑いの無いものである。一番直接的な方法は、各対象の様々な側面を測定し、各対象間の何らかの意味での類似性(あるいは非類似性)をそれらの間の「距離」として定義し、その距離の遠近のデータにもとづいて各測定対象をいくつかのクラスターに分ける方法である。しかし、既存のクラスタリング技法は多様であり(Everitt, Landau, Leese & Stahl, 2011)、また解析すべきデータに対して、それらのうち、どの技法を用いるのが最適であるかは必ずしも自明ではない。

さらに、多変量データが質的変数(カテゴリカル変数)で構成される場合には、そのように類似性(非類似性)の距離を定義すること自体が、必ずしも自明ではなく、統計的に数量化する何らかの正当化が必要である。まず、尺度水準の観点からは、名義尺度(カテゴリカル変数)や順序尺度の質的データにあった距離の定義がなされるべきである。これを考慮せず、質的変数を{0, 1}の 2 値変数(ダミー変数)に変換して、それから直ちに単純に距離（例.後述するマンハッタン距離等）を定義して「クラスター分析」を実行することもしばしば見られるようだが、各変数間の関係に内在する情報を十分に把握した数量化によって何らかの意味で尺度化を行い、これに基づいた距離を求めることがより望ましいであろう。

本研究では、質的変数の各変数間(各個体間)に内在する情報に適切な数量を与え、次元圧縮をしながら、それらの多次元的な構造を解明する手法として、「林の数量化Ⅲ類」(Hayashi, 1952; 林, 1993)の適用を考えた。

林の数量化Ⅲ類はカテゴリカルデータに対する主成分分析と考えられ、通常、固有値の大きい次元に対応する少数の次元までの多次元空間での対象の布置を考え、これを解釈することが多い。しかし、測定対象や変数が多数である大規模データにおいては、少数の次元だけでは十分な情報縮約がなされない場合がある。

そのような場合、数量化Ⅲ類の結果得られた固有値の比較的大きい次元までの空間において、対応する対象の座標から対象間の「距離」を計算し、距離行列を作成し、クラスター分析を実行することが、1 つの方法として考えられる。本稿では、この方法を「数量化Ⅲ類クラスタリング」と称し、これがしばしば試行される。

---

[1] ここで、「クラスター分析」という言葉を用いずに、統計的「分類」手法という語を用いたのはクラスター分析、「数量化Ⅲ類クラスタリング」だけではなく数量化Ⅲ類の変数や個体のプロットの考察も含めるためである。

一般に、各個別の解析対象群について「分類」を施す場合、アプリオリに万能な方法はない。本稿では、上述した 3 つの手法などを念頭に置いた上で、各種の言語データの分類について、試行錯誤を行い、各解析対象に適した選択や工夫を施し、総合的に判断していくことに努める。各統計的手法の形式的な結果ではなく、それらを試行錯誤していく過程の中で、各言語データの性質が浮かび上がってくることを期待する。

本章の以下の節では、主として、クラスター分析、数量化Ⅲ類など、本稿で活用する統計手法の基本を簡明にまとめておく。これらに関しては、すでに多数の文献や論文があるため、詳細は、関連する各文献を参照のこと。

2.2 クラスター分析

一般に、個体×変数の形で得られる多変量データにおいては、個体のクラスタリングと変数のクラスタリングがある。本研究では、主に個体のクラスタリングを念頭において記す。

クラスター分析には各種の方法があるが、大きく分けて、**階層的クラスター分析**と**非階層的クラスター分析**がある。階層的クラスター分析とは、個体間の類似度あるいは非類似度(距離)に基づいて、最も似ている個体から順次に集めてクラスターを作っていく方法である。階層的クラスター分析は、クラスターが作られていく様子が**樹形図（デンドログラム、dendrogram）**で示される(鄭·金, 2011, p.237)。しかし、階層的クラスター分析は、個体数が多いと計算量が膨大になり、大量のデータ解析には向いていない。大規模のデータセットのクラスター分析では非階層クラスター分析が用いられることが多い(鄭·金, 2011, p.245)。例えば、非階層クラスター分析の代表的な方法である k 平均法(k-means)は、階層性を考えず(あらかじめクラスター数[k]を指定して)分類する方法であり、計算量の観点から大規模データに対しても実行可能な計算量に収まることも特徴である。

以下にさらに、階層及び非階層クラスター分析について説明を加える。

2.2.1 階層的クラスター分析

すでに述べたように、階層的クラスター分析においては、主に 2 つの選択に迫られる。第一は距離の選択であり、第二はグルーピング手法の選択である。

2.2.1.1 距離の選択

個体iの変数tの値を$x_{it}\{i = 1,2, \dots n\}\{t = 1,2, \dots m\}$とすると、個体$k$と対象$l$の**ユークリッド距離(Euclidean distance)** d_{kl}は、以下のように定義される。

$$d_{kl} = \sqrt{\sum_{t=1}^{m}(x_{kt} - x_{lt})^2}$$

また、個体kと対象lの**マンハッタン距離(Manhattan distance)** d_{kl}は、以下のように定義される。

$$d_{kl} = \sum_{t=1}^{m} |x_{kt} - x_{lt}|$$

また、個体kと個体lの**相関距離(Correlation distance)** d_{kl}は、以下のように定義される。

$$d_{kl} = 1 - \frac{\sum_{t=1}^{m}(x_{kt} - x_{k\bullet})(x_{lt} - x_{l\bullet})}{\sqrt{\sum_{t=1}^{m}(x_{kt} - x_{k\bullet})^2}\sqrt{\sum_{t=1}^{m}(x_{lt} - x_{l\bullet})^2}}$$

ただし、$x_{k\bullet} = \frac{1}{m}\sum_{s=1}^{m} x_{ks}$, $x_{l\bullet} = \frac{1}{m}\sum_{s=1}^{m} x_{ls}$

以上のような個体間の距離d_{kl}のすべての組み合わせを算出することで距離行列$\boldsymbol{D} = \{d_{kl}\}$が求められる。

2.2.1.2 グルーピングの手法

一般にグルーピングとは以下のような手順である(佐藤, 2009)。

1. 初期状態として n 個の個体それぞれがクラスターをなしていると考える。したがってクラスターの個数 K は K=n とする。
2. 予め選択した、距離とグルーピングの手法によって最も距離の小さい対を求め、それを一つのクラスターに融合する。K を K-1 として、K>1 ならば 3. に進み K=1 ならばクラスタリングを終了する。
3. 新しく作られたクラスターと他のクラスターとの距離を、予め選択した距離とグルーピングの手法によって計算する。2.に戻る。

距離のデータに基づいて、対象群を各クラスターに分ける方法(グルーピング)には以下のようなバリエーションが存在する。例えば、階層的クラスター分析のグルーピングの手法にも、最短距離法、最長距離法(Sørensen, 1948)、ウォード法(Lee & Wilcox, 2014; Székely & Rizzo, 2005; Ward, 1963)、群平均法、重心法、などがある。これらの手法は Lance-Williams family (Lance & Williams, 1967; 佐藤, 2009)に属する。

最短距離法は、A と B のグループの間で、最も近い要素同士の距離を A と B の距離と定義し、グルーピングを行っていく方法である。

最長距離法は、A と B のグループの間で、最も遠い要素同士の距離を A と B の距離と定義し、グルーピングを行っていく方法である。

ウォード法は、A グループ内部の分散が最小に、A と B との間の分散が最大になるように、新しいクラスターA'を構成していく方法である。

群平均法は、A と B のグループの間で、互いのすべての要素同士の平均を A と B の距離と定義し、グルーピングを行っていく方法である。

重心法は、A のグループの重心と、B のグループの重心との距離を A と B

の距離と定義し、グルーピングを行っていく方法である。

クラスター分析のグルーピング手法については、Everitt, Landau, Leese & Stahl(2011]が詳しい。

これらの中で、最短距離法、最長距離法、ウォード法がしばしば用いられる方法であろう。本稿でも、主にこれらが用いられる。ただし、ウォード法については、Ward(1963)のオリジナル提案ではユークリッド二乗距離を用いているが、近年の研究(Lee & Wilcox, 2014; Székely & Rizzo, 2005)によって、ユークリッド距離やマンハッタン距離に基づいたウォード法も Lance-Williams family に属することが示されている。本稿では、これらの結果に基づき、ユークリッド距離とマンハッタン距離に基づいたウォード法を用いる。

これらのグルーピングの結果として、**樹形図(デンドログラム)**を得る。

2.2.2 非階層クラスター分析

非階層クラスター分析は、k-means (MacQueen, 1967; Steinhaus, 1957)や、k-medoids (Kaufman & Rousseeuw, 1987)や k-modes (Huang, 1997)などがある。

その中でも代表的なものとしてしばしば用いられる k-means(k-平均法)のアルゴリズムは、以下のようなプロセスで構成される。

1. クラスターの個数 k を、予めあたえる。
2. データの空間の中に、k 個の任意の点を発生させる。
3. 対象となる各点から k 個の点までの距離を計算し、k 個のうち最も近い点にラベル付けする。
4. 同じラベル付けをされた点の平均を新たな点として定める。(k 個の点が新たに更新されたと考えられる。)
5. ラベル付けが変わらなくなるまで、前述の 3-4 のプロセスを繰り返す。

非階層クラスター分析の特徴として、階層的な分類とは異なり、データをあらかじめ指定された個数(k 個)のクラスターに「分ける」ということがある。一般に、階層的クラスター分析で得られる樹形図は、「樹形図をどこで切るべきか」、「何個のクラスターがあると考えるべきか」という点は明示的ではない。非階層クラスター分析は、クラスターの個数を予め与えることでこの問題を克服しようとしている。ただし、k を適切な値にあらかじめ合理的に定めることは非常に難しい。Rousseeuw (1987)などにおいては、シルウェット(Silhouette)という概念を用いることを提案している。非階層クラスター分析に関する詳細は、佐藤(2009)を参照されたい。

本稿で扱う各言語データの分析においては、上記のように分類すべきクラスターの数を予め定めることは恣意的になりかねないという懸念があり、非階層クラスター分析の使用は避けた。ただし各言語データに関して、確実な知見が蓄積されていき、それらの基づき厳格なモデルの弁別が求められるような段階では、非階層クラスター分析の活躍の場はあり得よう。

2.2.3 階層的クラスター分析と非階層クラスター分析の選択の基準

階層クラスター分析と非階層クラスター分析の選択の基準は、2.2.1 節及び 2.2.2 節で述べたような単なるアルゴリズムの違いだけではなく、分析対象となるデータがどのように生成されたかということを考慮することで階層クラスター分析と非階層クラスター分析のどちらを選ぶべきかが明らかになることがある。例えば、人間の移動に伴う遺伝のデータは一つの共通の祖先から分岐していくモデルを考えることが自然であるため、階層クラスター分析を適用することがより適切であろう。人間の移動に付随する言語のデータについては、一つの共通の祖先から分岐していくモデルと、分岐した言語同士の接触によって、言語が変化するモデルの両方を考慮する必要がある。よって、分析対象となる言語の歴史的な背景を考慮しながら、前者のモデルである階層クラスター分析と後者のモデルである非階層クラスター分析を使い分けることが適切であろう。一方、本研究で扱う言語の変数や文学作品に関するクラスタリングについては、むしろ階層性を伴わない非階層クラスター分析を適用することが妥当な場合もあろう。

これらの違いを認識した上で、階層クラスター分析と非階層クラスター分析を使い分けることで、以下のような研究を行うことも可能であろう。

例えば、言語データに階層クラスター分析を適用し、得られた言語の樹形図が既存の言語学における樹形図とは異なる場合には、それらの言語においては水平伝播のような非階層的な要因が働いていると考えることができる。よって非階層クラスターを適用することによって、水平伝播の構造を明らかにすることができるだろう。

また、方言の研究においても、階層クラスター分析を適用した際に、既存の言語学における樹形図と一致する樹形図を与える単語群は、その言語を話す民族の移動とともに広まったものと考えることができるが、そのような系統性を示さない単語にはその民族が持つ特別な事情(水平伝播など)があると考えることができるだろう。

本論文では、主に 3 章と 5 章でクラスタリング技法を使っているが、どちらも階層クラスタリングである。源氏物語の成立論を統計科学的に再検討した 3 章では、距離とグルーピング手法を選択すれば、樹形図が一意に定まることを優先したため、非階層クラスタリングではなく、階層クラスタリングを用いた。アイヌ語の方言関係を扱った 5 章では、経験上アイヌ語の方言データには階層性を認めるのが妥当であり、またアイヌ語に関しては水平伝播が基調ではないため、階層クラスタリングを用いた。

ただし、言語類型論のデータを扱った 4 章に関しては MCA の適用のみで、主な知見が得られるため、階層クラスタリングと非階層クラスタリングの選択については副次的なものである。

2.2.4 クラスタリング手法の選択の基準

上述のようにクラスタリング手法には様々なものが存在するが、齋藤·宿久

(2006, p199)によれば、クラスタリング手法を評価するには、大きく分けて以下の 3 つの観点がある。

1)入力データとクラスタリング結果の比較によるもの(適合性基準や非適合性基準による評価)
2)クラスタリング結果自体のある種の良さによるもの(さまざまな指標による評価)
3)同じデータを解析した複数のクラスタリング結果の比較によるもの

1)の適合性基準の観点から階層的クラスタリングを評価する指標として、2 乗誤差基準、ミンコフスキー基準、Cophen 係数、Spearman の順位相関係数、Kendall の順位相関係数、Goodman-Kruskal の順序連関係数などが存在する。非階層クラスタリングについては Milligan(1981)が詳しい。2)クラスタリング結果自体を評価するものとして、許容性(admissibility)の観点から正許容性、k 群構造許容性、完全構造許容性、正規許容性、凸許容性、対象複製許容性、クラスター複製許容性、クラスター削除許容性など様々な指標が提案されている。詳しくは、Fisher and Van Ness(1971)や Chen and Van Ness(1996)、Mirkin(1996)などが参考になる。ただし、齋藤･宿久(2006)によれば、クラスタリング手法においては、2)の観点のあらゆる意味で「最適な」手法は存在しないとされている。3)の同じデータを解析した複数のクラスタリング結果を比較する指標としては、一致係数、不一致係数、Rand 統計量、修正 Rand 統計量などがある。後述するように、クラスタリングの「良さ」とは、参照する外部の理論との整合性との試行錯誤の中で見出していくものであろうが、クラスタリングの結果そのものを評価する様々な手法や指標が提案されていることに留意する。石田･西尾･椿(2011)においては、再現性、均等性、外的基準、内的基準、解釈可能性といった指針が提示されている。本研究では石田･西尾･椿(2011)のいくつかの基準(均等性、外的基準、解釈可能性)を参照することになる。

2.3 林の数量化Ⅲ類

2.3.1 林の数量化Ⅲ類と関連手法について

一般に、林の数量化Ⅲ類(Hayashi's Quantification Method Ⅲ, QMⅢと略すことがある)(Hayashi, 1952)は、カテゴリカルデータ(質的変数に関する多変量データ)に対する主成分分析と考えられる。

林の数量化Ⅲ類は、その歴史的経緯から各国で調査研究や言語解析等々、異なる背景の中で独立して発明されたことで、例えば Fisher（1940）は最適化尺度法(Optimal Scaling)、フランスの Benzécri（1973）は多重対応分析(Multiple Correspondence Analysis, MCA と略す)、カナダのトロント大学の西里静彦(Nishisato, 2006)は双対尺度法(Dual Scaling)、オランダの Gifi（1990）は等質性分析(Homogeneity Analysis)など、異なる名称で呼ばれることがあるが、本質的には数学的には同等なものである(足立･村上, 2010)。

この林の数量化Ⅲ類は、質的変数からなる多変量のデータに適切な数量を与え、当該の解析対象の集合の各要素の間の内在的な関連構造の次元を抽出し、結果として多次元空間に対象間の類似性を表示する。この意味で、林の数量化Ⅲ類は質的変数の主成分分析(足立･村上, 2010)と考えられる。(主成分分析はPearson[1901]によって発明され、その後、Hotelling[1933]によって再発明された手法である)。

数量化Ⅲ類が適用されるデータの型はいくつかある。本研究では、3 章においては、頻度表(林の用語では相関表[Contingency Table])のデータを扱い、4 章では、アイテムカテゴリー型の変形したデータを扱い、5 章ではアイテムカテゴリー型(岩坪, 1987)や多重選択型(山田･西里, 1993)[2]のデータを扱う。数量化Ⅲ類という用語では最も典型的には、社会調査などで得られる複数の質問項目とその選択肢に相当するアイテムカテゴリー型データに適用されるが、実際には、数量化Ⅲ類という言葉はこれら 3 つのデータに全てに使われており、最適な数量を与える基準は次に述べるように同一である。他方、Benzécri(1973)らの用語では 3 章で扱う頻度表の分析には Correspondence Analysis(CA)の語が当てられ、4 章 5 章で扱うアイテムカテゴリー型データの分析には Multiple Correspondence Analysis(MCA)の語が当てられる。

数量化Ⅲ類は、例えばアイテムカテゴリー型データに対しては、各アイテムのそれぞれのカテゴリーの該当・非該当を表す 2 値データ(ダミーデータ)に変換し、データ行列を作り、行項目と列項目に数量(重み)を与える。その際に、行項目と列項目の相関が最大になるように、数量を付与するのである。この最大化問題はある種の固有値問題に帰着し、このとき最大化される相関係数の 2 乗が対応する次元の固有値と一致する。その固有値は主成分分析で固有値を寄与率の指標とするのと同様に当該次元の寄与の指標として用いられる。

林の数量化Ⅲ類の結果は、通常、固有値の大きい次元に対応する少数の次元までの多次元空間での対象の布置を考え、これが解釈されることが多い。しかし、測定対象もその変数も多数である大規模データにおいては、場合によっては、少数次元だけでは情報縮約されず、数量化Ⅲ類の布置の解釈はかなりの高次元まで考慮しなければならず、結果の解釈が困難となることがあり得ることに注意する必要がある。

数量化Ⅲ類の具体的な計算手続きについては、Hayashi(1952)が初期の提案であり、より幅広い方法の記述については林(1993)が詳しい。近年の文献では永田･棟近(2014)も言及がある。数量化Ⅲ類は、対応分析(CA)と同様に、頻度表のデータについても適応可能であるが、日本で一般的に使用されているソフトウェアでは、林の数量化Ⅲ類(QMⅢ)ではアイテムカテゴリー型データのみが扱える場合も多いようである。本稿では、両方のデータを扱えるソフトウェアとして、R の corresp を使用した。

---

[2] 5 章の付録で扱ったデータは、山田･西里(1993)で扱われている多重選択型データの特殊な形であるが数量化Ⅲ類における相関最大の基準は変わっていない。

2.3.2 Correspondence Analysis による定式化

ここでは数量化Ⅲ類と数学的に等しく且つ本研究で使用したソフトウェアに合わせて、CA(Correspondence Analysis)の定式化に基づいて、使用した手法について紹介する（Greenacre & Blasius, 2006, pp.12-14)。上に述べた行項目と列項目の相関最大化の基準と CA の定式化の同等性については例えば柳井(1994)も参照されたい。

行列 $\mathbf{M}$ を n カテゴリーを持つ行項目と p カテゴリーを持つ列項目の間のクロス表として与えられる頻度表行列であるとする。行列 $\mathbf{M}$ は$\{m_{ij}\}, i = 1, \dots, n$ かつ$j = 1, \dots, p$と書くことができる。最初に $\mathbf{M}$ を $\mathbf{P}\{p_{ij}\}$に変形することを考える($p_{ij} = m_{ij}/m$, mは行列 $\mathbf{M}$ の要素の総和である)。今、行列 $\mathbf{P}$ の行和を$p_{i\cdot}$ $(= m_{i\cdot}/m)$、列和を$p_{\cdot j}$ $(= m_{\cdot j}/m)$として、それぞれを、r_i と c_jと更に簡略に表記する。さらに、行列 $\mathbf{P}$ を行列 $\mathbf{S}\{s_{ij}\}$($s_{ij} = (p_{ij} - r_i\, c_j)/\sqrt{r_i\, c_j}$)に変形すると、行列 $\mathbf{S}$ は、

$$\mathbf{S} = \mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc^T})\mathbf{D_c^{-1/2}}$$

のように書くことができる。

ただし、 $\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_p \end{pmatrix}$,

$$\mathbf{D_r^{-1/2}} = \begin{pmatrix} \mathbf{1/\sqrt{r_1}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1/\sqrt{r_n}} \end{pmatrix}, \ \mathbf{D_c^{-1/2}} = \begin{pmatrix} \mathbf{1/\sqrt{c_1}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1/\sqrt{c_p}} \end{pmatrix}$$

行列 $\mathbf{S}$ の要素$\{s_{ij}\}$行列 $\mathbf{M}$ の行と列が独立であるという仮説で観測頻度と期待頻度との乖離を評価した量で、これの全ての要素にわたる二乗和が独立性の検定におけるカイ二乗統計量となる。換言すると行列 $\mathbf{S}$ の交互作用(特定カテゴリー同士の結びつきの強さ)の値と捉えることができる。

さらに、行列 $\mathbf{S}$ について、特異値分解を適用し、

$$\mathbf{S} = \mathbf{U\Lambda V^T}$$

を得る。$\mathbf{U}$の列を「左特異ベクトル」といい、 $\mathbf{V}$の列を「右特異ベクトル」といい、$\mathbf{U^TU = VV^T = I}$が満たされる。$\mathbf{\Lambda}$を特異値行列と呼ばれる対角行列である。

行列$\mathbf{U}$は行項目と列項目の交互作用の大きさを特徴づける行項目の各カテゴリーの重みを、$\mathbf{V^T}$は同じく列項目カテゴリーの重みを、行列 $\mathbf{S}$ の特異値分解によって与えている。

この特異値分解の特異値の大きい少数に対応する部分を取り出し、

$$\mathbf{S} \simeq \mathbf{U_s \Lambda_s V_s^T}$$

のような低ランクの行列の積で近似することを意図した方法であると言える。

実際には、**U**や**V**の各列の要素を何らかの尺度調整をしたものを、カテゴリーの特徴を表している座標として解釈に用いる。

具体的には、CA においては、行、列それぞれにおいて、特異値を反映させた座標と反映させない座標が考えられる。
それぞれ、

特異値を反映させた行の座標(principal coordinates of rows): $\mathbf{F} = \mathbf{D_r^{-1/2}U\Lambda}$

特異値を反映させない行の座標(standard coordinates of rows): $\mathbf{A} = \mathbf{D_r^{-1/2}U}$

特異値を反映させた列の座標(principal coordinates of columns): $\mathbf{G} = \mathbf{D_c^{-1/2}V\Lambda}$

特異値を反映させない列の座標(standard coordinates of columns):$\mathbf{B} = \mathbf{D_c^{-1/2}V}$

で与えられる。

以下の各章では実質的に特異値を反映させた $\mathbf{F}$ や $\mathbf{G}$ を用いるが、これらの値を「座標」またはより丁寧に「空間座標」と表記する(2.4 節参照)。

なお、上記の特異値分解における特異値λの 2 乗が数量化Ⅲ類における固有値に相当し、得られる次元の寄与の指標として用いられる。

2.4 「数量化Ⅲ類クラスタリング」

冒頭で述べたように、数量化Ⅲ類の結果を、比較的大きい固有値に対応する次元までの空間布置からデータの情報を解釈しようとする場合、対象となる変数があまり多数ではない場合や、少数の次元のみが寄与率(固有値)が高い場合には問題はないが、変数がかなり多く、寄与率(固有値)が多次元に分散している場合では「解釈」は容易でない。

そこで、本研究では「数量化Ⅲ類」の結果に対して、得られた対象群の多次元の座標の布置から得られる距離を計算し、クラスター分析を適用し、対象群の分類によって解釈を容易にすることを考える。これを、本稿では**「数量化Ⅲ類クラスタリング」**と称することにする。本研究で使用する MCA によって距離行列を求め Neighbor-Net を適用する方法もこの一つである。「数量化Ⅲ類クラスタリング」は Lebart(1994)、菅(2001)、Wen and Chien(2011)等々でも適用されている手法であるが、1.3 節で既述の通り言語データ解析における適用はあまりみられないようである。また、その他の場合も多くは、同じデータに対してこの手法と直接的なクラスター分析との比較や、数量化Ⅲ類との比較はなされていないようである。本論文では、各種の言語データ解析において、それらとの比較も行いつつ、当該言語データの性質と各統計手法との相補的な関係を浮き彫りにしたい。換言すると、当該のデータの分類には、十分な感度

の分類手法が必要であるが、感度が高すぎるとノイズを拾いすぎて、明快にクラスターを分離することは望めない。その感度と分離度とのバランスを見ながら、データ解析を進めることになるであろう。

「数量化Ⅲ類クラスタリング」を適用するに当たって、数量化Ⅲ類の布置から距離データを得るために、各対象の布置の何次元目までの座標を利用するかに関しては、様々な方法があり得る。たとえば、主成分分析でしばしば行われるように、固有値の累積寄与率が例えば 80%以上、90%以上になるまでの次元を利用する方法がある。また、固有値を降順に並べたプロット(スクリープロット)から、固有値の減少が緩やかになる一つ前の次元の座標を利用する方法もある(足立･村上, 2010)。

通常、数量化Ⅲ類の解は相対的である(単位が決まっていない)ために、各次元の分布は平均 0、分散 1 となる条件のもとで得られる。そこで、実際の座標を多次元空間の距離として利用するためには、固有値を考慮して再計算する必要があることに留意すべきである。

林(1993: 86)では、この問題について以下のように記述している。

“総平均は 0, 寸法は$x(u)$や$y(v)$の分散を 1 にしたり、あるいは、$y(v)$の人による分散を 1 にそろえたり、相関係数の 2 乗にそろえたりするなど、適宜考えればよい”(林, 1993: 86)

本研究では、n 次元目までの空間におけるユークリッド距離D_{ij}は以下のように与えられる。これは、林の記述に基づけば各次元の分散を相関係数の 2 乗にそろえることに対応する。ここでは、表記は 2.3 節の CA の説明と対応させ、2.3.2 節**A**の定義に基づきの下記の式で対象間の距離を求めている。

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \omega_k \left(a_{ik} - a_{jk}\right)^2} \ , \ \omega_k = \lambda_k^2 / \sum_{t=1}^{n-1} \lambda_t^2 .$$

λ_kは CA の定式化における k 次元目の特異値である。

(注意. a_{ik}は k 次元目の特異値に対応する解[特異ベクトル]のカテゴリーi の値である。n 個のカテゴリーの m 次元までの空間の布置を考えている。行項目のカテゴリーn が列項目のカテゴリーp よりも小さければ、得られる最大の次元は n-1 である。)

このω_kによる重み付けは数量化Ⅲ類の固有値で次元の重要度を評価して距離に反映させたものと言える。

同様に m 次元目までの空間におけるマンハッタン距離d_{ij}は以下のように与えられる。

$$d_{ij} = \sum_{k=1}^{m} \sqrt{\omega_k} \left|a_{ik} - a_{jk}\right| \ , \ \omega_k = \lambda_k^2 / \sum_{t=1}^{n-1} \lambda_t^2 .$$

本論文では以下の各章のデータ解析においては、R Core Team(2014)の「corresp」(Venable & Ripley, 2002)のコマンドを使って、得られた座標に上記の式を適用することによって、$\{d_{ij}\}$ を求めた。前述のように、本論文では、この $\{d_{ij}\}$ に基づいてクラスター分析を行う方法を「数量化Ⅲ類クラスタリング」とする。

2.5 クラスター分析に関わる統計手法の補足

2.5.1 Neighbor-Net とは

本節では、本論文 3 章、4 章、5 章で試行した Neighbor-Net 及びそれに関連して小野(2015b)で提案した MCA_Neighbor-Net について説明する。MCA_Neighbor-Net とは、「数量化Ⅲ類クラスタリング」と同様に 数量化Ⅲ類の結果の空間布置から得られる距離データに対して、Neighbor-Net を適用する方法である。この手順からわかるように MCA_Neighbor-Net はタンデムクラスタリングの一手法である。この方法を本論文では 3 章、4 章では補足的に用い、5 章では中心的に活用する。この方法で有益な知見がもたらされた事例は主に 5.6 節で紹介される。

Neighbor-Net (ここでは NN と略す)は本来、生物進化の系統樹の解析に用いられて開発されてきた（Bryant & Moulton, 2003)。生物の系統進化を解明するには、単純な樹状構造を扱うのでは不十分で、多数の樹状構造を同時に表現するネットワークの方が適切である。特に、データ解析の初期段階では当該の生物進化に関する知見が乏しい場合、推論のための方法ではなく、データの表現の方法として NN が用いられる。単純な樹状構造などでは、誤ったモデルでの推論となる危惧を避けるためである。本稿で取り扱う言語の伝搬変容や地理的に隣接する各言語（方言）間の解明なども、その生物進化の解明と、方法論としては類似の側面があり、NN の活用は有益と思える。

より具体的には、NN は、対象間の距離行列を入力データとして、緩やかに近隣のもの同士（類似の対象同士）を結びつける系統樹的なクラスターを作り上げるアルゴリズムである。NN は、クラスターの一部が重複（overlap）し、かつ階層とはならないクラスターの構造、一種の系統樹的なネットワークを構成する。通常のクラスター分析の手法を距離データ行列に適用した場合、得られる樹形図はデータについて、いわば「最も強い」構造を取り出したものである。しかし、NN を適用することで、データの「最も強い」構造だけではなく、いわば「2 番目以降に強い」構造も取り出すことが可能になる。

NN のプロセスには(1)分類対象を凝集するプロセス(2)距離行列を縮約するプロセス(3)各辺の長さを推定するプロセス、及び(4)実際にネットワークを描画するプロセスが存在する。(1)(2)(3)については、Bryant, Moulton and Spillner(2007)及び Huson and Bryant(2006)を参照されたい。また(4)については、Huson(1998)及び Wetzel(1995)を参照されたい。

図 1 に Neighbor-Net(NN)の解釈の仕方について簡単な概念図で説明する。

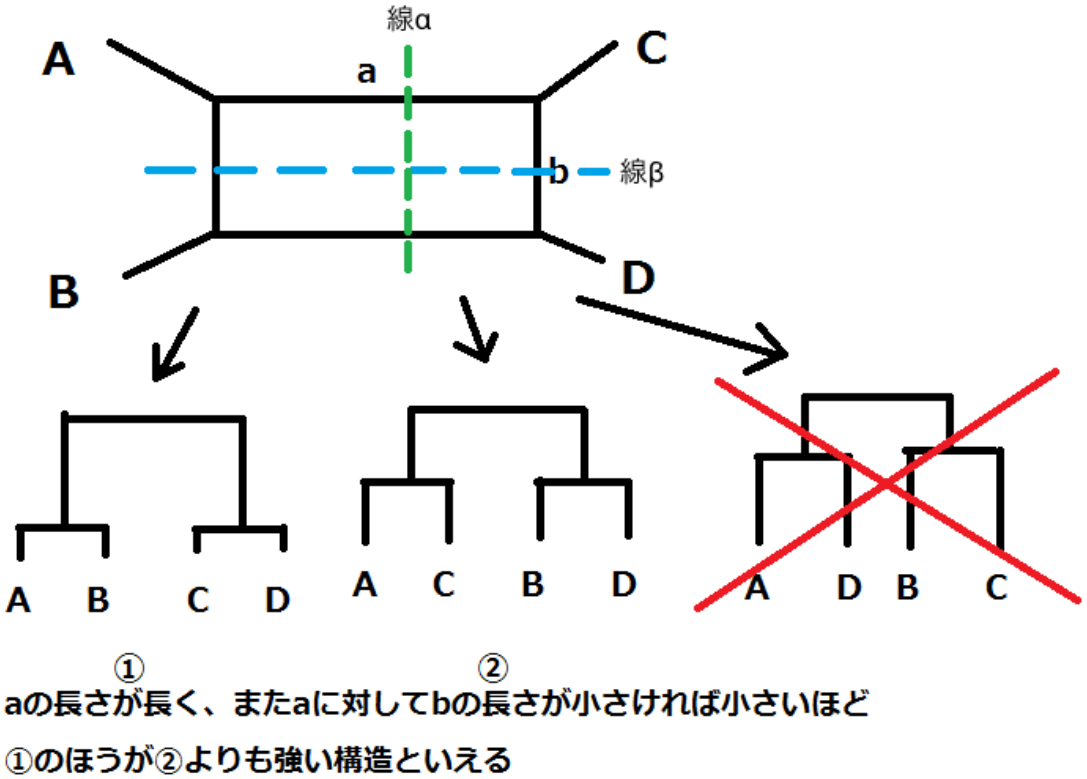


図 1. Neighbor-Net(NN)の解釈の仕方

NN の出力は図 1 のように、網の目状の(ネットワークのような)形をしている。このネットワークの図は線 α で切ることによって A と B がまとまり、C と D がまとまり最後に 2 つのクラスターがまとまる左側のデンドログラムと解釈することができる。次に線 β で切ることによって、A と C がまとまり、B と D がまとまり最後に 2 つのクラスターがまとまる真ん中のデンドログラムと解釈することができる。しかし、A と D がまとまり、B と C がまとまり最後に 2 つのクラスターがまとまる右側のデンドログラムとは解釈することができない。また、a の長さが長く、また a に対して b の長さが小さければ小さいほど、左側のクラスターの構造が真ん中のクラスターの構造よりも強い構造と解釈することができる。

NN の図を解釈する際には、まず、(1)図 1 の a に対応する長い辺を最初に探し、長い辺を線 α によって切断し、得られるクラスター構造から解釈を行う。次に、(2) (1)の過程で得られた切断線 α と交わる切り方(図 1 の線 β に対応する)のうち辺の長さの長い(図 1 の b に対応する)構造から解釈を行う。以上のような手続きによって図を解釈する。実際には、対象の数が多数に及び複雑な網の目状の図が構成されるため、上記の方針で得られた多数の切り方の可能性を同時に考察する必要が生じる。

2.5.2　Box-Cox 変換

クラスタリング等において、各変数の比率データから距離を計算する方法もあろうが、場合によってはその比率データに分散安定化変換の一種である Box-Cox 変換(Box & Cox, 1964)を適用し、それを距離データとしてクラスター分析を施すような方法も考えられる（3 章に例示される）。

Box-Cox 変換では、変換前のデータを X とし、変換後のデータを Y とするとしたとき以下のような関係が成立する。ここで d は分析者が設定するパラメータである。

$$Y = \begin{cases} \dfrac{X^d - 1}{d} (d \neq 0) \\ \log X \quad (d \to 0) \end{cases}$$

Box-Cox 変換がデータ分析に際してどのような性質を持つかについては Osborne(2010)が詳しい。特に d=0.5 の場合は平方根変換、d =0.0（0.0 に漸近）の場合は対数変換と呼ばれる。元のデータに変換を施す方策は各種考えられるが、それらの是非は当該のデータへの適用の試行錯誤や、既存の知見などとの総合的検討からなされるべきであろう。

第 3 章 源氏物語成立論の統計科学的再考察

― 助動詞の出現頻度に基づく 54 帖の分類[1]

3.1 はじめに

本章は源氏物語の作者推定に関する記念碑的研究である村上·今西(1999)を統計科学的に再検討した小野(2015a)を中心に論じる。村上·今西(1999)は源氏物語における 54 帖の各帖における 21 の助動詞の割合に対して、数量化Ⅲ類を適用し、得られた 2 次元の布置から、源氏物語の全体の帖を 4 つのグループに分け、それらの執筆順序について示唆的な結論をもたらしている。本章では提示しないが、クラスター分析を村上·今西(1999)のデータに直接適用して得られる樹状図では当該の 4 グループの分類が明瞭ではない結果を得るので、これが村上・今西(1999)が数量化Ⅲ類を適用する動機になったものと推定される。

一般に、様々な言語データに対して適用する統計手法に関しては、例えば安本·野崎(1976)に詳述されている中で適用されている方法は、語彙×言語(方言)に関するカテゴリカルデータから、相関係数を測るもの(Kroeber & Chrétien, 1937; Kroeber & Chrétien, 1939)や、相関係数の有意性を検定するもの(Chrétien, 1943)や、クラスター分析を行うもの(Asai, 1974)などがある。しかし、それらは当該のカテゴルカルデータの尺度水準（名義尺度であり、必ずしも間隔尺度をなさない）の性質を特に考慮しているわけではなく、この点で、村上・今西（1999）の数量化Ⅲ類の適用は意味のあるものと思える。

しかし、さらに、筆者が村上·今西(1999)の数量化Ⅲ類の解釈をより深く考察するために、数量化Ⅲ類の 2 次元の布置及び 5 次元の布置から計算される 2 つの距離データに対してクラスター分析を適用した結果、より厳密な方法をとっているにも関わらず、村上・今西が主張する 4 分類はむしろ不明瞭となった。そこで、本章では大別して 2 つの工夫を導入した。第一に、素データに対する変数変換の適用、第二に、タンデムクラスタリングの適用である。これらの工夫の組み合わせによる様々な分析手法を試した結果、分散安定化変換の一種である平方根変換を施し、変換後の値をクラスター分析することによって、村上·今西(1999)で推測された源氏物語の成立論に関する結果を示唆する樹形図が作成された。得られた樹形図は、村上·今西(1999)の推論とは矛盾しないが、より厳密に言えば、高々2 グループの成立順序を示唆するものであった。つまり、当該のデータによる統計的解析だけでは、村上・今西（1999）のように 4 分類を明示するのは困難であると思われる結果を得た。しかし、これは村上・今西（1999）の推論を全面的に否定するものではなく、彼らの推論をより確実にするには、当該のデータの解析だけでは必ずしも十分ではなく、既存の言語学的

[1] 本章は、雑誌「計量国語学」の 29 巻 8 号(pp.296-313)の小野(2015a)を大幅に加筆修正したものである。

知見を含め、新たに他の角度から眺めたデータを含めた解析を展開するなど、総合的に考察しなければならないであろう。

さらに本章では、クラスター分析の結果「混在」していると判断された帖について、Neighbor-Net を適用することによって、「弱い構造」まで考慮した時に、それらの帖が「真に混在している帖」なのか、それとも「真に混在しているとまではいえない帖(『境界事例』とここでは呼ぶ)」なのか追究した結果、松風(18)、初音(23)、夕霧(39)について既存の文献学的分類の再検討を提言した。何らかの参照できる外的基準が与えられているデータを分析する際に、クラスター分析の結果と外的基準が合致しない場合、合致しない例について Neighbor-Net を適用し、検討するという本研究の方針は今後のクラスター分析を用いた探索的研究の羅針盤となるだろう。

本章では、上述の解析結果について以下の各節で詳述する。なお、3.9 節は 3.6 節までの結果を補足する材料としての副次的な分析結果をまとめてある。

3.2 研究の背景

源氏物語は全 54 帖にわたる長編小説で、成立後千年を過ぎている。紫式部(970?~1014?)が書いたとされ、平安時代の貴族生活を背景に光源氏の恋と栄華、そして平安朝における貴族社会の生態を描き出し、我が国の古典の最高峰と目され、諸外国にも広く翻訳され、古くから研究がなされてきた。

源氏物語 54 帖は全体の構成の観点から以下の 3 部に分けるのが通説になっている。(池田, 1951)

第 1 部：(巻 1)「桐壺」～(巻 33)「藤裏葉」
第 2 部: (巻 34)「若菜上」～(巻 41)「幻」
第 3 部: (巻 42)「匂宮」～(巻 54)「夢浮橋」

「源氏物語が如何にして成立したか」については今日まで様々な議論が繰り返され、例えば、今西・室伏(2010)によれば、

> どうも、『源氏物語』成立論が不毛な水かけ論に堕してしまったのは、各研究者の用語や概念や思考方法が食い違っていたからであるように思えてなりませんでした。多くの『源氏物語』研究者が、以下の四つの概念をゴチャマゼにしたまま議論をしているのです。
>
> P、諸巻の成立順序(各巻は、どういう順序で執筆されたか)
> Q、諸巻の配列（各巻をどのように並べるのが妥当か）
> R、年立（物語内の時間は、どのように進行しているか）
> S、作者（諸巻の作者は、一人か二人か、それ以上か）
>
> この四点は、それぞれ別個に考察されるべき問題です。
>
> （今西･室伏，2010,pp.18-19）

たとえば、上記の S(作者)に関していえば、後半の 45 帖から 54 帖までの「宇治十帖」を別作者とするものが有名である。この宇治十帖別作者説については、

古くから一条兼良の「花鳥余情」、一条冬良の「世諺問答」などで、紫式部の娘である大弐三位が作者であるという説がある。これに対しては、池田(1951)や大野(1984)は、否定的な見解を示している。この説について文体の計量的な分析を試みたのは、「宇治十帖の作者—文章心理学による作者推定」(安本, 1957)や「文体統計による作者推定—源氏物語、宇治十帖作者について」(安本, 1958)などが嚆矢であり、源氏物語の 1 頁あたりの、和歌、直喩、声喩、色彩語、心理描写文、句点の数、および源氏物語の各巻から千字選んで、(各巻の最初の五百字と最後の五百字の) 品詞分類を行い、各々の名詞、用言、助詞、助動詞、品詞のあわせて 11 個について、宇治十帖の 10 帖と、それ以外の 44 帖との間で統計的検定により両者の作者が同じとは言い難いと結論づけられている。これに対して、新井(1997)は五十音図の子音行列と母音列の頻度データに基づいて、宇治十帖別作者説を否定している。近年の研究では、例えば、Tsuchiyama and Murakami(2013)では、名詞、動詞、形容詞、形容動詞、副詞、助動詞、助詞に対しての主成分分析の適用や、ランダムフォレストという手法を用いて特徴量を抽出した上で、主成分分析を適用する分析を行っている。結果は宇治十帖別作者説に対して否定的なものであった。

一方、前述の P(成立順序)に関して、「源氏物語」が現在みられる巻序で成立したものでないという可能性を最初に指摘したのは、1922 年に発表された、和辻哲郎の「源氏物語について」(後に『日本精神史研究』に所収) である。和辻は、「帚木」巻の冒頭部の記述について分析し、以下のように述べている。

> (略) ...すなわち周知の題材の上にまず短い『源氏物語』が作られ、それに後からさまざまの部分が付加せられたと見るのである。が、いずれであるにしても、とにかく現存の『源氏物語』が桐壺より初めて現在のままに序を追うて書かれたものでないことだけは明らかだと思う。
>
> (和辻, 1926, pp209-210)

その後、「源氏物語執筆の順序」(阿部, 1939)、「源氏物語成立攷」(玉上, 1940)などを経て、「源氏物語の研究」(武田, 1954)において、前半部 33 巻のうち、長編的要素からなる紫の上系 17 帖と短編的性格の強い玉鬘系 16 帖との異質性を根拠に、玉鬘系 16 帖は紫の上系 17 帖が成立した後に執筆され、現行の順に挿入されたものであるという、源氏物語成立論が展開されるに至った。

武田は「源氏物語の研究」(武田, 1954)において、各巻の登場人物に関して表 1 を示した。一見して分かるように、玉鬘系に出てくる人物は紫の上系の話には登場せず、紫の上系 17 帖が成立した後に、玉鬘系 16 帖が執筆されたという一つの根拠となっている。さらに、武田は「源氏物語の研究」(武田, 1954)に収められた別稿「源氏物語の最初の形態再論」において、紫の上系 17 帖の成立の後に、玉鬘系 16 帖が執筆されたと考えられる根拠について 11 点をあげている。それらのうち、中心となる 4 点を引用する。(引用文中で使われていた旧字はすべて新字に直した)

一、 第一部三十三帖中紫上系十七帖だけで連続統一をもったものとして完結した物語である。
二、 玉鬘系十六帖は一見バラバラのように見えるが全体を通じて脈絡があり、紫上系とは別の統一を持って居る。
三、 玉鬘系の巻々の事件·人物共に紫の上系の物語上に痕跡を与えず、紫上系は玉鬘系より独立して居り、三十三帖中玉鬘系十六帖を除き去っても何等の支障を来さない。
四、 それだけで連続統一を持つ紫上系の物語の巻々の所々に玉鬘系の巻々が入って居る為、紫上系の物語を切断して、無理に割り込ませた形になって居り、紫上系から玉鬘系、玉鬘系から又紫上系へのうつりに、不自然さがある。

(今西・室伏, 2010, pp.215-216)

武田説は多くの賛同者を集めたが、他方で激しい批判も受け、論争は現在も続いている。

表 1:源氏物語での紫の上系と玉鬘系の登場人物の分布を示した表(武田, 1954, p10 より抜粋)

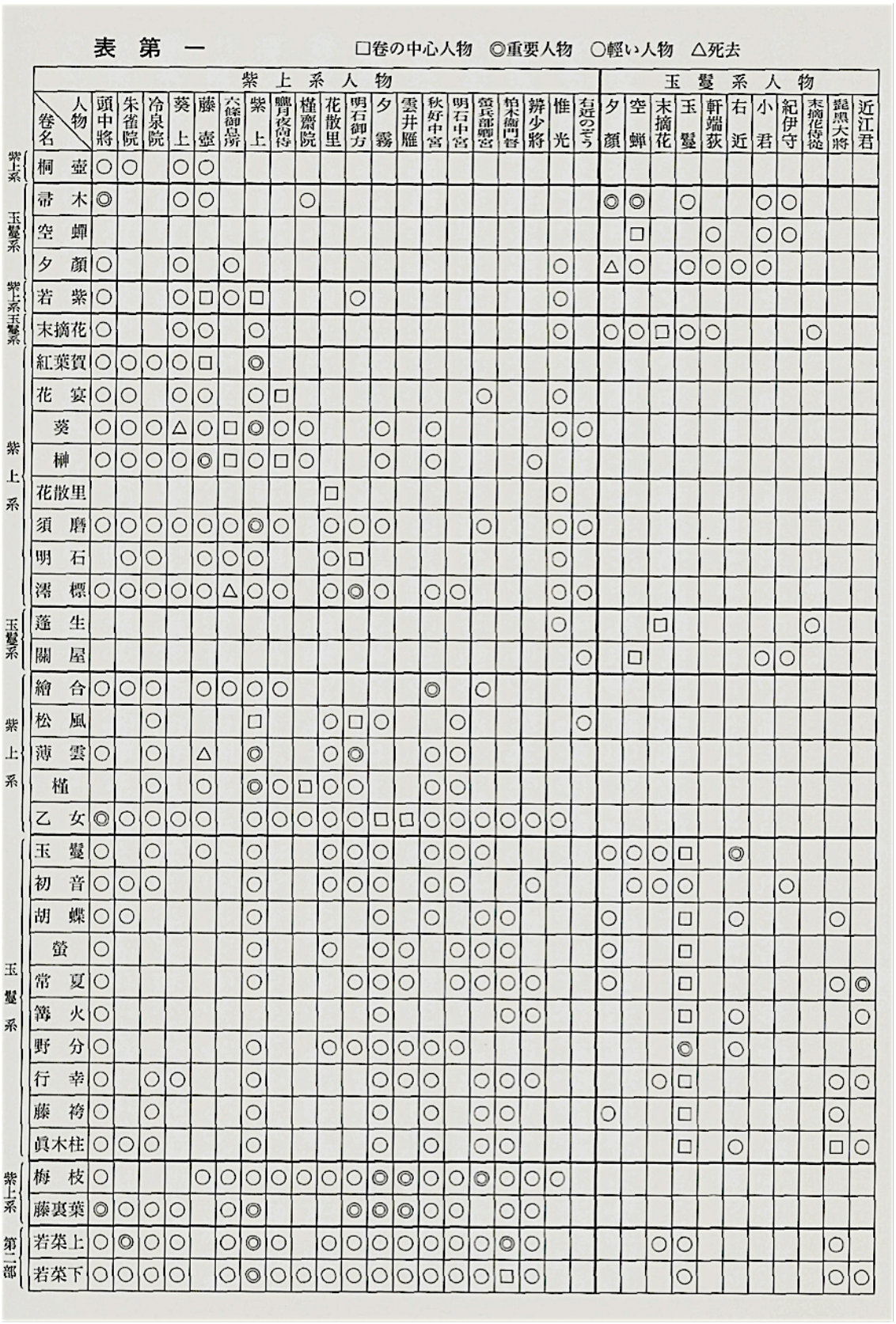


表　第　一　　□巻の中心人物　◎重要人物　○軽い人物　△死去

	巻名＼人物	紫上系人物																					玉鬘系人物									
		頭中将	朱雀院	冷泉院	葵上	藤壺	六條御息所	紫上	朧月夜尚侍	權齋院	花散里	明石御方	夕霧	雲井雁	秋好中宮	明石中宮	螢兵部卿宮	鉢末衛門督	辨少將	惟光	右近のぞう	夕顔	空蟬	末摘花	玉鬘	軒端荻	右近	小君	紀伊守	末摘花侍從	髭黒大將	近江君
紫上系	桐壺	○	○		○	○																										
玉鬘系	帚木	◎			○	○				○												◎	◎		○			○	○			
	空蟬																						□			○		○	○			
	夕顔	○			○		○													○		△	○		○	○	○	○				
紫上系玉鬘系	若紫	○			○	□	○	□				○								○												
	末摘花	○			○	○		○												○		○	○	□	○	○				○		
	紅葉賀	○	○	○	○	□		◎																								
	花宴	○	○		○	○		○	□								○			○												
紫上系	葵	○	○	○	△	○	□	◎	○	○			○		○					○	○											
	榊	○	○	○	○	◎	□	○	□	○			○		○				○													
	花散里										□									○												
	須磨	○	○	○	○	○	○	◎	○		○	○	○				○			○	○											
	明石		○	○		○	○	○			○	□								○												
	澪標	○	○	○	○	○	△	○	○		○	◎	○		○	○				○	○											
玉鬘系	蓬生																			○				□					○			
	關屋																				○		□				○	○				
紫上系	繪合	○	○	○		○	○	○	○						◎		○															
	松風			○				□			○	□	○			○					○											
	薄雲	○		○		△		◎			○	◎			○	○																
	槿			○		○		◎	○	□	○	○			○	○																
	乙女	◎	○	○	○	○		○	○	○	○	○	□	□	○	○	○	○	○	○												
玉鬘系	玉鬘	○		○		○		○			○	○	○		○	○	○					○	○	○	□		◎					
	初音	○	○	○				○			○	○	○		○	○			○				○	○	○			○				
	胡蝶	○	○					○					○		○		○	○				○			□		○				○	
	螢	○						○			○		○	○		○	○	○				○			□							
	常夏	○						○				○	○	○		○	○	○	○			○			□						○	◎
	篝火	○											○					○	○						□		○					○
	野分	○						○			○	○	○	○	○	○									◎		○					
	行幸	○		○	○			○					○	○	○		○	○	○					○	□						○	○
	藤袴	○		○				○					○		○		○	○				○			□						○	
	眞木柱	○	○	○				○					○		○		○	○							□		○				□	○
紫上系	梅枝	○				○	○	○	○	○	○	○	◎	◎	○	○	◎	○	○	○												
	藤裏葉	◎	○	○	○		○	◎				◎	◎	◎	○	○		○	○													
第二部	若菜上	○	◎	○	○		○	◎	○		○	○	○	○	○	○	○	◎	○					○	○						○	
	若菜下	○	○	○	○		○	◎	○	○	○	○	○	○	○	○	○	□	○						○						○	○

3.3 先行研究-村上･今西(1999)について

3.3.1 村上･今西(1999)の概要

3.2 節で述べた研究史の上でも、「P. 成立順序」に関して、計量的な分析を試みた記念碑的論文と言えるものが、村上･今西(1999)である[2]。村上･今西(1999)では、全帖においての出現頻度の多い上位 21 語の助動詞[3]を取り上げ、各 54 帖での出現頻度をもとに、分析を行った。具体的には、村上･今西(1999)のデータは、i 帖の j 番目助動詞の出現数をm_{ij}としたとき、i 帖総語数で、m_{ij}を割ったa_{ij}である。

村上･今西(1999)は、源氏物語 54 帖を以下のように分類し、紫の上系を A グループ、玉鬘系を B グループ、第 2 部及び匂宮三帖を C グループ、宇治十帖を D グループとした。

A　紫の上系　　：1，5，7，8，9，10，11，12，13，14，17，18，19，20，21，32，33
B　玉鬘系　　：2，3，4，6，15，16，22，23，24，25，26，27，28，29，30，31
C　第二部　　：34，35，36，37，38，39，40，41
C　匂宮三帖　　：42，43，44
D　宇治十帖　　：45，46，47，48，49，50，51，52，53，54

村上･今西(1999)では、54 帖各帖の 21 の助動詞の総語数に対する出現頻度に対して、数量化Ⅲ類を適用し、得られた 2 次元の布置図から、主に以下の 4 つを観察している。

1. A グループ（紫の上系）と B グループ（玉鬘系）はあまり重なっておらず、これは武田宗俊の源氏物語成立論と関係が示唆される。
2. C グループ(第二部と匂宮三帖)と D グループ(宇治十帖)とは重なりがあまりない。
3. 全体の布置を見ると、A グループ（紫の上系）と C グループ(第二部と匂宮三帖)とが重なり、B グループ（玉鬘系）と D グループ(宇治十帖)とが重なっている。
4. A グループ（紫の上系）より D グループ(宇治十帖)の方があとに書かれた

---

[2]村上･今西(1999)によれば、この分析を行うために、「源氏物語大成」(池田, 1985) をもとにし、源氏物語 54 巻の全文を単語に分割した上で、品詞コード等の数量分析に必要な情報をつけた約 37 万 6 千語のデータベースを作成し、このデータベースをもとに分析を行っている。古文データをコンピュータで形態素解析することが可能な今日と異なり、当時はすべて手作業でデータを作成しており、この意味でも村上･今西(1999)は記念碑的論文である。

[3] 具体的には「ず」、「む」、「たり」、「けり」、「なり」、「り」、「ぬ」、「き」、「べし」、「つ」、「る」、「す」、「めり」、「さす」、「らむ」、「らる」、「じ」、「けむ」、「まじ」、「まし」、「まほし」の 21 語である。

という前提に立てば、執筆の順序は、A グループ（紫の上系）→C グループ(第二部と匂宮三帖)→ B グループ（玉鬘系）→D グループ(宇治十帖)と推定される。

特に 4 番目の推論は、紫の上系成立後に玉鬘系が挿入されたという、武田(1954)の説と合致している。

3.3.2 村上·今西(1999)の潜在的課題

しかし、村上・今西(1999)の研究は次のような潜在的な問題を抱えている。

先に述べた源氏物語に関する文献学的な分類が絶対的なものであるならば、{A},{B},{C},{D}の 4 群の平均値について明瞭な差が表れると考えられる。

図 4 は尺度をそろえるために 21 の助動詞を Z 変換した上で、4 群の平均値をプロットしたものである。

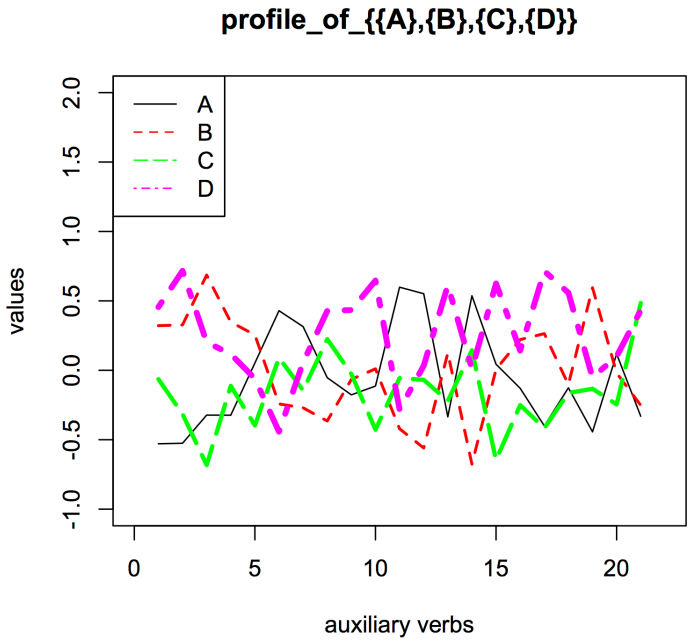


図 1: 21 の助動詞を Z 変換し{A},{B},{C},{D}の 4 群について、平均値をプロットしたもの

しかし、Z 変換した変数の値からは、グループの差は不明瞭であるため、単に変数の情報を観察していているだけでは不十分なことがわかる。そこで、次に、{A},{B},{C},{D}を外的基準として判別分析による分類を試みた。しかし、基本的な{A},{B},{C},{D}の 4 群判別では正判別率(方法は leave-out-one cross validation[LOOCV]を用いた) が 54%程度であった。また、4 つのグループのいくつかのグループを併合した 2 群、3 群判別も試行した[4]が、正判別率は 75%から 88%程度で芳しくなかった。もっとも正判別率が良かった{A},{B,C,D}にお

[4] これは 2 群判別が 7 通り、3 群判別が 6 通りの計 13 通りの場合がある。

いても、文献学的な知見とはあまり整合しておらず、判別分析の結果から何らかの結論を導くことは難しい。

よって、本研究では文献学の分類を絶対としないが、参照する枠組みとし、探索的追究を目的としてクラスター分析を適用する方針をとる。

さらに、村上・今西(1999)では、前節で述べたように、数量化Ⅲ類を活用し、A グループ、B グループ、C グループ、D グループの大まかな関係を把握した。ただし、彼らは 2 次元目までの結果に基づいて考察している。筆者はこうした数量化Ⅲ類の布置のグループ分類について考察をより深めるために、数量化Ⅲ類で得られた布置にクラスター分析を試行した[5]。次元の決定には、たとえば因子分析でみられるように固有値が一定の値以上のものを選ぶ考え方(Greenacre & Blasius, 2006) や、固有値のスクリープロットから、固有値の減少が緩やかになる点の一つ前の次元までを採用する考え方（足立・村上, 2010）などがあるが、本稿では現時点の一つの選択肢として後者を選択した。足立・村上(2010)の考え方に従い、固有値のスクリープロット（図 2）を観察すると、第 2 次元目までと第 3 次元目以下では大きな段差があり、さらに、5 次元目までと第 6 次元目以下でも段差がある。よって、2 次元目までの空間座標、及び 5 次元目までの空間座標を取り上げることとした。村上・今西(1999)では 2 次元目までの座標を採用している。2 次元目までの累積寄与率は 40%で、5 次元目までの累積寄与率は 65%であった。

[5] 以降統計的な分析や図の出力のためのソフトウェアとして、R(2014)を使用した。具体的には数量化Ⅲ類には corresp を、クラスター分析には dist 及び stats を使用した（corresp は、多重対応分析 MCA のソフトウェアであるが、MCA と数量化Ⅲ類は数学的には同等である）。

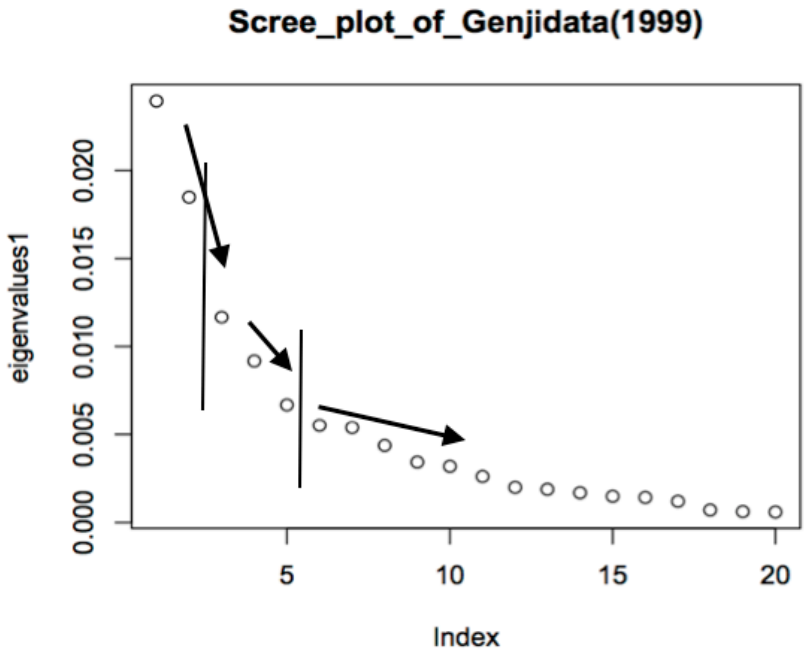


図 2: 村上・今西(1999)のデータに対して数量化Ⅲ類を適用し、固有値をプロットした結果. 本研究では 2 次元目までの座標だけではなく、6 次元目から固有値の減少が緩やかになることから、クラスター分析には 5 次元目までの座標も用いた。

まず数量化Ⅲ類の結果得られた 2 次元目までの空間座標からユークリッド距離を用い[6]、ウォード法によるクラスター分析を適用した。樹形図は図 3 に示すように、B 以外の区別が全く出来ていない。

次に、累積寄与率を高めて 5 次元まで空間座標（累積寄与率 65%）から得られる距離に、クラスター分析(ユークリッド距離、ウォード法)を適用した結果を図 4 に示す。A、B、C、D の区別が全くできていないことが観察される。

[6] この距離の計算では、2.4 に記したように座標軸のスケールを各次元の固有値の大きさに対応させた。以下同様

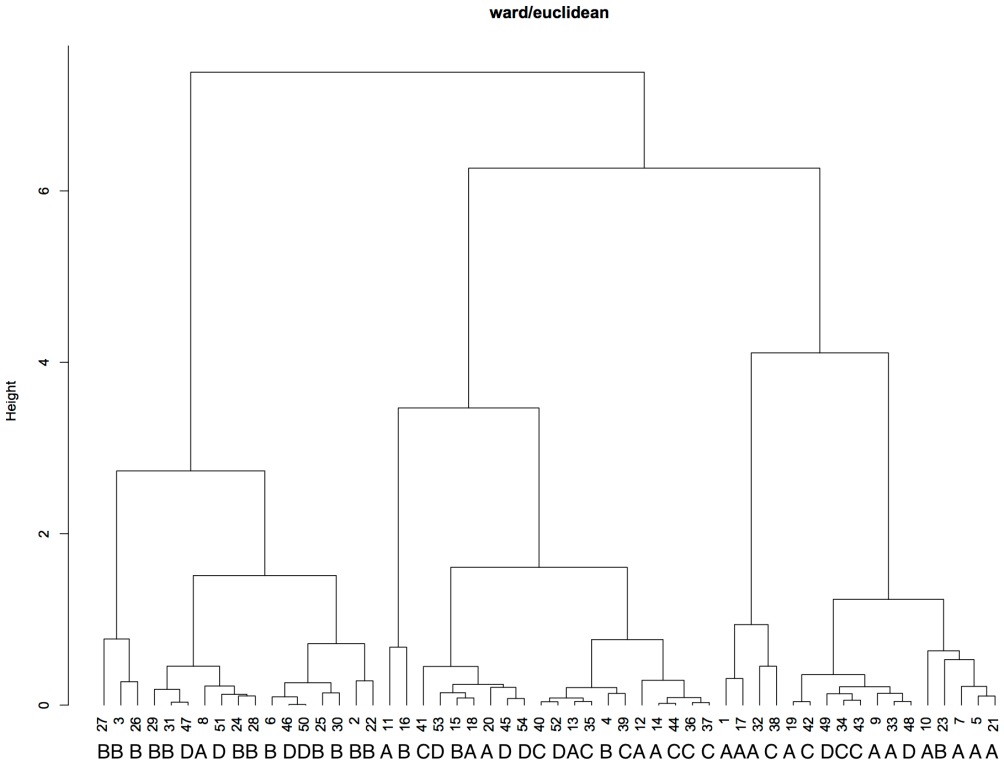


図 3: 数量化Ⅲ類の 2 次元目までの座標をもとにクラスター分析(ユークリッド距離、ウォード法)を行った樹形図.
左のクラスターに B がまとまっているが､右の 2 つのクラスターでは A,C,D の判別が難しい。

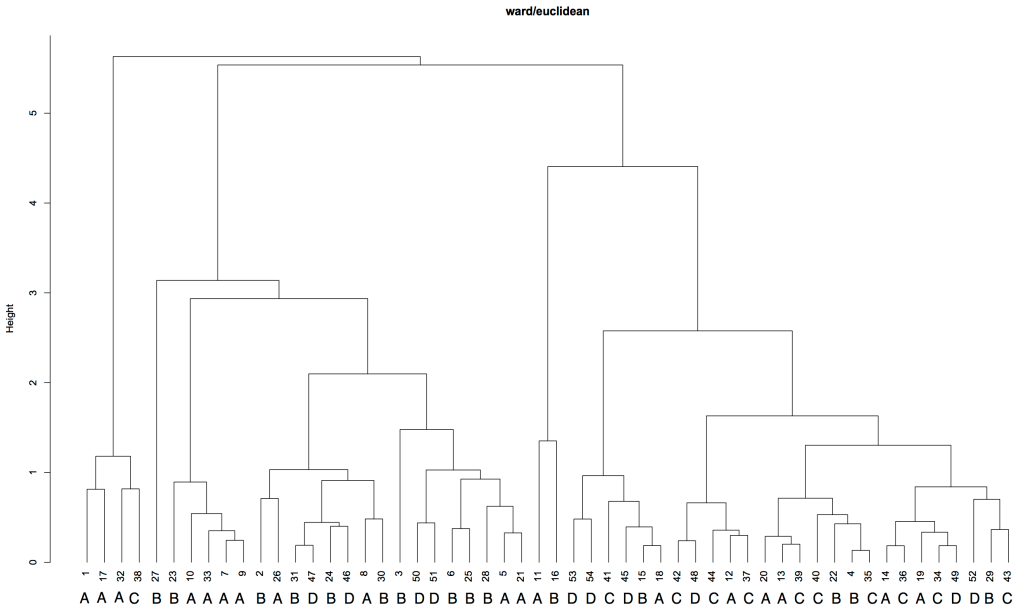


図 4: 数量化Ⅲ類の 5 次元目までの座標をもとにクラスター分析(ユークリッド距離、ウォード法)を行った樹形図.数グループに分類して解釈することが難しい。

先に示した、村上・今西(1999)の 4 番目の帖の成立順序に関する主張を積極的に支持するためには、そもそもクラスター分析の時点で{A},{B},{C},{D}の分類が明瞭である必要がある。ところが、{A},{B},{C},{D}各グループのプロファイルの結果、判別分析の結果及び数量化Ⅲ類の結果得られた 2 次元の布置や 5 次元の布置を利用したクラスター分析の結果には{A},{B},{C},{D}のグループの存在を明瞭に示すものはなく、それらの情報だけから、源氏物語の成立順序を推測するには限界がありそうである。ここで、クラスター分析の方法のいくつか、例えば距離の選択においては、ユークリッド距離またはマンハッタン距離、またグルーピングの手法では最短距離法、最長距離法、またはウォード法のすべての組み合わせを試みたが、いずれも同様の結果である。それらの結果には、そもそも村上・今西(1999)が扱うような 4 グループのような少数の分類からはかなり外れたものも少なくない[7]。その中では、ユークリッド距離でウォード法を用いた場合が、相応に、分類の是非を論じられるような結果をもたらしている印象である[8]。

以上のような予備的な研究から、村上・今西(1999)の源氏物語の助動詞のデータによる、{A}→{C}→{B}→{D}という成立論に関する主張は、助動詞のプロファイルや判別分析、数量化Ⅲ類から得られた布置へのクラスター分析など、従来からよく使われる手法とその単純な組み合わせによる一定の客観的な手続きを持って導き出すことが難しいという問題を抱えていることがわかる。

3.3.3 本研究の前提

本研究では、村上・今西(1999)の「助動詞」の情報を用いたデータから、どこまで源氏物語の成立論に貢献できるかを追究する。村上・今西(1999)では、なぜ「助動詞」の情報に注目するかについて次のように述べられている。

> “助動詞を分析対象に取り上げたのは、まず、名詞、動詞、形容詞等の他の品詞と比べて異なり語数が少ないからである。また自立語(名詞、動詞、形容詞等)が文章の意味内容に関与する語であるのに対して、付属語である助動詞は文章の意味内容ではなく陳述、すなわち意味内容の統合に関わる品詞として、物語の語り口と密接な関連を有するからである。”
>
> (村上・今西, 1999, pp774-775)

さらに、現代日本語の文体研究における記念碑的研究である Jin and Murakami (1993)においては読点の前の文字の割合が重要な変数となっているが、後の研究によって、読点の前の文字の割合と助動詞の割合は密接な関係

---

[7] 例えば、付録の図 28、図 29 に最短距離法による結果を示しているが、そうしたものの極端な例である。

[8] 後に示す条件 4 の一部に相当する。付録では図 13、図 14 を参照。

があることがわかっている。よって本研究では村上・今西(1999)の古典日本語の文体分析に助動詞の情報が有効であるということを前提に研究を進めていく。

3.4 本研究の目的

村上・今西(1999)のデータを用いて、村上・今西(1999)の「執筆の順序は、A グループ（紫の上系）→C グループ(第二部と匂宮三帖)→ B グループ（玉鬘系）→D グループ(宇治十帖)と考えられる」という主張を検討する。本研究では、「解釈」が数量化の結果の図を眺めてアドホックに行うのではなく、一定の手続きとして確定したクラスター分析によって得られた樹形図によって、より客観的に結論が得られるように諸方法を試みる。

その際に注意すべきこととして、以下の 2 点について論じる。第一に、3 章付録表 1 に載せたように変数の分散が異なっていることがクラスタリングに影響を与えることを考慮して変数変換を行うべきかということ、第二に、3.3.3 節で論じたようにタンデムクラスタリングを行うべきかということである。

3.5 分析方法

3.5.1 2 つの工夫

前述 3.3 の試行の結果を経て、本研究では主に 2 つの点での工夫を考えた。第一点は、変数変換の適用であり、第二点はタンデムクラスタリングの適用についてである。

まず、一つ目の工夫として、村上・今西(1999)の出現頻度データ(54 帖 × 21 変数)にそのままクラスター分析を適用するのではなく、より扱いやすい形にしてから、クラスター分析を施したらどうかと考えた。具体的には、分散安定化変換(Hocking, 2013)の一種である Box-Cox 変換(Box & Cox, 1964; Osborne, 2010)の中から、平方根変換と対数変換を考慮した。

本研究において、分散安定化変換を行い、それらの変換値にクラスター分析を施すことが妥当であろうと推察した理由は以下の通りである。クラスター分析を適用するために距離行列を計算する際に、導かれた距離が、分散安定化変換を施すことによって、距離行列が分散の大きい変数の影響を大きく受けてしまうという効果が除かれる。具体的には、分散安定化変換（本研究では Box-Cox 変換）を適用した後にクラスター分析を行うことは、各 54 帖の変数である各助動詞の分布の分散が一定であることにより各助動詞の影響が均一化し、その上変換後の分布が正規分布に近づくという利点が生じる。

村上・今西(1999)のデータは、i帖のj番目助動詞の出現数をm_{ij}としたとき[9]、i帖総語数で、m_{ij}を割ったa_{ij}である[10]。本研究では、a_{ij}について変数変換の有無の観点から、(1)無変換のa_{ij}(以下 Raw data と称する)と(2)a_{ij}に対数変換を施

[9] ここでのデータに関する記法は、2 章の 2.3 節に従った。

[10] これは、総語数の影響を取り除くためには必要な処理であると考えられる。よって、村上・今西(1999)のデータを扱うにあたっては、総語数の影響を取り除いたデータであるということを留意して、分析にあたることが適切であろう。

したデータと、(3)a_{ij}に平方根変換を施したデータのそれぞれに対し正規分布モデルを当てはめた場合の BIC の値を比較した。結果、(1)a_{ij}と(3)a_{ij}に平方根変換を適用したデータを比較することとした。これらの詳細については 3 章付録を参照されたい。

2 つ目の工夫として、タンデムクラスタリングの適用について考慮した。本研究では、主に数量化Ⅲ類(QMⅢ)によるタンデムクラスタリングのみを検討したが、3.9.1 においては主成分分析によるタンデムクラスタリングについても考慮した。タンデムクラスタリングを適用する際には次元数の選択についても考慮した。前処理で QMⅢを適用する場合には、図 2(条件 4)や付録図 6(条件 5)の固有値のスクリープロットを参照し、累積寄与率は 40%、65%、80%を目安に 3 通りの条件を設定した。前処理として主成分分析を含めたタンデムクラスタリングを検討した結果は 3.9.1 に記す。

3.5.2 クラスター分析の方法

上記の 2 つの工夫に共通する点としては、クラスター分析における距離の選択とグルーピングの技法の選択がある。距離の選択としてはユークリッド距離またはマンハッタン距離の 2 通り、グルーピングの方法としてウォード法と最長距離法の 2 通りを採用し、距離とグルーピングの組み合わせ 4 通りの全ての方法の結果を比較した[11]。

3.5.3 分析条件のまとめ

これらの点を考慮した結果、具体的には、2 つの工夫(データ変換の有無とタンデムクラスタリングの有無)により、前処理として、Raw data、QMⅢ、平方根変換、平方根変換＋QMⅢの 4 条件を比較することとした。平方根変換+QMⅢは前処理としてデータに平方根変換を施し、変換した値に QMⅢを適用することを意味する(ここでは、3.9 節で他の条件と合わせて結果を整理する関係で、Raw data を条件 1、QMⅢを条件 4、平方根変換を条件 2、平方根変換＋QMⅢを条件 5 とする)。さらに、QMⅢと平方根変換＋QMⅢについては、タンデムクラスタリングの次元数の選択としてそれぞれ 3 条件を考えたため、前処理とタンデムクラスタリングに関しては合計 8 条件を検討した。さらに、クラスター分析の際の距離の選択(ユークリッド距離かマンハッタン距離か)とグルーピングの手法の選択(ウォード法か最長距離法か)に関する 4 条件を考慮すると、合計で 32 条件についてデンドログラムを得た。

3.5.4 一致度による結果の評価

これらのデンドログラムについて、本研究では「一致度」という指標を用いて比較を行った。「一致度」は村上・今西の 4 グループの再現性の指標として次のように試行的に定義し計算した。

一致度: デンドログラムを見て例えばAを一番多く含むクラスターをAのクラスターとして、そのクラスターが実際に A を含んでいる率を

[11]最短距離法についても検討したが、村上･今西[1999]の 4 分類から逸脱し、均等性(石田･西尾･椿, 2011)からも適切ではないのでここでは示さない。

「一致度」[12]として計算する。B, C, D の場合も同様(ただし、あるクラスターが A を一番含むことと、その他のクラスターと比較して A の比率が一番高い場合が一致し無い場合があることに留意する必要がある)。A の「一致度」と B の「一致度」、C の「一致度」、D の「一致度」をすべて合計したものを全体の一致度とする。

ただし、均等性(石田･西尾･椿, 2011)の観点から妥当なクラスター化が再現されなかったケースでは一致度は計算しなかった。また、「一致度」自体は、石田他(2011)における外的基準との整合性の確認の一種と言える。

3.6 結果

まず、QMⅢの適用に当たっては、源氏物語 54 帖の多次元空間における布置を把握するためにプロットの一覧を作成した。3 章の付録図 1 に、条件 4 における 5 次元までの座標空間の各 2 次元平面の布置(座標軸のスケールは各次元の固有値の大きさに対応させた)を載せた。3 章の付録図 2 に、条件 5 における 5 次元までの座標空間の各 2 次元平面の布置を載せた。3.5 節で述べたようにここではデータ変換、前処理については、Raw data、QMⅢ、平方根変換、平方根変換＋QMⅢの 4 条件の一致度を比較する。

以上の 4 条件(次元数を含めると 8 通りの組み合わせ)について 4 種のクラスター分析を適用した結果の一致度の一覧が表 4 である[13]。ここで表示する一致度は基準とするグループを{A},{B},{C},{D}の 4 つとした場合と、{A,C},{B,D}のように 2 グループずつをまとめ 2 つとした場合の 2 通りを計算している。[14]

表 4 を一見して分かるように、基準とするグループを 4 つとした場合の一致度は概して高くない。ここでは、基準とするグループを 2 つとした場合の結果を中心に記述する。

一致度を比較した結果、平方根変換を施しクラスター分析を適用した結果が最も一致度が良かった。本節ではそれらの結果について主に検討していく。ただし、統計的には他にも様々な分析手法(主成分分析など)が考えられ、この点については 3.9 節で追加的に述べる。

村上･今西(1999)の助動詞の出現頻度データ(54 帖×21 変数)に対して平方根変換を適用したものにユークリッド距離で距離行列を求め、ウォード法によってクラスタリングを行って得られた樹形図が図 5、マンハッタン距離で距離行列を求め、ウォード法によってクラスタリングを行って得られた樹形図が図 6、ユークリッド距離で距離行列を求め、最長距離法によってクラスタリングを行って得られた樹形図が図 7、マンハッタン距離で距離行列を求め、最長距離法

---

12 ここで判別分析での用語の「精度」という言葉を用いないのは本章におけるクラスター分析においては、明確な外的基準を持たないためである。

13 参考までに、ユークリッド二乗距離を用いる本来のウォード法も検討したが、大差はなかった。

14 その他、考えられるパターンによる一致度の計算も行ったが、結果は{A,C},{B,D}の場合よりも悪かった。

によってクラスタリングを行って得られた樹形図が図 8 である。

以下の考察では、図 5 から図 8 に共通して最も左に登場する 2 帖、すなわち極端に総語数が少ない(11)花散里と(16)関屋は除いて考える。

表4: 各手法でのクラスタリングの一致度の一覧表.附録の図10から図18参照。

	付録図番号	前処理	用いた次元 (累積寄与率)	距離	ｸﾗｽﾀﾘﾝｸﾞ 手法	A,B,C,Dの 一致度[%]	(A,C)と(B,D)の 一致度[%]
条件1	3章付録図11	Raw data	21	Euclid	W	46.15(24/52)	75.93(41/54)
			21	Manhattan	W	49.02(25/51)	85.19(46/54)
			21	Euclid	D	32.69(17/52)	63.46(33/52)
			21	Manhattan	D	38.46(20/52)	64.71(33/51)
条件4	3章付録図13	QMIII	2(40.0%)	Euclid	W	48.15(26/54)	79.63(43/54)
			2(40.0%)	Manhattan	W	44.44(24/54)	75.93(41/54)
			2(40.0%)	Euclid	D	40.00(18/45)	55.56(25/45)
			2(40.0%)	Manhattan	D	39.22(20/51)	58.33(24/48)
	3章付録図14		5(65.9%)	Euclid	W	51.28(20/39)	62.00(31/50)
			5(65.9%)	Manhattan	W	35.42(17/48)	61.11(33/54)
			5(65.9%)	Euclid	D	44.68(21/47)	70.21(33/47)
			5(65.9%)	Manhattan	D	38.46(20/52)	60.87(28/46)
	3章付録図15		20	Euclid	W	53.19(25/47)	58.33(28/48)
			20	Manhattan	W	33.33(16/48)	50.00(25/50)
			20	Euclid	D	39.22(20/51)	54.90(28/51)
			20	Manhattan	D	44.23(23/52)	69.57(32/46)
条件2	図5(3章付録図10)	平方根変換	21	Euclid	W	52.94(27/51)	92.31(48/52)
	図6		21	Manhattan	W	56.86(29/51)	90.38(47/52)
	図7		21	Euclid	D	47.06(24/51)	75.00(30/40)
	図8		21	Manhattan	D	60.00(27/45)	72.73(32/44)
条件5	3章付録図16	平方根変換 +QMIII	2(38.7%)	Euclid	W	47.06(24/51)	74.07(40/54)
			2(38.7%)	Manhattan	W	43.14(22/51)	76.92(40/52)
			2(38.7%)	Euclid	D	42.00(21/50)	76.47(39/51)
			2(38.7%)	Manhattan	D	31.91(15/47)	58.14(25/43)
	3章付録図17						
			5(68.8%)	Euclid	W	43.14(22/51)	53.85(28/52)
			5(68.8%)	Manhattan	W	45.45(20/44)	55.77(29/51)
			5(68.8%)	Euclid	D	49.02(25/51)	72.55(37/51)
			5(68.8%)	Manhattan	D	44.23(23/52)	68.63(35/51)
	3章付録図18		20	Euclid	W	53.19(25/47)	63.83(30/47)
			20	Manhattan	W	45.10(23/51)	68.63(35/51)
			20	Euclid	D	52.27(23/44)	54.90(28/51)
			20	Manhattan	D	39.22(20/51)	64.71(33/51)

平方根変換+QMIIIは平方根変換を先にRawデータに適用し、変換した値にQMIIIを適用したことを示す。またクラスタリング手法のWはWard法、Dは最長距離法を示す。一致度の()の中は、分母が一致度の計算の際に考慮した帖の数、分子が実際に一致した数を意味する。

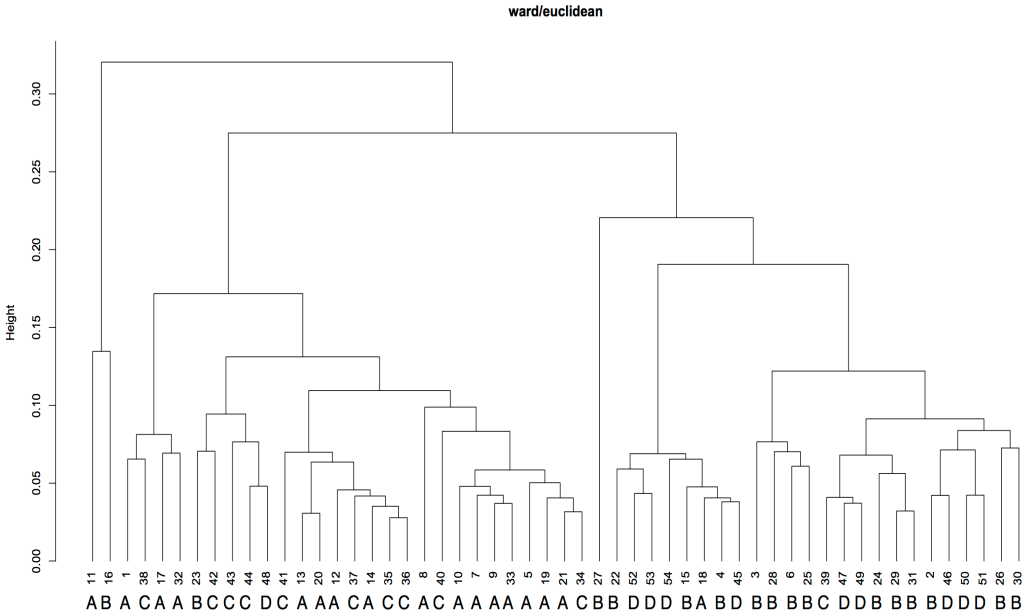


図 5:前処理として平方根変換を採用し、クラスター分析(ユークリッド距離，ウォード法)で得られた樹形図.左のクラスターに A と C が、右のクラスターに B と D が集まっている。

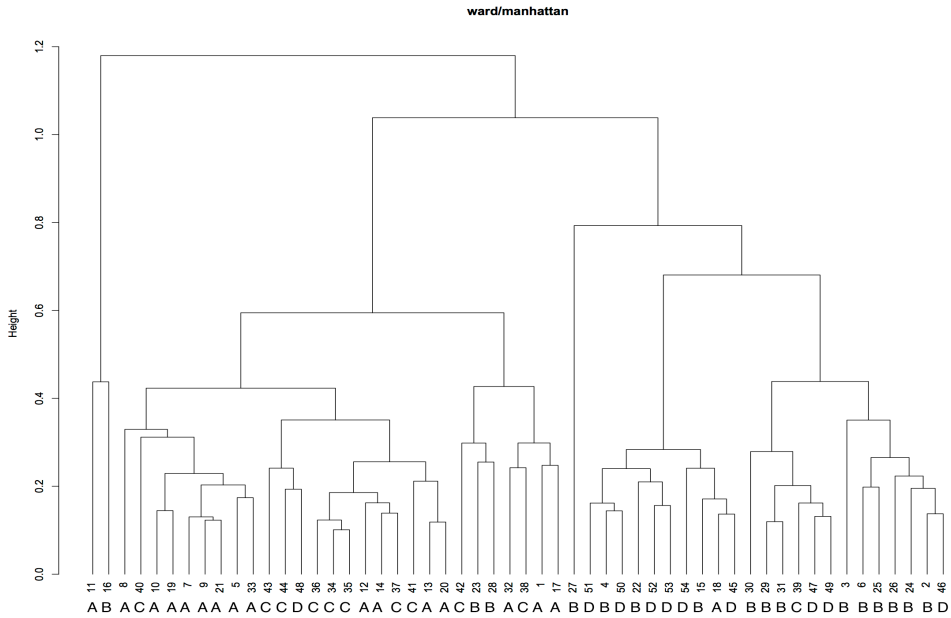


図 6: 前処理として平方根変換し、クラスター分析(マンハッタン距離，ウォード法)で得られた樹形図.左のクラスターに A と C が、右のクラスターに B と D が集まっている

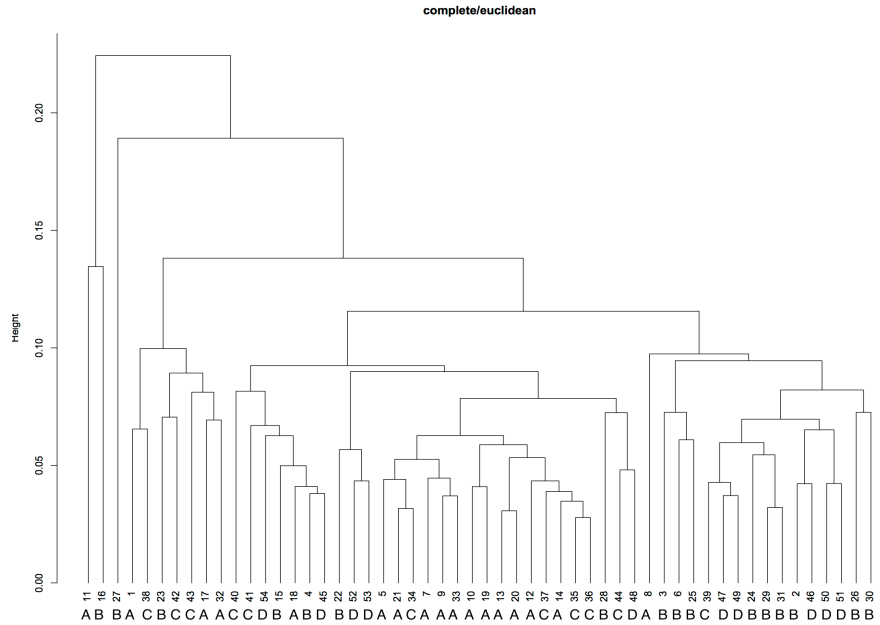


図 7:前処理として平方根変換し、クラスター分析(ユークリッド距離，最長距離法)で得られた樹形図.右のクラスターに B と D が集まり、それ以外に A と C が集まっている。

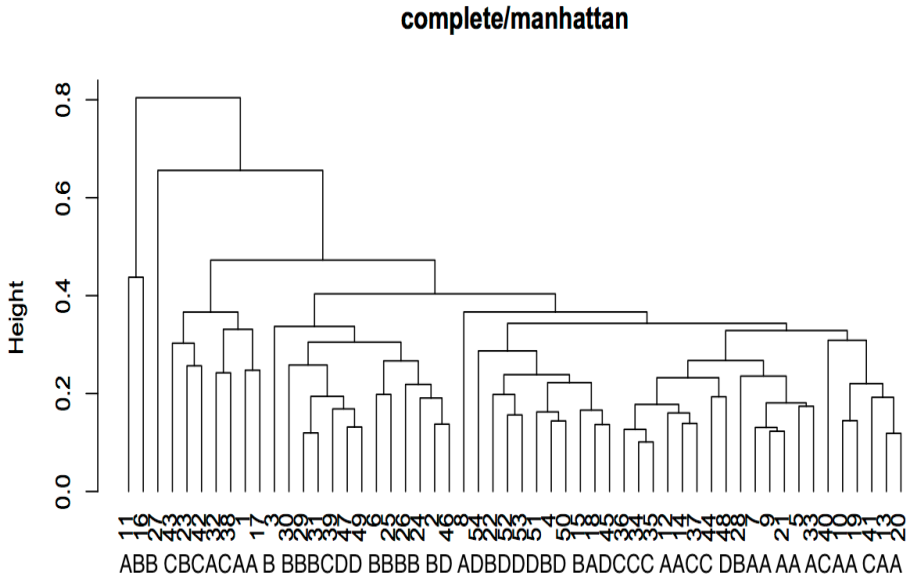


図 8: 前処理として平方根変換し、クラスター分析(マンハッタン距離，最長距離法)で得られた樹形図.最も右のクラスターに A と C が集まり、その左のクラスターに B と D が集まっている。

村上・今西(1999)の助動詞出現頻度データに対して平方根変換を適用した値に、クラスター分析(ユークリッド距離，ウォード法)を施した図 5 に関しては、左のクラスターに A と C が、右のクラスターに B と D が集まっている。その中で、左のクラスターに B の(23)初音と D の(48)早蕨が、逆に、右側のクラスターに A の(18)松風 C の(39)夕霧が混在している。

また、平方根変換を適用した値をマンハッタン距離とウォード法でクラスター分析を行った図 6 に関しては、左のクラスターに A と C が、右のクラスターに B と D が集まっている。その中で、左のクラスターに B の(23)初音、(28)野分と D の(48)早蕨が、逆に、右側のクラスターに A の(18)松風、C の(39)夕霧が混在している。

平方根変換を適用した値をユークリッド距離と最長距離法でクラスター分析を行った図 7 に関しては、最も右のクラスターに B と D が、それ以外のクラスターに A と C が集まっている。その中で、左のクラスターに、B の(4)夕顔、(15)蓬生、(22)玉鬘、(23)初音、(27)篝火、 (28)野分と D の(45)橋姫、(48)早蕨、(52)蜻蛉、(53)手習、(54)夢浮橋が、逆に右側のクラスターに A の(8)花宴、C の(39)夕霧が混在している。

最後に、平方根変換を適用した値をマンハッタン距離と最長距離法でクラスター分析を行った図 8 に関しては、最も右のクラスターに A と C が、その左のクラスターに B と D が集まっていると考え、残りのクラスターは除いて考えると、最も右のクラスターに B の(4)夕顔、(15)蓬生、(22)玉鬘、(28)野分、D の(45)橋姫、(48)早蕨、(50)東屋、(51)浮舟、(52)蜻蛉、(53)手習、(54)夢浮橋が、その左のクラスターに C の(39)夕霧が混在している。

通常、クラスター分析の結果は、分析に用いる距離、および手法によって大きな影響をうけるが、図 5、図 6、図 7、図 8 の結果から、源氏物語 54 巻は大まかに一貫して｛A,C｝と｛B,D｝とに分かれ、一致度(表 4 参照)は図 5 では 92.31%、図 6 では 90.38%、図 7 では 75.00%、図 8 では 72.34%であることから、クラスター分析の結果得られた｛A,C｝と｛B,D｝という分類は比較的安定していると考えられる。もともと、村上・今西(1999)は「A→C→B→D」の執筆順を示唆しており、その 4 分類 A,B,C,D は、上のレベルのクラスター{A,C}と｛B,D｝の 2 分類をさらに分割したものとみなすことができる。クラスター分析の手法の細かい選択に依らず、結局｛A,C｝と｛B,D｝の 2 分類は再現された。

3.7 考察

表 4 から、図 5 と図 6 は、｛A, C｝,{B, D}の分類に関して高い一致度を示すことがわかる(図 5 では 92.31%を、図 6 では 90.38%を示している)。この 2 グループへの分類が安定的であることを前提として、執筆順序について議論すれば次のようになる。狭義の統計分析だけからは、執筆順序について指摘することはできないが、A よりも D の方が後に書かれたと仮定すれば、少なくとも{A,C}が書かれた後に｛B,D｝が書かれたということになり、これは、紫の上 17 帖の成立の後に独立して玉鬘 16 帖が書かれたとする武田説(武田, 1954)を支持するものである。

ただし、村上・今西(1999)で示唆されたような「A→C→B→D」の成立順序を明瞭に確認するものではない。なぜなら、{A,C},{B,D}の一致度が良かった図 5、図 6 においても、図 5 の左側のクラスターでは A と C とを分けることは難し

く、同様に、右側のクラスターではBとDとを分けることは難しいからである（{A},{B},{C},{D}の各クラスターの一致率は、それぞれ、図5で52.94%と図6で56.86%であり、分類の一致度はよくない）。

さらに、表4から、A,B,C,Dの分類に関する一致度はどの手法においても60.00%を超えることはなかった。したがって、P（成立順序）の問題として図5、図6をとらえると、執筆時代によって、助動詞の使い方(各助動詞の出現頻度)が変化すると仮定するならば、源氏物語は「{A,C} → {B,D}」の順に書かれたと推論することは可能であるが、AとC、BとDの執筆順序を判別することは、それ以前にAとC、BとDの間の分類が明瞭ではないことから、当該のデータの解析だけからは結論しがたいようである。

他方で、今までは2章の今西・室伏(2010)の引用で紹介したように、P(成立順序)の問題としてこのクラスターを考察したが、実際にはS(作者)の問題としても、図5をとらえ直すことができる。S(作者)の問題としてとらえると、助動詞の使い方が古典の作者の特徴をあらわしていると仮定するならば、源氏物語には{A,C}を書いた作者と{B,D}を書いた作者の少なくとも二人を想定することができる。もっとも、これは一人の作者の助動詞の使い方が経年効果で変化した結果、二人の作者がいるようにみえるという可能性も考慮しなければならない。これについては、村上(2002)の川端康成作品の中の文体の変化に関する分析が参考になる。川端の作品では、戦前の作品は読点の前の文字が「と」である比率が高く、逆に戦後の作品は読点の前の文字が「し」である比率が高く、経年変化による文体の変化がみられる(村上 2002, p56).

さらに、図5において、(18)松風、(23)初音、(39)夕霧、(48)早蕨の4帖が、さらに図6では、(18)松風、(23)初音、(28)野分、(39)夕霧、(48)早蕨の5帖がなぜ混在したのか検討が必要である。助動詞だけのデータでは情報量が足りないと考えるべきか、それとも今までの成立論では見落とされていた何かを物語っているのか、今後の文献学における学術的研究を待ちたい。この点について、統計学の観点からは、(18)松風、(23)初音、(39)夕霧について既存の文献学的枠組みを再考する必要があると本研究では考える。詳細については、3.9.2節にて議論する。

3.8 今後の課題

本章では、村上・今西(1999)の源氏物語の各帖の各助動詞の出現頻度のデータに対して、分散安定化変換の一種である平方根変換を適用し、さらに変換された値に対してクラスター分析を適用した。これにより、村上・今西(1999)では、数量化Ⅲ類の結果得られた布置の「解釈」という形でしか見いだせなかった源氏物語の成立論に関する考察を、樹形図による明確な形で提示することができた。この結果は、村上・今西(1999)における「A→C→B→D」という成立順序の推察とは矛盾しないが、本稿の分散安定化変換（Box-Cox変換）を施したデータにクラスター分析を適用した結果では、より統計学的に手の込んだ手法を用いて、「{A,C} → {B,D}」という成立順序を確認したが、AとC、BとDのそ

れぞれの執筆順序についてはむしろ判別が難しいことを示した。

本研究では、Box-Cox 変換に関して、BIC の計算で、21 の助動詞のうち、多数の助動詞で採用された平方根変換を採用したが、このように変数ごとに変換の候補が分かれたときに、その中で多数のものを採用するべきかについても今後の研究が必要であろう。さらに、本稿では対数変換と平方根変換の比較を試みたが、Box-Cox 変換式の d は 0(対数変換)、0.5(平方根変換)以外にもさまざまな値をとり得るので、どのような d の値が適切かも今後の検討課題である。

分散安定化変換は、これまでは実用上はむしろ回帰分析における不均一分散を解消するための手段として用いられることが多く、クラスター分析に用いられるのは例えば DNA の解析(Zwiener, Frisch & Binder, 2014)の際などのようであり、人文科学においては筆者の知る限りほとんど見られない。今回の分析から、何らかの変数変換がデータの持つ隠れた特性を引き出すことがあることが分かったので、その適用範囲と限界についても探っていくことも意味があると思える[15]。今後は、より綿密に、文献学的知見と統計学的手法の融合による、相補的な考察が深められるべきである。

ただし、本章では階層的クラスター分析を適用したが、本章で扱った源氏物語の成立論における問題の性質はむしろ非階層クラスター分析の適用が適している可能性に注意する必要がある。本研究では、非階層クラスター分析が初期値によって異なるクラスターを生成してしまうこと(初期値依存)から、階層クラスター分析を採用した。非階層クラスター分析の適用は今後の課題である。

また、本研究にはデータに関して以下のような限界点があることにも留意する必要がある。第一に、源氏物語の写本から助動詞のデータを得る際に、ある語がどの助動詞に分類されるかということに関して、専門家の中においても意見が分かれることがあるという点である。さらに、本研究で扱ったデータは各帖の助動詞の出現頻度を各帖の総語数で割った比率の値を用いているが、源氏物語においては、各帖の総語数に差があるため、比率の値のゆらぎに差がある。例えば花散里では総語数が 724 語に対して、若菜上では総語数が 20196 語であるので、比率推定値の標準誤差としては $\sqrt{\frac{20196}{724}} \simeq 5$倍程度の開きがある。よって、源氏物語研究において村上・今西(1999)のデータを用いて、本研究よりもさらに汎用的なモデルを立てることは意味があると思われるが、上記の限界を分析の際に常に意識しなければ統計学的分析のみが先行し、文献学的現実と乖離した結果となるだろう。さらに、本研究では村上・今西(1999)に従い、26 の助動詞のうち 21 の助動詞を分析に用いたが、助動詞の変数選択という視点も今後の研究には必要と考えられる。

---

[15] もっとも、安易なデータ変換はデータの扱いをより困難にすることも我々は肝に銘じておかなければならないだろう。たとえば,正規分布にしたがうあるデータ X と別の正規分布にしたがうあるデータ Y の比の値(X/Y)は、コーシー分布に従う。コーシー分布は平均と分散および高次のモーメントが存在しない。

3.9 統計的補足

本論の展開が複雑であったため、本節では、2 つの点について補足を行う。第一点は、主成分分析の結果を加え、「タンデムクラスタリング」に関する結果をより深く考察することである。第二点は、クラスター分析の結果、既存の文献学的知見と参照し、「混在」していると判断された帖について、Neighbor-Net を適用することによって、「最も強い」構造以外の構造まで考慮した上で、「真に混在している」と考えられる帖について追究することである。

3.9.1 「タンデムクラスタリング」に関する補足

まず、最初に本節では、データ変換、前処理について Raw data(条件 1)、平方根変換(条件 2)、Z 変換(条件 3)、QMⅢ（条件 4)、平方根変換＋QMⅢ（条件 5)、主成分分析(共分散行列)（条件 6)、平方根変換＋主成分分析(共分散行列)（条件 7)、主成分分析(相関行列)(条件 8)の 8 条件の一致度を比較する。3.6 と同様に、平方根変換＋QMⅢは前処理としてデータに平方根変換を施し、変換した値に QMⅢを適用することを、平方根変換＋主成分分析は、前処理としてデータに平方根変換を施し、変換した値の分散共分散行列に基づいて主成分分析を実行したことをそれぞれ意味する。その後のクラスター分析については、距離の選択としてユークリッド距離またはマンハッタン距離の 2 通り、グルーピングの方法としてウォード法と最長距離法の 2 通りを採用し、距離とグルーピングの組み合わせ 4 通りの全ての方法の結果を比較した。

タンデムクラスタリングの事前分析における QMⅢや主成分分析等の適用にあたっては、源氏物語 54 帖の多次元空間における布置を把握するためにプロットの一覧も作成した。3 章の付録図 1 に条件 4 における 5 次元までの座標空間の各 2 次元平面の表示(座標軸のスケールは各次元の固有値の大きさに対応させた。以下同様)を、3 章の付録図 2 に条件 5 における 5 次元までの座標空間の各 2 次元平面の表示を、3 章付録図 3 に条件 6 における 4 次元までの座標空間の各 2 次元平面の表示を、3 章付録図 4 に条件 7 における 6 次元までの座標空間の各 2 次元平面の表示を、3 章付録図 5 に条件 8 における 3 次元までの座標空間の各 2 次元平面の表示をそれぞれ載せた。

さらに、前処理で QMⅢや主成分分析を適用する場合には、3.3 節の図 2(条件 4)の他に 3 章付録 2 の付録図 6(条件 5)、付録図 7(条件 6)、付録図 8(条件 7)、付録図 9(条件 8)の固有値のスクリープロットを参照し、累積寄与率は 40%、65%、80%を目安に 3 通りの条件を設定した。

以上の 8 条件(次元数を含めると 19 通り)について 4 種のクラスター分析を適用した結果の一致度の一覧が表 5 である。ここで表示する一致度は基準とするグループを A,B,C,D の 4 つとした場合と{A,C},{B,D}のように 2 つのグループずつをまとめ 2 つとした場合の 2 通りを計算している(条件 1,2,4,5 は表 4 より再掲載)。本章の付録においては、主に各条件におけるユークリッド距離とウォード法を適用した樹形図の分析結果を載せる(付録の図 10 から図 27 参照)。最短距離法については表 4 と同様に村上･今西[1999]の 4 分類から逸脱し過ぎている

ので表 5 には結果を示さないが、参考までに付録図 28、付録図 29 に樹形図を例示した。

これら 8 条件の一致度を比較した動機は以下の通りである。

条件 1 においては各助動詞の分散の情報を分析に用いているが、条件 2 においては分散の情報を補正するとともに分布を正規分布に近づけている。さらに、条件 3 においては分散の情報を完全に取り除いている。よって、条件 1～3 を比較することで、各助動詞の分散の情報がクラスタリングに与える影響を測る。さらに、条件 6～条件 8 までは、条件 1～3 それぞれを主成分分析で次元圧縮した結果に対応する。条件 1～3 の結果と条件 6～8 の結果を比較することで、2 章で述べたタンデムクラスタリングが、村上･今西(1999)のデータで有効かどうかを考察する。

条件 4 と条件 5 では、数量化Ⅲ類による次元圧縮が村上･今西(1999)のデータにどのように影響を与えるかを考察するために設けた。(数量化Ⅲ類は、データの値が正であるような頻度表を想定しているため、Z 変換＋QMⅢは行わなかった)

統計的な分析結果として表 5 から、以下のようなことが観察できる。

第一に、2 章で述べたタンデムクラスタリングを行うべきか、ということを考えると、条件 1 と条件 4、条件 2 と条件 5 を比較すると QMⅢの場合には累積寄与率をあげると一致度はむしろ単調に悪くなり、上位の少数の次元の解が明確によいことがわかる。これは、主成分分析では、個体の変数への反応を変数間の相関係数(共分散)に集約し、変数の空間布置として解を求めるのに対して、QMⅢでは、個体の変数選択パターンから個体と変数の相関に着目し、個体の空間布置と変数の空間布置を同時に求める、という次元圧縮の解法に違いがあり、この特徴によることが考えられる。また、条件 1 と条件 6、条件 2 と条件 7、条件 3 と条件 8 を比較すると主成分分析の場合は QMⅢとは異なり累積寄与率をあげると一致度は単調に悪くなるとは言えないが、タンデムクラスタリングがうまくいっているとも言えない。よって、第一点において共通することは、村上･今西(1999)の助動詞の比率のデータにおいては、全般的にタンデムクラスタリングはうまくいっておらず、Raw データや変数変換したデータにクラスター分析を適用した結果の方が一致度は良いということである。

第二に、仮にタンデムクラスタリングを行う場合、次元数の選択による結果の違いを見るため条件 1 と条件 6、条件 1 と条件 8 を比較すると、Raw データと、Raw データに Z 変換を適用した場合には、主成分分析を行って累積寄与率をあげても一致度は単調に高まっていくとは言えない。よってある次元をもって次元縮約を行い、クラスター分析を行った方がよいことが分かる。また、第一点で述べたように、条件 1 と条件 4 を比較すると、QMⅢを Raw データに適用する場合でも、寄与率もあげても一致度は単調に高まっていくとは言えず、最初の 2 次元をもって次元縮約を行い、クラスター分析を行った方が良い。

表5: 各手法でのクラスタリングの一致度の一覧表.

	付録図番号	前処理	用いた次元 (累積寄与率)	距離	ｸﾗｽﾀﾘﾝｸﾞ 手法	A,B,C,Dの 一致度[%]	(A,C)と(B,D)の 一致度[%]	分類の再検討を必要とする帖
条件1	3章付録図11	Raw data	21	Euclid	W	46.15(24/52)	75.93(41/54)	
			21	Manhattan	W	49.02(25/51)	85.19(46/54)	18, 23, 28, 39
			21	Euclid	D	32.69(17/52)	63.46(33/52)	
			21	Manhattan	D	38.46(20/52)	64.71(33/51)	
条件3	3章付録図12	Z変換	21	Euclid	W	46.30(25/54)	85.19(46/54)	18,23,39
			21	Manhattan	W	50.98(26/51)	68.52(37/54)	
			21	Euclid	D	48.00(24/50)	84.62(44/52)	
			21	Manhattan	D	43.75(21/48)	54.90(28/51)	
条件4	3章付録図13	QMIII	2(40.0%)	Euclid	W	48.15(26/54)	79.63(43/54)	23, 48, 49
			2(40.0%)	Manhattan	W	44.44(24/54)	75.93(41/54)	
			2(40.0%)	Euclid	D	40.00(18/45)	55.56(25/45)	
			2(40.0%)	Manhattan	D	39.22(20/51)	58.33(24/48)	
	3章付録図14		5(65.9%)	Euclid	W	51.28(20/39)	62.00(31/50)	
			5(65.9%)	Manhattan	W	35.42(17/48)	61.11(33/54)	
			5(65.9%)	Euclid	D	44.68(21/47)	70.21(33/47)	
			5(65.9%)	Manhattan	D	38.46(20/52)	60.87(28/46)	
	3章付録図15		20	Euclid	W	53.19(25/47)	58.33(28/48)	
			20	Manhattan	W	33.33(16/48)	50.00(25/50)	
			20	Euclid	D	39.22(20/51)	54.90(28/51)	
			20	Manhattan	D	44.23(23/52)	69.57(32/46)	
条件2	図5(3章付録図10)	平方根変換	21	Euclid	W	52.94(27/51)	92.31(48/52)	18, 23, 39
	図6		21	Manhattan	W	56.86(29/51)	90.38(47/52)	無し
	図7		21	Euclid	D	47.06(24/51)	75.00(30/40)	
	図8		21	Manhattan	D	60.00(27/45)	72.73(32/44)	
条件5	3章付録図16	平方根変換 +QMIII	2(38.7%)	Euclid	W	47.06(24/51)	74.07(40/54)	
			2(38.7%)	Manhattan	W	43.14(22/51)	76.92(40/52)	
			2(38.7%)	Euclid	D	42.00(21/50)	76.47(39/51)	
			2(38.7%)	Manhattan	D	31.91(15/47)	58.14(25/43)	
	3章付録図17							
			5(68.8%)	Euclid	W	43.14(22/51)	53.85(28/52)	
			5(68.8%)	Manhattan	W	45.45(20/44)	55.77(29/51)	
			5(68.8%)	Euclid	D	49.02(25/51)	72.55(37/51)	
			5(68.8%)	Manhattan	D	44.23(23/52)	68.63(35/51)	
	3章付録図18		20	Euclid	W	53.19(25/47)	63.83(30/47)	
			20	Manhattan	W	45.10(23/51)	68.63(35/51)	
			20	Euclid	D	52.27(23/44)	54.90(28/51)	
			20	Manhattan	D	39.22(20/51)	64.71(33/51)	
条件6	3章付録図19	PCA(VCov)	2 (48.3%)	Euclid	W	51.85(28/54)	79.63(43/54)	18, 23, 28, 48
			2 (48.3%)	Manhattan	W	42.59(23/54)	75.93(41/54)	
			2 (48.3%)	Euclid	D	50.00(26/52)	61.11(33/54)	
			2 (48.3%)	Manhattan	D	46.15(24/52)	73.08(38/52)	
	3章付録図20		4 (70.6%)	Euclid	W	50.00(26/52)	81.48(44/54)	18, 23, 24, 28
			4 (70.6%)	Manhattan	W	48.89(22/45)	75.93(41/54)	
			4 (70.6%)	Euclid	D	46.94(23/49)	69.23(36/52)	
			4 (70.6%)	Manhattan	D	44.44(24/54)	81.40(35/43)	
	3章付録図21		6 (81.1%)	Euclid	W	50.00(26/52)	75.93(41/54)	
			6 (81.1%)	Manhattan	W	50.00(26/52)	64.81(35/54)	
			6 (81.1%)	Euclid	D	48.00(24/50)	57.69(30/52)	
			6 (81.1%)	Manhattan	D	33.33(17/51)	51.92(27/52)	
条件7	3章付録図22	平方根変換+ PCA(VCov)	2 (42.1%)	Euclid	W	55.77(29/52)	68.52(37/54)	
			2 (42.1%)	Manhattan	W	47.06(24/51)	68.52(37/54)	
			2 (42.1%)	Euclid	D	52.94(27/51)	56.86(29/51)	
			2 (42.1%)	Manhattan	D	42.86(21/49)	58.82(30/51)	
	3章付録図23		6 (72.5%)	Euclid	W	56.25(27/48)	78.85(41/52)	18, 23, 24, 28, 35, 40
			6 (72.5%)	Manhattan	W	52.94(27/51)	86.27(44/51)	18, 23, 39
			6 (72.5%)	Euclid	D	39.58(19/48)	56.86(29/51)	
			6 (72.5%)	Manhattan	D	59.09(26/44)	52.27(23/44)	
	3章付録図24		8 (80.8%)	Euclid	W	45.10(23/51)	80.77(42/52)	18, 23
			8 (80.8%)	Manhattan	W	52.94(27/51)	79.63(43/54)	無し
			8 (80.8%)	Euclid	D	47.06(24/51)	64.71(33/51)	
			8 (80.8%)	Manhattan	D	42.86(21/49)	66.67(34/51)	
条件8	3章付録図25	PCA(R)	3 (43.2%)	Euclid	W	51.92(27/52)	83.33(45/54)	39
			3 (43.2%)	Manhattan	W	50.00(26/52)	81.48(44/54)	18,39
			3 (43.2%)	Euclid	D	50.98(26/51)	80.56(29/36)	
			3 (43.2%)	Manhattan	D	54.90(28/51)	82.05(32/39)	
	3章付録図26		8 (73.8%)	Euclid	W	46.00(23/50)	85.19(46/54)	18,23,39
			8 (73.8%)	Manhattan	W	48.98(24/49)	79.63(43/54)	18, 39
			8 (73.8%)	Euclid	D	56.52(26/46)	58.82(30/51)	
			8 (73.8%)	Manhattan	D	45.83(22/48)	70.59(36/51)	
	3章付録図27		10 (81.5%)	Euclid	W	51.92(27/52)	85.53(45/52)	8, 18, 23
			10 (81.5%)	Manhattan	W	56.86(29/51)	75.93(41/54)	
			10 (81.5%)	Euclid	D	36.54(19/52)	57.78(26/45)	
			10 (81.5%)	Manhattan	D	59.46(22/37)	68.09(32/47)	

平方根変換+QMIIIは平方根変換を先にRawデータに適用し、変換した値にQMIIIを適用したことを示す。またクラスタリング手法のWはWard法、Dは最長距離法を示す。一致度の()の中は、分母が一致度の計算の際に考慮した帖の数、分子が実際に一致した数を意味する。

3.9.2 Neighbor-Net による「混在」した帖の分析

前節では、様々な手法(変数変換、タンデムクラスタリング等)をクラスター分析において試し、試行錯誤した結果を「一致度」の観点から表 5 に示したが、クラスター分析を用いている以上、視覚化された構造はデータのもっとも強い構造である。よって、「混在」していると判断された帖が、弱い構造を考慮してもなお混在していると判断される帖(これを「真に混在している帖」と呼ぶ)なのか、それとも弱い構造を考慮すると、必ずしも「混在」しているとは言い切れない帖(「境界事例」とここでは呼ぶ)なのか、ということについてふるいにかける必要がある。

そこで、本節ではデータの強い構造だけではなく弱い構造も抽出することのできる Neighbor-Net (Huson & Bryant, 2006)を用いることによって、「真に混在している帖」について追究した。

Neighbor-Net から得られた図に対して、ヒューリスティックな図の切り方を幾つか試すことによって、「真に混在している帖」はそれらの切り方をしてもなお「混在」している帖と考えることができ、「境界事例」はそれらの切り方によって、「混在」の状態が解消される帖のことを指す。分析例について図 9 に示す。

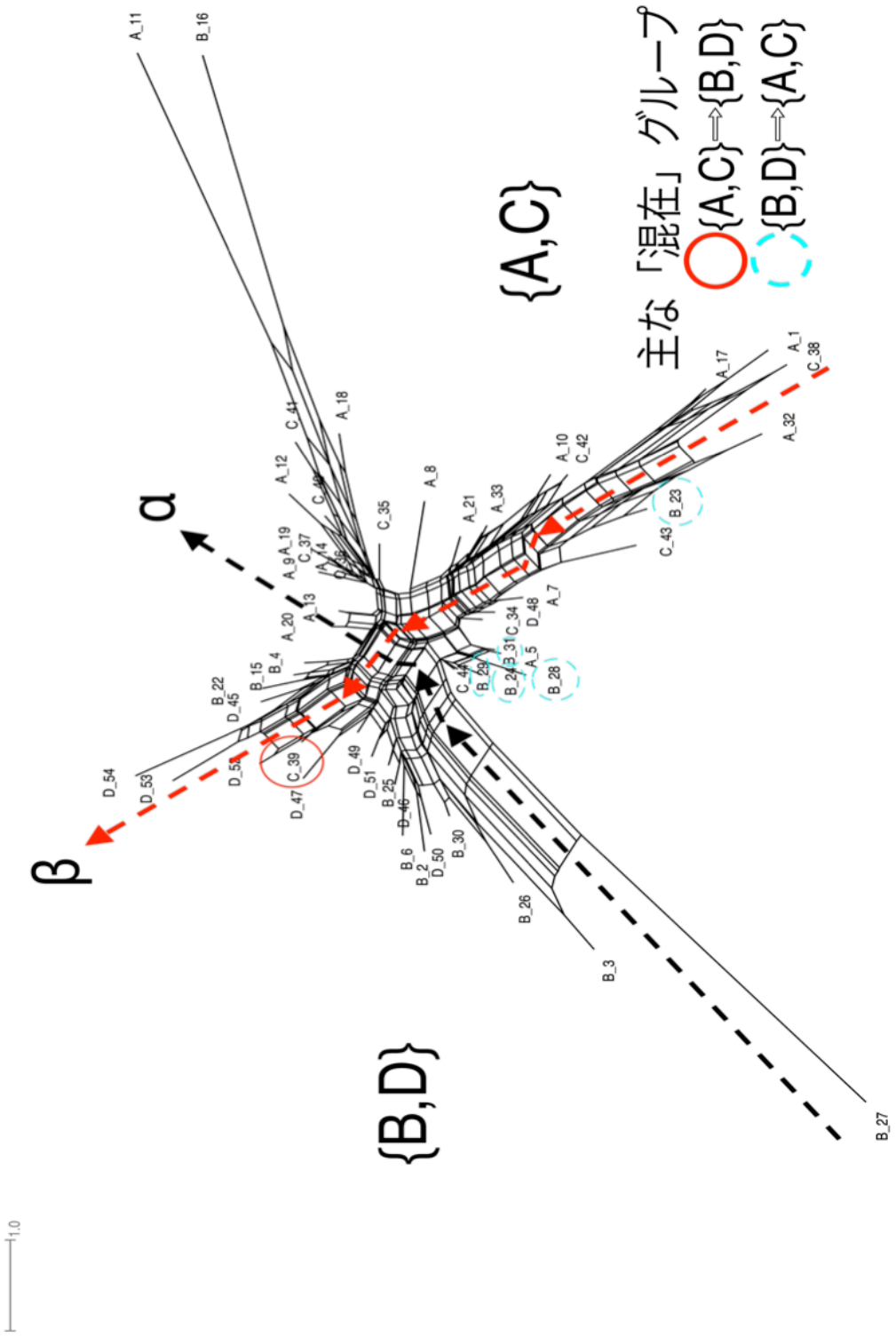


図 9: 村上・今西(1999)の助動詞のデータについて相関行列を用いた主成分分析を適用し、第 3 次元目までの座標をもとにユークリッド距離で距離行列を計算し、Neighbor-Net を適用した結果。

図 9 は村上・今西(1999)のデータに相関行列に基づいた主成分分析を適用し、第 3 次元目までの座標をもとにユークリッド距離で距離行列を計算し、Neighbor-Net を適用した結果である。ウォード法を適用したクラスター分析において主に、{A,C}と文献学的には分類されていたもので、{B,D}に混在していたものを◯で、{B,D}と文献学的には分類されたいたもので、{A,C}に混在していたものを点線の◯で、それぞれ表した。Neighbor-Net による分析では、クラスター分析の結果に対応する切り方は図 9 においてαの切り方である。しかし、図 9 をβの切り方で切ることによって、23 帖、24 帖、28 帖、29 帖、31 帖、48 帖は混在の状態ではなくなり、{B,D}グループに属する。よって、これら 6 帖は「境界事例」と考えられる。ただし、βの切り方でも◯で囲った 39 帖は混在状態を解消することができず、「真に混在している帖」と考えられる。

表 5 で「一致度」が高かった 16 条件について Neighbor-Net を行い、「真に混在している帖」を検討したところ、8 帖:1 回(花宴)、18 帖:12 回(松風)、23 帖:11 回(初音)、24 帖:2 回(胡蝶)、28 帖:4 回(野分)、35 帖:1 回(若菜下)、39 帖:8 回(夕霧)、40 帖:1 回(御法)、48 帖:2 回(早蕨)、49 帖:1 回(宿木)という結果となった。この結果から、1)松風(18)は{A}に属していると既存の文献学的研究では考えられているが、本研究の結果からは{B,D}グループに属している可能性を検討する必要がある。2)初音(23)は{B}に属していると既存の文献学的研究では考えられているが、本研究の結果からは{A,C}グループに属している可能性を検討する必要がある。3)夕霧(39)は{C}に属していると既存の文献学的研究では考えられているが、本研究の結果からは{B,D}グループに属している可能性を検討する必要がある。特に夕霧については、文献学的研究においても{B,D}グループに属するのではないかということが、半世紀前に藤井(1966)において主張されているため、本研究から得られた源氏物語成立論に関する示唆は、文献学的にも無根拠なものではない。以上の 3 点が助動詞だけの情報を使った統計学的な研究から文献学へ提言できる最も確実な提案と言える。

３章付録

３章付録 1: 平方根変換と対数変換の比較

本節では村上・今西(1999)データに対して予備分析として変数変換(Box-Cox 変換に属する 2 つの方法)に関する詳細を示す。Box-Cox 変換とは以下の式で表される。X が元のデータであり、Y が変換後のデータである。

$$Y = \begin{cases} \dfrac{X^d - 1}{d} \ (d \neq 0) \\ \log X \ (d \to 0) \end{cases}$$

上式の d としてよく用いられるのは、d→0 の場合(つまり対数変換$Y = \log X$)、もしくは d=0.5(つまり平方根変換$Y = \frac{\sqrt{X}-1}{2}$、ただし本研究では、定数項と定数倍を除いた$Y = \sqrt{X}$として計算した)の場合があり、以降では、実際のデータの性質も参考にしながら、対数変換と平方根変換のどちらが村上・今西(1999)のデータに適しているかを検討する。

永田(2005, pp.53-54)によれば、分散安定化変換において、対数変換を採用する場合は、元のデータの標準偏差と平均が一定の比になっている必要がある。また、平方根変換を採用する場合には元のデータの分散と平均が一定の比になっている必要がある。

以下、3 章付録表 1 に村上・今西(1999)のデータの基本的な統計量を示す。

3 章付録表 1: 村上・今西(1999)の源氏物語の各帖の助動詞の出現頻度に関する統計量一覧.

助動詞	平均	分散	分散と平均の比	標準偏差と平均の比
「ず」	0.0145223	0.0000032	0.0002232	0.1239665
「む」	0.0117019	0.0000154	0.0013140	0.3350919
「たり」	0.0116281	0.0000085	0.0007302	0.2505890
「けり」	0.0098042	0.0000069	0.0007006	0.2673210
「なり」	0.0096402	0.0000067	0.0006989	0.2692625
「り」	0.0091697	0.0000074	0.0008028	0.2958800
「ぬ」	0.0083970	0.0000039	0.0004603	0.2341199
「き」	0.0078275	0.0000123	0.0015688	0.4476903
「べし」	0.0075489	0.0000035	0.0004616	0.2472695
「つ」	0.0036193	0.0000025	0.0006966	0.4387062
「る」	0.0040273	0.0000024	0.0005859	0.3814197
「す」	0.0032710	0.0000036	0.0011002	0.5799591
「めり」	0.0025044	0.0000009	0.0003491	0.3733650
「さす」	0.0018256	0.0000009	0.0004797	0.5125997
「らむ」	0.0017003	0.0000005	0.0002892	0.4124435
「らる」	0.0017326	0.0000004	0.0002550	0.3836249

「じ」	0.0012774	0.0000003	0.0002681	0.4581634
「けむ」	0.0012763	0.0000005	0.0003602	0.5312360
「まじ」	0.0013328	0.0000009	0.0006929	0.7210261
「まし」	0.0010765	0.0000007	0.0006874	0.7990600
「まほし」	0.0005415	0.0000001	0.0002391	0.6644674

3 章付録表 1 の右 2 列の値を比較しても明らかなように、標準偏差と平均の比と、分散と平均の比とを比較してもどちらがより安定しているかは、判然としない。

ここで、源氏物語 54 帖の任意の二つの i 帖と j 帖の、21 の助動詞の出現頻度によるユークリッド距離d_{ij}を、i 帖の k 番目の助動詞の各帖の総語数に対する相対出現頻度値をa_{ik}として、以下のように定義する。

$$d_{ij} = \sqrt{\sum_{k=1}^{21}\left(a_{ik} - a_{jk}\right)^2} \qquad \{i, j = 1 \cdots 54\}$$

助動詞の出現頻度分布のデータにクラスター分析を施す場合は、この助動詞間の距離データ$\{d_{ij}\}$が用いられる。しかし、3.5 節で述べた工夫の第一点として、a_{ik}などを対数変換か平方根変換することを考える。

助動詞の 54 帖にわたる出現頻度値の分布、合計 21 の分布について、対数変換と平方根変換を適用し、正規分布により近い変換を与えたものをそれぞれの助動詞の分布として選び、対数変換がより多く選ばれれば、対数変換を全体のデータに適用することとする。同様に平方根変換がより多く選ばれれば、平方根変換を全体のデータに適用することとする。

変換した助動詞の分布の正規分布への当てはまりの良さに関しては、R の mclust のパッケージの中にある、データを正規分布にフィッティングさせる関数 densityMclust を利用し、指標としてベイズ情報量基準(Bayesian Information Criterion)(Schwarz, 1978)を用いた[16]。モデルの良さを表す指標である BIC が小さい値を与えた変換が統計学的には望ましい変換といえる[17]。

$$BIC = -2 \cdot (\text{最大対数尤度}) + (\text{自由度}) \cdot \log(\text{サンプルサイズ})$$

である。

[16] 以下 BIC と略す。

[17] R の mclust のパッケージの中では、BIC の符号の向きが逆である。そのため、通常は BIC が最小のものを選べば良いが、mclust を使った計算では最大のものを選ばなければならない。本稿では煩雑さをさけるため、mclust の BIC の符号の向きを逆転させ、通常通り、解釈ができるようにした。R の mclust で BIC を求める際には留意する必要がある。

ある助動詞の分布を対数変換し正規分布を当てはめた場合と、平方根変換し正規分布を当てはめた場合の BIC は直接比較することはできない。なぜならば、BIC の中の項をなす尤度の尺度が対数変換と平方根変換では変わってしまっているからである。詳細については、竹内編(1989)などを参照。本稿の 3 章付録表 2 では、対数変換の尺度と平方根変換の尺度を変換前の尺度に合わせるように補正を行った BIC を記す[18]。以下、3 章付録表 2 がにそれぞれの助動詞の分布に対して、変換前、対数変換後、平方根変換後のデータに正規分布モデルを当てはめて BIC を比較したものである[19]。

3 章付録表 2: 源氏物語の各助動詞の出現頻度の分布に対数変換と平方根変換を適用した際の BIC の値.

助動詞	変換前の BIC	対数変換の BIC	平方根変換の BIC
「ず」	721.07	723.38	721.79
「む」	801.49	807.93	801.86
「たり」	773.08	768.43	769.92
「けり」	761.63	757.50	757.97
「なり」	738.85	735.65	736.72
「り」	765.37	768.98	765.03
「ぬ」	730.57	733.02	730.71
「き」	791.61	793.78	787.18
「べし」	724.98	724.12	729.77
「つ」	707.11	710.43	710.84
「る」	703.93	695.46	696.98
「す」	707.58	698.22	703.78
「めり」	650.32	663.84	655.59
「さす」	640.00	640.80	639.82
「らむ」	616.71	620.16	621.88
「らる」	613.46	619.86	613.37
「じ」	596.25	603.60	596.78
「けむ」	604.79	608.33	605.54
「まじ」	612.80	620.39	618.00
「まし」	626.28	608.63	615.78
「まほし」	537.05	533.61	531.26

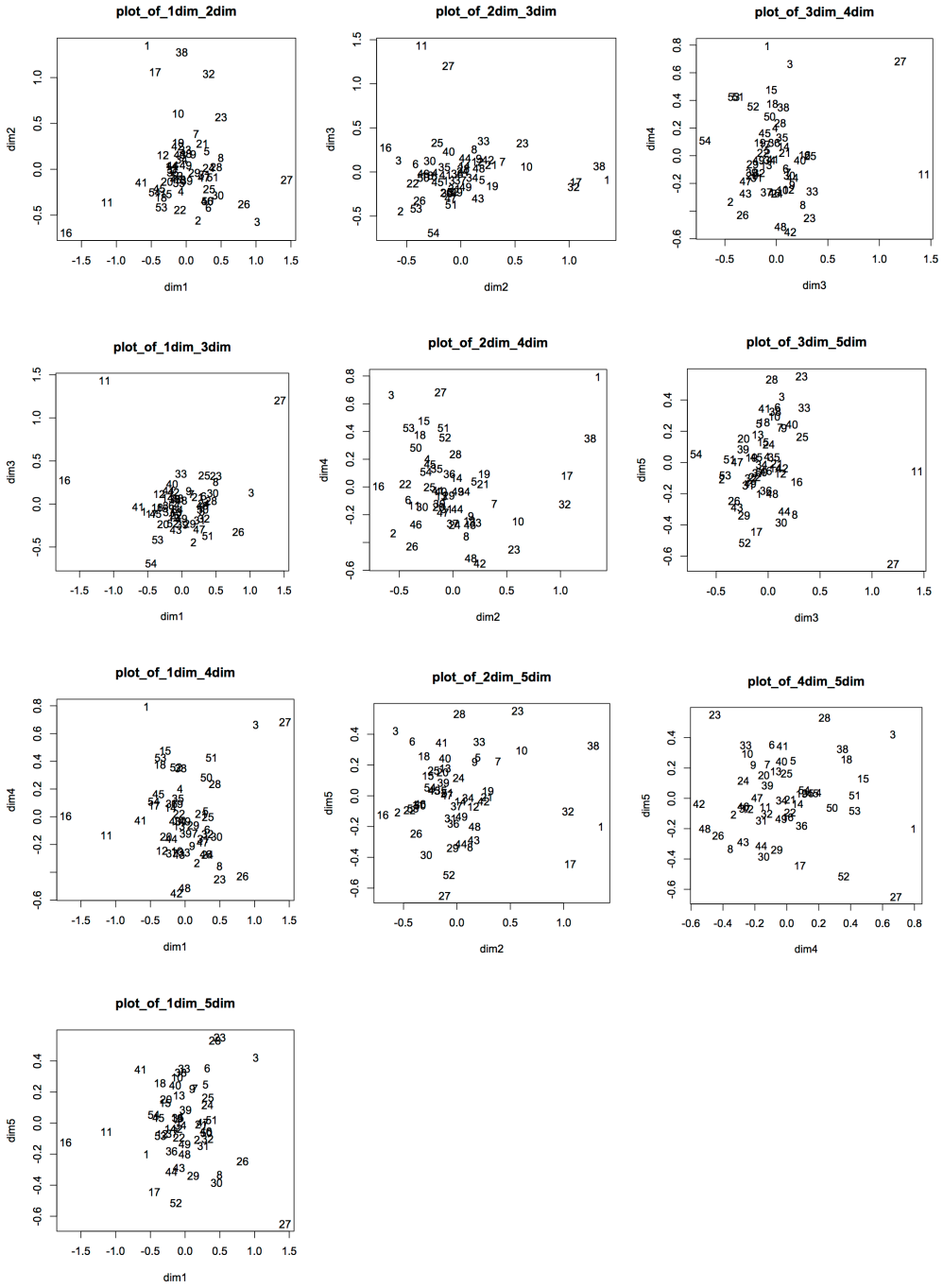
[18] 補正項の具体的な計算については、Burnham and Anderson(2002)を参考にした。

[19] 本研究では、a_{ij}を 10 万語あたりの値に直したものに、densityMclust を適用し、BIC を計算した。ただし、a_{ij}がゼロの値をとる場合には、j 番目の助動詞の値の中で最小の値の二分の一の値を便宜的に代入して BIC を計算した。実際のクラスター分析に使った値はこのようなゼロセルの処理をしていないデータである。

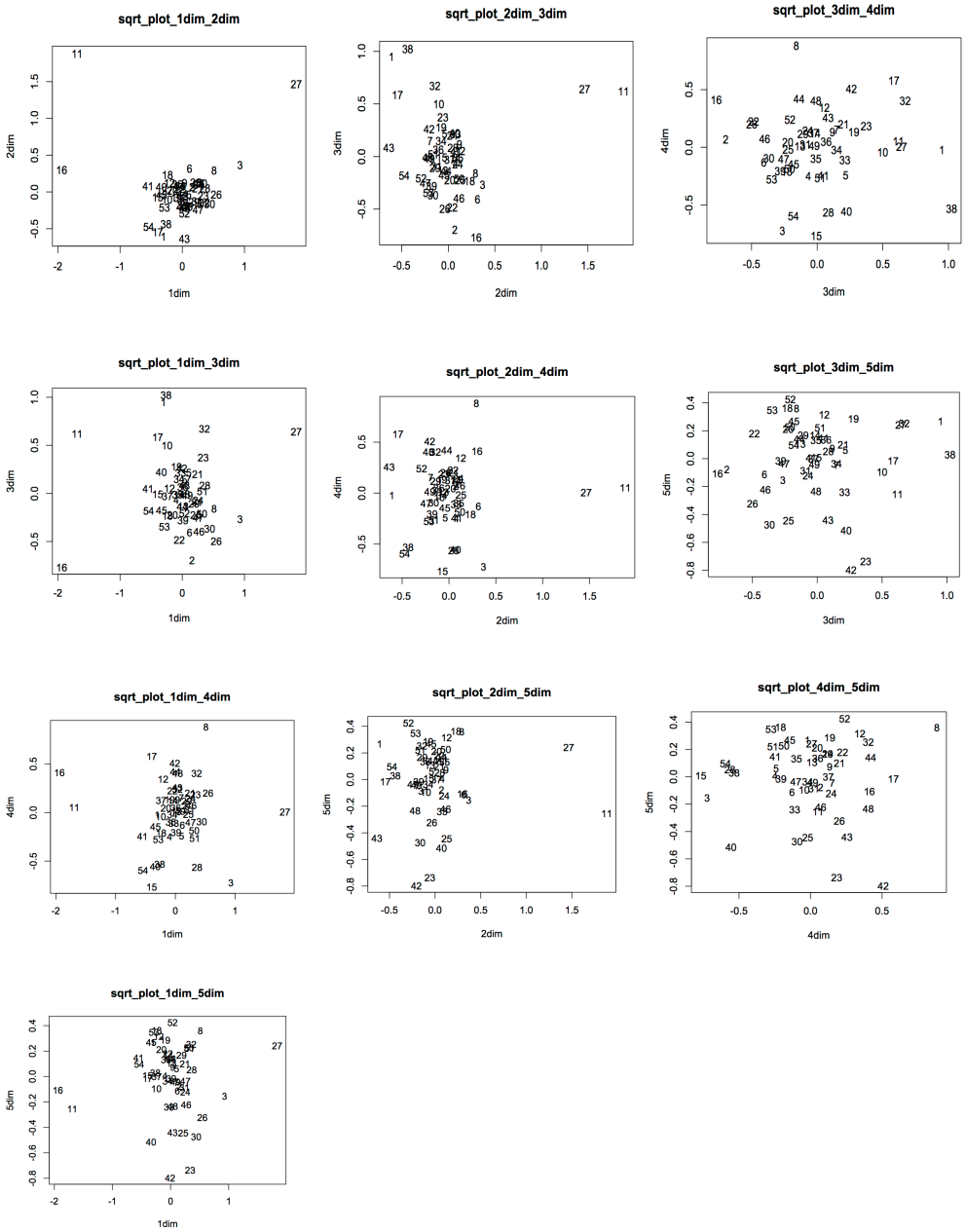
3 章付録表 2 の結果から、21 の助動詞のうち 12 の助動詞で、平方根変換が対数変換よりも適し、21 の助動詞のうち 13 の助動詞で変換前が対数変換よりも適し、21 の助動詞のうち 11 の助動詞で平方根変換が変換前よりも適しているという結果を得た。よって、本稿においては全体のデータに平方根変換を適用することとした。ただし、BIC は原理的には候補となっている条件の BIC のうち最も小さいものを選べばよいが、3 章付録表 2 の結果を見ると、変換前と対数変換と平方根変換の BIC の差はわずかなものが多い。

３章付録 2: タンデムクラスタリングを用いた際の事前分析に関する資料

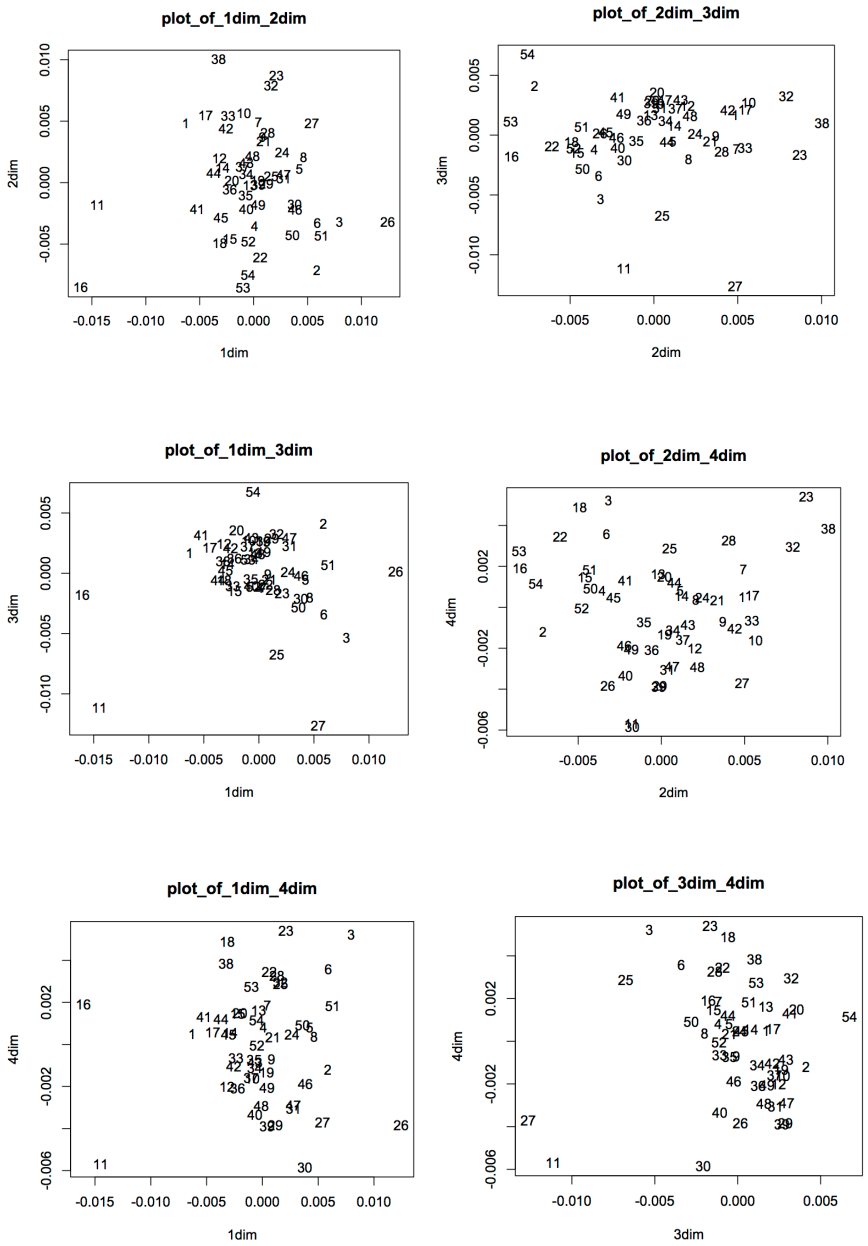
本節では、タンデムクラスタリングを用いる際に、事前分析としての数量化Ⅲ類や主成分分析における、個体のスコアのプロットや次元数選択のスクリープロットに関わる結果を提示する。



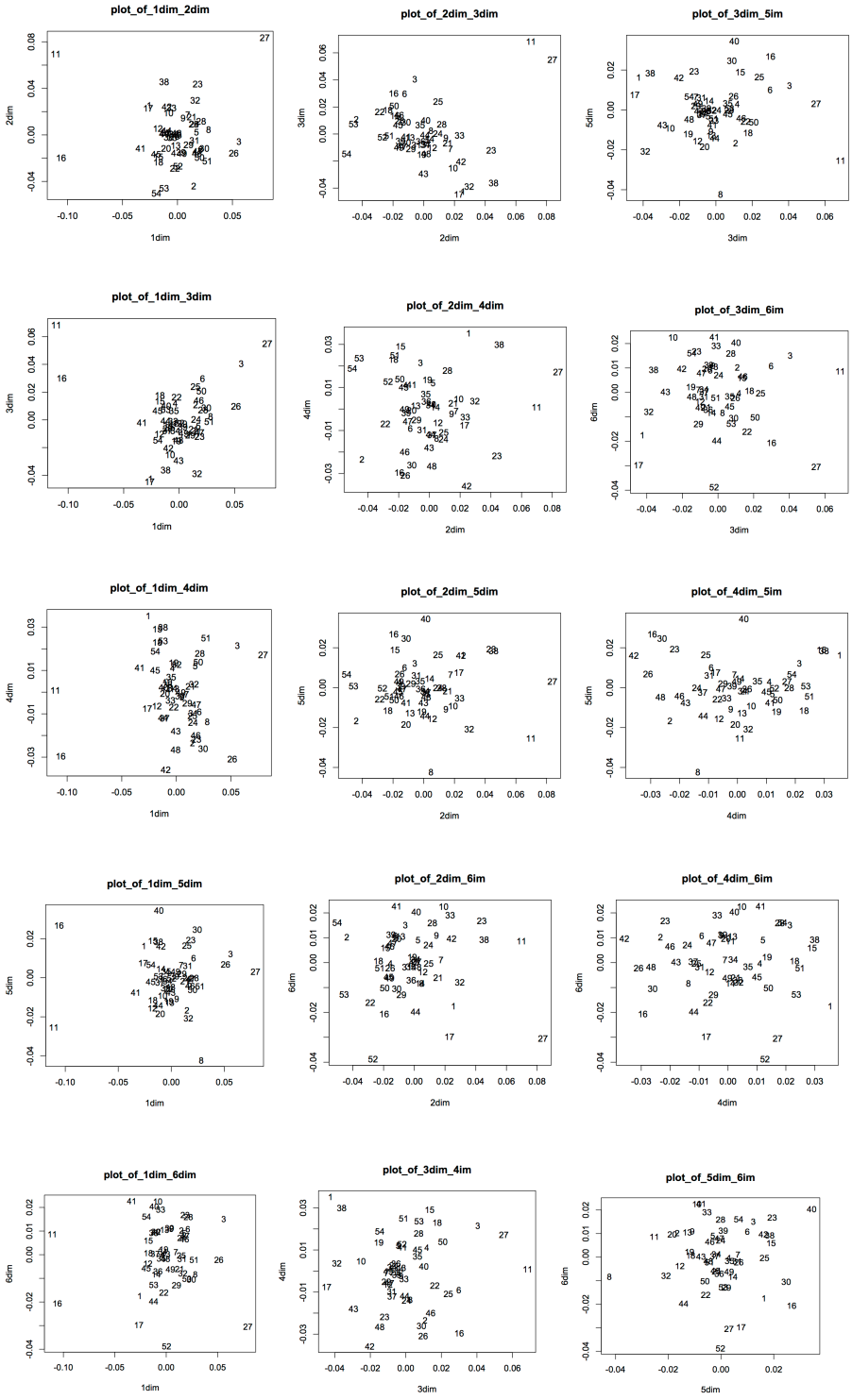
3 章付録図 1: Raw data に数量化Ⅲ類を適用した場合の、5 次元までの座標空間の各 2 次元の平面表示(座標軸のスケールは各次元の固有値の大きさに対応させた)



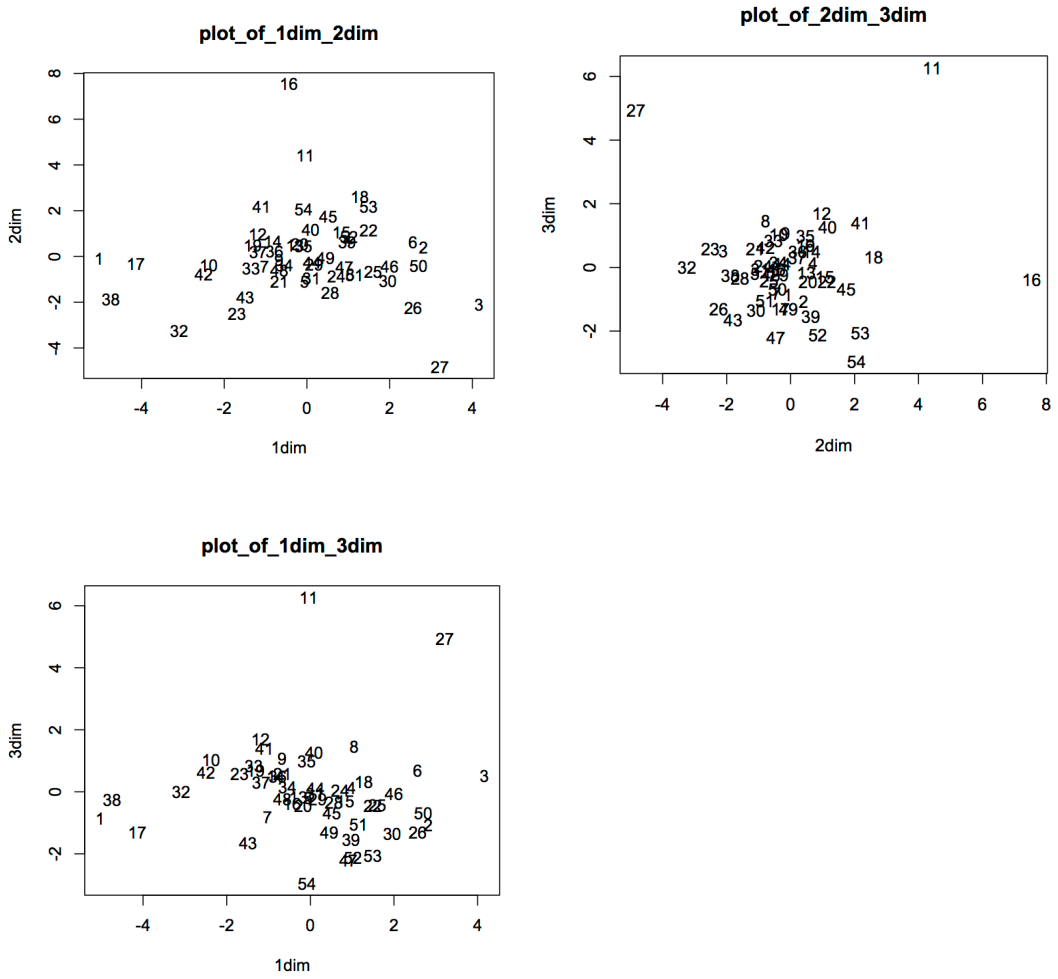
3 章付録図 2:Raw data に平方根変換を施した後に数量化Ⅲ類を適用した場合の、5 次元までの座標空間の各 2 次元の平面表示(座標軸のスケールは各次元の固有値の大きさに対応させた)



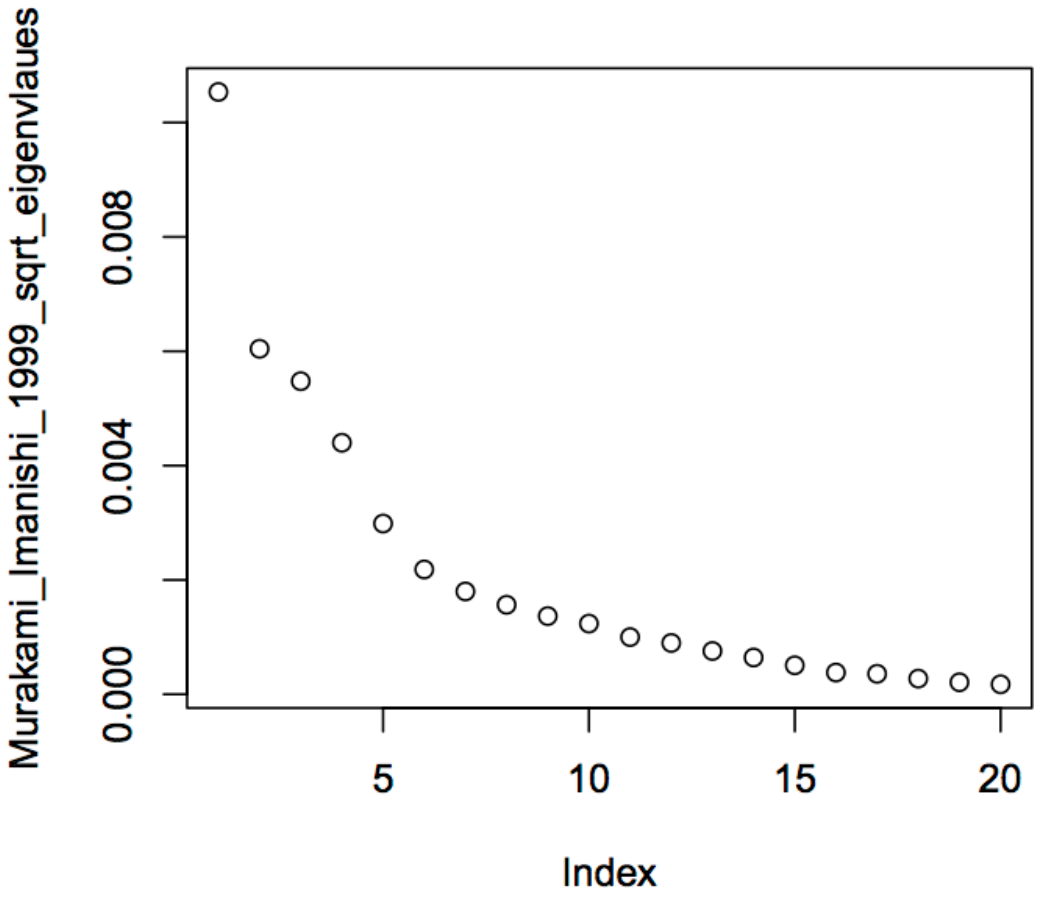
3 章付録図 3:Raw data の分散共分散行列をもとに主成分分析を適用し、4 次元までの座標空間の各 2 次元の平面表示したもの



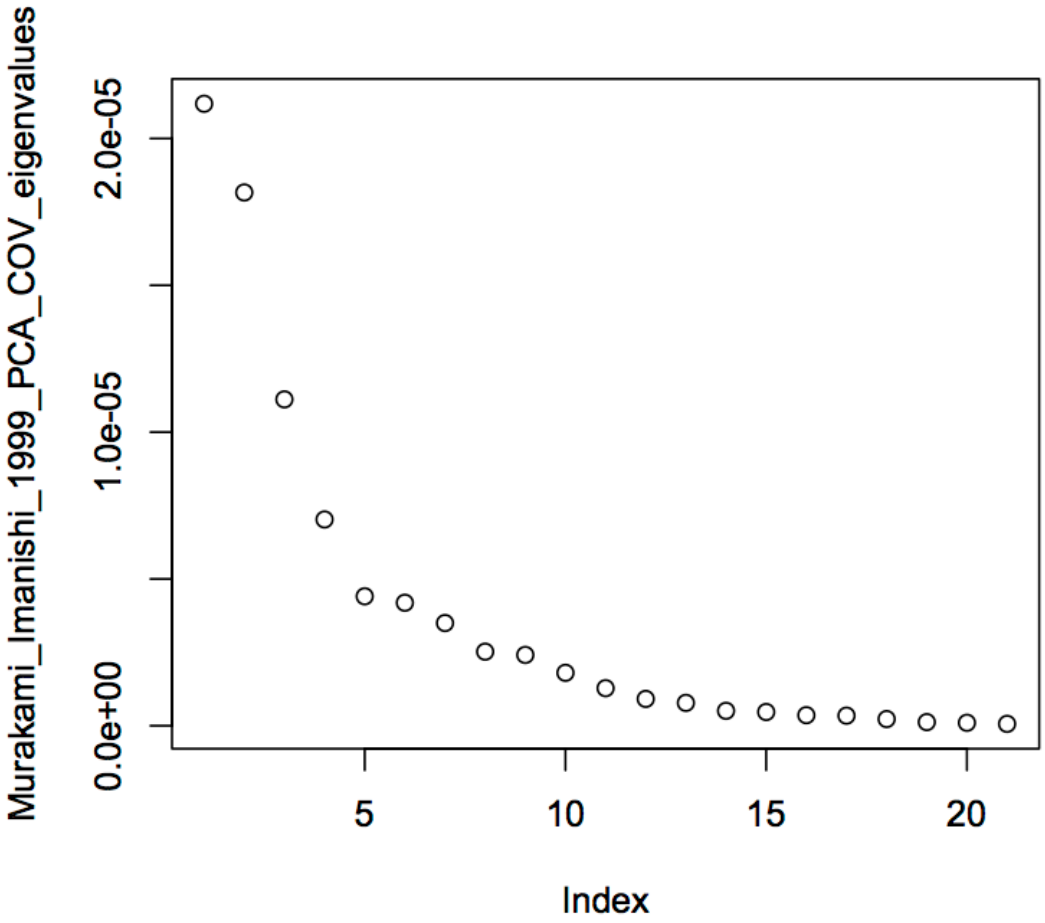
3 章付録図 4:Raw data に平方根変換を施した値の分散共分散行列をもとに主成分分析を適用し、6 次元までの座標空間の各 2 次元の平面表示したもの



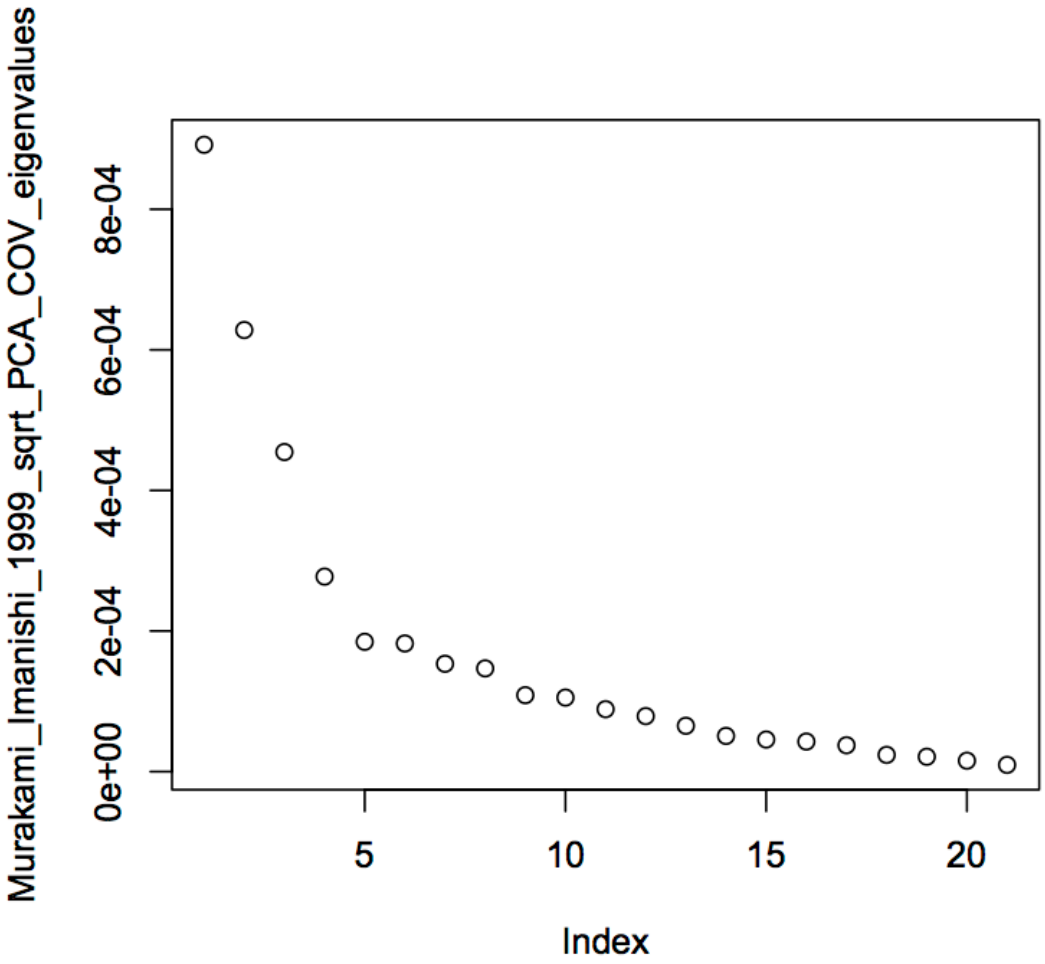
3 章付録図 5: Raw data の相関行列をもとに主成分分析を適用し、3 次元までの座標空間の各 2 次元の平面表示したもの



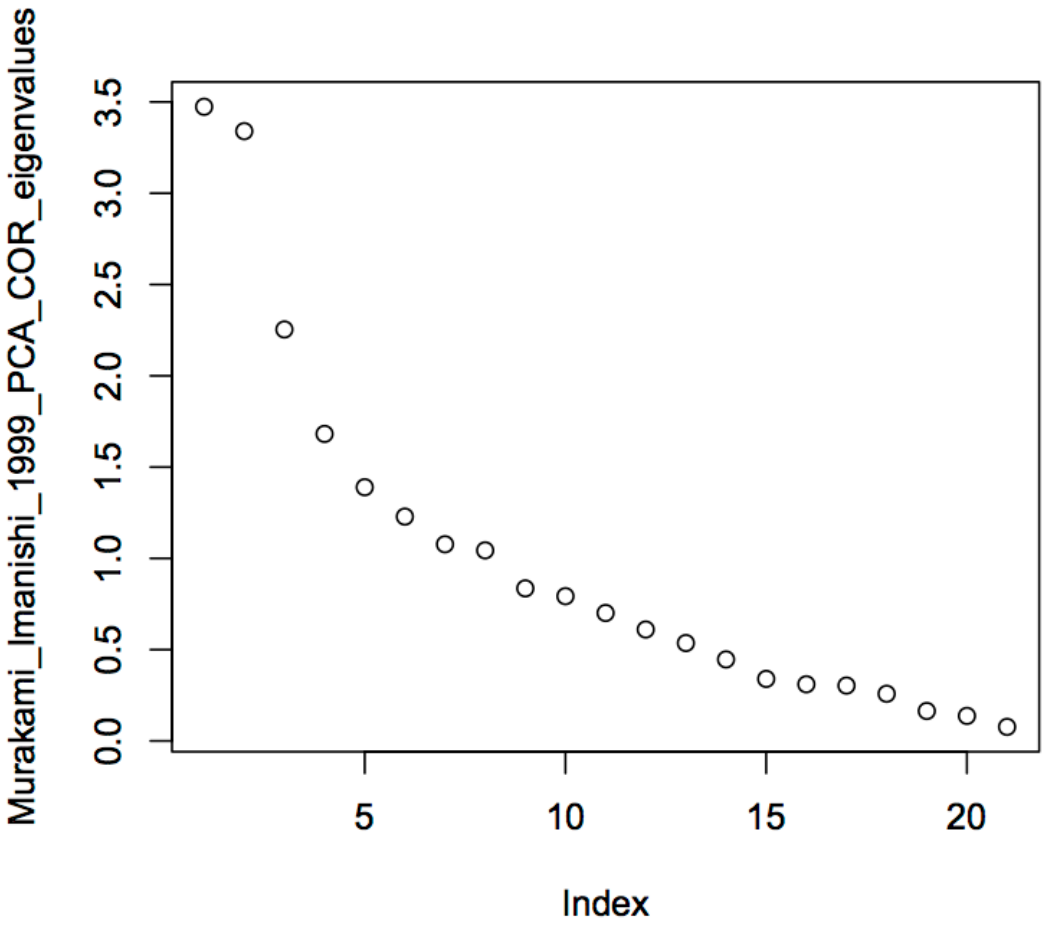
3 章付録図 6:Raw data に平方根変換を施した後に数量化Ⅲ類を適用し得られた固有値のスクリープロット。表 5 の条件 5 の次元数を検討するために用いた。



3 章付録図 7:Raw data の分散共分散行列をもとに主成分分析を適用し得られた固有値のスクリープロット。表 5 の条件 6 の次元数を検討するために用いた。



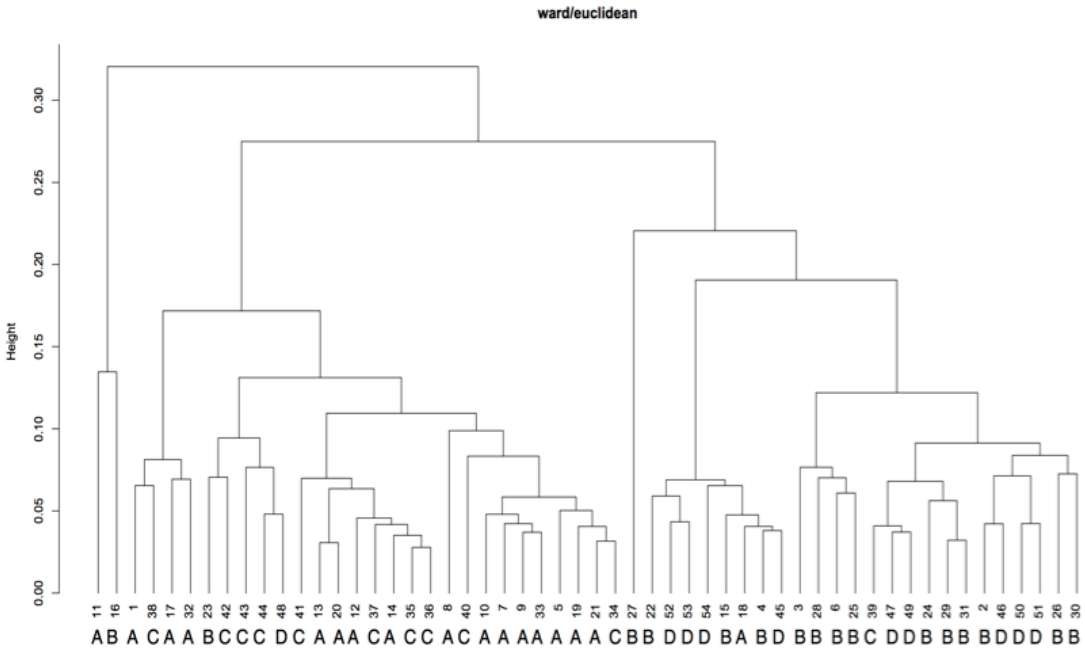
3 章付録図 8:Raw data に平方根変換を施した値の分散共分散行列をもとに主成分分析を適用し得られた固有値のスクリープロット。表 5 の条件 7 の次元数を検討するために用いた。



3 章付録図 9: Raw data の相関行列をもとに主成分分析を適用し得られた固有値のスクリープロット。表 5 の条件 8 の次元数を検討するために用いた。

３章付録 3: クラスタリングの結果の「一致度」による評価

次に、本文中では掲載しなかった村上･今西(1999)のデータにクラスター分析を適用したものを載せる。主に、ユークリッド距離とウォード法を適用した樹形図の分析を載せた(3 章表 4 参照)が、例えば、図 10 は平方根変換を施した値にユークリッド距離とウォード法に基づいてクラスタリングをした結果、図 11 は、村上・今西(1999)の素データにユークリッド距離とウォード法に基づいてクラスタリングをした結果である。以下、図 10 から図 27 は表 4、表 5 と対応させて参照されたい。なお、図 28、図 29 のように最短距離法を用いた結果はすべてクラスタリングがうまくいかなかった。「一致度」の定義については 3.5.4 節を参照されたい。



3 章付録図 10: 3 章本文図 5 再掲。Raw data に平方根変換を施した値からユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.15 周辺で切り、クラスターを左から 1,2,3,4,5,6 と番号付けると、

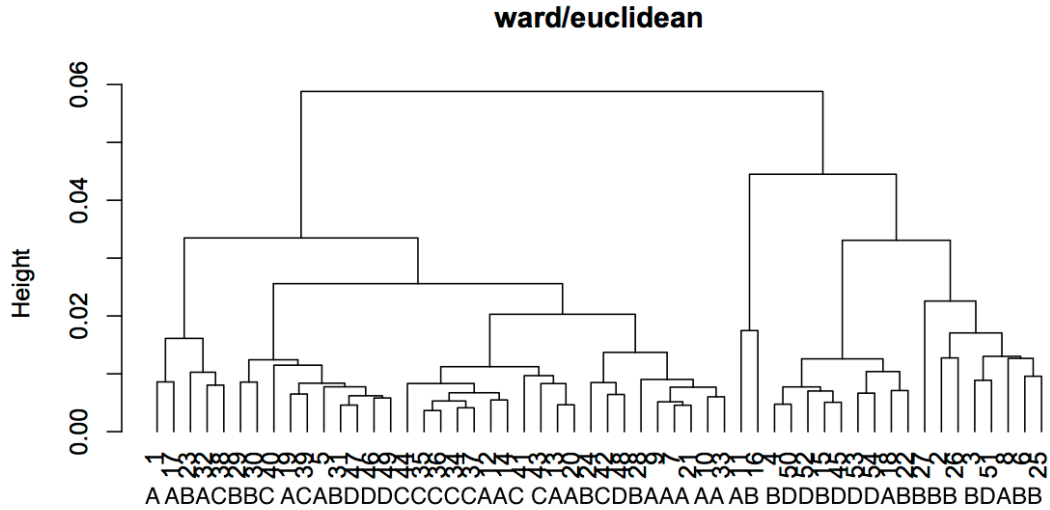
	1	2	3	4	5	6
A	1	3	12	0	1	0
B	1	0	1	1	3	10
C	0	1	9	0	0	1
D	0	0	1	0	4	5

このとき一致度は、
A(3): B(6): C(2): D(5)=12/23: 10/16: 1/4: 4/8, 全体として 27/51=52.94

さらにデンドログラムを高さ 6.5 周辺で切り、クラスター1 を除いて考えると

	2,3	4,5,6
A,C	25	2
B,D	2	23

このとき一致度は、
(A,C)(2,3): (B,D)(4,5,6)=25/27: 23/25,全体として 48/52=92.31



3 章付録図 11:Raw data からユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.03 周辺で切り、クラスターを左から 1,2,3,4,5 と番号付けると、

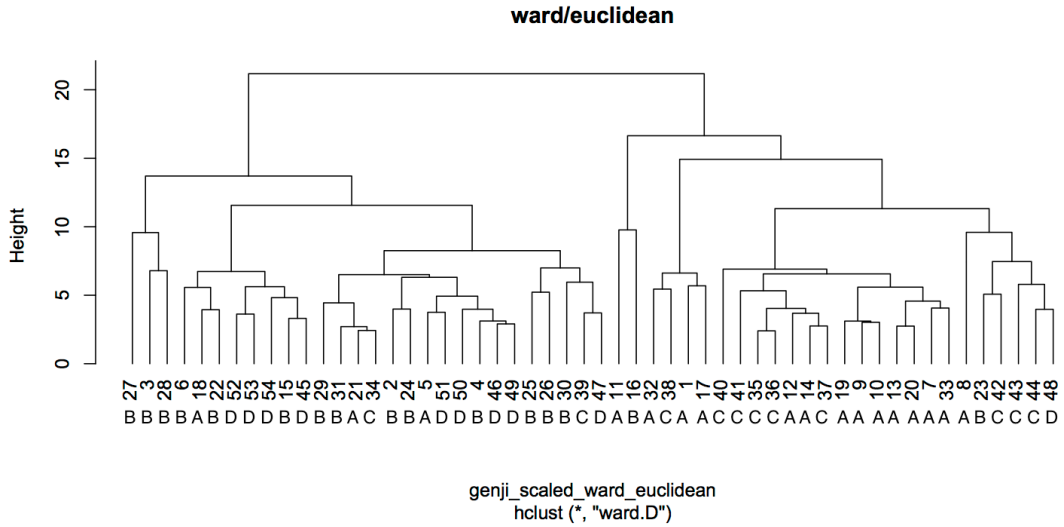
	1	2	3	4	5
A	3	11	1	1	1
B	1	5	1	3	6
C	1	10	0	0	0
D	0	4	0	5	1

このとき一致度は、
A(1): B(5): C(2): D(4)=3/5: 6/8: 10/30: 5/9,全体として 24/52=46.15

さらに、デンドログラムを高さ 0.05 周辺で切ると、

	1,2	3,4,5
A,C	25	3
B,D	10	16

このとき一致度は、
(A,C)(1,2): (B,D)(3,4,5)= 25/35: 16/19,全体として 41/54=75.93



3 章付録図 12:Raw data に Z 変換を施した値からユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 15 周辺で切り、クラスターを左から 1,2,3,4 と番号付けると、

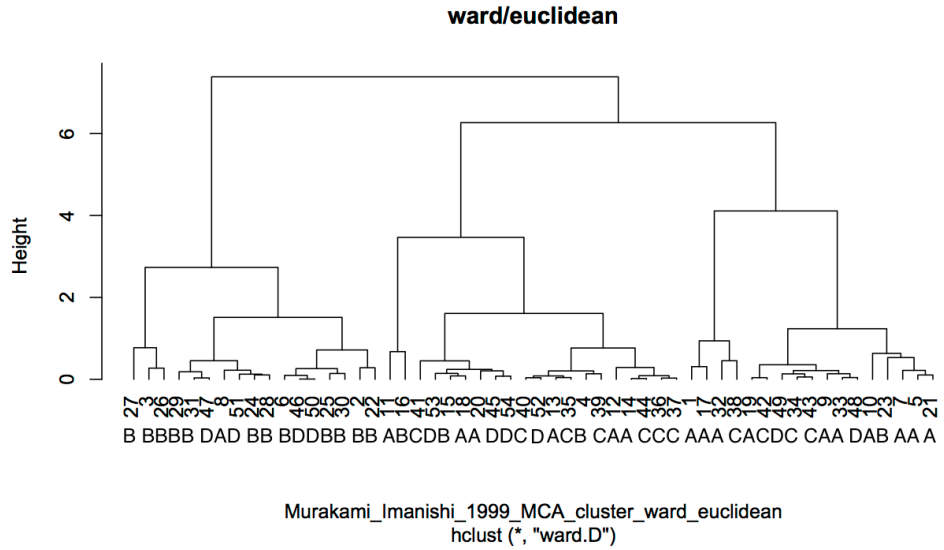
	1	2	3	4
A	3	1	3	10
B	14	1	0	1
C	2	0	1	8
D	9	0	0	1

このとき一致度は、
A(3): B(1): C(4): D(2)=3/4: 14/28: 8/20: 0/2,全体として 25/54=46.30

さらに、デンドログラムを高さ 18 周辺で切ったとき、

	1	2,3,4
A,C	5	23
B,D	23	3

このとき一致度は、
(A,C)(2,3,4): (B,D)(1)= 23/26: 23/28,全体として 46/54=85.19



3 章付録図 13: Raw data に数量化Ⅲ類を適用し、2 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 4 周辺で切り、クラスターに左から 1,2,3,4 と番号をつけたとき、

	1	2	3	4
A	1	6	3	7
B	12	3	0	1
C	0	7	1	3
D	4	4	0	2

このとき、一致度は

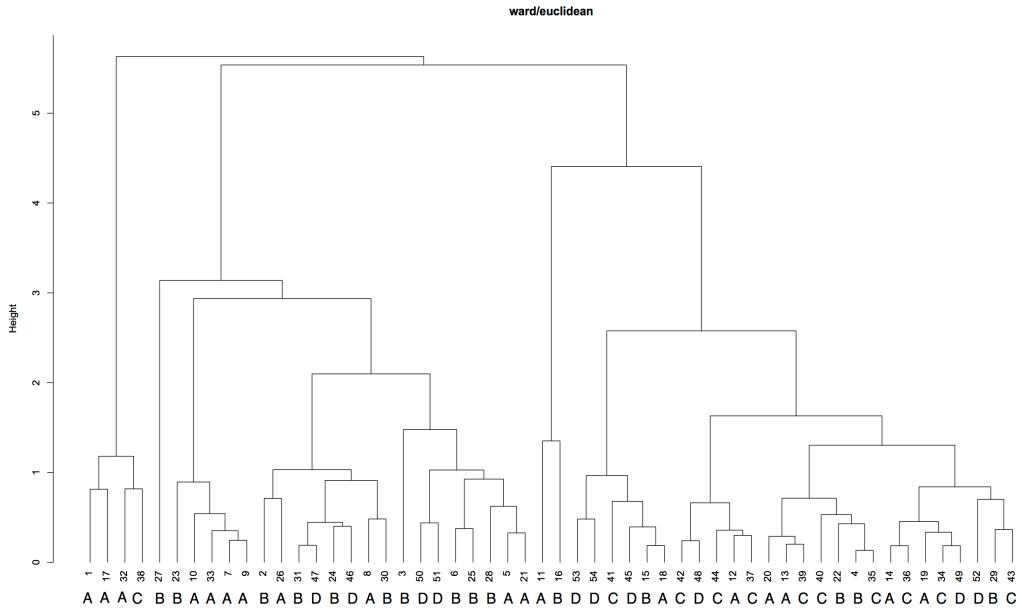
A(4):B(1):C(2):D(3)= 7/13: 12/17: 7/20: 0/4, 全体としては 26/54=48.15

さらに、デンドログラムを高さ 6.5 周辺で切ったとき、

	1	2,3,4
A,C	1	27
B,D	16	10

このとき、一致度は

(A,C)(2,3,4): (B,D)(1)= 27/37: 16/17, 全体としては 43/54=79.63



3 章付録図 14: Raw data に数量化Ⅲ類を適用し、5 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 2 周辺で切り、クラスターを左から 1,2,3,4,5,6,7,8 と番号付けたとき、

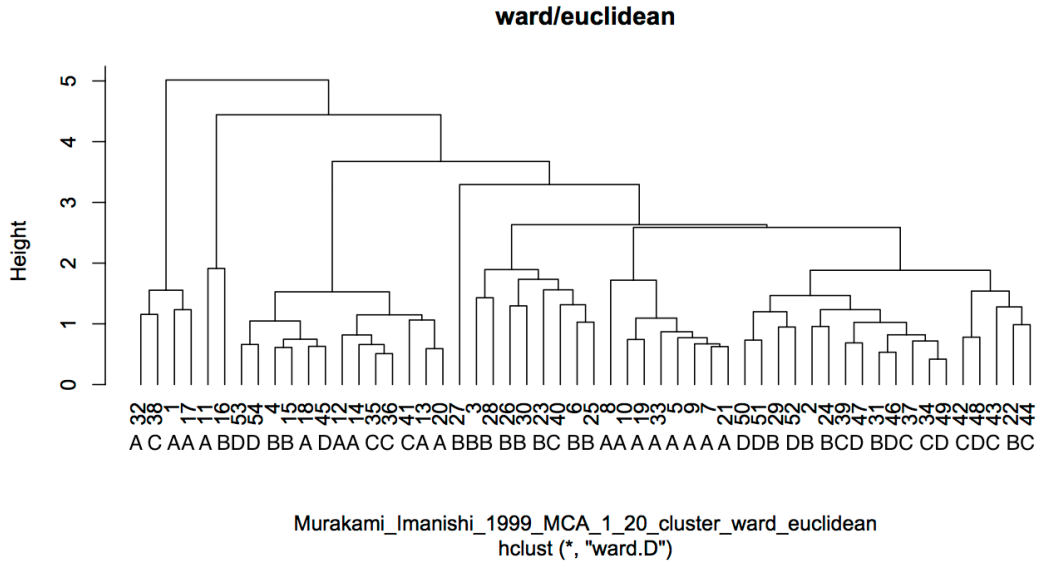
	1	2	3	4	5	6	7	8
A	3	0	4	2	2	1	1	5
B	0	1	1	4	4	1	1	3
C	1	0	0	0	0	0	1	9
D	0	0	0	2	2	0	3	3

このとき一致度は、
A(3): B(5): C(8): D(7)= 4/5: 4/8: 9/20: 3/6, 全体として 20/39=51.28

さらにデンドログラムを高さ 5 周辺できり、1 のクラスターを除いて考えると、

	2,3,4,5	6,7,8
A,C	8	17
B,D	14	11

このとき一致度は、
(A,C)(6,7,8): (B,D)(2,3,4,5)=17/28: 14/22,全体として 31/50=62.00



3 章付録図 15: Raw data に数量化Ⅲ類を適用し、20 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 2.2 周辺で切り、クラスターを左から 1,2,3,4,5,6,7 と番号付けると、

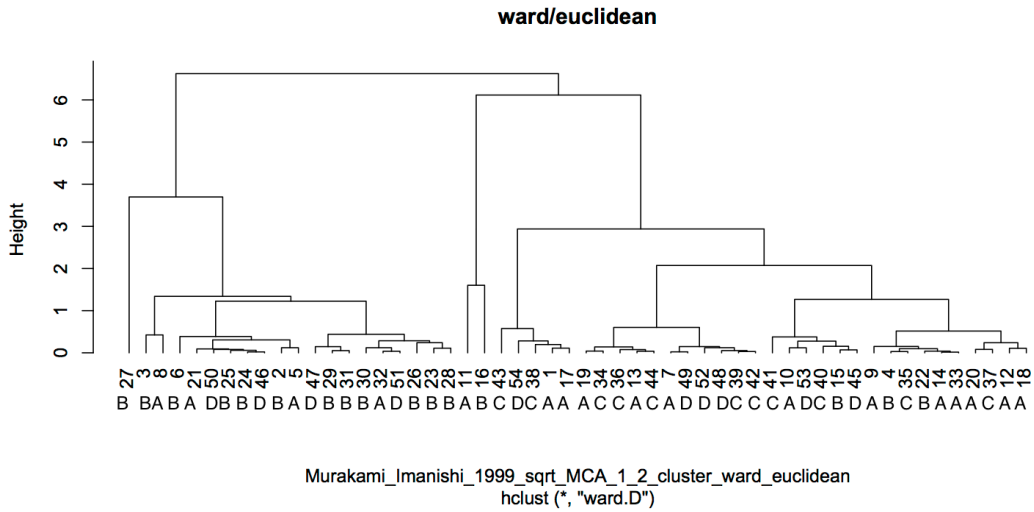
	1	2	3	4	5	6	7
A	3	1	5	0	0	8	0
B	0	1	2	1	7	0	5
C	1	0	3	0	1	0	6
D	0	0	3	0	0	0	7

このとき一致度は、
A(6): B(5): C(3): D(7)= 8/8: 7/8: 3/13: 7/18, 全体として 25/47=53.19

さらに、デンドログラムを高さ 3.2 周辺で切り、1,2 のクラスターを除いて考えると、

	3	4,5,6,7
A,C	8	15
B,D	5	20

このとき一致度は、
(A,C)(3): (B,D)(4,5,6,7)= 8/13: 20/35,全体として 28/48=58.33



3 章付録図 16: Raw data に平方根変換を施した値に数量化Ⅲ類を適用し、2 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 1.5 周辺で切り、クラスターを左から 1,2,3,4,5,6 と番号付けると、

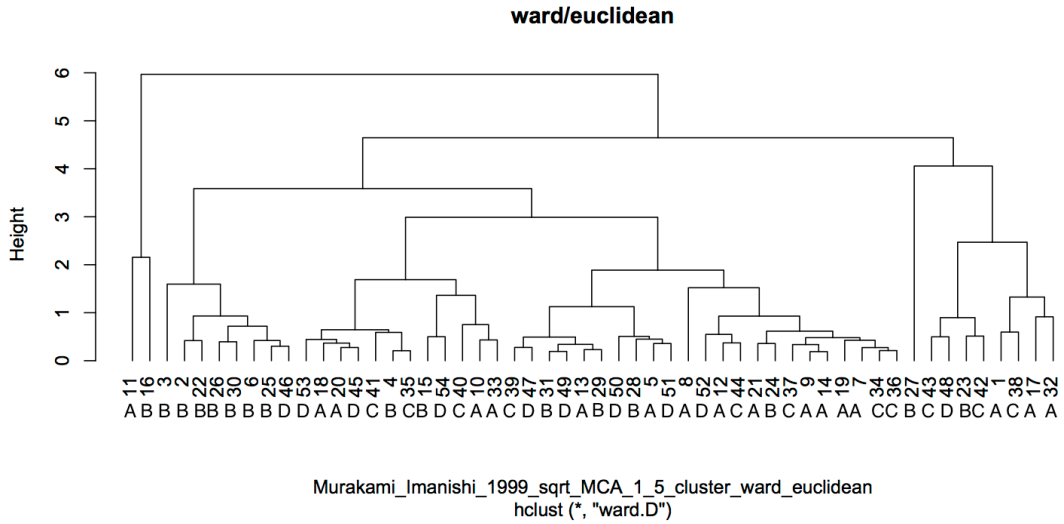
	1	2	3	4	5	6
A	0	4	1	2	3	7
B	1	11	1	0	0	3
C	0	0	0	2	5	4
D	0	4	0	1	3	2

このとき一致度は、
A(6): B(2): C(5): D(4)=7/16: 11/19: 5/11: 1/5, 全体として 24/51=47.06

さらにデンドログラムを高さ 6.5 周辺で切ると、

	1,2	3,4,5,6
A,C	4	24
B,D	16	10

このとき一致度は、
(A,C)(3,4,5,6): (B,D)(1,2)=24/34: 16/20,全体として 40/54=74.07



3 章付録図 17: Raw data に平方根変換を施した値に数量化Ⅲ類を適用し、5 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 2.5 周辺で切り、クラスターを左から 1,2,3,4,5,6 と番号付けると、

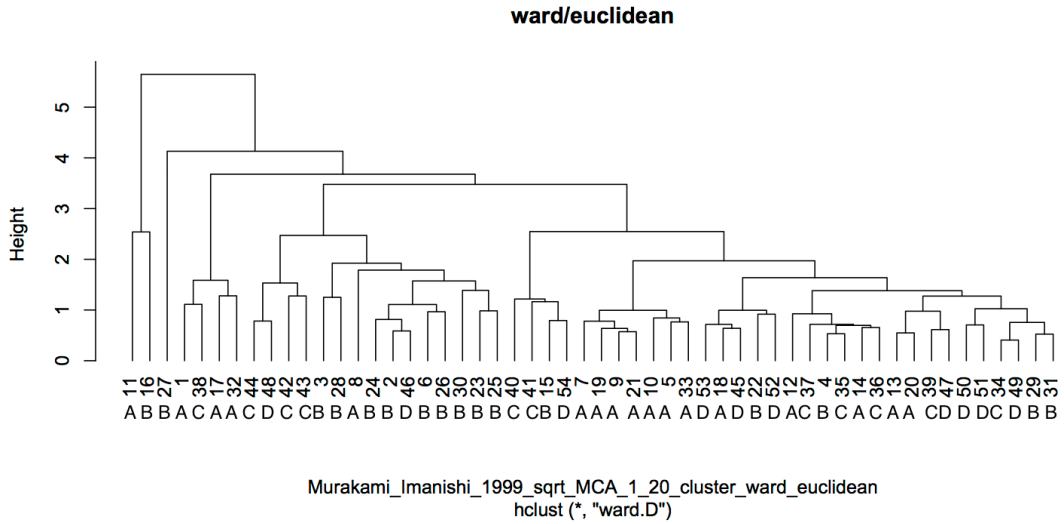
	1	2	3	4	5	6
A	1	0	4	9	0	3
B	1	7	2	4	1	1
C	0	0	3	5	0	3
D	0	1	3	5	0	1

このとき一致度は、
A(4): B(2): C(6): D(3)= 9/23: 7/8: 3/8: 3/12, 全体として 22/51=43.14

さらに、デンドログラムを高さ 4.2 周辺で切り、1 クラスターを除いて考えると、

	2,3,4	5,6
A,C	21	6
B,D	22	3

このとき一致度は、
(A,C)(5,6): (B,D)(2,3,4)= 6/9: 22/43, 全体として 28/52=53.85



3 章付録図 18: Raw data に平方根変換を施した値に数量化Ⅲ類を適用し、20 次元目までの座標を採用し、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 2.0 周辺で切り、クラスターを左から 1,2,3,4,5,6,7,8 と番号付けると、

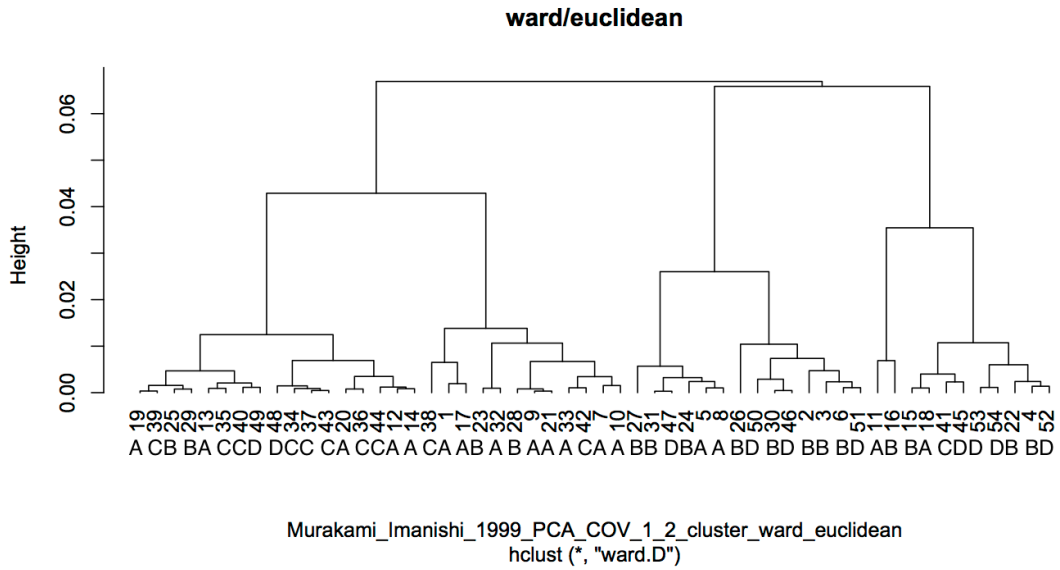
	1	2	3	4	5	6	7	8
A	1	0	0	3	0	1	0	12
B	0	1	1	0	0	9	1	4
C	0	0	0	1	3	0	2	5
D	0	0	0	0	1	1	1	7

このとき一致度は、
A(8): B(6): C(5): D(7)= 12/28: 9/11: 3/4: 1/4,全体として 25/47=53.19

さらにデンドログラムを高さ 3 周辺で切り、1,2,3,4 のクラスターを除いて考えると、

	5,6	7,8
A,C	4	19
B,D	11	13

このとき一致度は、
(A,C)(7,8): (B,D)(5,6)= 19/32: 11/15,全体として 30/47=63.83



3 章付録図 19: Raw data に主成分分析(PCA)を適用(分散共分散を使用)し、2 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.04 周辺で切り、クラスターに左から 1,2,3,4 と番号をつけたとき、

	1	2	3	4
A	5	8	2	2
B	2	2	8	4
C	8	2	0	1
D	2	0	4	4

このとき、一致度は

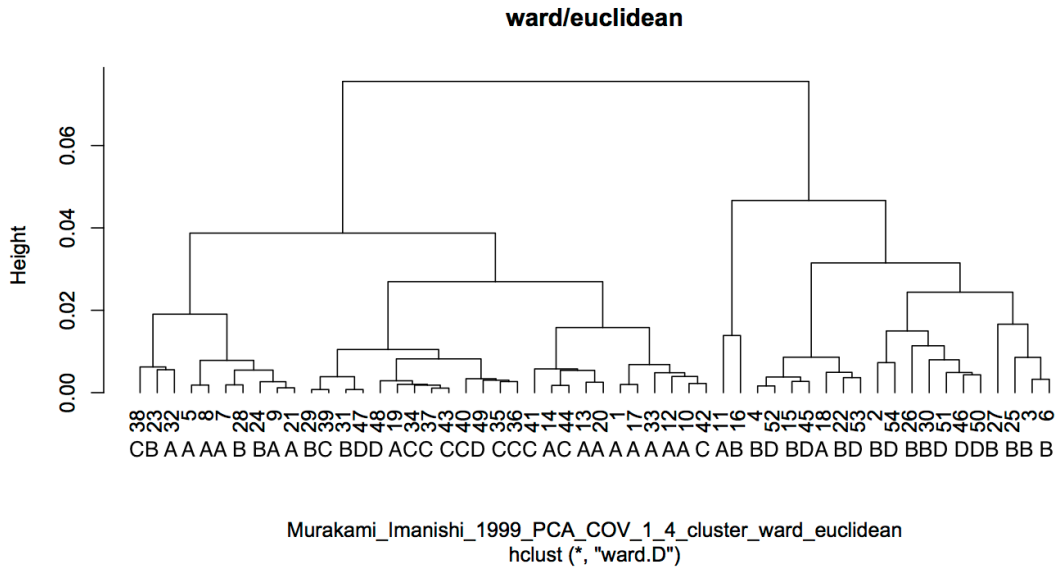
A(2):B(3):C(1):D(4)= 8/12: 8/14: 8/17: 4/11, 全体としては 28/54=51.85

さらに、デンドログラムを高さ 0.07 周辺で切ったとき、

	1,2	3,4
A,C	23	5
B,D	6	20

このとき、一致度は

(A,C)(1,2): (B,D)(3,4)= 23/29: 20/25, 全体としては 43/54=79.63



3 章付録図 20: Raw data に主成分分析(PCA)を適用(分散共分散を使用)し、4 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.03 周辺で切り、クラスターに左から 1,2,3,4,5 と番号をつけたとき、

	1	2	3	4	5
A	6	9	1	1	0
B	3	2	1	3	7
C	1	10	0	0	0
D	0	3	0	3	4

このとき、一致度は

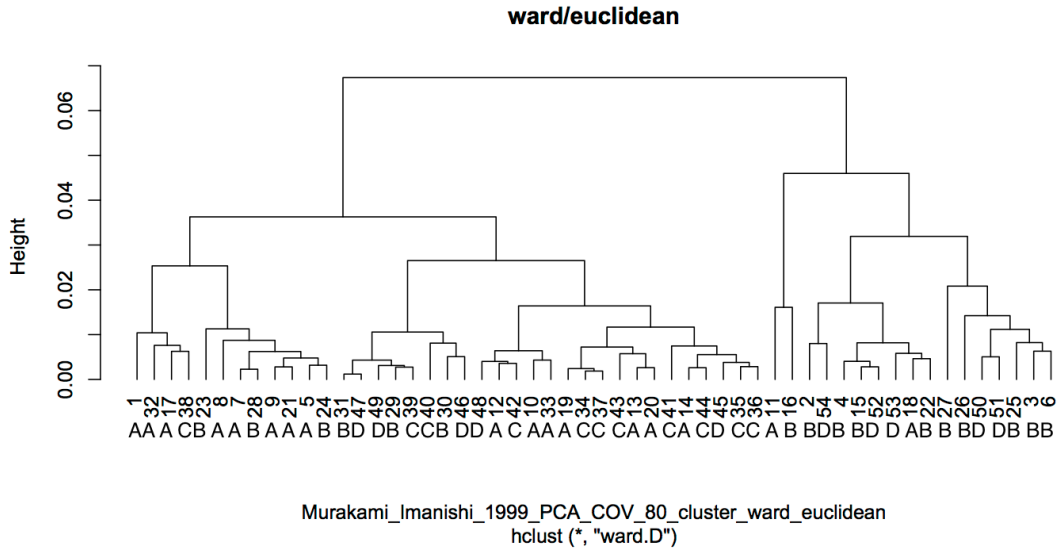
A(1):B(5):C(2):D(4)= 6/10: 7/11: 10/24: 3/7, 全体としては 26/52=50.00

さらに、デンドログラムを高さ 0.06 周辺で切ったとき、

	1,2	3,4,5
A,C	26	2
B,D	8	18

このとき、一致度は

(A,C)(1,2): (B,D)(3,4,5)= 26/34: 18/20, 全体としては 44/54=81.48



3 章付録図 21: Raw data に主成分分析(PCA)を適用(分散共分散行列を使用)し、6 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.03 周辺で切り、クラスターを左から 1,2,3,4,5 と番号付けると、

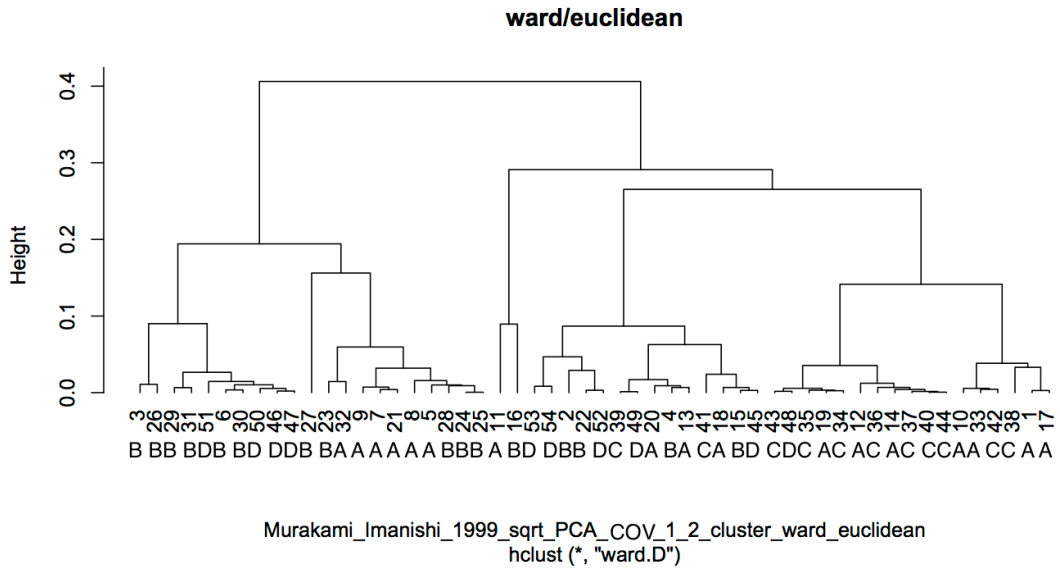
	1	2	3	4	5
A	8	7	1	1	0
B	3	3	1	4	5
C	1	10	0	0	0
D	0	5	0	3	2

このとき一致度は、
A(1): B(5): C(2): D(4)=8/12: 5/7: 10/25: 3/8,全体として 26/52=50.00

さらに、デンドログラムを高さ 0.06 周辺で切ると、

	1,2	3,4,5
A,C	26	2
B,D	11	15

このとき一致度は、
(A,C)(1,2): (B,D)(3,4,5)= 26/37: 15/17,全体として 41/54=75.93



3 章付録図 22: Raw data に平方根変換を施した値に主成分分析(PCA)を適用(分散共分散を使用)し、2 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.18 周辺で切り、クラスターを左から 1,2,3,4,5 と番号付けると、

	1	2	3	4	5
A	0	6	1	3	7
B	6	5	1	4	0
C	0	0	0	2	9
D	4	0	0	5	1

このとき一致度は、

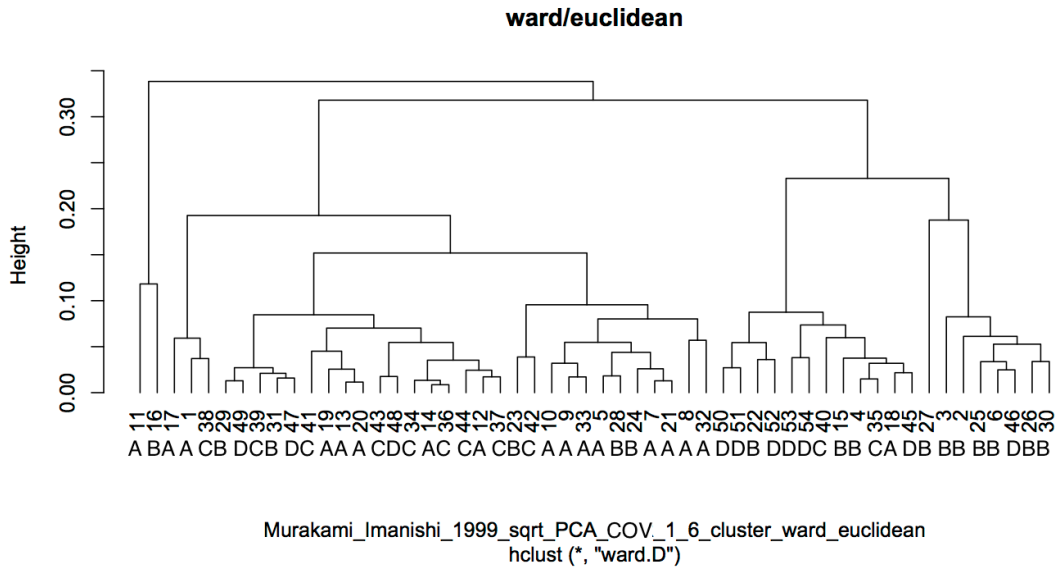
A(2): B(1): C(5): D(4)=6/11: 9/10: 5/14: 9/17,全体として 29/52=55.77

さらに、デンドログラムを高さ 0.35 周辺で切ると、

	1,2	3,4,5
A,C	6	22
B,D	15	11

このとき一致度は、

(A,C)(3,4,5): (B,D)(1,2)= 22/33: 15/21,全体として 37/54=68.52



3 章付録図 23: Raw data に平方根変換を施した値に主成分分析(PCA)を適用(分散共分散を使用)し、6 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.15 周辺で切り、クラスターを左から 1,2,3,4,5,6,7 と番号付けると、

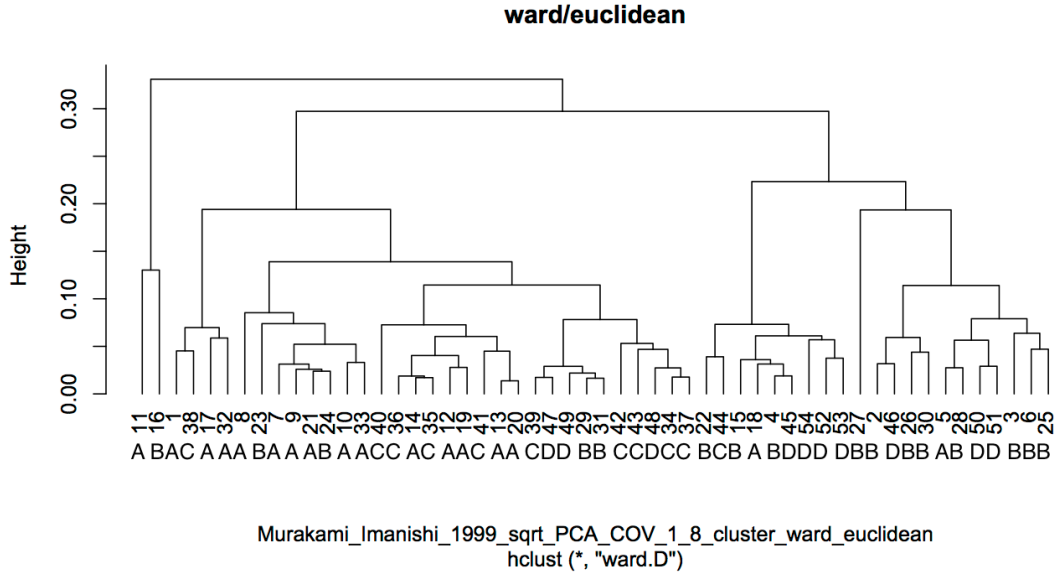
	1	2	3	4	5	6	7
A	1	2	5	8	1	0	0
B	1	0	2	3	3	1	6
C	0	1	7	1	2	0	0
D	0	0	3	0	6	0	1

このとき一致度は、
A(4): B(7): C(3): D(5)= 8/12: 6/7: 7/17: 6/12, 全体として 27/48=56.25

さらに、デンドログラムを高さ 0.3 周辺で切り、1 のクラスターを除いて考えると、

	2,3,4	4,5,6,7
A,C	24	3
B,D	8	17

このとき一致度は、
(A,C)(2,3,4): (B,D)(5,6,7)= 24/32: 17/20,全体として 41/52=78.85



3 章付録図 24: Raw data に平方根変換を施した値に主成分分析(PCA)を適用(分散共分散を使用)し、8 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 0.15 周辺で切り、クラスターを左から 1,2,3,4,5,6,7 と番号付けると、

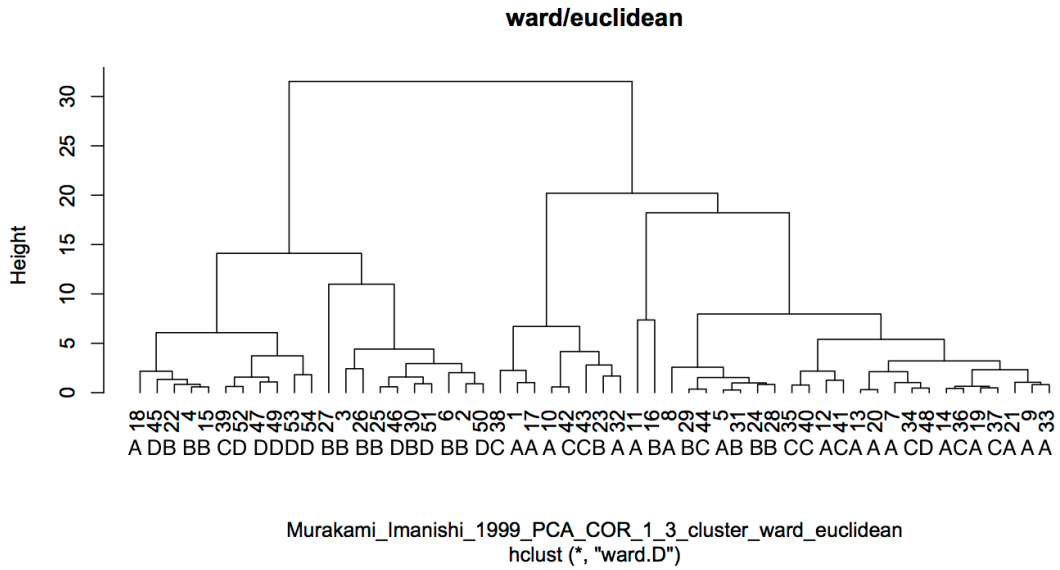
	1	2	3	4	5	6
A	1	3	11	1	0	1
B	1	0	4	3	1	7
C	0	1	9	1	0	0
D	0	0	3	4	0	3

このとき一致度は、
A(3): B(6): C(2): D(4)= 11/27: 7/11: 1/4: 4/9, 全体として 23/51=45.10

さらに、デンドログラムを高さ 0.25 周辺で切り、1 のクラスターを除いて考えると、

	2,3,	4,5,6
A,C	24	3
B,D	7	18

このとき一致度は、
(A,C)(2,3): (B,D)(4,5,6)= 24/31: 18/21,全体として 42/52=80.77



3 章付録図 25: Raw data に主成分分析(PCA)を適用(相関行列を使用)し、3 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 12.5 周辺で切り、クラスターを左から 1,2,3,4,5 と番号付けると、

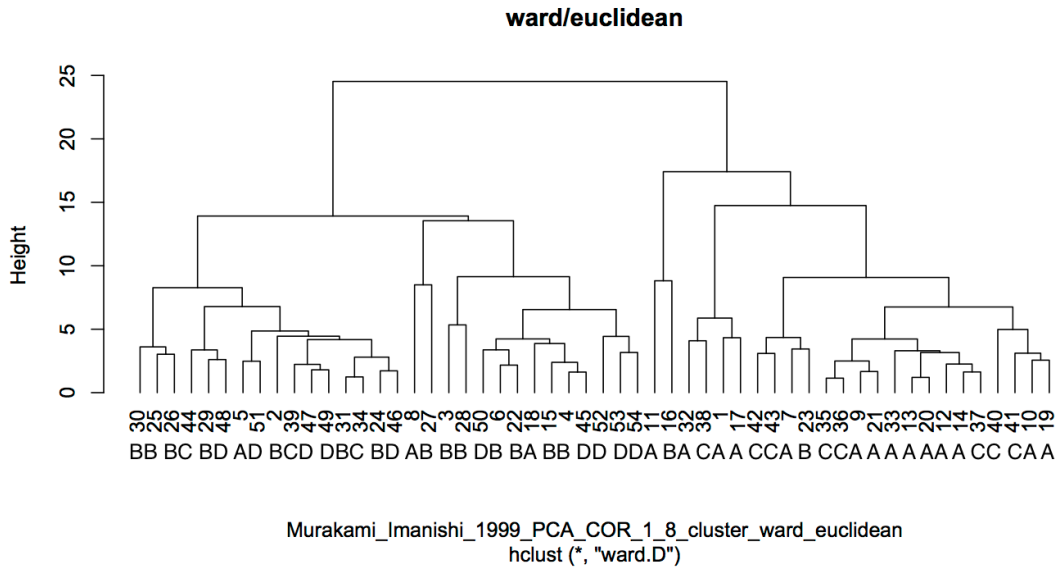
	1	2	3	4	5
A	1	0	4	1	11
B	3	7	1	1	4
C	1	0	3	0	7
D	6	3	0	0	1

このとき一致度は、
A(5): B(2): C(3): D(1)=11/23: 7/10: 3/8: 6/11,全体として 27/52=51.92

さらに、デンドログラムを高さ 25 周辺で切ると、

	1,2	3,4,5
A,C	2	26
B,D	19	7

このとき一致度は、
(A,C)(3,4,5): (B,D)(1,2)= 26/33: 19/21,全体として 45/54=83.33



3 章付録図 26: Raw data に主成分分析(PCA)を適用(相関行列を使用)し、8 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 10 周辺で切り、クラスターを左から 1,2,3,4,5,6 と番号付けると、

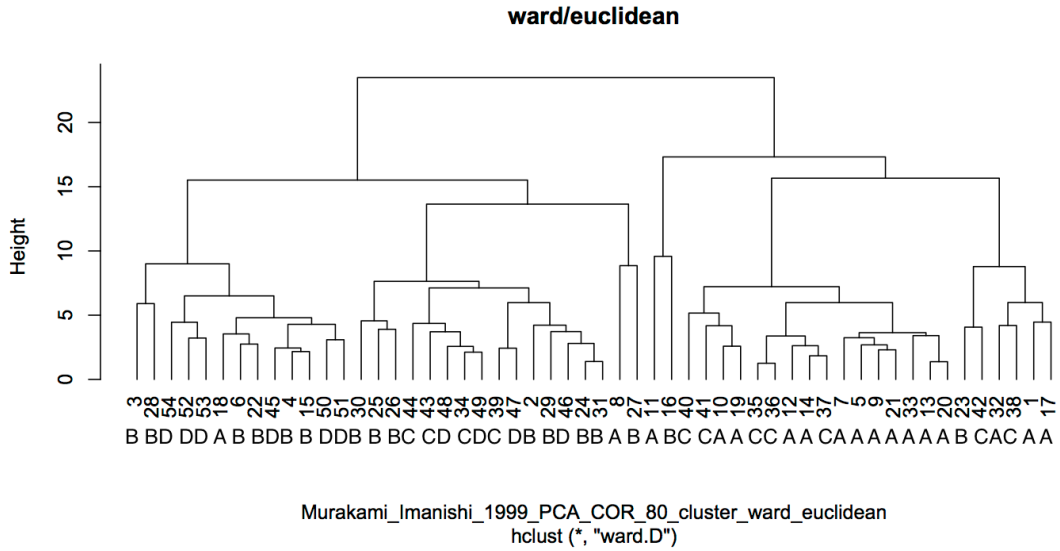
	1	2	3	4	5	6
A	1	1	1	1	3	10
B	7	1	6	1	0	1
C	3	0	0	0	1	7
D	5	0	5	0	0	0

このとき一致度は、
A(6): B(1): C(5): D(3)=10/18: 7/16: 1/4: 5/12,全体として 23/50=46.00

さらに、デンドログラムを高さ 20 周辺で切ると、

	1,2,3	4,5,6
A,C	6	22
B,D	24	2

このとき一致度は、
(A,C)(4,5,6): (B,D)(1,2,3)= 22/24: 24/30,全体として 46/54=85.19



3 章付録図 27: Raw data に主成分分析(PCA)を適用(相関行列を使用)し、10 次元目までの座標から、ユークリッド距離に基づき距離行列を計算し、ウォード法によってクラスタリングを行った結果

デンドログラムを高さ 15 周辺で切り、クラスターを左から 1,2,3,4,5 と番号付けると、

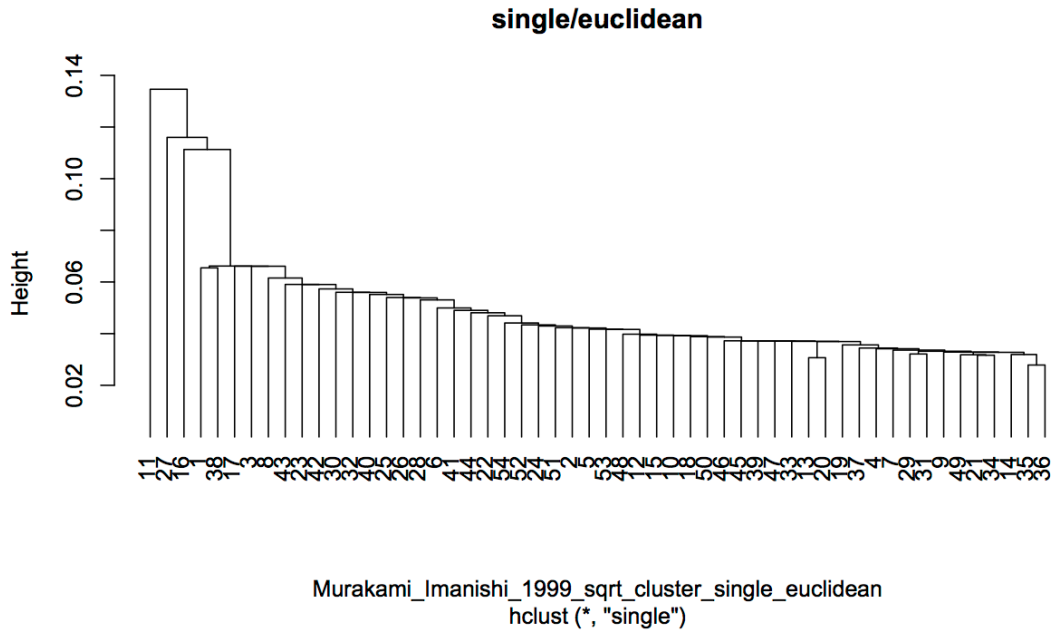
	1	2	3	4	5
A	1	1	1	11	3
B	6	8	1	0	1
C	0	4	0	5	2
D	6	4	0	0	0

このとき一致度は、
A(4): B(2): C(5): D(1)=6/13: 8/17: 11/16: 2/6,全体として 27/52=51.92

さらに、デンドログラムを高さ 20 周辺で切りクラスター3 を除いて考えると

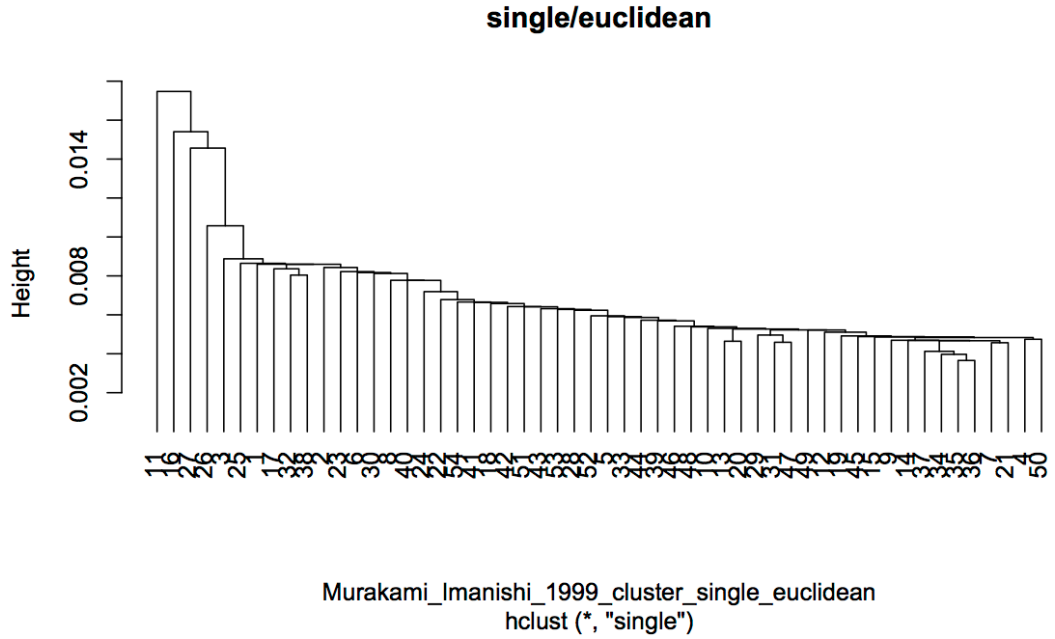
	1,2	4,5
A,C	6	21
B,D	24	1

このとき一致度は、
(A,C)(4,5): (B,D)(1,2)= 21/22: 24/30,全体として 45/52=85.53



3章付録図28: Raw dataに平方根変換を施した値からユークリッド距離に基づき距離行列を計算し、最短距離法によってクラスタリングを行った結果

このときデンドログラムはクラスターをほとんど形成していないため一致度は計算できなかった。



3 章付録図 29: Rw data からユークリッド距離に基づき距離行列を計算し、最短距離法によってクラスタリングを行った結果

このときデンドログラムはクラスターをほとんど形成していないため一致度は計算できなかった。

第 4 章 言語類型論データへの多重対応分析の適用

ー語順規則と言語の同時分類

4.1 はじめに

本章では言語類型論に多変量解析による統計的アプローチを導入した一つの先駆的研究である Tsunoda ,Ueda and Itoh(1995a)の研究を再検討する。Tsunoda et al. (1995a)は 129 言語の 19 の語順に関する変数の計 87 のカテゴリーについてのデータベース(角田 1991; Tsunoda, Ueda & Itoh, 1995b)に対してクラスター分析を適用した。1990 年代にこのようなデータベースを独力で作成したこと自体画期的なことであるが、さらに、通常の言語類型論の研究が言語の分類に注目するのに対して、彼らの研究の革新的な点は言語の分類とともに変数の分類をも試みた点である[1]。さらに、上田・伊藤(1995)では、主成分分析を使うことによって変数間の潜在構造を明らかにしている。しかし、彼らの研究は尺度水準の観点からいくつか問題点があった。彼らは、129(言語)×87(カテゴリー)のデータを作成しているが、そこでは各言語の 19 の語順のそれぞれの行和が 1 になるように、角田の言語学的な知見をもとに重み付けされ離散値(0.0 や 0.4, 0.6, 1.0)が与えられている(表 4 参照)が、これらは高々順序尺度であり、仮に非線形の効果が考えられるのならば、名義尺度である。すなわち、上田・伊藤(1995)が用いた間隔尺度を前提とした主成分分析ではデータの構造を十分に捉えているとは言えない可能性がある。そこで本研究では Tsunoda et al.(1995b)のデータに(多重)対応分析を適用することによって、Tsunoda et al.(1995a)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、統計的分析を行った。

結果として主に 3 つの点を指摘することができた(4.6 節)。第一に、Tsunoda et al. (1995a)の主な主張である無側置詞言語(Adpositionless)の言語は語順の観点からは後置詞言語(Postpositional)の言語と同じように振る舞うという結果は再考を要することがわかった。Tsunoda et al. (1995b)のデータに多重対応分析を適用した結果、1 次元目と 3 次元目の値を 2 次元平面で見ることによって、無側置詞言語(Adpositionless)と後置詞言語(Postpositional)と前置詞言語(Prepositional)は 3 つの独自のクラスターを成しているということが観察された。第二に、言語をクラスタリングした結果から、彼らは側置詞が分類の指標としては最も優れているということを指摘しているが、我々の分析では確かに側置詞が分類の指標としては最も優れていたが、同程度の指標として「主語、目的語、動詞の順序」及び、「所有格と目的語の順序」が考えられることを指摘した。第三に、変数の分析から、2 つの分類が重要であることがわかってきた。一つは、名詞句内で「主要部前置型」であるか「主要部後置型」であるかとい

[1] このような手法について Cysouw(2005)では、「reversed approach」と評している。

うことと、もう一つは節レベルで「主要部前置型」であるか「主要部後置型」であるかということである(詳しくは本論を参照)。さらに、興味深い点として、名詞句内であっても「所有格と目的語の順序」だけは、節レベルの変数と同じ振る舞いをすることがわかった。この現象については、Whitman and Ono (2017, to appear)において、通言語学的な観点から説明が試みられている。

考察では次の 2 点を指摘した(4.7 節)。まず、我々が Tsunoda et al. (1995b)のデータベースを、近年作られたより大きな言語類型論のデータベースである *WALS* (The World Atlas of Language Structure) Online (Dyer & Haspelmath, 2013)と比較し精査したところ、無側置詞言語(Adpositionless)と判断されたもので *WALS* では後置詞言語(Postpositional)に分類されているものが多くあることがわかった。すなわち、Tsunoda et al. (1995a)が導いた「無側置詞言語(Adpositionless)の言語は語順の観点からは後置詞言語(Postpositional)の言語と同じように振る舞うという結果」の一部はデータベースの誤りによる可能性であることが示唆された。次に、我々は、Tsunoda et al. (1995b)のデータベースの言語のサンプルが系統的にも地域的にも偏っていることを指摘した。系統的は、インド・ヨーロッパ語族の言語が極端に多く、地域的にはアフリカの言語が極端に少なくヨーロッパの言語が極端に多い。これは、網羅的な言語のデータベースである Ethnologue(Lewis, Gray & Charles, 2013)と比較しても明らかである。

以上のように、Tsunoda et al. (1995b)のデータベースやそれを用いた Tsunoda et al. (1995a)の結果は、我々の分析により限界や再考を要する点が幾つか散見された。しかし、1990 年代にこれほどのデータベースを一人の言語学者の力によって構築し、さらには統計分析を適用したという角田たちの先駆性はやはり革新的な試みといえ、Tsunoda et al. (1995a)は記念碑的な研究である。今後は、系統的にも地域的にもよりバランスのとれたデータベースをもとに、言語類型論のデータに対して統計的なアプローチを取る研究(仮にこれを「Statistical Typology」という)を前進させていくことが望まれる。

本章では、上述の解析結果について以下の各節で詳述する。

4.2 研究の背景

言語類型論とは言語における普遍的な構造を主に追究する学問のことであるが、初期の言語類型論は現在の言語類型論とは異なるものであった。1800 年代、Friedrich von Schlegel や Wilhelm von Humboldt らに代表される研究は、形態論を主な対象とし、それぞれの言語の形態論的類型に価値判断が介在しやすかった。例えば、ある理想型に対して、ドイツ語は中国語より適合しているために、ドイツ語は中国語より優れた言語であり、故に、ドイツ思想は中国思想より優れているといった論理である(リンゼイ 1997, 大堀･古賀･山泉訳, 2006: 27)。その後 20 世紀になって Ferdinand de Saussure、Leonard Bloomfield やプラハ学派などの人々によって言語学自体が大きな転換をした。このような流れの中で、言語類型論に革新をもたらしたのが、Joseph Greenberg である。

Greenberg の研究によって、「どのような種類の言語があるか」という問いから「諸言語にはどのような構造があるか」という問いに言語類型論の主題が変化した。さらに、彼は諸言語の構造の中に「ある言語に X があるなら、Y もまた見られる」という「含意的普遍性」というプラハ学派にあった考え方を明示的に打ち立てた(リンゼイ 1997, 大堀･古賀･山泉訳, 2006: 30)。統計的には、この「含意的普遍性」という概念は「相関」の概念に近いものであり(ただし、「含意的普遍性」は Y ならば X ということは意味していない点で「相関」とは異なる)、統計学の手法を適用する下地ができたといえよう。ただし、Greenberg (Greenberg, 1966)の研究は革新的なものであったが、同時に言語類型論者に課題を突き付けた。即ち、それは言語のサンプルの問題である。Greenberg (1966)は当時としては大きい 30 の言語サンプルを使い、サンプルの偏りを避けようとしたが、Dryer (1989)や Hawkins (1983)が指摘するように、それには実際は偏りがあり、妥当なサンプルの構築の重要性が指摘された。

この中で、言語類型論のデータに対して統計的なアプローチを取る研究(Statistical Typology)が、近年になって発展してきた。しかし、それらの研究の多くが、言語類型論の研究で以前から問題になっていた「研究の対象をどのようにサンプリングするべきか」(Maslova, 2000)ということなど、質的な研究を再検討することが主な課題であった。詳しくは Bickel (2008)が詳しい。このような研究史の中にあって Tsunoda et al.(1995a)は先駆的な研究である。

4.3 先行研究 Tsunoda, Ueda and Itoh (1995a)について

Tsunoda et al.(1995a)は、129 の言語に関する 19 の語順に関する 87 の変数に関するデータベース(角田 1991; Tsunoda, Ueda & Itoh, 1995b)に対してクラスター分析を適用した。以下、表 1 に 129 の言語の一覧を、表 2 に 19 の語順の一覧、表 3 に 19 の語順の 87 の変数についての一覧をそれぞれ載せる。さらに、表 4 に本研究で使用した Tsunoda et al. (1995b)の実際の数値データの一部を載せる。また、4 章付録に 129 言語の位置を記した世界地図を載せる。

表1. 角田 (1991)の129の言語一覧。X117はChoctawとChickasawの両方を含む

language.ID	language.name	language.ID	language.name	language.ID	language.name
X1	Japanese	X44	Swahili	X87	Tol
X2	Korean	X45	Haya	X88	Highland_Chontai
X3	Mongolian	X46	Tamil	X89	Walapai
X4	Evenki	X47	Kannada	X90	Southeastern_Pomo
X5	Turkish	X48	Burushaski	X91	Eastern_Pomo
X6	Mari	X49	Tibetan	X92	Mam
X7	Hungarian	X50	Mizo	X93	Ixil
X8	Finnish	X51	Burmese	X94	Quiche
X9	Abkhaz	X52	Mandarin_Chinese	X95	Polomchi
X10	Adyghe	X53	Thai	X96	Rabinal_Achi
X11	Kabardian	X54	Lao	X97	Cakchiquel
X12	Avar	X55	Cambodian	X98	K/ekchi/
X13	Georgian	X56	Vietnamese	X99	Jacaltec
X14	Russian	X57	Malay	X100	Tojolabal
X15	Polish	X58	Indonesian	X101	Chontai_Mayan
X16	Czech	X59	Tagalog	X102	Chorti
X17	Bulgarian	X60	Ilokano	X103	Copala
X18	Serbo-Croatian	X61	Kapampangan	X104	Isthumus_Zapotec
X19	Swedish	X62	Bikol	X105	Pipil
X20	Norwegian	X63	Palauan	X106	Nahuatl
X21	Danish	X64	Chamorro	X107	Yaqui
X22	German	X65	Tongan	X108	Papago
X23	Dutch	X66	Samoan	X109	Hopi
X24	English	X67	Niuean	X110	Chemehuevi
X25	Irish	X68	Maori	X111	Comanche
X26	Welsh	X69	Warrungu	X112	Luiseno
X27	Breton	X70	Kalkatungu	X113	Kiowa
X28	French	X71	Diyari	X114	Navajo
X29	Portuguese	X72	Aiyawarra	X115	Sarcee
X30	Spanish	X73	Waripiri	X116	Slavey
X31	Italian	X74	Djaru	X117	Choctaw
X32	Rumanian	X75	Kuniyanti	X118	Yuchi
X33	Modern_Greek	X76	Amuesha	X119	Omaha-Ponca
X34	Persian	X77	Jaqaru	X120	Dakota
X35	Panjabi	X78	Aymara	X121	Blackfoot
X36	Hindi	X79	Guarani	X122	Atikamekw
X37	Bengali	X80	Urubu-Kaapor	X123	Sahptin
X38	Basque	X81	Canela	X124	Nez_Perce
X39	Egyptian_Arabic	X82	Piraha	X125	Coast_Tsimshian
X40	Modern_Heberw	X83	Hixkarynana	X126	Gitksan
X41	Tigrinya	X84	Apalai	X127	Eskimo
X42	Hausa	X85	Quechua	X128	Chukchi
X43	Yoruba	X86	Tuyucan	X129	Nivkh

表 2. 語順に関する 19 変数の内容(日本語、英語による例示). 角田(1991)の付録から抜粋

No.	項目	日本語	英語
1	S, O と V	SOV 等	SVO
2	名詞と側置詞	＋	－
3	所有格と名詞	＋	＋, －
4	指示詞と名詞	＋	＋
5	数詞と名詞	＋	＋
6	形容詞と名詞	＋	＋
7	関係節と名詞	＋	－
8	固有名詞と普通名詞	＋	－, ＋
9	比較の表現	＋	－
10	本動詞と助動詞	＋	－
11	副詞と動詞	V より前	様々
12	副詞と形容詞	＋	＋, －
13	疑問の印	文末	無し
	一般疑問文での S, V の倒		
14	置	無し	有る
15	疑問詞	平叙文	文頭
	特別疑問文での S, V の倒		
16	置	無し	有る
17	否定の印	動詞語尾	V の直後
18	条件節と主節	＋	＋, －
19	目的節と主節	＋	－

注) ここでは日本語が主としてとるカテゴリーを＋とし、日本語が主としてとるカテゴリーと反対のカテゴリーを－とした。＋と－の両方の値をとる場合には、＋、－を併記した。

表3.19の語順に関する変数の87カテゴリーの内容. Tsunoda, Ueda and Itoh (1995b)より作成

語順	変数(語順のカテゴリー)		日本語	英語
1 S, OとV				
	X1_1	SOV	○	
	X1_2	SVO		○
	X1_3	OSV		
	X1_4	VSO		
	X1_5	OVS		
	X1_6	VOS		
2 名詞と側置詞				
	X2_1	後置詞 (+)	○	
	X2_2	前置詞 (−)		○
	X2_3	無側置詞		
	X2_4	その他		
3 所有格と名詞				
	X3_1	所有格-名詞 (+)	○	○
	X3_2	名詞-所有格 (−)		○
	X3_3	名詞を挟む		
	X3_4	その他		
4 指示詞と名詞				
	X4_1	指示詞-名詞 (+)	○	○
	X4_2	名詞-指示詞 (−)		
	X4_3	名詞を挟む		
	X4_4	無し		
	X4_5	その他		
5 数詞と名詞				
	X5_1	数詞-名詞 (+)	○	○
	X5_2	名詞-数詞 (−)		
	X5_3	名詞を挟む		
6 形容詞と名詞				
	X6_1	形容詞-名詞 (+)	○	○
	X6_2	名詞-形容詞 (−)		
	X6_3	名詞を挟む		
	X6_4	その他		
7 関係節と名詞				
	X7_1	関係節-名詞 (+)	○	
	X7_2	名詞-関係節 (−)		○
	X7_3	その他		
8 固有名詞と普通名詞				
	X8_1	固有名詞-普通名詞 (+)	○	○
	X8_2	普通名詞-固有名詞 (−)		○
9 比較の表現				
	X9_1	A-B-比較級 (+)	○	
	X9_2	A-比較級-B (−)		○
	X9_3	無し		
	X9_4	その他		
10 本動詞と助動詞				
	X10_1	本動詞-助動詞 (+)	○	
	X10_2	助動詞-本動詞 (−)		○
	X10_3	無し		
11 副詞と動詞				
	X11_1	Vより前	○	○
	X11_2	Vより後		○
	X11_3	文頭		○
	X11_4	文末		○
	X11_5	その他		○

右側の2列は日本語と英語がどのカテゴリーに該当するかを例示している。

語順	変数（語順のカテゴリー）		日本語	英語
12 副詞と形容詞				
	X12_1	副詞-形容詞 (+)	○	○
	X12_2	形容詞-副詞 (−)		○
	X12_3	その他		
13 疑問の印				
	X13_1	文末	○	
	X13_2	文頭		
	X13_3	無し		○
	X13_4	(文頭)より2番目		
	X13_5	Vの直後(VとSの間)		
	X13_6	質問の焦点の直後		
	X13_7	動詞の語尾		
	X13_8	文末から2番目		
	X13_9	動詞の接頭辞		
	X13_10	Vより前		
	X13_11	否定の印の直後		
	X13_12	助動詞の直前		
	X13_13	文を挟む		
	X13_14	その他		
14 一般疑問文でのS, Vの倒置				
	X14_1	無し	○	
	X14_2	有る		○
15 疑問詞				
	X15_1	平叙文式	○	
	X15_2	文頭		○
	X15_3	文末		
	X15_4	Vの直後		
	X15_5	Vの直前		
	X15_6	2番目		
16 特別疑問文でのS, Vの倒置				
	X16_1	無し	○	
	X16_2	有る		○
17 否定の印				
	X17_1	動詞語尾	○	
	X17_2	Vより前(Vの直前)		
	X17_3	文頭		
	X17_4	Vより後(Vの直後)		○
	X17_5	動詞接頭辞		
	X17_6	Vを挟む		
	X17_7	否定の焦点の直後		
	X17_8	否定の焦点の食前		
	X17_9	否定の焦点を挟む		
	X17_10	否定助動詞		
	X17_11	文末		
	X17_12	2番目		
	X17_13	3番目		
18 条件節と主節				
	X18_1	条件説-主節 (+)	○	○
	X18_2	主節-条件説 (−)		○
19 目的節と主節				
	X19_1	目的節-主節 (+)	○	
	X19_2	主節-目的説 (−)		○

表 4. 本研究で使用した Tsunoda et al. (1995b)の数値データの一部

Language	X1_1	X1_2	X1_3	X1_4	X1_5	X1_6	…	X5_1	X5_2	X5_3
X1.Japanese	1	0	0	0	0	0	…	1	0	0
X2.Korean	0.7	0	0.3	0	0	0	…	1	0	0
X3.Mongolian	0.7	0	0.3	0	0	0	…	1	0	0
X4.Evenki	1	0	0	0	0	0	…	1	0	0
X5.Turkish	1	0	0	0	0	0	…	1	0	0
X6.Mari	1	0	0	0	0	0	…	1	0	0
X7.Hungarian	0	1	0	0	0	0	…	1	0	0
X8.Finnish	0	1	0	0	0	0	…	1	0	0
X9.Abkhaz	1	0	0	0	0	0	…	0.6	0.4	0
X10.Adyghe	1	0	0	0	0	0	…	0	1	0
X11.Kabardian	0.6	0	0.4	0	0	0	…	0	1	0
X12.Avar	1	0	0	0	0	0	…	1	0	0
X13.Georgian	0.6	0.4	0	0	0	0	…	1	0	0
X14.Russian	0	1	0	0	0	0	…	1	0	0
X15.Polish	0	1	0	0	0	0	…	1	0	0
X16.Czech	0	1	0	0	0	0	…	1	0	0
X17.Bulgarian	0	1	0	0	0	0	…	1	0	0
X18.Serbo-Croatian	0	1	0	0	0	0	…	1	0	0
X19.Swedish	0	0.7	0	0	0.3	0	…	1	0	0
X20.Norwegian	0	0.7	0	0	0.3	0	…	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮	⋮
X60.Ilokano	0	0	0	1	0	0	…	**0**	**0**	**0**
X61.Kapampangan	0	0	0	0.6	0	0.4	…	1	0	0
X62.Bikol	0	0	0	0.6	0	0.4	…	0.6	0.4	0

注:X60.Ilokano の X5_1, X5_2, X5_3 の行和が 0 であるのは、値が欠測値であることを意味する。

4.1 で述べたように、1990 年代にこのようなデータベースを独力で作成したこと自体画期的なことであるが、さらに、通常の言語類型論の研究が言語の分類に注目するのに対して、Tsunoda et al. (1995a)の革新的な点は言語の分類と

ともに変数の分類をも試みた点である。

まず、彼等は Tsunoda et al. (1995b)のデータにクラスター分析を適用して得た言語の樹形図について、その樹形図のクラスターを最もよく説明することができる(単)変数を選び出した。結果、彼等は 2 つの知見を得ている。第 1 に、変数の中で「名詞と側置詞の順序」が言語に関する樹形図を最もよく説明すること。第 2 に、「名詞と側置詞の順序」の中でも、言語が「前置詞言語(Prepositional)」か、それ以外(「後置詞言語(Postpositional)」と「無側置詞言語(Adpositionless)」)に分かれることである。特に、Tsunoda et al. (1995a)では、今まで位置づけの難しかった「無側置詞言語(Adpositionless)」が「後置詞言語(Postpositional)」と語順の観点からは同じ振る舞いを示すという結果は言語類型論の観点から価値があると強調している。

4.4 本研究の目的

Tsunoda et al. (1995a)では、4.3 で述べたように 129(言語)×87(変数)のデータを作成しているが、そこでは各言語の 19 の語順のそれぞれの行和が 1 になるように、角田の言語学的な知見をもとに重み付けされた離散値(0.0 や 0.4, 0.6, 1.0)が与えられている。この型のデータを仮に変形アイテムカテゴリー型データと呼んでおく。Tsunoda et al. (1995a)は 19 の各次元が独立していると仮定してクラスター分析を適用しているが、この分析手法には 2 つの問題がある。

第 1 に、Tsunoda et al. (1995a)の目的は変数間の潜在構造を捉えることでもあった。しかし、単なるクラスター分析では、変数間の内在的なパターン(相関関係)を十分に捉えることができない。よって、変数間の内在構造を考慮した多変量解析や主成分分析の適用が望ましい。

第 2 に、Tsunoda et al. (1995b)のデータは高々順序尺度であり、仮に非線形の効果が考えられるのならば、名義尺度である。すなわち、Tsunoda et al. (1995a)のクラスター分析や、それを発展させた上田・伊藤(1995)が用いた間隔尺度を前提とした主成分分析ではデータの構造を十分に捉えていない可能性がある。よって、名義尺度データに対する主成分分析と考えられる多重対応分析(MCA)を適用することが妥当と考えられる。

本研究の目的は、Tsunoda et al. (1995b)のデータに、分析手法として MCA を適用することで、Tsunoda et al. (1995a)の結果を再検討するとともに、それらの結果から言語学的に新たな知見を得ることである。

4.5 分析方法

本章では、Tsunoda et al. (1995b)のデータに MCA を適用し、得られた少数の次元の布置に基づいて言語および変数を解釈し(4.6.1 節)、そこでの解釈を補強するためにクラスター分析を適用した。具体的には、データの型が頻度表データと変形アイムカテゴリー型データであるという違いはあるものの、2 章で述べた「数量化Ⅲ類クラスタリング」の一種である(4.6.2 節)。

クラスタリングの手法については、ウォード法(Ward, 1963)を用い、用いる距離としては、ユークリッド距離及びマンハッタン距離の 2 種を試行した。本来のウォード法はユークリッド二乗距離を用いたものであるが、ユークリッド距離、及びマンハッタン距離を用いた手法は Székely and Rizzo (2005)や Lee and Wilcox (2014)によって Lance-Williams family (Lance & Williams, 1967)に属することが示されている。

本章では、MCA には R(2014)の「MASS」パッケージ(Venables & Ripley, 2002)の「corresp」のコマンドを、クラスター分析には「stats」パッケージの「dist」および「hclust」のコマンドを用いた。

なお、4.9 節の統計的補足では、通常のクラスター分析では捉えることのできないデータの「2 番目以降に強い」構造を捉えるべく、Neighbor-Net 及び MCA_Neighbor-Net を適用した結果を紹介している。

4.6 結果

4.6.1 多重対応分析の結果

まず、図 1 に MCA を適用して得られた固有値のスクリープロットを載せる。

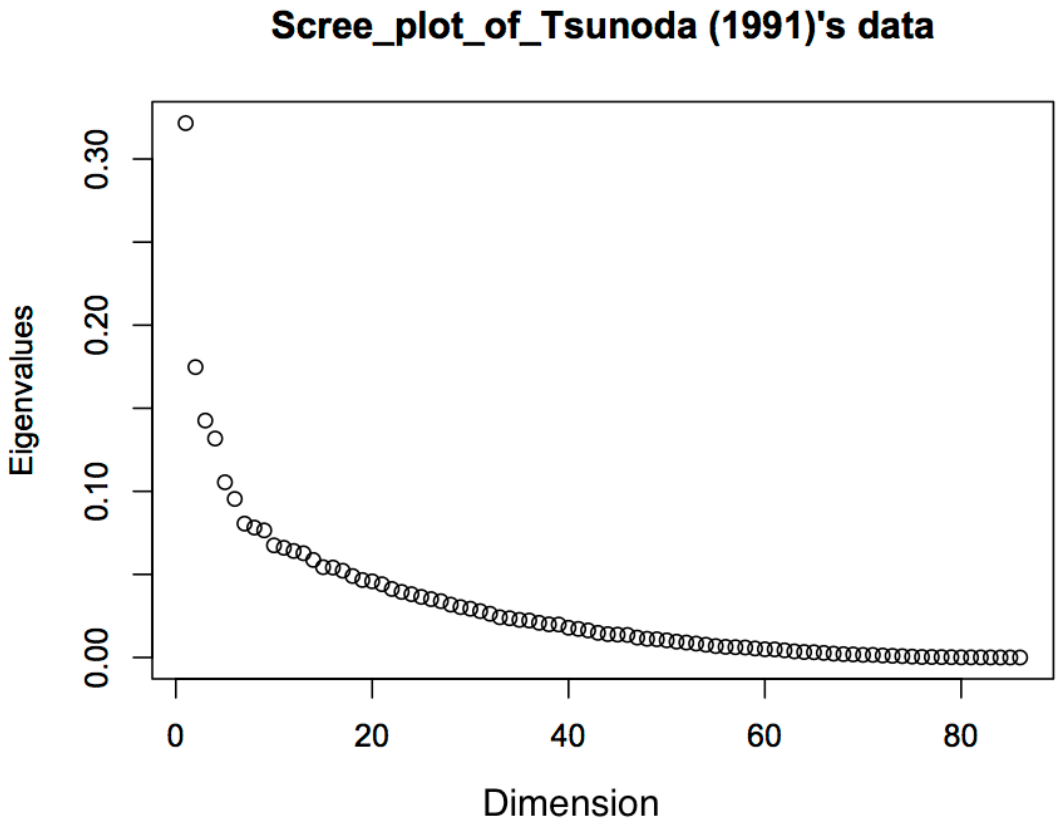


図 1. MCA のスクリープロット。

図 1 のスクリープロットからは、MCA の結果得られた布置について、何次元目までに注目するべきか明瞭ではない(1 次元目を除いて)。そこで、本研究では、まず 3 次元目までの座標について分析した。さらに、固有値の減少が緩やかになる 1 つ前の次元(6 次元目)までの座標にクラスター分析を適用して、MCA の結果を分かりやすく視覚化する。

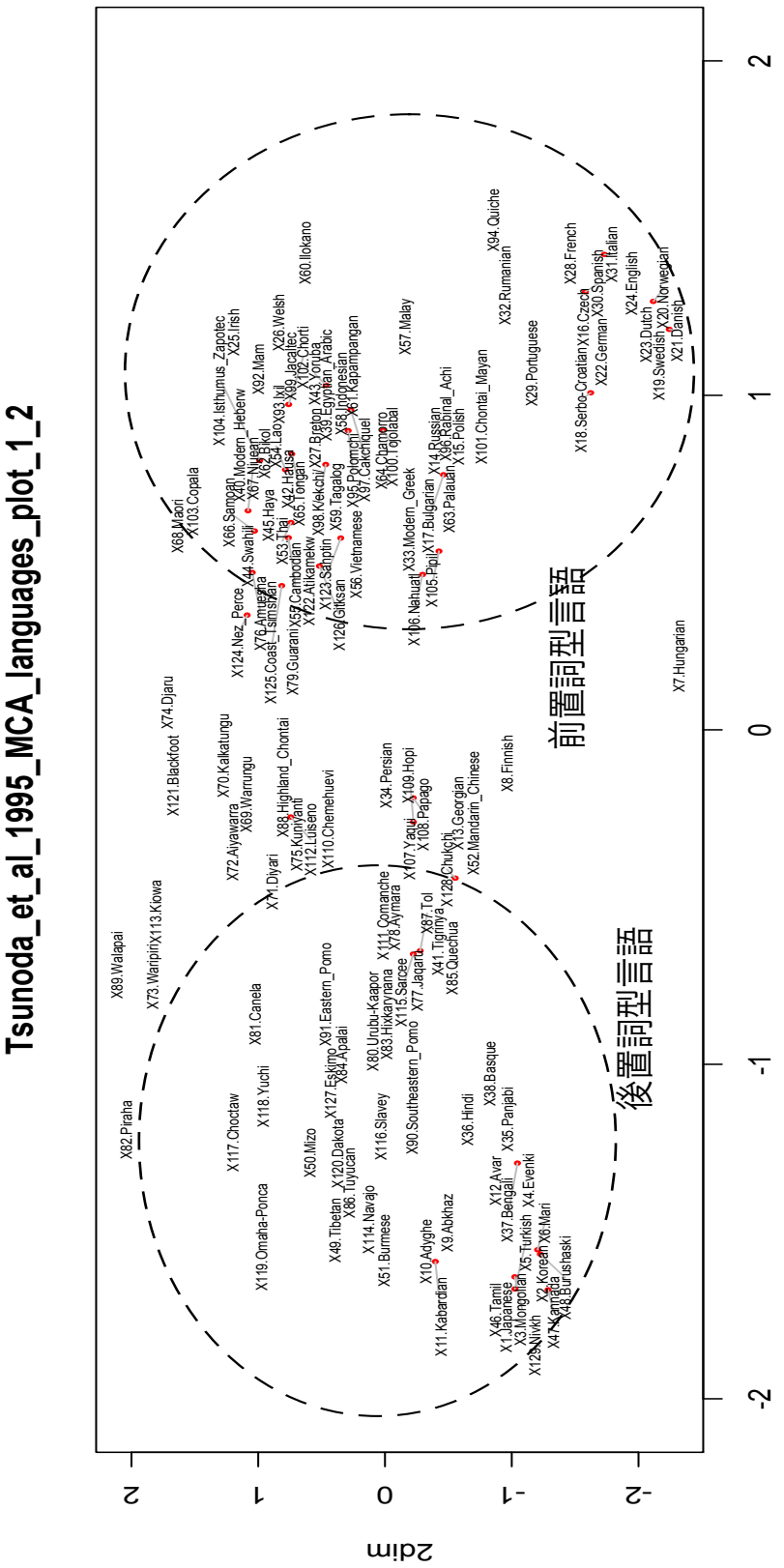


図 2-(a). MCA の結果得られた 129 の言語の 1 次元目と 2 次元目上の布置

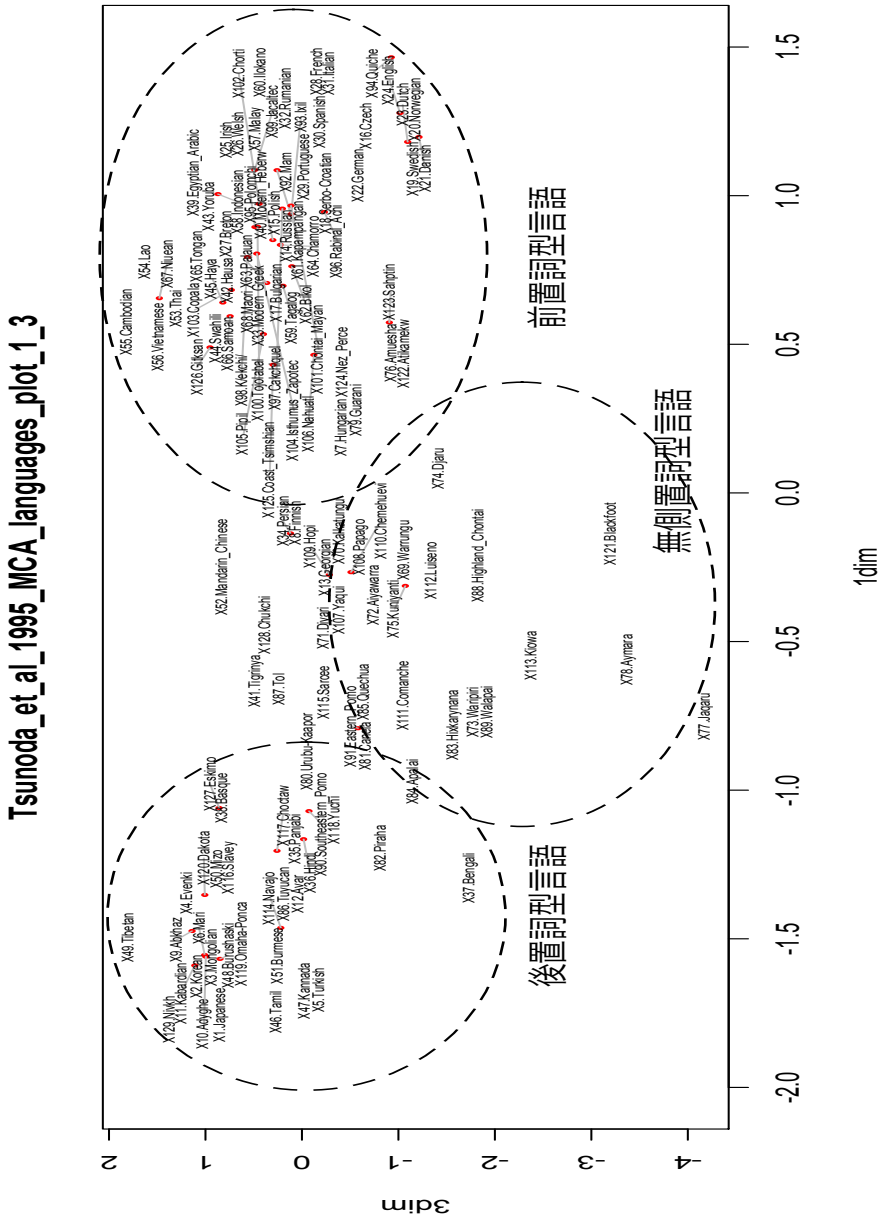


図 2-(b). MCA の結果得られた 129 の言語の 1 次元目と 3 次元目上の布置

図 2-(a)(b)は、MCA の結果得られた 129 の言語の 1 次元目と 2 次元目の平面、1 次元目と 3 次元目の平面である。3 次元目までの累積寄与率は 24.07%であった。図 2-(a)においては、左側に後置詞型言語が右側に前置詞型言語が集まっていることが分かる。図 2-(b)においては左側に後置詞型言語が右側に前置詞型言語が集まり、さらに、3 次元目に無側置型言語が、後置詞型言語と前置詞型言語とは別にクラスターをなしていることが分かる。

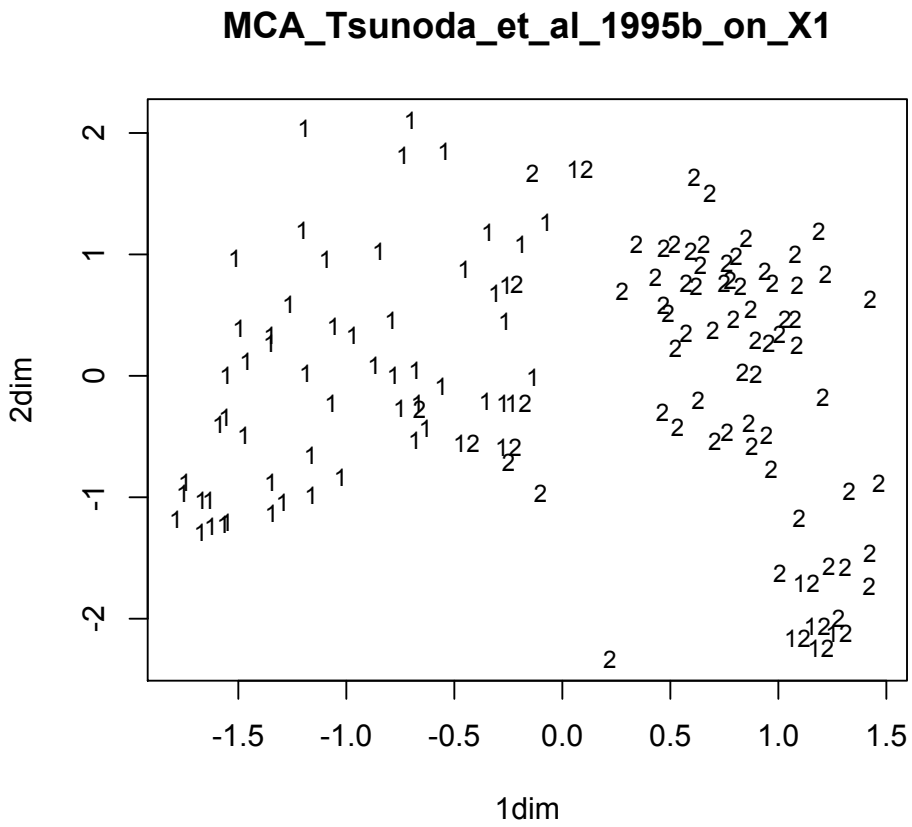


図 3-(a). MCA の結果得られた 129 の言語の 1 次元目と 2 次元目上の布置。X1 において SOV、OVS、OSV をとる言語を OV の値をとる言語として「1」とし、SVO、VSO、VOS をとる言語を VO の値をとる言語を「2」と表示した。さらに、複数のカテゴリーの値をとるものは全ての値を表示した。以下同様

図 3-(a)(b)、図 4-(a)(b)、図 5-(a)(b)はそれぞれ本研究で注目した X1、X2、X3 の値を主要部後置型、主要部前置型の観点から整理し、視覚化したものである。具体的には図 3-(a)(b)では X1 について主要部後置型として SOV、OVS、OSV を、OV をとる「1」としてまとめ、主要部前置型として SVO、VSO、VOS を、VO をとる値として「2」としてまとめた。図 4-(a)(b)では X2 について主要部後置型として後置詞を「1」、主要部前置型として前置詞を「2」、無側置詞の言語を「3」、「その他」を「4」とした。図 5-(a)(b)では X3 について、主要部後置型として「所有格-目的語」をとる言語を「1」、主要部前置型として「目的語-所有格」をとる言語を「2」、「名詞を挟む」を「3」、「その他」を「4」とした。

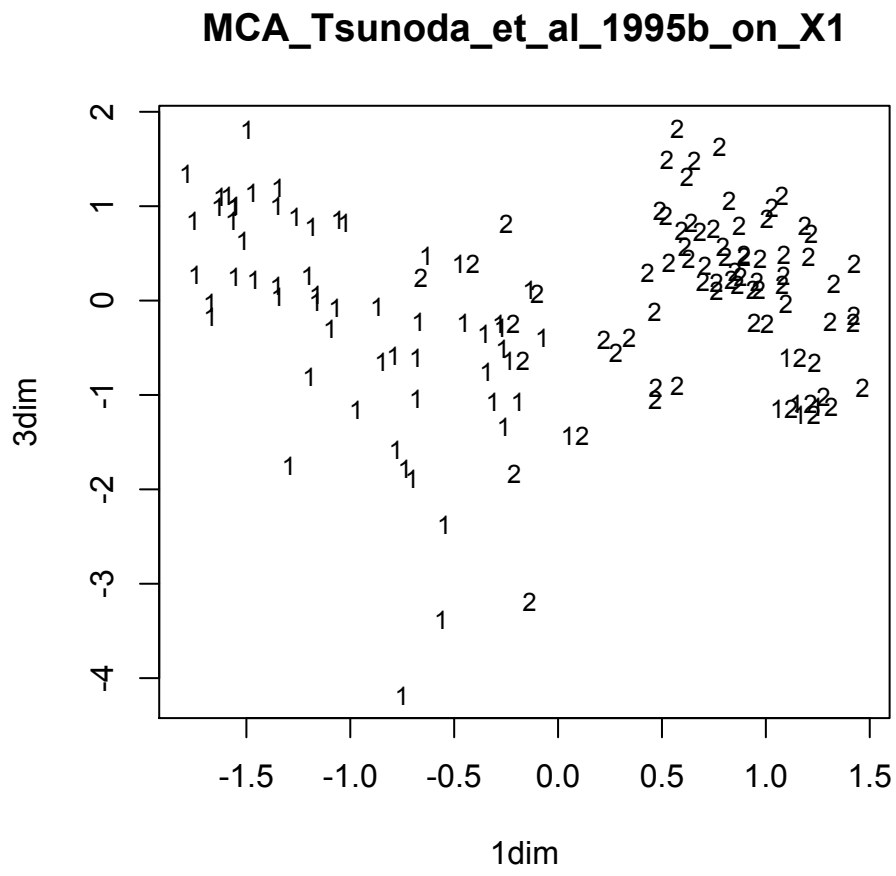


図 3-(b). MCA の結果得られた 129 の言語の 1 次元目と 3 次元目上の布置。X1 に関して，OV の値をとる言語を「1」とし、VO の値をとる言語を「2」とした

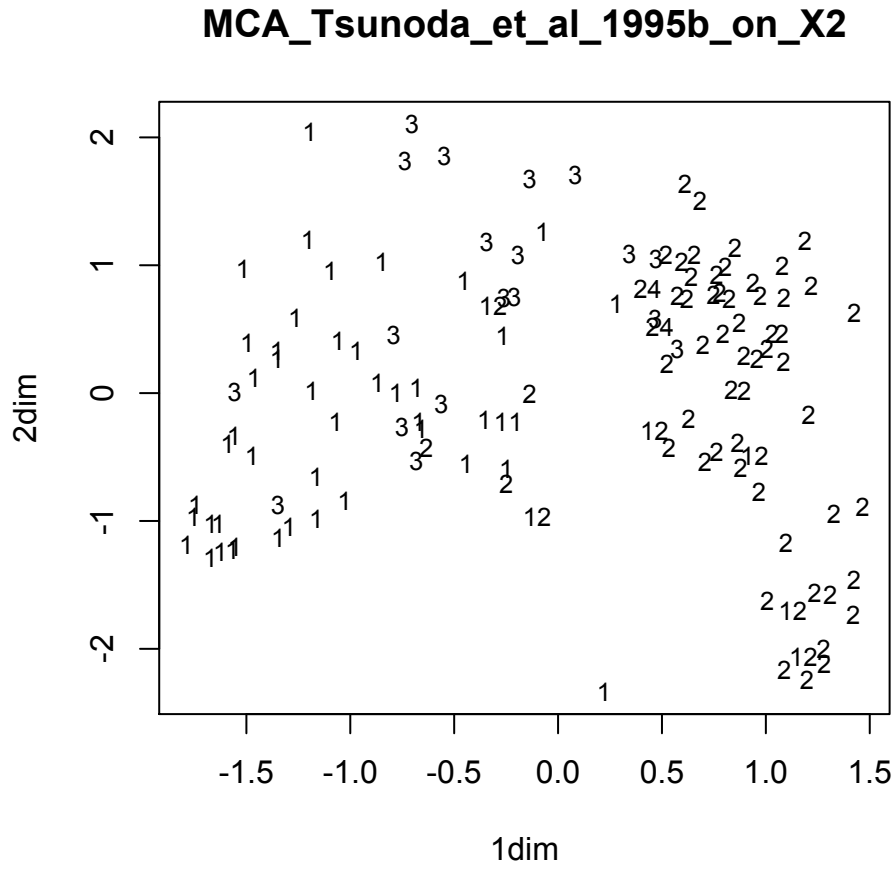


図 4-(a). MCA の結果得られた 129 の言語の 1 次元目と 2 次元目上の布置。X2 に関して，後置詞の値をとる言語を「1」、前置詞の値をとる言語を「2」、無側置詞の値をとる言語を「3」、「その他」の値をとる言語を「4」とした。

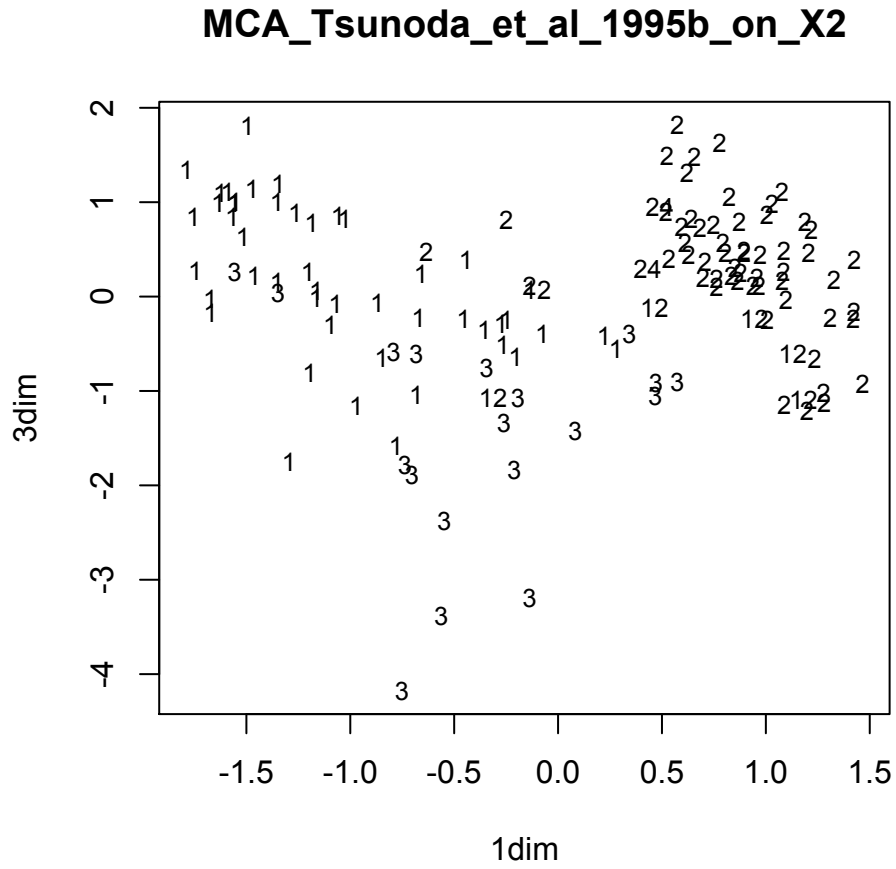


図 4-(b). MCA の結果得られた 129 の言語の 1 次元目と 3 次元目上の布置。X2 に関して，後置詞の値をとる言語を「1」、前置詞の値をとる言語を「2」、無側置詞の値をとる言語を「3」、「その他」の値をとる言語を「4」とした。

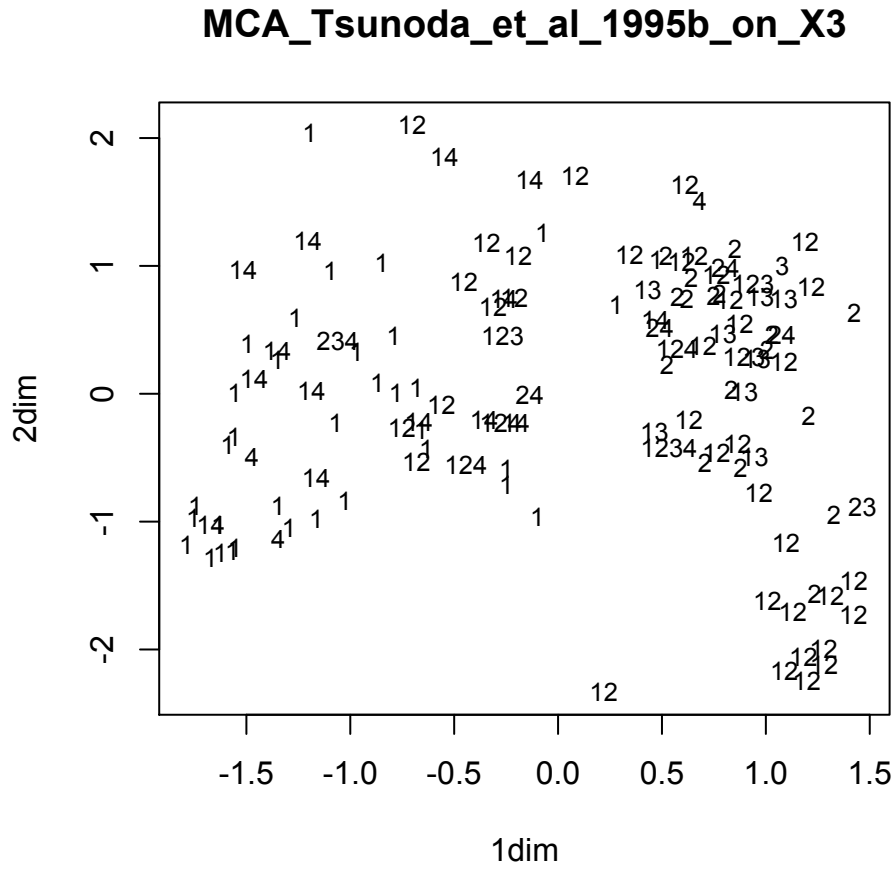


図 5-(a). MCA の結果得られた 129 の言語の 1 次元目と 2 次元目上の布置。X3 に関して,「所有格-目的語」の値をとる言語を「1」、「目的語-所有格」の値をとる言語を「2」、「名詞を挟む」の値をとる言語を「3」、「その他」の値をとる言語を「4」とした。

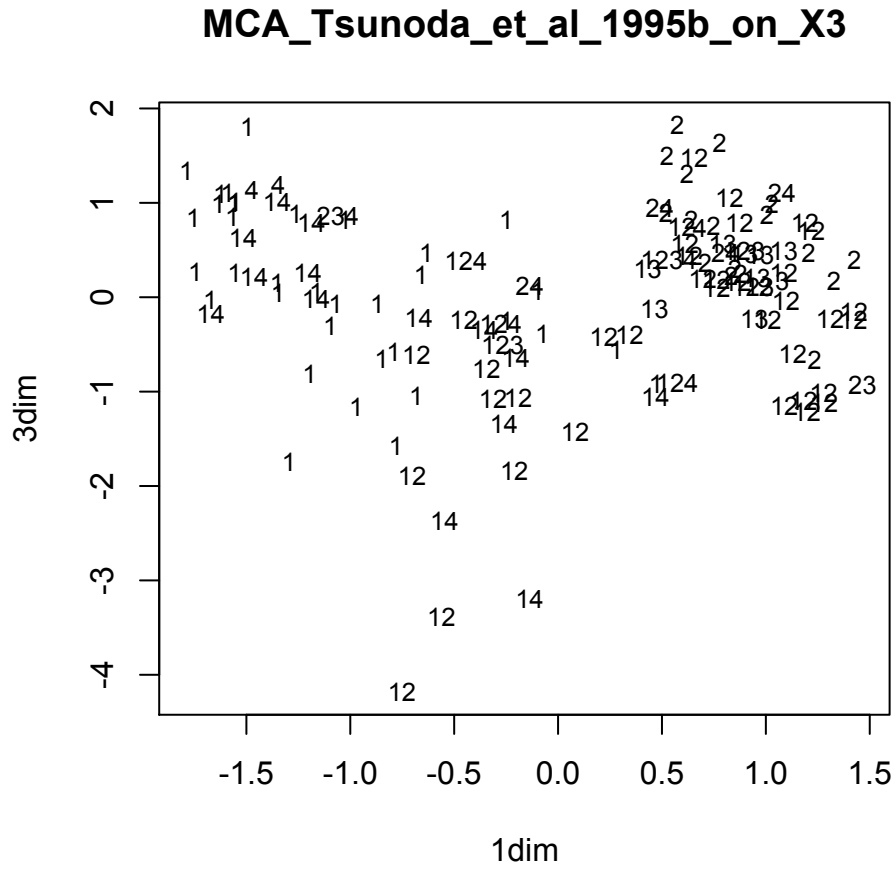


図 5-(b). MCA の結果得られた 129 の言語の 1 次元目と 3 次元目上の布置。X3 に関して,「所有格-目的語」の値をとる言語を「1」、「目的語-所有格」の値をとる言語を「2」、「名詞を挟む」の値をとる言語を「3」、「その他」の値をとる言語を「4」とした。

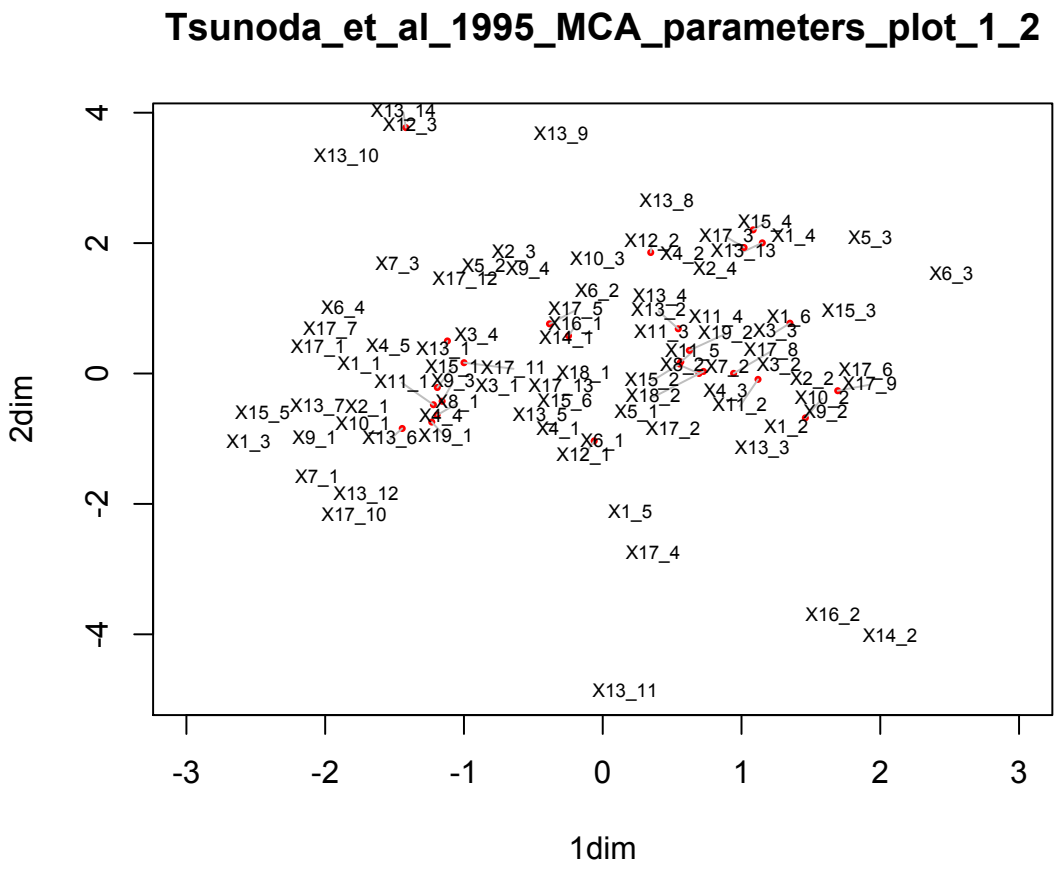


図 6-(a). MCA の結果得られた 87 の変数の 1 次元目と 2 次元目上の布置

図 6-(a)(b)は 87 変数の 1 次元と 2 次元上の布置、ならびに 87 変数の 1 次元と 3 次元上の布置を示した。さらに、図 6-(c)(d)では、本研究では注目した X1_1、X1_2、X2_1、X2_2、X2_3、X3_1、X3_2、X4_1、X4_2、X5_1、X5_2、X6_1、X6_2 の布置に注目することによって節レベルの主要部前置型、主要部後置型と名詞句レベルの主要部前置型、主要部後置型の構造を観察することができる。

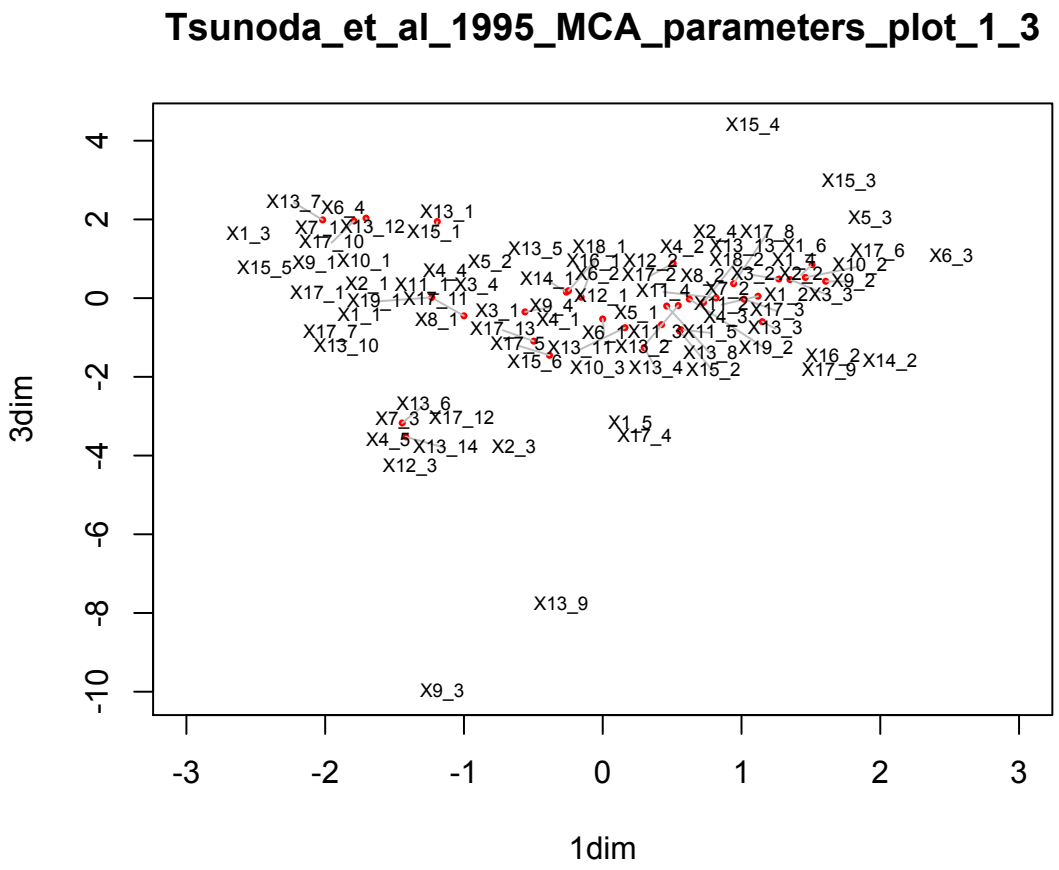


図 6-(b). MCA の結果得られた 87 の変数の 1 次元目と 3 次元目上の布置

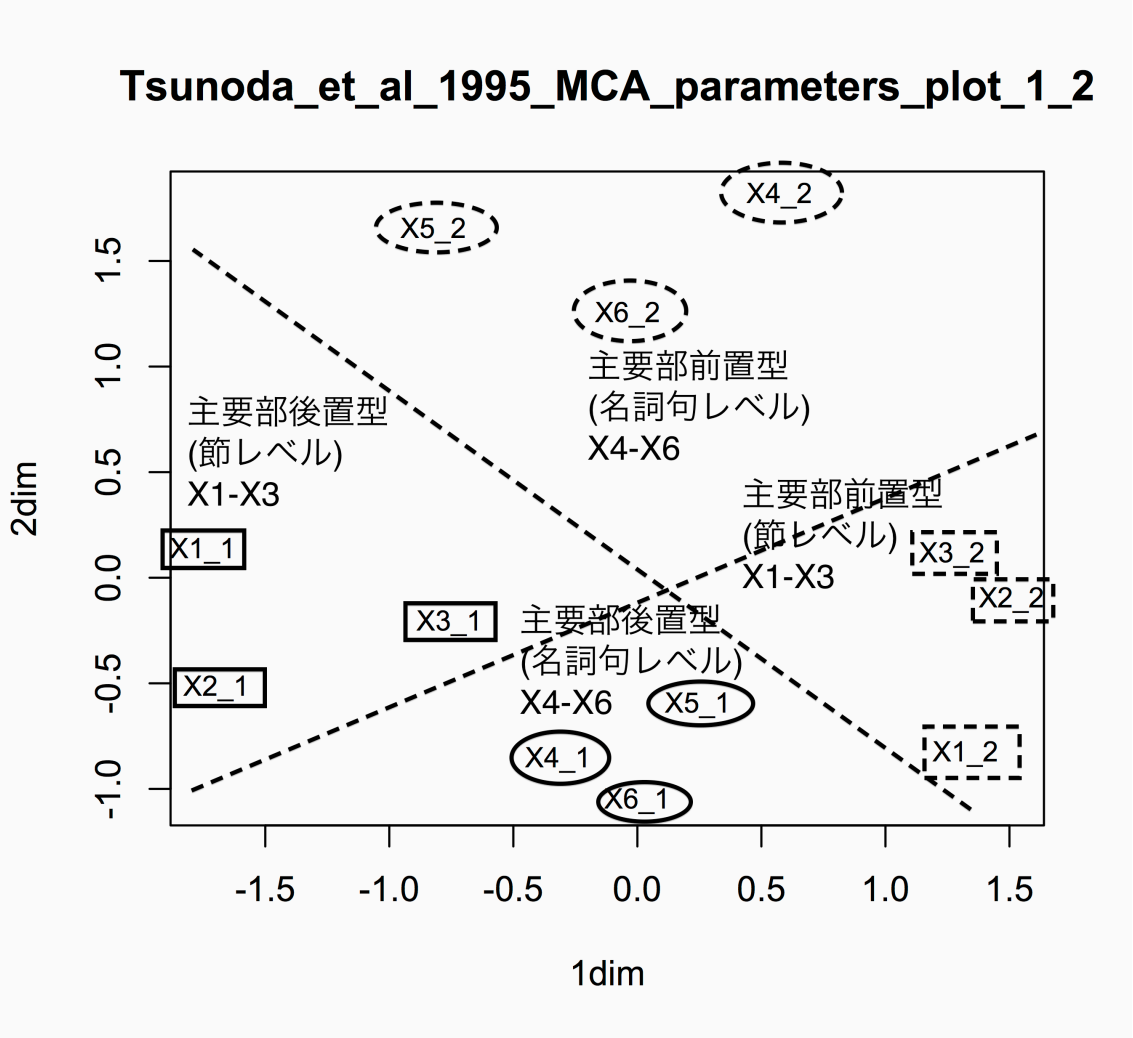


図 6-(c). 図 6-(a)の中で、本研究が注目した変数(X1_1, X1_2, X2_1, X2_2, X3_1, X3_2, X4_1, X4_2, X5_1, X5_2, X6_1, X6_2)の布置

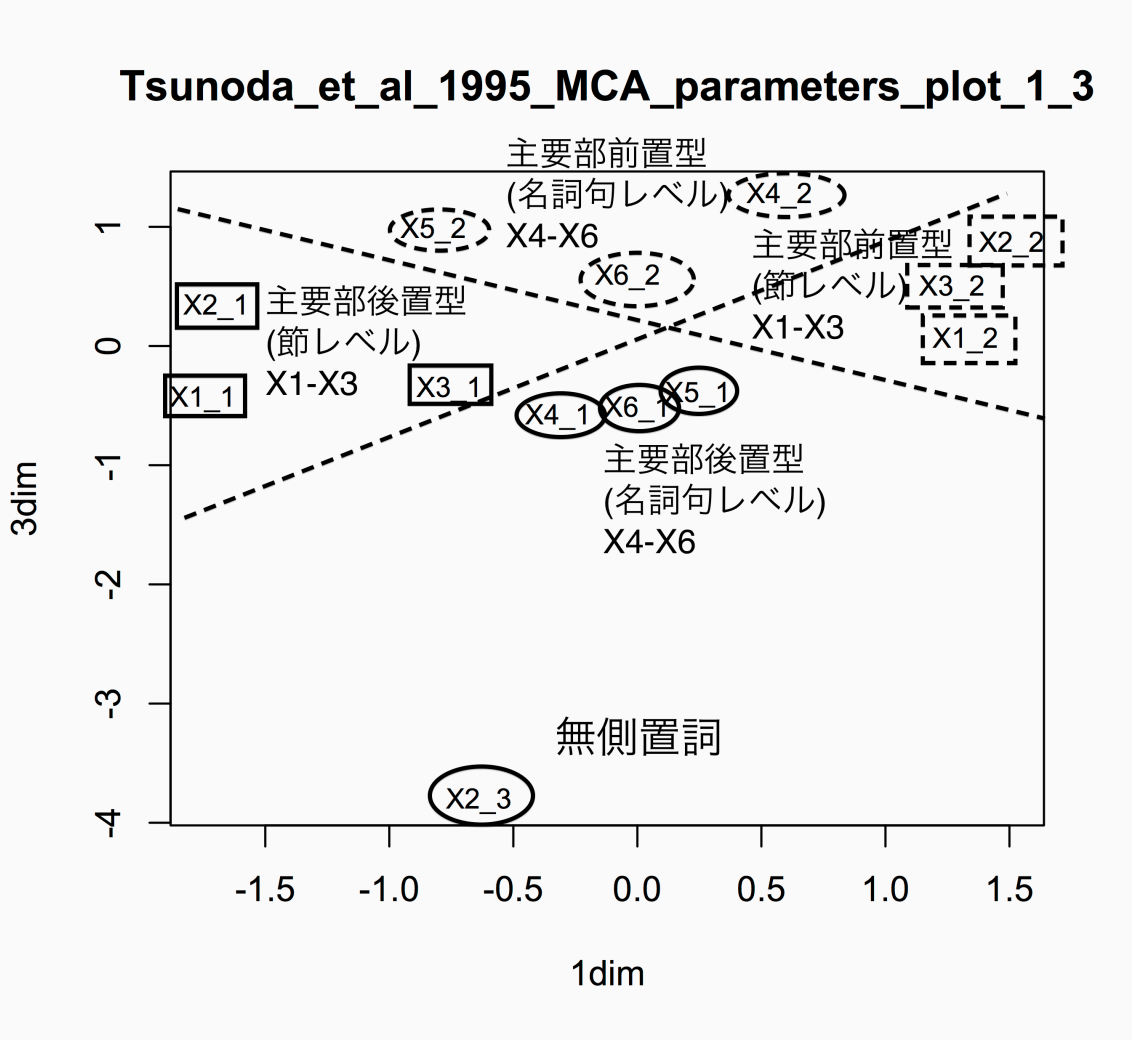


図 6-(d). 図 6-(b)の中で、本研究が注目した変数(X1_1, X1_2, X2_1, X2_2, X3_1, X3_2, X4_1, X4_2, X5_1, X5_2, X6_1, X6_2)と無側置詞(X2_3)の布置。

図 2-(a)(b)、図 4-(a)(b)及び図 6-(a)(b)(c)(d)から、以下の 2 点が明らかになった。第一に、図 2-(a)(b)及び図 4-(a)(b)から、129 の言語は後置詞型言語、無側置詞型言語、前置詞型言語の 3 つのクラスターをなしている。これらの 3 つのクラスターの相対的な位置は、図 6(d)における後置詞(X2_1)と X2_3(無側置詞)と X2_2(前置詞)の相対的な位置関係と対応していることが分かる。このことは、Tsunoda et al. (1995a)の主要な知見である「無側置詞型言語が語順の観点からは後置詞型言語と同じ振る舞いをする」ということが再考を要することを示唆する。MCA の結果からは、これらの 3 つのタイプの言語はそれぞれ独立したクラスターをなすと考えた方が自然である。

次に、本研究が注目した変数(X1_1, X1_2, X2_1, X2_2, X3_1, X3_2, X4_1, X4_2, X5_1, X5_2, X6_1, X6_2)の座標を示すと、

主要部後置　　主要部前置
X1_1[SOV]=(-1.75, 0.13, -0.45)　　X1_2[SVO]=(1.32, -0.83, 0.05)
X2_1[後置詞]=(-1.70, -0.52, 0.33)　　X2_2[前置詞]=(1.51, -0.11, 0.85)
X3_1[所有格-名詞]=(-0.76, -0.22, -0.37)　X3_2[名詞-所有格]=(1.27, 0.11, 0.48)
X4_1[指示詞-名詞]=(-0.32, -0.86, -0.59)　X4_2[名詞-指示詞]=(0.57, 1.81, 0.97)
X5_1[数詞-名詞]=(0.24, -0.60, -0.41)　　X5_2[名詞-数詞]=(-0.82, 1.64, 0.97)
X6_1[形容詞-名詞]=(-0.00, -1.06, -0.53)　X6_2[名詞-形容詞]=(-0.04, 1.25, 0.58)

これらの座標値および図 6-(c)(d)から、主要部前置型(X1_2, X2_2, X3_2, X4_2, X5_2, X6_2)と主要部後置型(X1_1, X2_1, X3_1, X4_1, X5_1, X6_1)の変数に分かれることが明らかになった。さらに、その中でも節レベル(X1_2, X2_2, X3_2/ X1_1, X2_1, X3_1)と名詞句内レベル(X4_1, X5_1, X6_1/ X4_2, X5_2, X6_2)の変数に分かれる。しかし、節レベル(X1_2, X2_2, X3_2/ X1_1, X2_1, X3_1)の変数の中には、本来名詞句内レベルと考えられる「所有格と名詞(X3_2, X3_1)」が一貫して混在している。この現象については、*WALS* を分析した筆者の研究(Whitman & Ono, 2017, to appear)でも観察された。その中では言語変化の観点からこの現象の説明を試みているが、本章では統計的な側面が主眼であるため立ち入らない。

4.6.2 数量化Ⅲ類クラスタリングの結果

ここでは、前節で得られた MCA の結果をより分かりやすくするために 2 章でのべた数量化Ⅲ類クラスタリングを適用する。数量化Ⅲ類クラスタリングの適用に際して、筆者は図 1 のスクリープロットから、6 次元目までの布置を用いることとした。6 次元目までの累積寄与率は 36.61%であった。さらに、数量化Ⅲ類クラスタリングの結果と、Tsunoda et al. (1995a)のようにクラスター分析を Tsunoda et al. (1995b)に直接適用した結果を比較した。

4.6.2.1 129 言語の分析の結果

MCA の結果、6 次元目までの言語の布置からユークリッド距離に基づき距離行列を計算しウォード法によってクラスター分析を実行した結果を図 7-(a)(b)に、MCA の結果から、6 次元目までの言語の布置からマンハッタン距離に基づ

きウォード法を実行した結果を図 8-(a)(b)に、Tsunoda et al. (1995b)の素データから、ユークリッド距離に基づきウォード法によってクラスター分析を実行した結果を図 9-(a)(b)に、Tsunoda et al. (1995b)の素データからマンハッタン距離に基づきウォード法によってクラスター分析を実行した結果を図 10-(a)(b)にそれぞれ示す。図 7 から図 10 は(a)(b)を合わせて一つの大きなデンドログラムをなす。

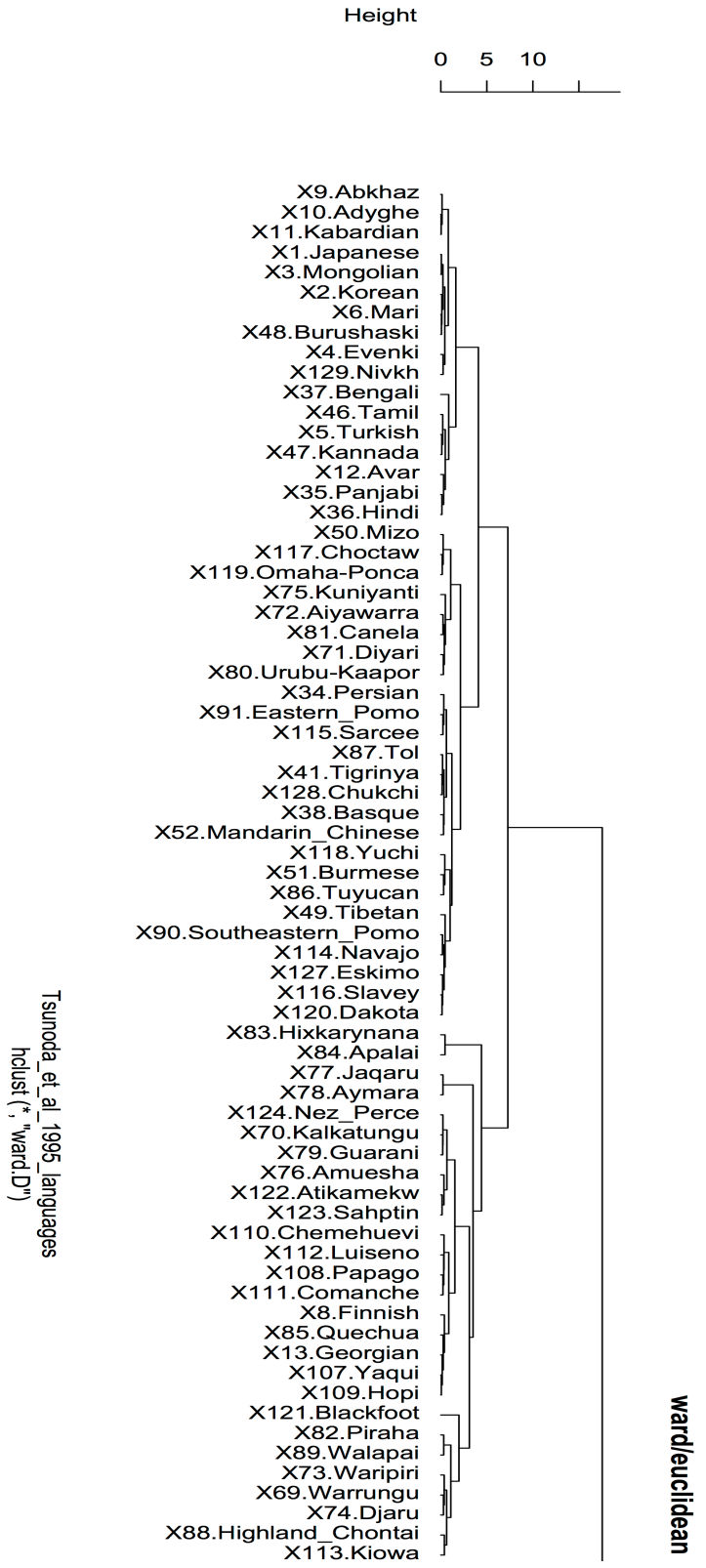


図 7-(a). MCA の 6 次元目までの布置からユークリッド距離に基づき距離行列を計算しウォード法によって言語のクラスター分析を実行した結果(その 1)

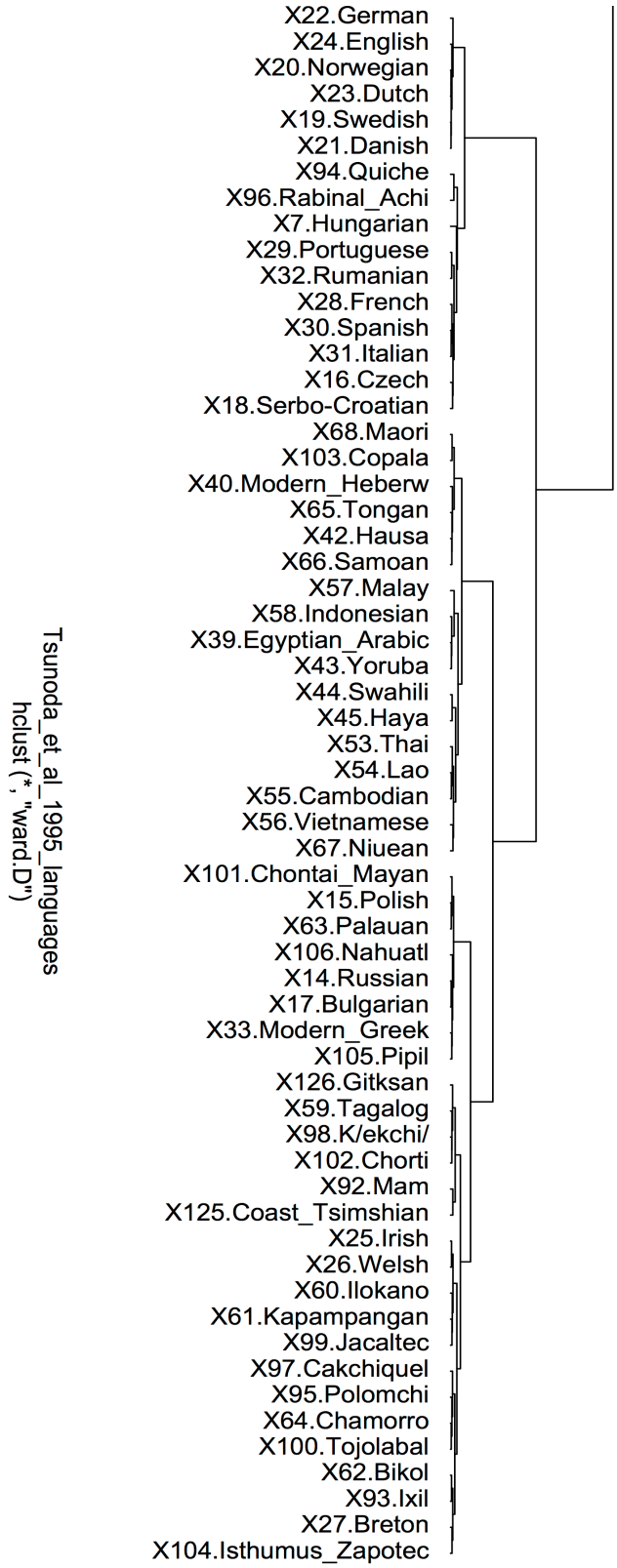


図 7-(b). MCA の 6 次元目までの布置からユークリッド距離に基づき距離行列を計算しウォード法によって言語のクラスター分析を実行した結果(その 2)

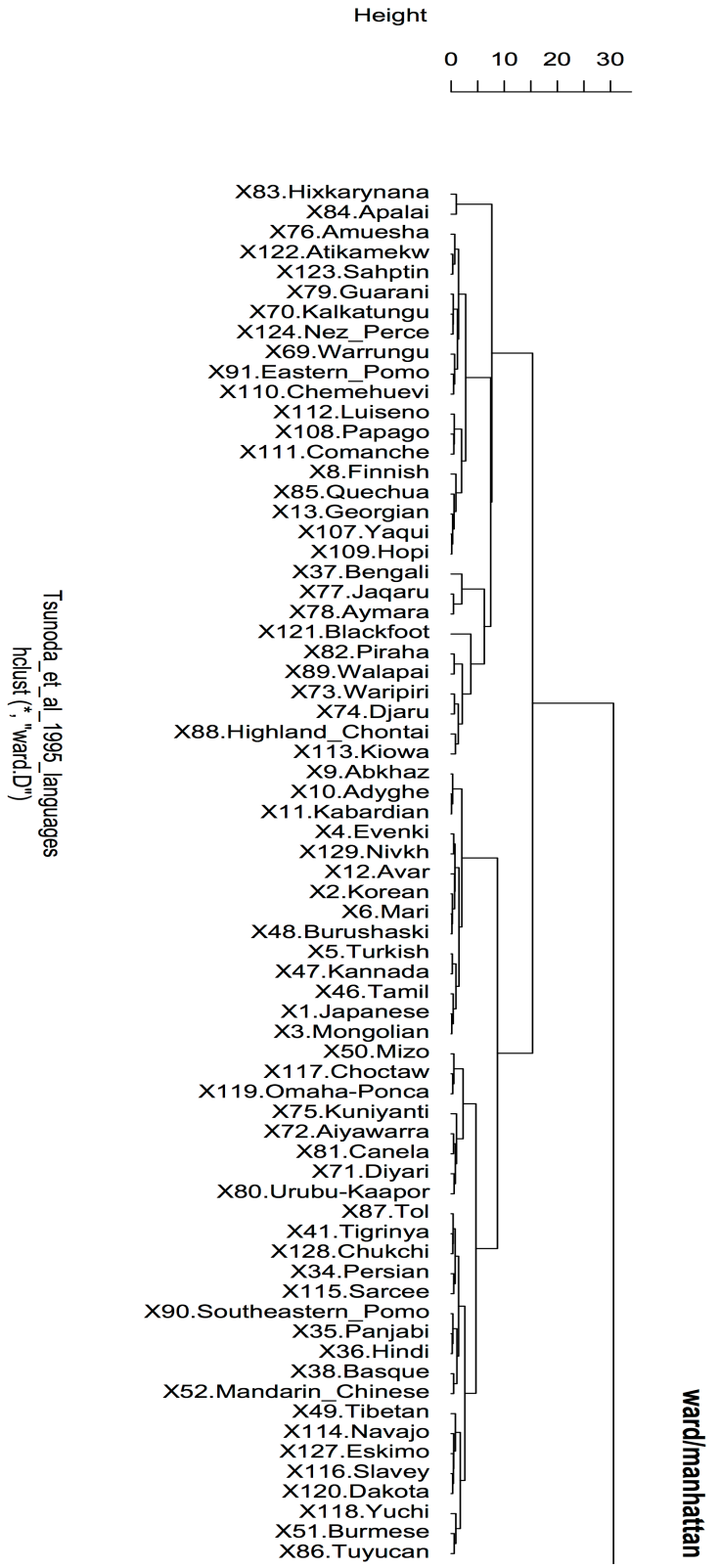


図 8-(a). MCA の 6 次元目までの布置からマンハッタン距離に基づき距離行列を計算しウォード法によって言語のクラスター分析を実行した結果(その 1)

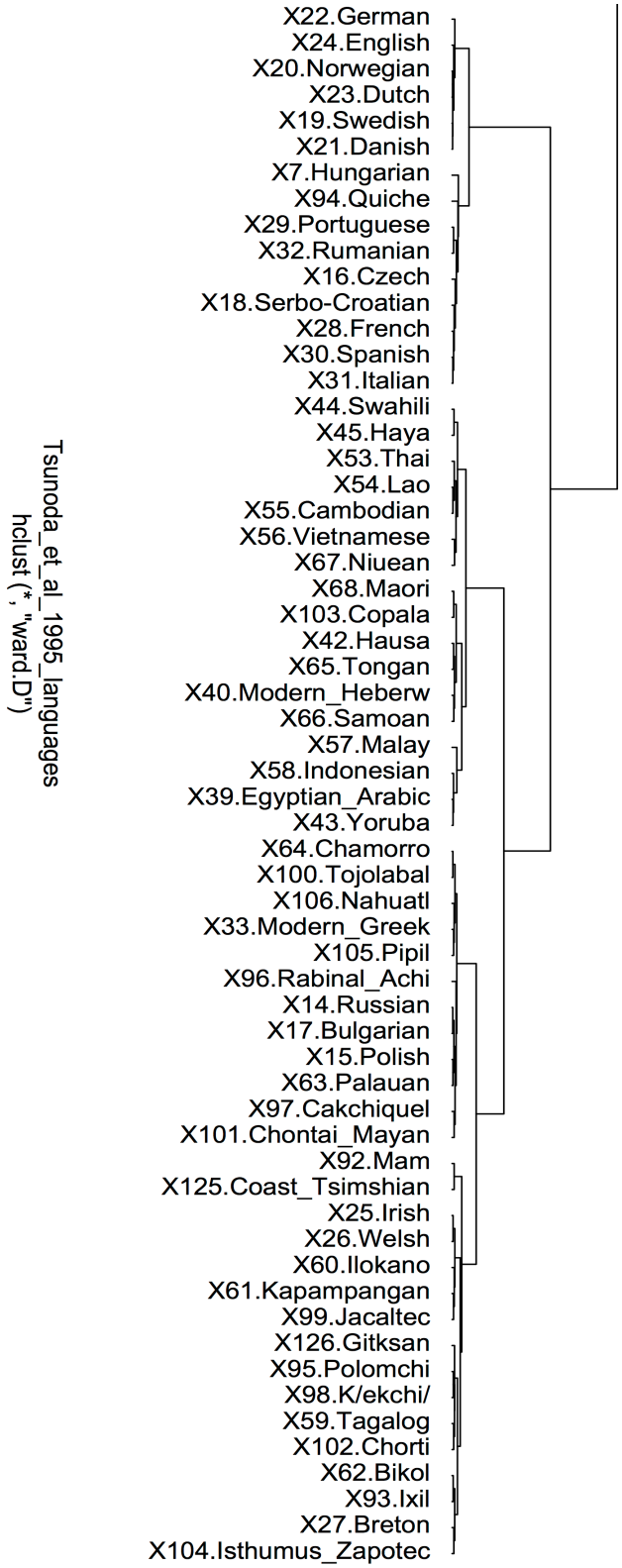


図 8-(b). MCA の 6 次元目までの布置からマンハッタン距離に基づき距離行列を計算しウォード法によって言語のクラスター分析を実行した結果(その 2)

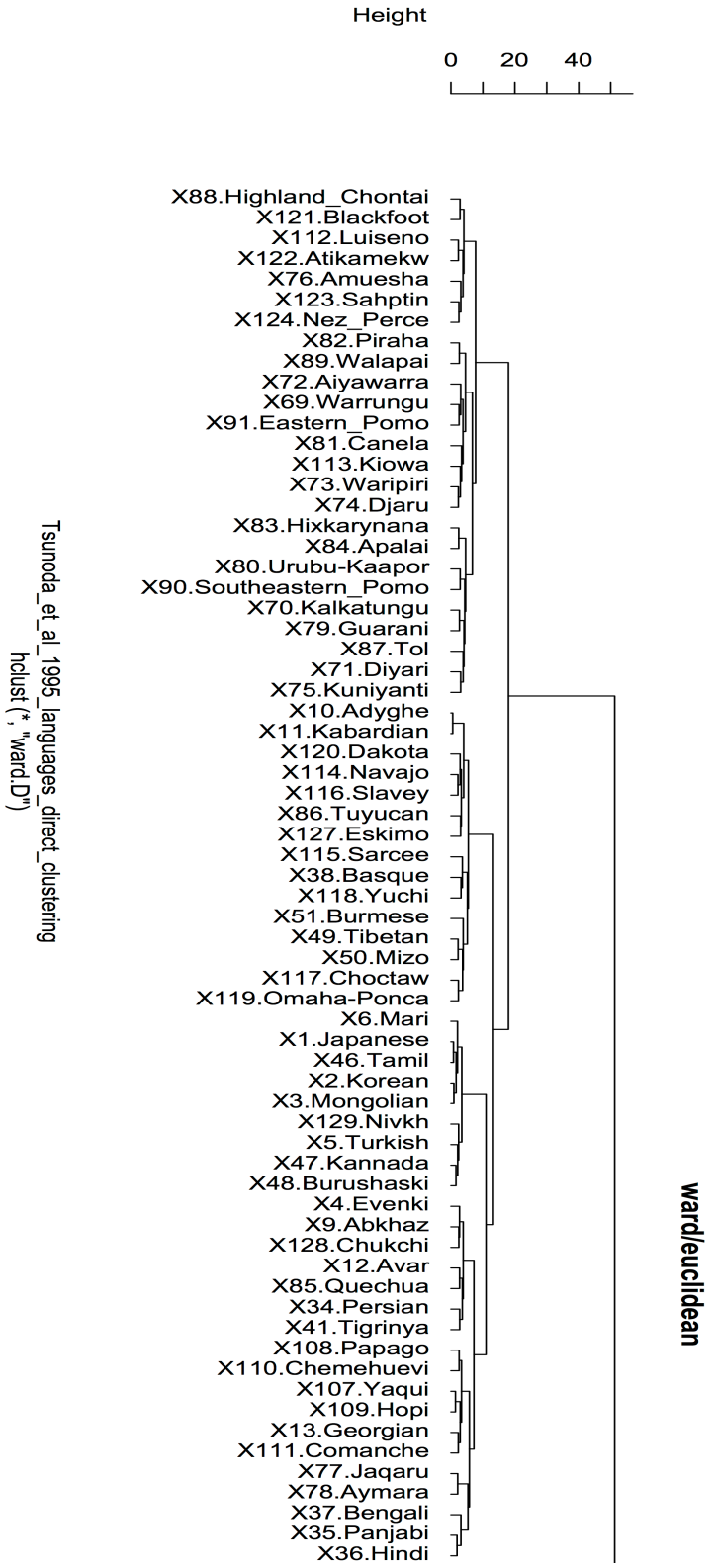


図 9-(a).素データの全変数を用いてユークリッド距離に基づき距離行列を計算し、ウォード法によって言語のクラスター分析を実行した結果(その 1)

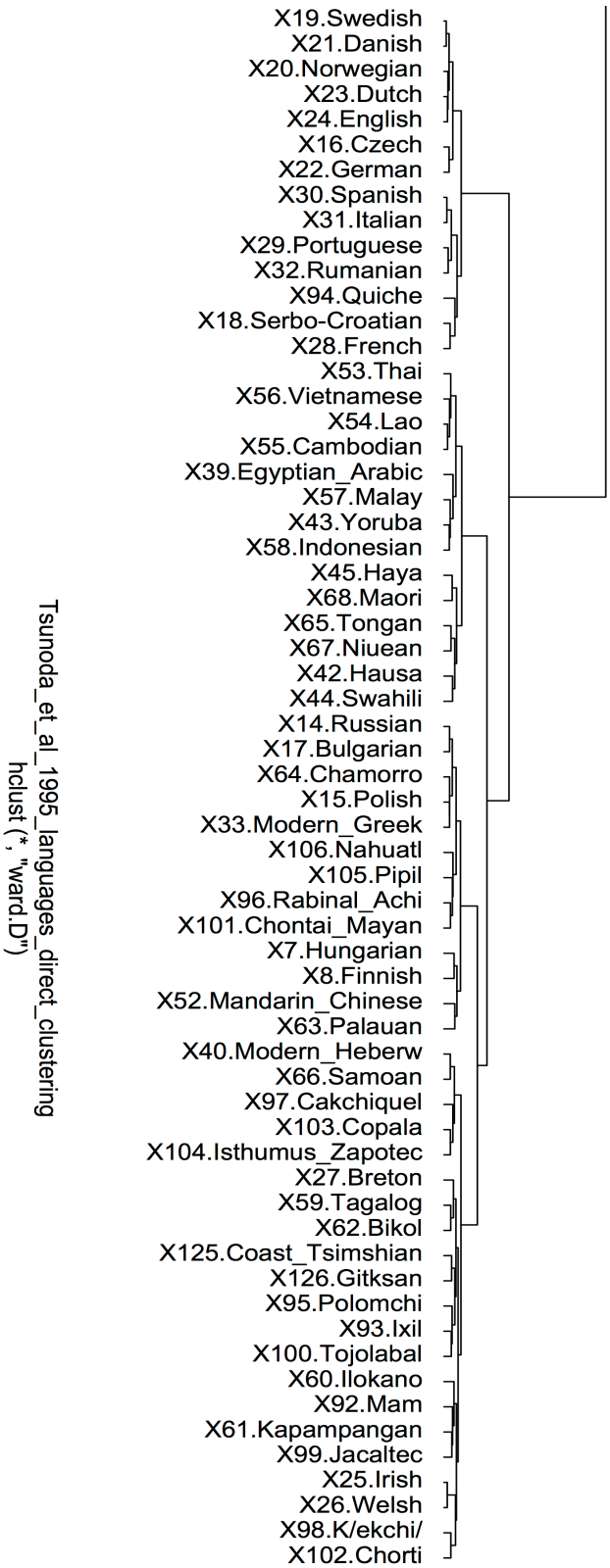


図 9-(b). 素データの全変数を用いてユークリッド距離に基づき距離行列を計算し、ウォード法によって言語のクラスター分析を実行した結果(その 2)

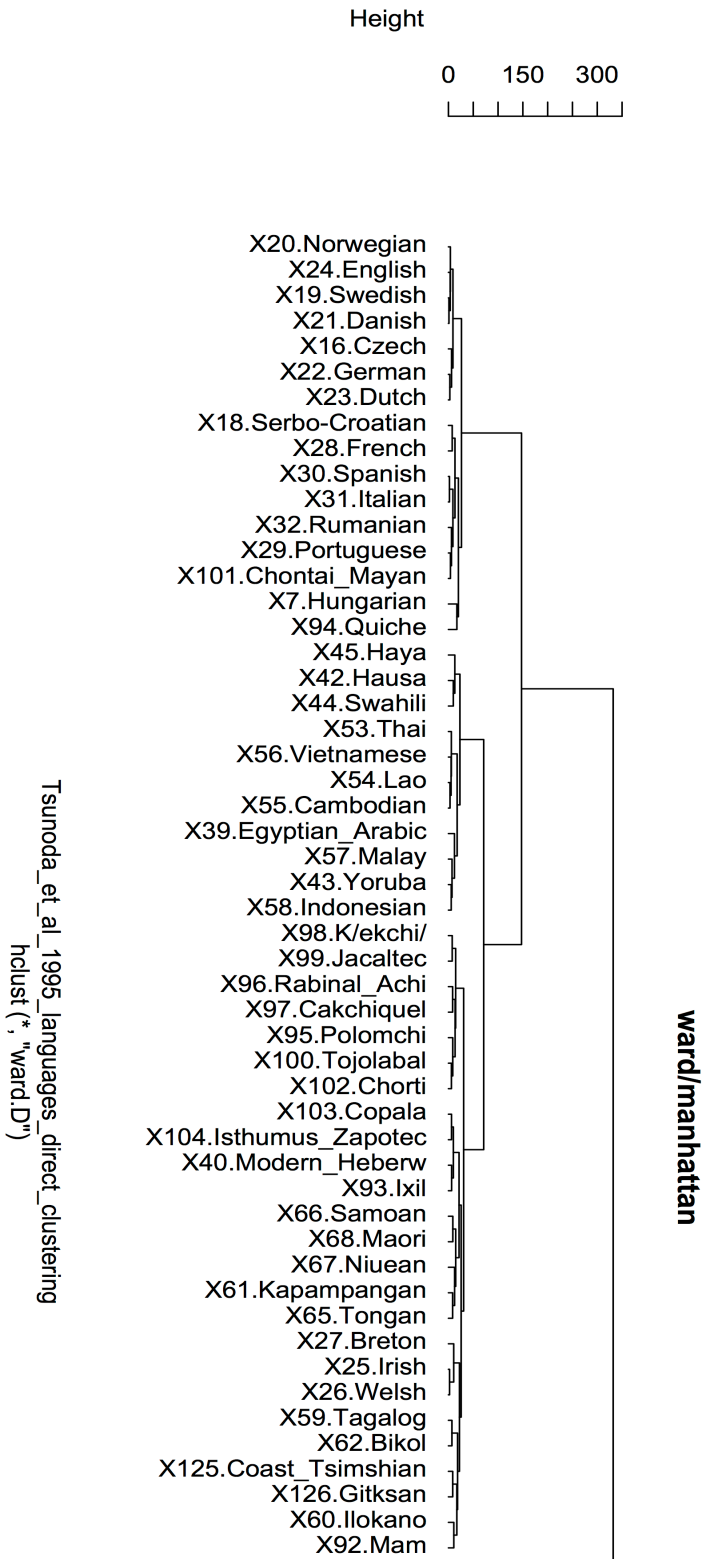


図 10-(a). 素データの全変数を用いてマンハッタン距離に基づき距離行列を計算し、ウォード法によって言語のクラスター分析を実行した結果(その 1)

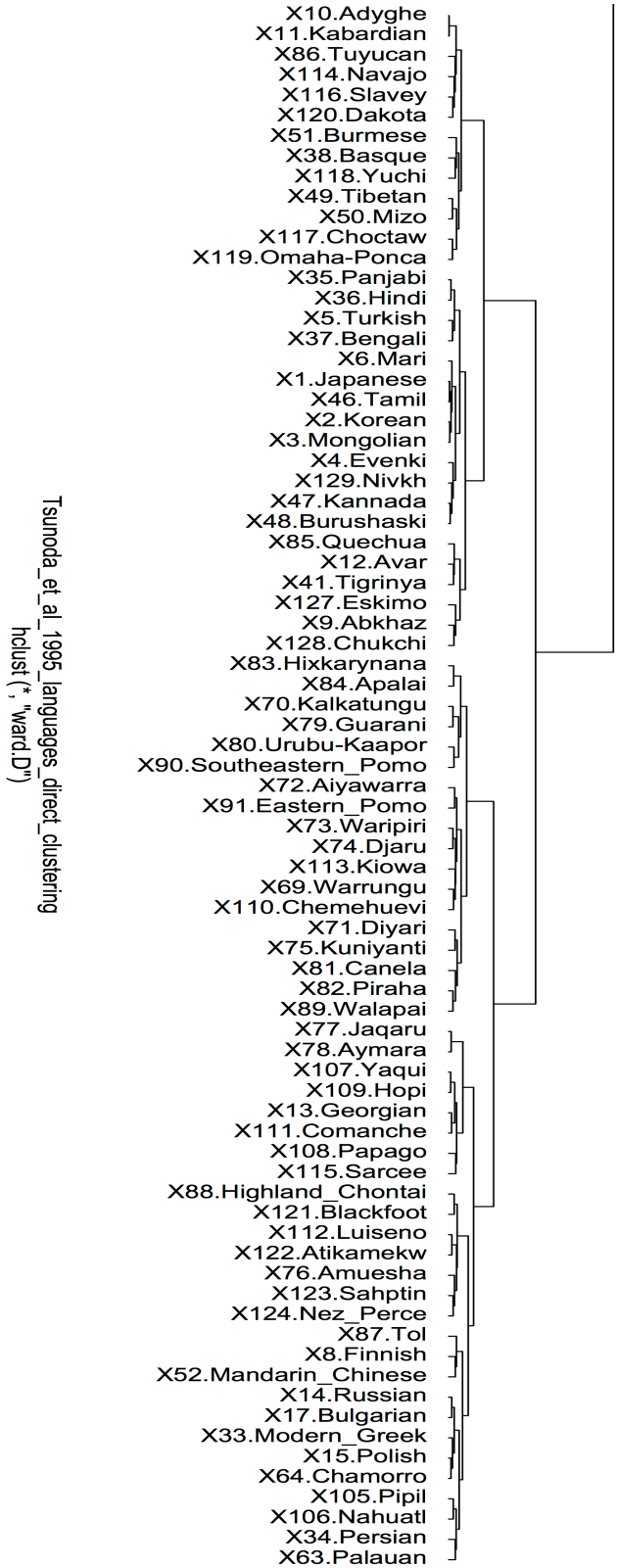


図 10-(b). 素データの全変数を用いてマンハッタン距離に基づき距離行列を計算し、ウォード法によって言語のクラスター分析を実行した結果(その 2)

図 7-(a)(b)、図 8-(a)(b)、図 9-(a)(b)、図 10-(a)(b)を見ると、129 の言語は、MCA の適用に関わらず、大きく 2 つのグループに分かれている。これらの 2 つのグループを 4.6.1 節で注目した主要部後置型、主要部前置型の(単)変数の 2 つの主カテゴリーに対して、2×2 のクロス表を作成し(2 つの主カテゴリー以外に属する言語は無視した)、対角セルの和の最大値をセルの合計値で割った値を図 7-(a)(b)に関して示す。

主要部後置型	vs.	主要部前置型	説明率
X1_1 [SOV]	vs.	X1_2 [SVO]	91.51%
X2_1 [後置詞] and X2_3 [無側置詞]	vs.	X2-2 [前置詞]	96.36%
X3_1 [所有格-名詞]	vs.	X3_2 [名詞-所有格]	87.10%
X4_1 [指示詞-名詞]	vs.	X4_2 [名詞-指示詞]	64.80%
X5_1 [数詞-名詞]	vs.	X5_2 [名詞-数詞]	45.24%
X6_1 [形容詞-名詞]	vs.	X6_2[名詞-形容詞]	52.34%

図 8 についても、大きな 2 つのグループに関しては図 7-(a)(b)と全く同じ分類が得られた。このことは、MCA の結果に対するクラスター分析の適用が、使用する距離に対してロバストな結果をもたらすことを意味する。

これらの結果から、Tsunoda et al. (1995a)の知見の通り X2[名詞と側置詞]が単変数としてはクラスターの 2 分類を最もよく説明できている。しかし、実際には、X2 だけではなく、X1[S, O と V]と X3[所有格と名詞]もほぼ同程度にクラスターの 2 分類を説明できている。このことから、Tsunoda et al. (1995a)の「側置詞が分類の指標としては最も優れている」という知見には注意を要することが明らかになった。

4.6.2.2 87 変数の分析の結果

次に、MCA の結果、6 次元目までの変数の布置からユークリッド距離に基づき距離行列を計算しウォード法によってクラスター分析を実行した結果を図 11 に、MCA の結果から、6 次元目までの言語の布置からマンハッタン距離に基づきウォード法を実行した結果を図 12 に、Tsunoda et al. (1995b)の素データからユークリッド距離に基づきウォード法によってクラスター分析を実行した結

果を図 13 に、Tsunoda et al. (1995b)の素データからマンハッタン距離に基づきウォード法によってクラスター分析を実行した結果を図 14 にそれぞれ示す。本研究で特に注目した X1_1, X1_2, X2_1, X2_2, X2_3, X3_1, X3_2, X4_1, X4_2, X5_1, X5_2, X6_1, X6_2 については頭に”〇”を付け加えた。

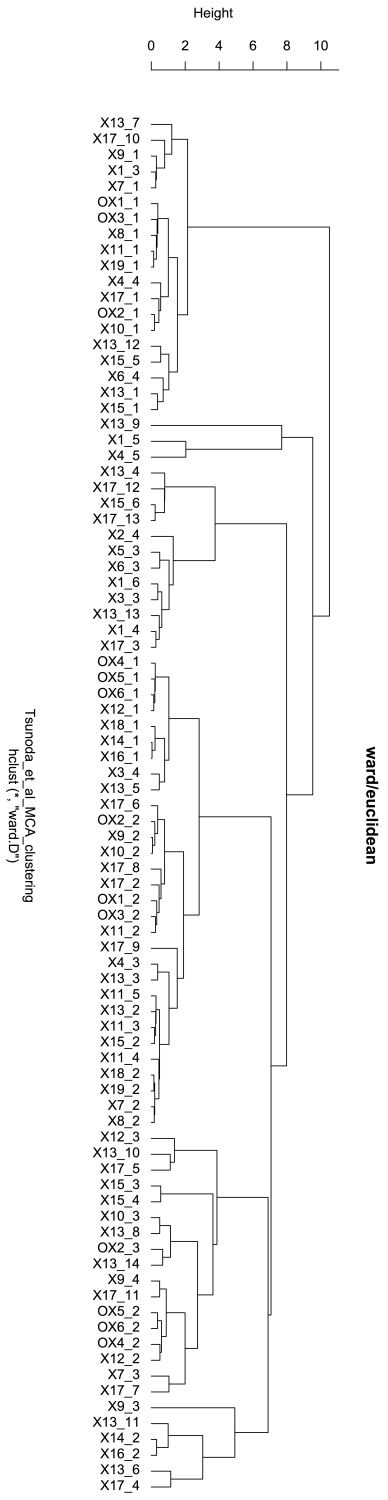


図 11. MCA の 6 次元目までの布置からユークリッド距離に基づき距離行列を計算しウォード法によって変数のクラスター分析を実行した結果。

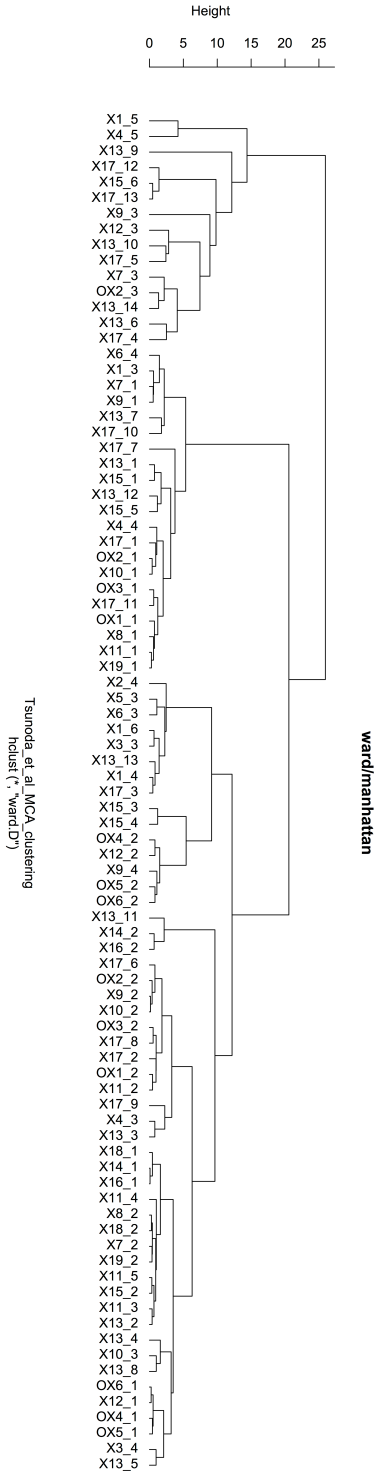


図 12. MCA の 6 次元目までの布置からマンハッタン距離に基づき距離行列を計算しウォード法によって変数のクラスター分析を実行した結果。

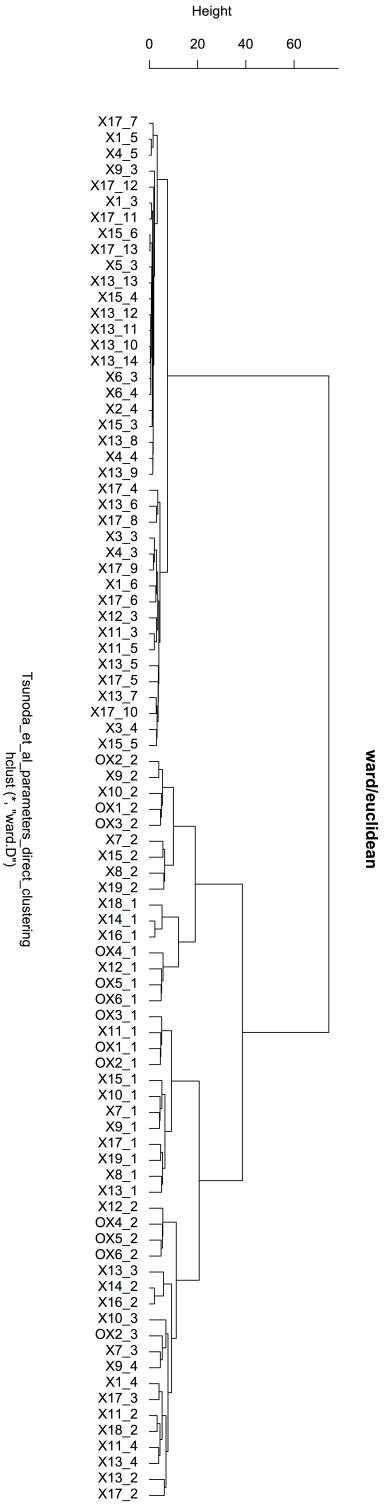


図 13. 素データの全変数を用いてユークリッド距離に基づき距離行列を計算しウォード法によって変数のクラスター分析を実行した結果。

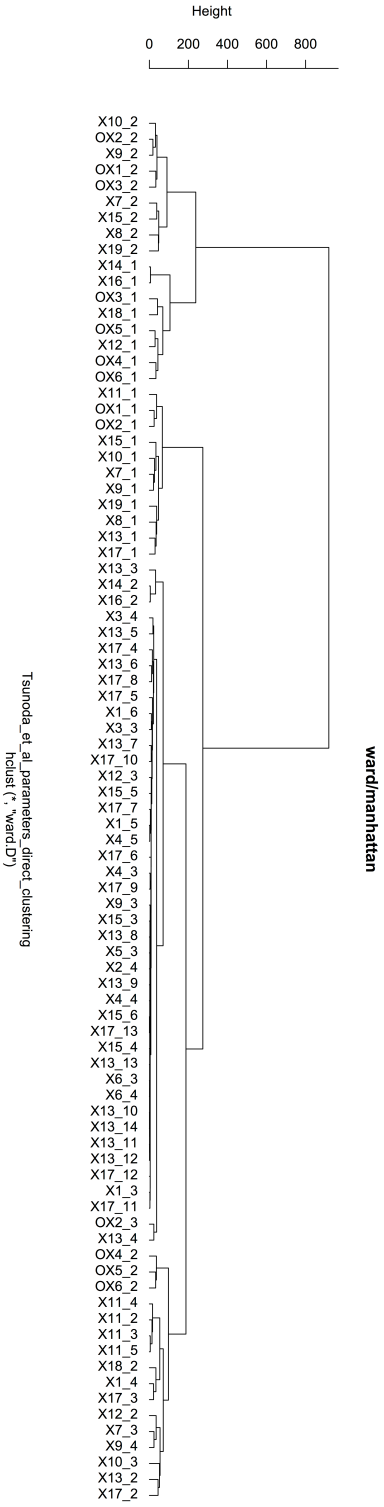


図 14. 素データの全変数を用いてマンハッタン距離に基づき距離行列を計算しウォード法によって変数のクラスター分析を実行した結果。

図 11 から図 13 について、一貫して 4 つの変数のクラスターが確認される。1) X1_1, X2_1, X3_1 からなる節レベルの主要部後置型の変数と名詞句内レベルの[所有格-名詞](主要部後置型)の変数のグループ、2) X1_2, X2_2, X3_2 からなる節レベルの主要部前置型の変数と名詞句内レベルの[名詞-所有格](主要部前置型)の変数のグループ、3)X4_1, X5_1, X6_1 からなる X3_1 以外の名詞句内部の主要部後置型の変数のグループ、4) X4_2, X5_2, X6_2 からなる X3_2 以外の名詞句内部の主要部後置型の変数のグループである(ただし、図 14 においては X3_1 と X4_1, X5_1, X6_1 が同じクラスターをなしていることには注意する必要がある)。

しかしながら、図 11 から図 13 の変数のクラスター分析からは、伝統的な言語類型論の知見とはやや異なった結果が観察される。図 11 と図 12 においては、主要部前置型の X1_2, X2_2, X3_2 のグループと主要部後置型の X4_1, X5_1, X6_1 が 1 つのクラスターに存在している。さらに、図 13 においては主要部後置型の X1_1, X2_1, X3_1 と主要部前置型の X4_2, X5_2, X6_2 が 1 つのクラスターに存在している。

従来の知見と違いがある中でも MCA の結果得られた図 6-(c)(d)と同様に、クラスター分析の結果においても 1)主要部前置型と主要部後置型という分類と 2)X3[所有格と名詞]以外の名詞句内部の変数かそれ以外かという分類は安定して観察された。

最後に、Tsunoda et al. (1995a)の「無側置詞型言語が語順の観点からは後置詞型言語と同じ振る舞いをする」という知見は注意を要する。図 11 から図 14 について一貫して X2_3[無側置詞]のカテゴリーは、主要部後置型のカテゴリーと同じクラスターをなしていない。先行研究とのこのような違いについての言語学的な含意の考察は、本論文の範囲を超えるが、言語類型論の立場からは興味深い洞察をもたらす可能性がある。

4.7 考察

4.7.1 本研究で得られた知見について

本研究の主要な結果は以下の 3 点である。第一に、Tsunoda et al. (1995a)の結論である「無側置詞型言語が語順の観点からは後置詞型言語と同じ振る舞いをする」ことは再考を要する。MCA の布置及び数量化Ⅲ類クラスタリングの結

果から、後置詞型言語、前置詞型言語、無側置詞型言語を 3 つの独立した類型として考えることがより適切といえる。第二に、Tsunoda et al. (1995a)の「X2[名詞と側置詞]が単変数としては最もよい分類の指標である」という主張には注意する必要がある。本研究でも確かに、X2 は単変数としては最も優れた分類の指標であったが、同程度に良い分類の指標として X1[S, O と V]と X3[所有格と名詞]が挙げられた。同じような結果は *WALS* を使った Albu (2006)でも得られている。 第三に、変数の分類において 1) X3[所有格と名詞]以外の名詞句内レベルの主要部前置型と主要部後置型の分類と 2)節レベルと X3[所有格と名詞]での主要部前置型と主要部後置型の分類が発見された。本来、名詞句内レベルの変数である X3[所有格と名詞]がなぜ節レベルの変数と同じクラスターを形成しているかについては Whitman and Ono (2017, to appear)を参照されたい。言語類型論の立場からは、この発見自体が言語学にとって重要であるだけではなく、角田が独自の言語学知見をもって収集分析し数値化したデータベースを MCA で分析した本研究の知見と、Tsunoda et al. (1995b)とは異なったデータベースである *WALS* のデータベースに MCA を適用した結果である Whitman and Ono (2017, to appear)が同じ知見に至ったことも大変重要な結果である。なぜならば、上記の知見が言語間の変数のパターンとして頑強なものであることを示唆しているためである。

4.7.2 本研究の知見の背景について

本研究は、先行研究である Tsunoda et al. (1995a)や Whitman and Ono (2017, to appear)とはやや異なった知見が得られた。本節では、そのような結果が得られた背景について考察する。

第一に、Tsunoda et al. (1995b)のデータベース上の誤分類が「無側置詞型言語が語順の観点からは後置詞型言語と同じ振る舞いをする」という Tsunoda et al. (1995a)の知見を生んだ可能性がある。Tsunoda et al. (1995b)で「無側置詞型言語」と分類されている 19 の言語のうち、Avar, Burmese, Alyawarra, Aymara, Quechua (Imambura, Huallaga), Walapai, and Luiseño の 8 つの言語は *WALS* においては「後置詞型言語」に分類されており、*WALS* と Tsunoda et al. (1995b)の両方で「無側置詞型言語」と分類されているのは Blackfoot と Kiowa の 2 言語だけである。それ以外の Tsunoda et al. (1995b)で「無側置詞型言語」と分類されている言語は *WALS* では NA である。「無側置詞型言語が語順の観点からは後置詞型言語と同じ振る舞いをする」という知見はこのことに

起因する可能性がある。ただし、そのようなデータの問題がありながらも本研究の MCA の結果は、後置詞型言語、前置詞型言語、無側置詞型言語を 3 つの独立した類型を示唆していることから、今後さらなる言語学的、統計学的な追究を行う必要がある。

第二に、Tsunoda et al. (1995b)で選ばれた言語のサンプルには系統的にも地域的にも偏りがあった。129 言語の系統的な分類を以下に示す。(各系統に属する言語の数は、系統名の後ろに付け加えた)

129 言語の系統的な分類

(1)系統不明: 6 (日本語等),
(2) アルタイ語族: 3,
(3) ウラル語族: 3,
(4) 北西コーカサス語族: 3,
(5) 北東コーカサス語族: 1,
(6) 南コーカサス語族: 1,
(7) 印欧語族: 24,
(8) アフロ・アジアティック語族: 4,
(9) ニジェール・コルドファニアン語族: 3,
(10) ドラビダ語族: 2,
(11) シナ・チベット語属: 4,
(12) カム・タイ語族:2,
(13) モン・クメール語族: 2,
(14) オーストロネジア語族: 12,
(15) パマ・ニュンガン語族: 2,
(16) ブナバン語族: 1,
(17) アラワカン語族: 1,
(18) ハケ語族: 2,
(19) トゥピ・グアラニ語族: 2,
(20) ジエ語族: 1,
(21) ムラ語族: 1,
(22) カリブ語族: 2,
(23) ケチュア語族: 1,
(24) トゥカノアン語族: 1,
(25) ホカン語族: 5,
(26) マヤ語族: 11,
(27) オト・マングエアン語族: 2,
(28) ウト・アズテカ語族: 8,
(29) カイオワ・タノアン語族: 1,
(30) アサパスカン語族: 3,
(31) スー語族: 2,
(32) ムスコギアン語族: 1,
(33) アルゴンキアン語族: 2,
(34) サハプティアン語族: 2,
(35) ツィムシアン語族: 2,
(36) エスキモー・アリュート語族: 1,
(37) チュコトゥコ・カムチャトゥカン語族: 1.

Ethnologue (Lewis, Gray & Charles, 2013)によれば、現在話者が生きている言語のうち 30.2%はアフリカに、14.9%はアメリカ大陸に、32.4%はアジアに、4%はヨーロッパに、18.5%は太平洋に存在している。しかし、Tsunoda et al.

(1995b)においては、アフリカの言語は 4.2%しか含まれないのに対して、ヨーロッパの言語は全体の 20.2%を占める。よって、Tsunoda et al. (1995a)は系統的にも地域的にも非常に偏ったデータベースに基づいていることが分かる。今後の課題としては、より系統的にも地域的にもバランスのとれたデータベースを用いて統計的言語類型論の研究を進めることが挙げられる。実際、筆者は Whitman and Ono (2017, to appear)において *WALS* に MCA を適用し、同様の検討をしている。

4.8 今後の課題

MCA を使った Statistical Typology の研究として、今後の研究課題としては 2 つの方向性が考えられる。第一に変数の連関について、本研究は探索的なものであったがより仮説検証に適した MCA に関連した統計手法を応用することによって変数間の連関について、既存の言語類型論における仮説を統計的に検証することが考えられる。

第二に、本研究では MCA とクラスタリングとの併用のなかで、言語のクラスタリングに寄与している変数を探すことによって、言語類型論においてより本質的な変数を絞り込むことがある程度可能であった(MCA では上位の数次元を考察することによってそのことが達成された)。しかし、本質的な変数を取り除いたときに(主に言語の)クラスターにどのような特徴が現れるかということを検証することも大きな課題となろう。

従来からの質的な言語類型論における洞察に基づいた仮説を、MCA に関連した仮説検証的な手法を用い統計的な立場から検討を行う相補的なアプローチが、今後の言語類型論の実質科学的な発展に寄与することが期待される。

4.9 統計的補足：Neighbor-Net 及び MCA_Neighbor-Net の適用

本節では、Tsunoda et al. (1995b)のデータに対して Neighbor-Net を適用することによって、クラスター分析では捉えることのできなかったデータの 2 番目以降に強い構造を視覚的に捉えることで、言語学的に有益な情報を引き出すことを目的とする。図 15 は Tsunoda et al. (1995b)のデータに対して、MCA を実行し得られた距離行列に対して Neighbor-Net を適用したものである(次元数は MCA クラスターと同様に 6 次元とした)。図 16 は Tsunoda et al. (1995b)のデータからユークリッド距離に基づいて距離行列を計算し得られた距離行列に

対して、Neighbor-Net を適用したものである。図 15、図 16 はそれぞれ X1 に関して、図 17、図 18 はそれぞれ X2 に関して、図 19、図 20 はそれぞれ X3 に関してラベリングしている。

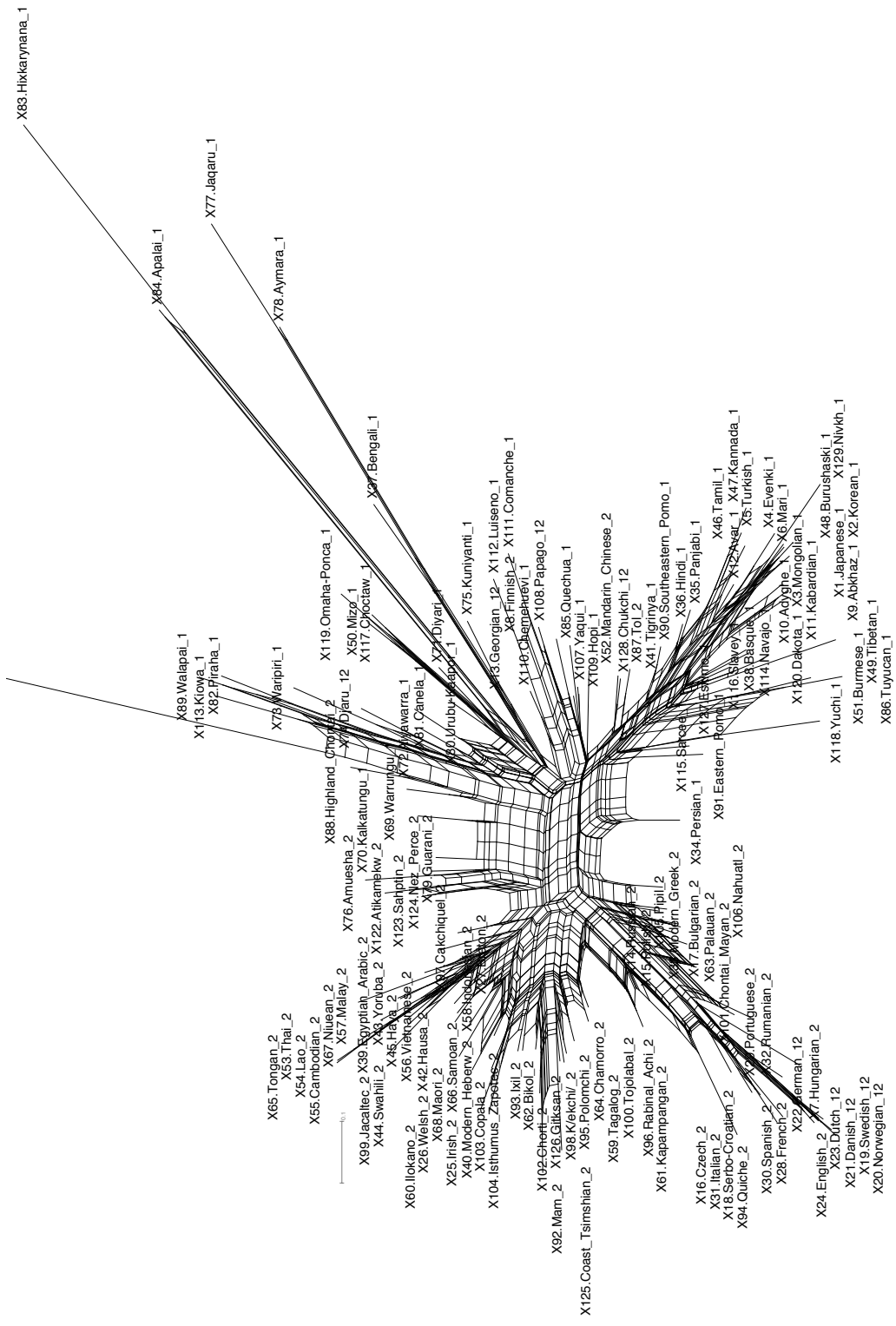


図 15. Tsunoda et al. (1995b)のデータに対して、MCA_Neighbor-Net を適用したもの。X1 についてラベリングしている。Blackfoot については、枝が長かったため省略した。

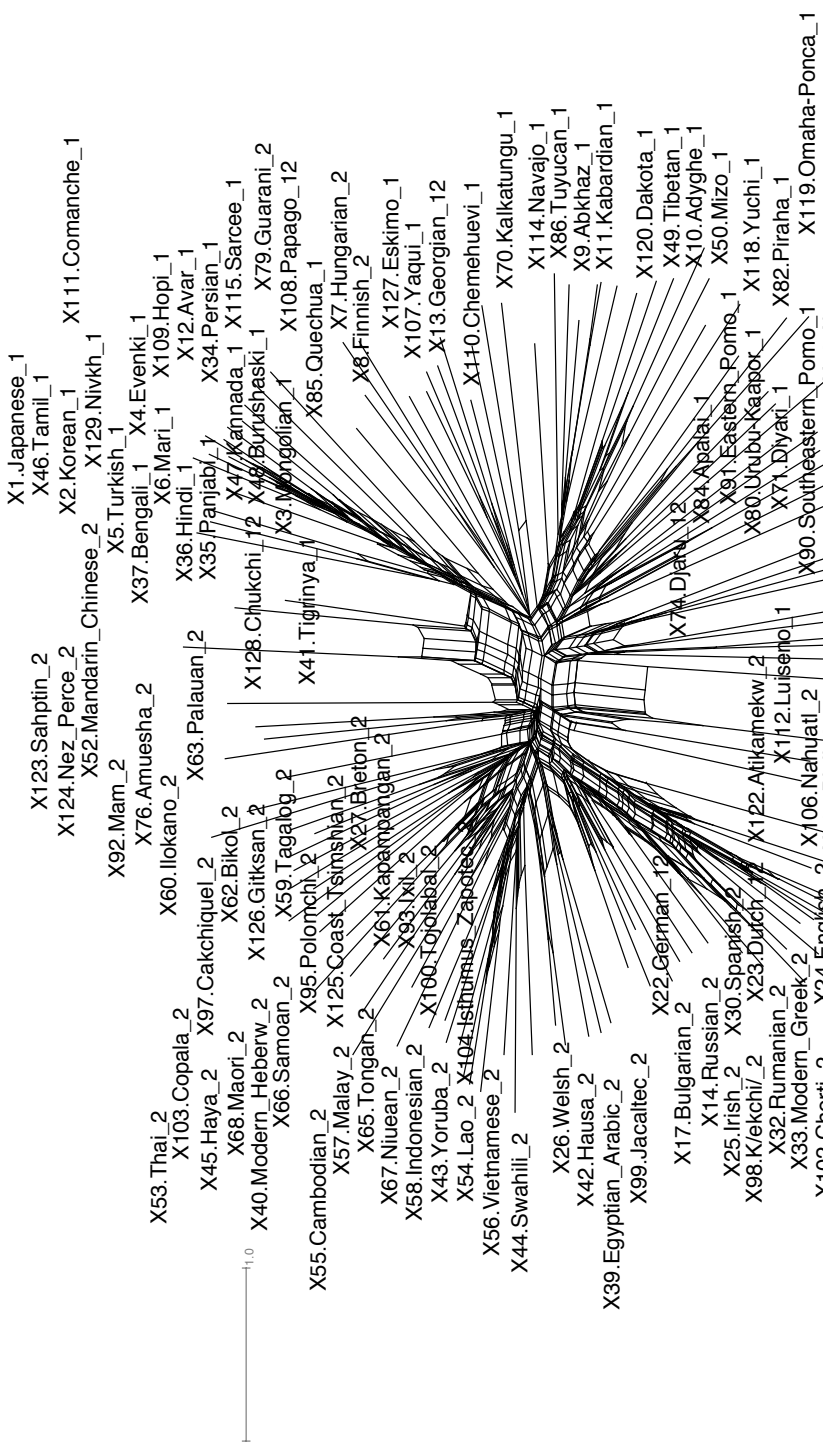


図 16. Tsunoda et al. (1995b)のデータに対して、Neighbor-Net を適用したもの。X1 についてラベリングしている。

図 17. Tsunoda et al. (1995b)のデータに対して、MCA_Neighbor-Net を適用したもの。X2 についてラベリングしている。Blackfoot については、枝が長かったため省略した。

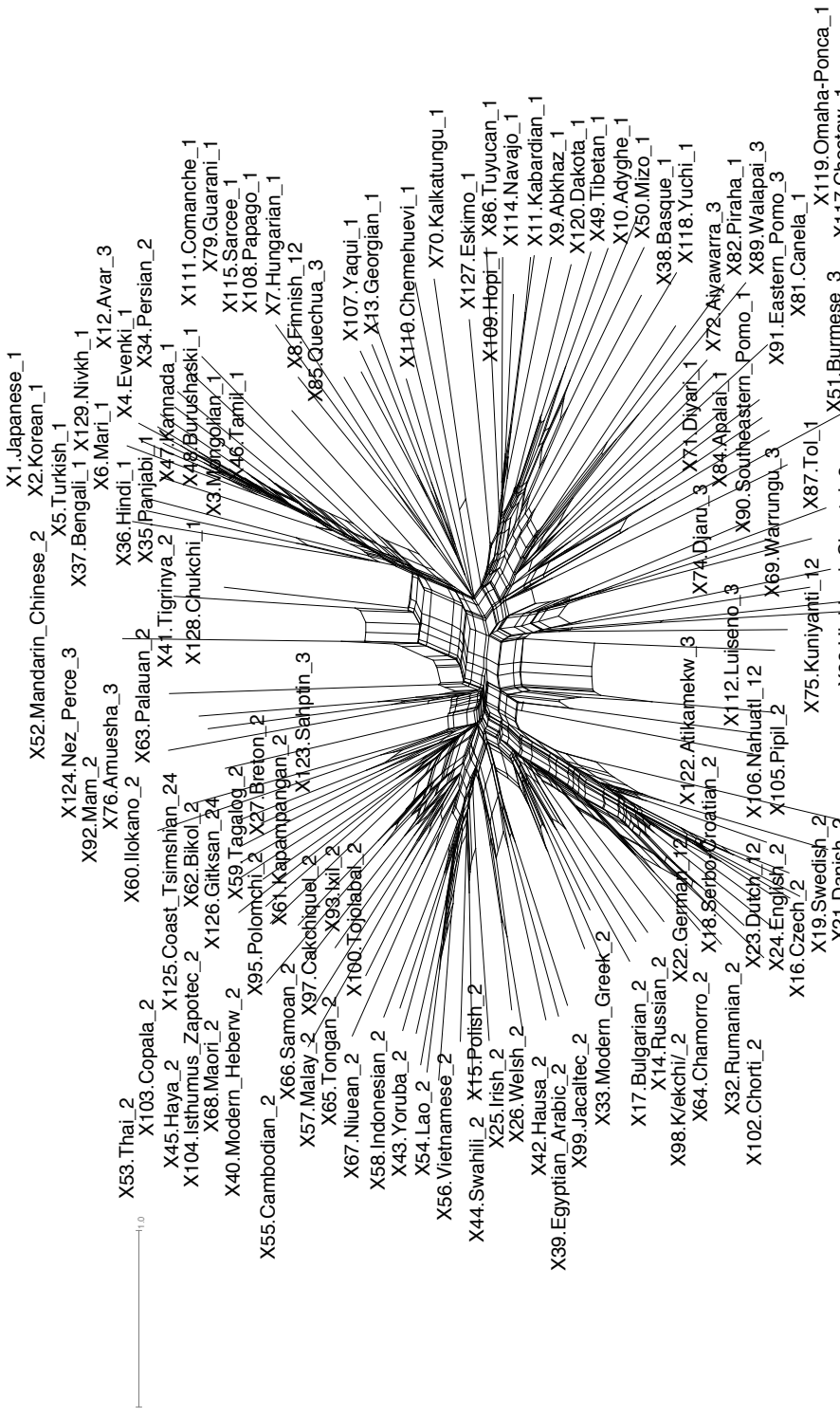


図 18. Tsunoda et al. (1995b)のデータに対して、Neighbor-Net を適用したもの。X2 についてラベリングしている。Blackfoot については、枝が長かったため省略した。

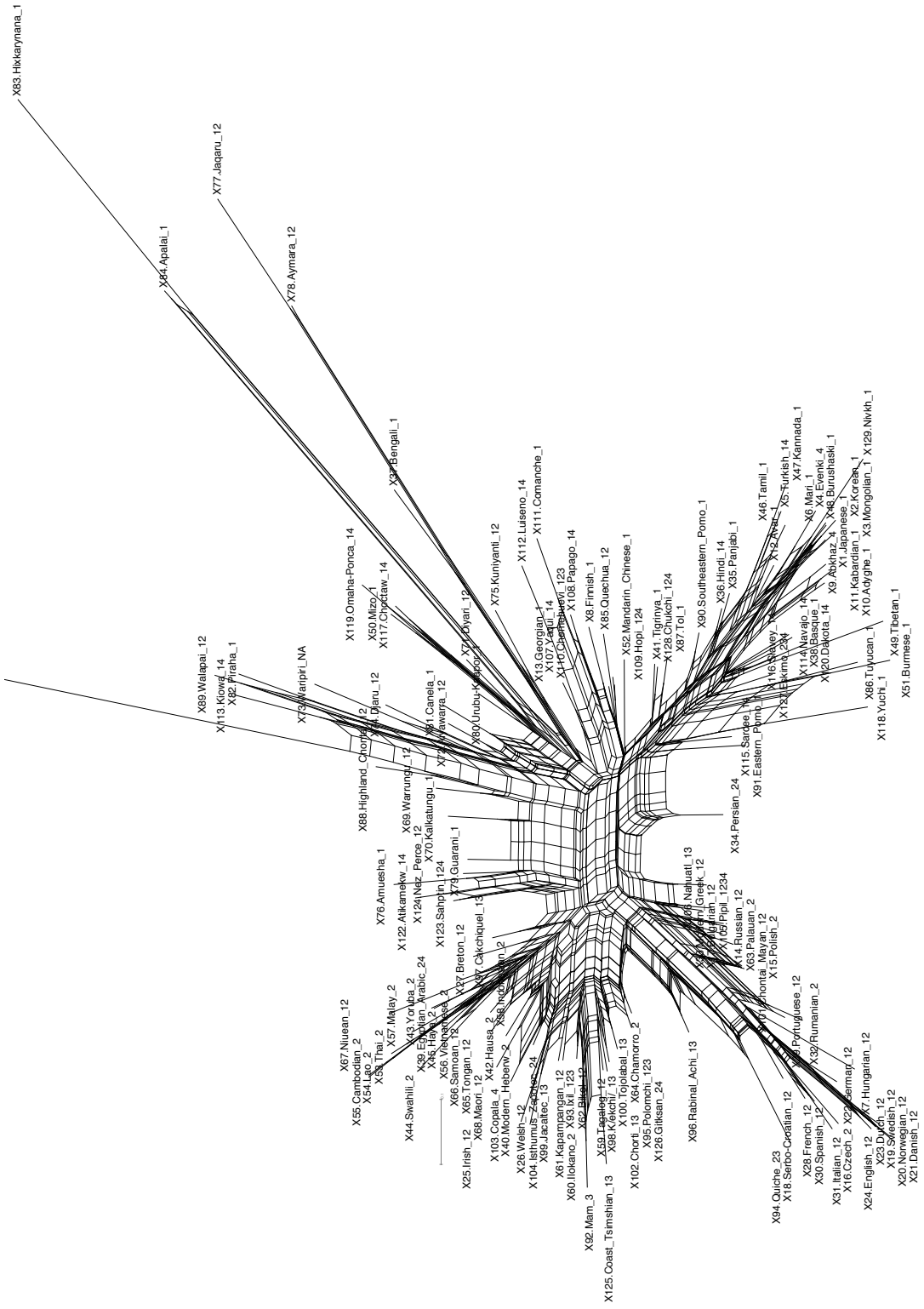


図 19. Tsunoda et al. (1995b)のデータに対して、Neighbor-Net を適用したもの。X3 についてラベリングしている。Blackfoot については、枝が長かったため省略した。

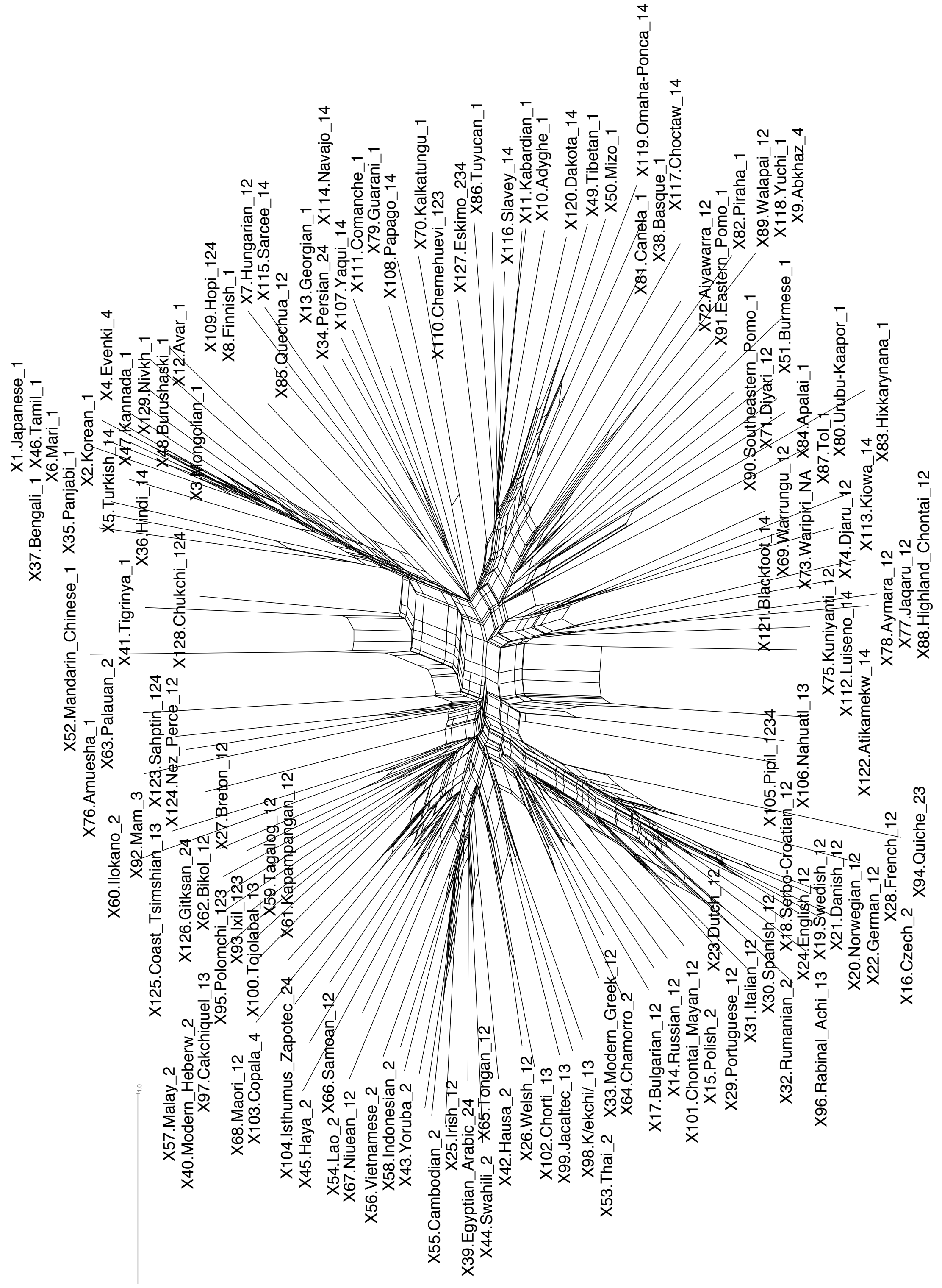


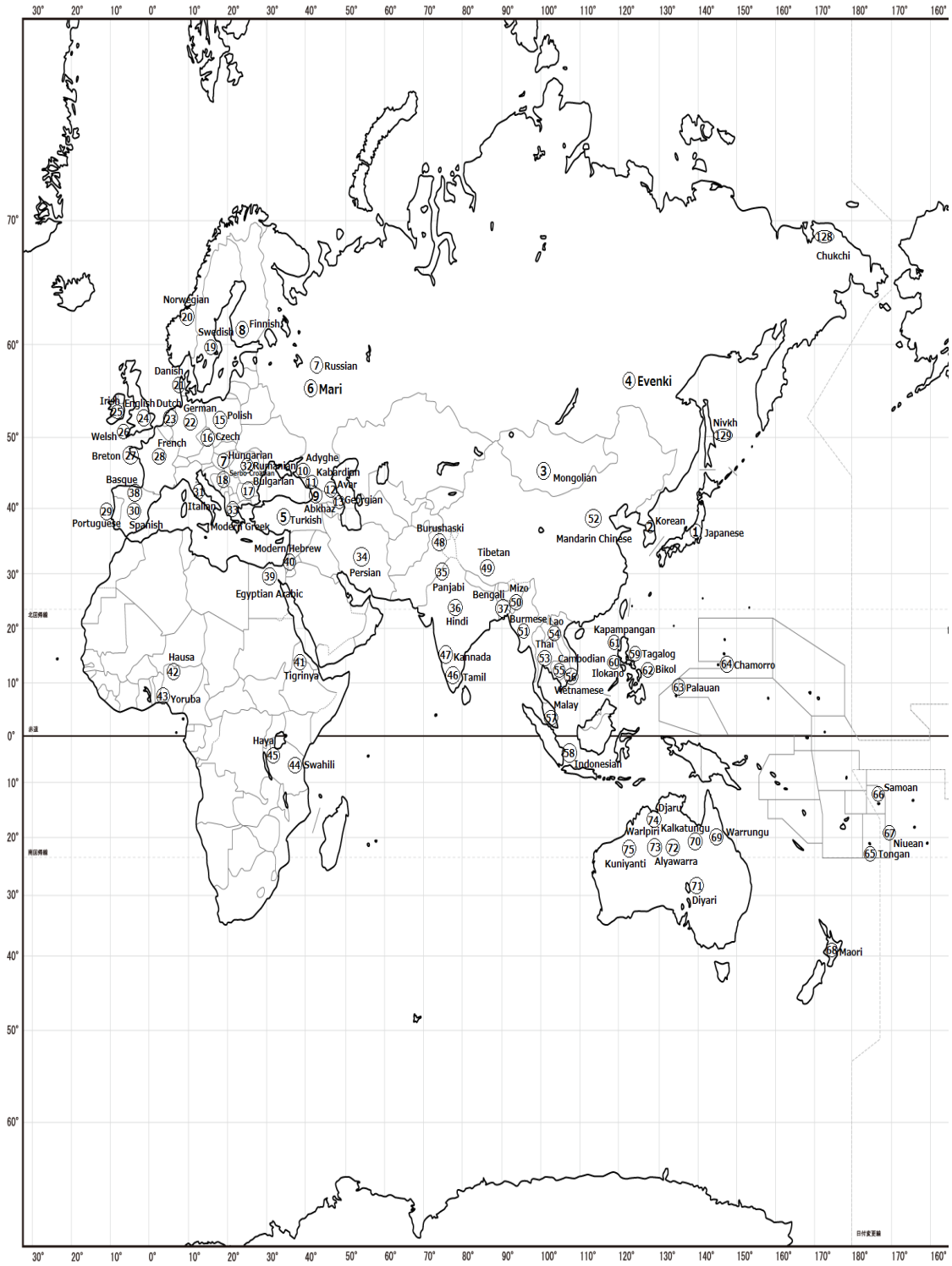
図 20. Tsunoda et al. (1995b)のデータに対して、Neighbor-Net を適用したもの。X3 についてラベリングしている。

図 15、図 16 においては、はっきりと図の上側に OV 型言語が集まり、図の下側に VO 型言語が集まっている。

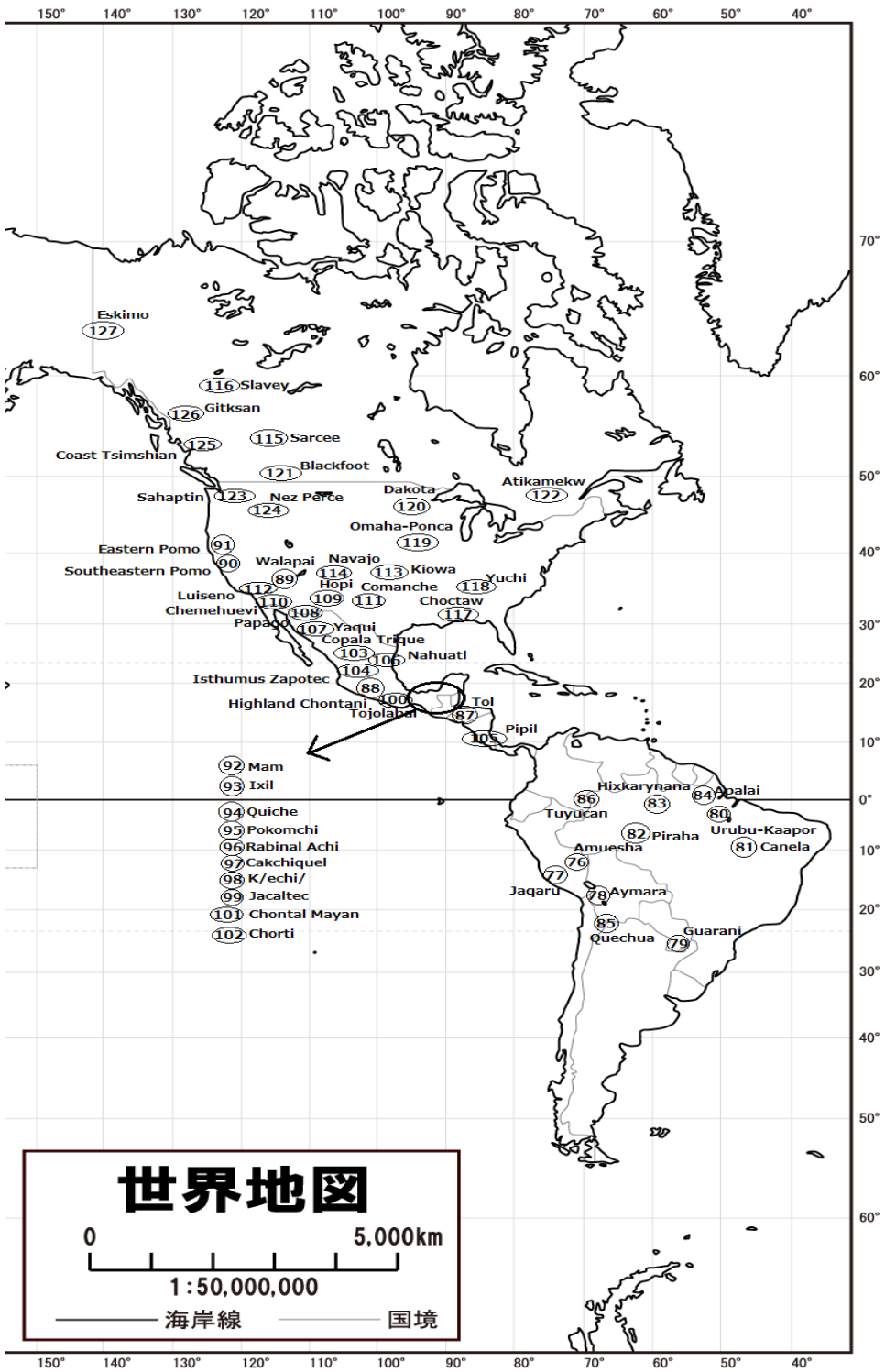
図 17、図 18 においてもは図の上側に後置詞型言語(1)が集まり、図の下側に前置詞型言語(2)が集まり、それらの中間に無側置言語(3)が集まっている。

ただし、図 19、図 20 においては「所有格−目的語」型言語が上側に「目的語−所有格」型言語が下側に集まっているものの、X1 及び X2 に関する図ほど、明確には分かれていない。このことは、X1 及び X2 に対して X3 の説明率がやや低いことを反映していると考えられる。

このように、4.6.2 節で「数量化Ⅲ類クラスタリング」を適用したデンドログラムの観察では 129 の言語が大きく 2 分類される構造が強調されるが、MCA_Neighbor-Net や Neihgbor-NetAnalysis では、4.6.1 節の MCA の結果の注意深い観察で得られる 3 分類がより柔軟に図に表現されている。今後の課題として、この図を言語学的知見により詳細に検討することで言語類型論における新たな知見を発見することが期待される。



4 章付録図 1. 4 章で取り上げた 129 の言語の位置を示した世界地図


150° 140° 130° 120° 110° 100° 90° 80° 70° 60° 50° 40°
70° 60° 50° 40° 30° 20° 10° 0° 10° 20° 30° 40° 50° 60°
Eskimo 127
116 Slavey
126 Gitksan
125 Coast Tsimshian
115 Sarcee
121 Blackfoot
Sahaptin 123 Nez Perce 124
Dakota 120
Atikamekw 122
Omaha-Ponca 119
Eastern Pomo 91
Southeastern Pomo 90
Walapai 89
Navajo 114
113 Kiowa
Yuchi 118
112 Hopi
Comanche 111
Luiseno 110
109
Chemehuevi 108
Papago 107
Choctaw 117
Yaqui
Copala Trique 103
106 Nahuatl
104
Isthumus Zapotec 88
Highland Chontani 100
Tojolabal 87
Tol
Pipil 105
92 Mam
93 Ixil
94 Quiche
95 Pokomchi
96 Rabinal Achi
97 Cakchiquel
98 K/echi/
99 Jacaltec
101 Chontal Mayan
102 Chorti
Hixkaryana 86
83
Palai 84
Tuyuca 80
82 Piraha
Urubu-Kaapor
Amuesha
81 Canela
76
77
Jaqaru
78 Aymara
85
Quechua
Guarani 79
世界地図
0
5,000km
1:50,000,000
海岸線
国境

5 章 アイヌ語方言の分類に関する再検討
—基礎語彙の同根性データに基づいて[1]

5.1 はじめに

本章はアイヌ語方言に関する基礎語彙の同根性データを収集、分析した服部・知里(1960)を統計科学的に再検討する。服部·知里(1960)は昭和 30 年から 31 年にかけて、北海道 13 地点、樺太 6 地点の計 19 地点のアイヌ語の方言について、大規模な調査を行っている。これは、日本における、基礎語彙統計学的研究の嚆矢であり、また今日までに多くのアイヌ語の話者が亡くなっており、アイヌ語の資料の保存という観点からも記念碑的な論文である。彼らは、得られたデータから、各方言の対どうしの類似度を定義し、アイヌ方言全体の関係について推測を行っている。残念ながら、その類似度の定義はアドホックな重みづけの感は否めないし、他方で、おそらく計算機の性能などに関する当時の時代的制約上、自由に統計処理を行える環境にはなかったであろう。

その後、計算機環境の改善もあり、Asai(1974)は服部·知里(1960)のデータに基づき、アイヌ語の方言に関するクラスター分析をおこなっている。しかし、これもデータの扱い方について統計学的に改善の余地があるように思われる。例えば、服部·知里(1960)のデータがカテゴリカルデータであるという特徴を考慮せず、方言間の類似度をアドホックな重みづけで計算していることが挙げられる。また、服部·知里(1960)において、アイヌ語の方言 A の単語 c と方言 B の単語 c ’の同根性について(何らかの理由で)「判断することができない」と分類される場合が少なくないが、そのようなデータを除外して解析しているのは、相応の貴重な情報が活用できていないのではないかと考えられる。

さらに、近年の Lee and Hasegawa(2013)の研究でも、服部·知里(1960)のデータをもとに、ベイズ法を用いたアイヌ方言の系統樹推定を行っているが、服部·知里(1960)のデータがカテゴリカルデータであるという特徴を考慮せず、また「同根性を判断することができない」と分類されたデータを除外している点は、Asai(1974)と同様であり、改善の余地がありそうである。

本章では、服部·知里のデータを統計的により適切と思われる取り扱いをしてクラスター分析を試み、言語地理学的な観点からもより妥当な結果を追究する。この中で、多重対応分析(MCA)とクラスター分析を組み合わせた方法や、

[1]本章は、雑誌「北方人文研究」の 8 巻(pp.25-41)の小野(2015b)を加筆修正したものである。

Neighbor-Net を組み合わせた MCA_Neighbor-Net と筆者が称する方法も適用し、今までのクラスター分析では見落とされていたアイヌ語における方言周圏論的構造を明らかにすることも試みた。

「同根性を判断することができない」と分類されたデータを分析に取り入れ、MCA を適用した上で求めた距離行列にクラスター分析を施すことにより、樺太の 6 つの方言が東海岸と西海岸に分かれた。このことは樺太に南北に山脈が走り、海上の沿岸交通を主としているという地理学的知見と一致している。

さらに、MCA の結果得られた距離行列に対して、データの「最も強い」構造のみを取り出すクラスター分析ではなく、「2 番目以降に強い」構造を取り出せる Neighbor-Net を適用すること(MCA_Neighbor-Net)によって、中川(1996)で指摘されていた沙流・千歳を中心とした方言周圏論的構造、沙流・千歳・樺太型及び西蝦夷型・東蝦夷型を確認することができた。これらの結果は、先行研究がデータの「最も強い」構造のみを取り出すクラスター分析を用いていたことによって、明らかにされていなかった「2 番目以降に強い」構造を MCA_Neighbor-Net によって視覚化できたことを意味する。

さらに、これらの結果と、「同根性を判断することができない」と分類されたデータを除外したデータに MCA を適用し得られた距離行列に Neighbor-Net 分析を適用した結果を比較したところ、沙流・千歳を中心とした方言周圏論的構造を確認することはできたが、沙流・千歳・樺太型及び西蝦夷型・東蝦夷型は確認されなかった。このことは、本研究で扱った服部・知里(1960)のデータにおいては、「同根性を判断することができない」と分類されたデータが有意味な情報を持っていることを示唆した。

本章では、上述の解析結果について以下の各節で詳述する。

5.2 研究の背景

ここでは、本章の分析に至った地理的背景について簡単に述べておく。

図 1 に、アイヌ語が以前話されていたカムチャッカ半島、千島列島及び樺太、また現在アイヌ語が話されている北海道の全体図を図 1(地理院地図, 2016)に示す。次に、北海道の全体図を図 2-(a)(地理院地図, 2016)に、樺太の地形図を図 2-(b)に示す(浅井, 1933)。さらに、図 3 には、服部・知里(1960)で収集され、本章の分析対象とした樺太方言、及び北海道方言に対応する地図を服部・知里(1960)より引用する。

図 1. アイヌ語が以前話されていたカムチャッカ半島，千島列島及び樺太，また現在アイヌ語が話されている北海道の全体図(地理院地図, 2016)

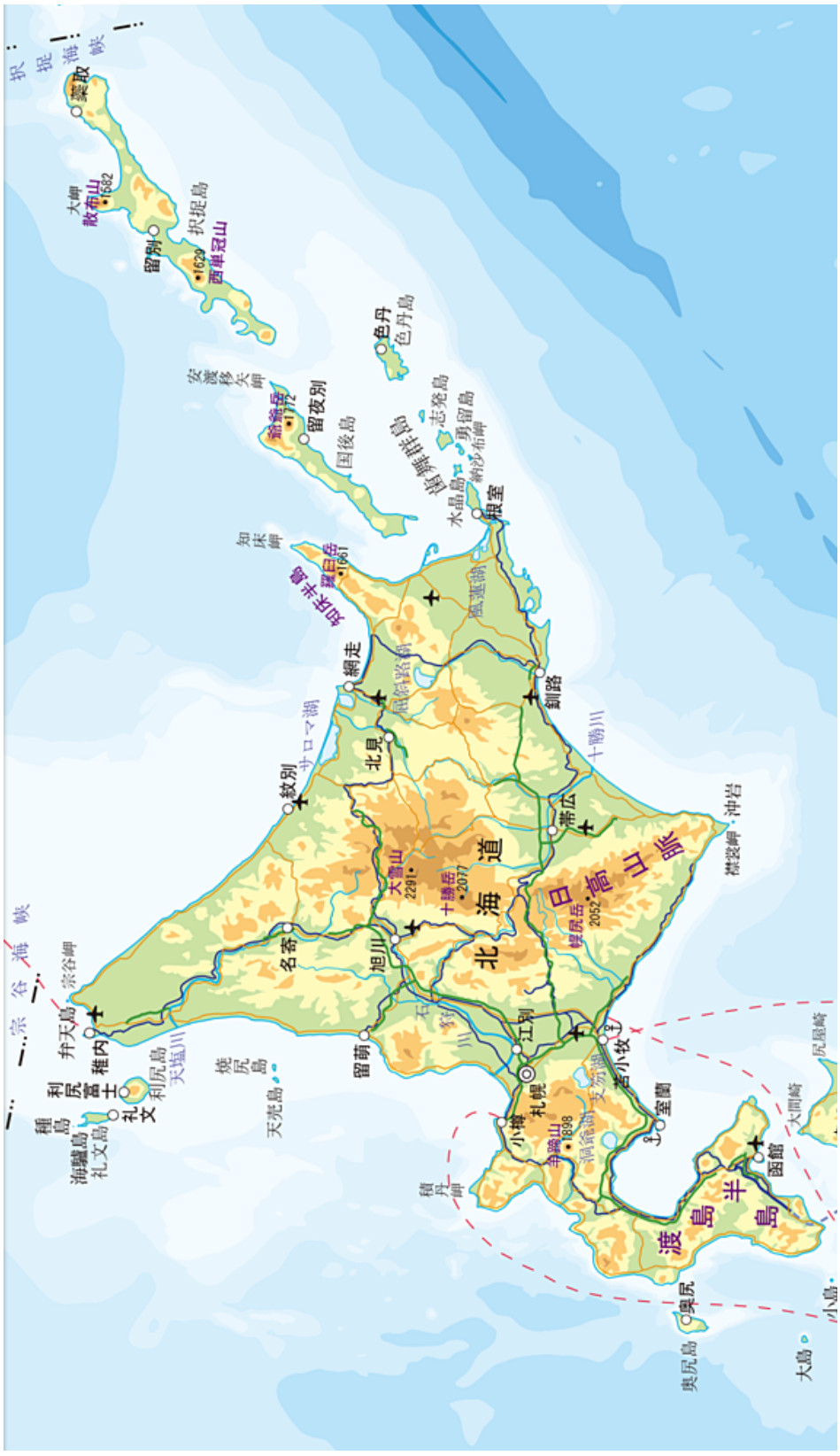


図 2-(a). 北海道の全体図(地理院地図, 2016)

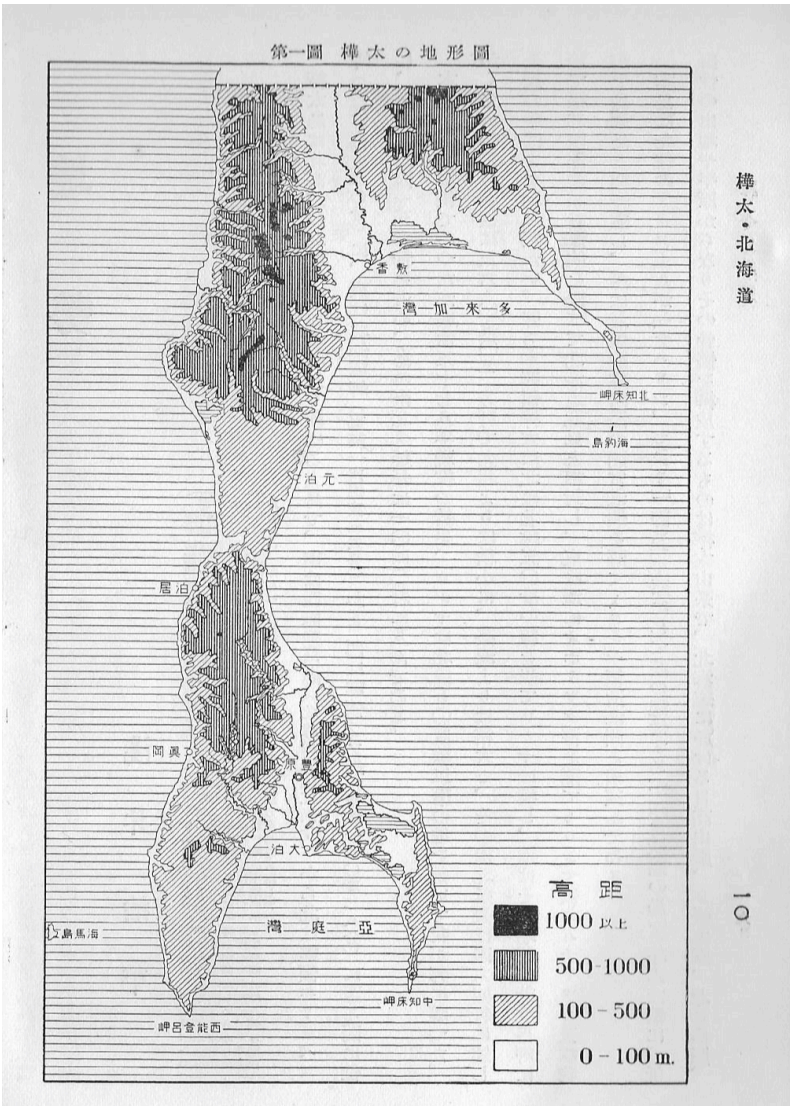


図 2-(b). 樺太の地形図(浅井, 1933)

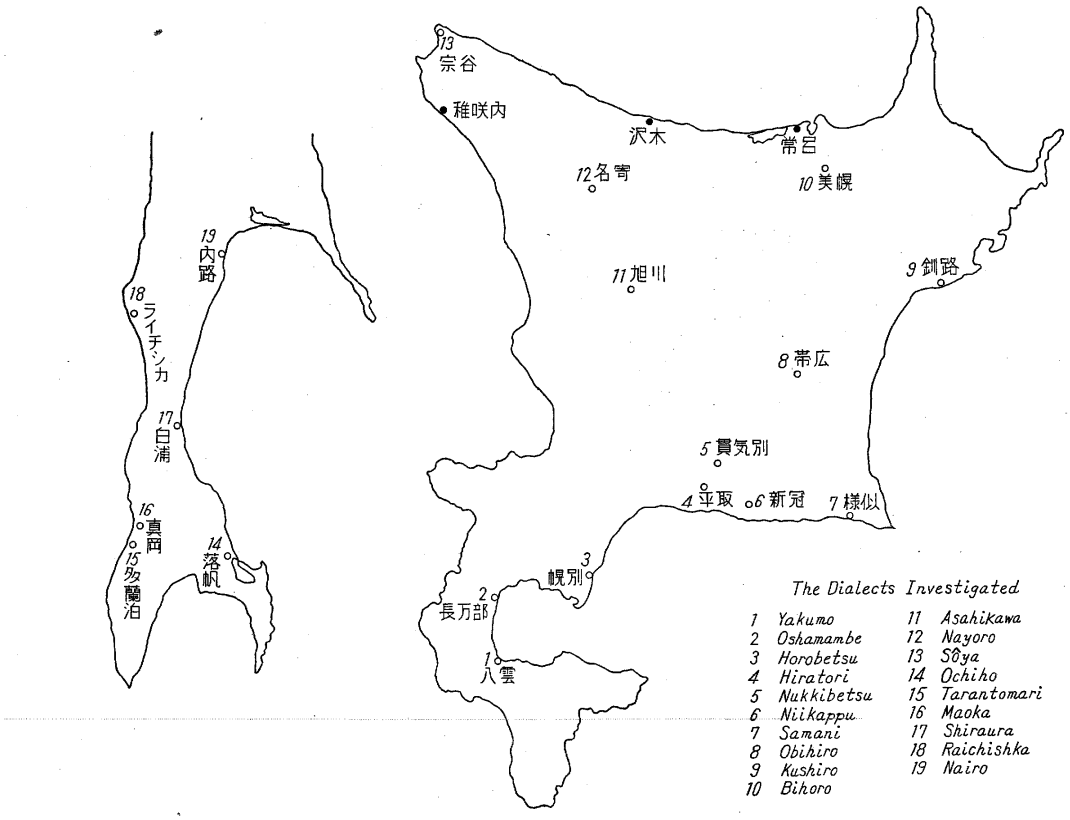


図 3. 分析の対象とした樺太方言、及び北海道方言に対応する地図(服部・知里, 1960)

南樺太には図 2-(b)の通り、南北に山脈が通っており，浅井(1933)によれば交通も主に沿岸の海路である。南樺太の東海岸と西海岸について、浅井(1933)の記述が参考になる(旧仮名遣い及び旧字は新仮名遣い及び新字に筆者が直した)。

“海岸は概して単調で多来加、亜庭の二湾と能登呂、中知床、北知床、の三半島以外には水平的肢節は頗る貧弱である。殊に西海岸の如気は鵜城火山群が造る鈍き一突起の単調を破る以外、二三の鈍い膨起が緩い弧を描くのみで殆ど南北に走ると称して差支なく、只本島海岸の特色たる段丘が殊に中央部によく発達し、その下には小平野を展開し、小湾を控へて鰊漁の小中心をなし本島漁場の優秀地たる自然の条件を与えている。(中略)東海岸は南方中知床岬から愛郎岬までは山岳が岸に迫って沿岸は絶壁でなければ段丘が直ちに海に迫っている”

(浅井, 1933 pp.25-26)

5.3 節に詳述するが、先行研究においてはこれらの地理的事情が言及されていないのではないかと筆者は考えた。

特に、樺太に南北の山脈が走り、沿岸には山が迫っている地形であり、交通は主に沿岸交通を主としているという浅井(1933)の記述から、樺太の方言は東海岸と西海岸にわかれるのではないか、推測した。

したがって、本研究の目指すところは統計的により適切な手法を用いながら、言語地理学的により妥当な解釈を与える解を追究することである。

5.3 先行研究について

5.3.1 先行研究の概要

5.3.1.1 服部·知里(1960)

服部·知里(1960)は、Swadesh(1955)の基礎語彙 200 語をもとに独自の言語学的な知見に基づき実際に調査する基礎語彙を選択し、19 の地点の語彙の同根性について、以下の表 1 のような基準にのっとり、調査データをまとめている。

表 1: 服部·知里(1960)での語彙の同根性に関する基準。服部·知里(1960, p.309)より作成

＋	語根が互に対応し合うもの
―	語根が互に対応しないもの
±	比較さるべき形態素が、一方または両方の方言に二つ(以上)あり、そのうち一つどうしの関係が＋で、他方の関係が―であると認められる場合。被調査者が、問題の二つ(以上)の単語の使用法の違いについて詳しく且つ正確な報告ができ、その結果意義素が正確に記述されている場合には、そのうちどちらかを採るべきか明らかなのが普通だが、主としてこれらの条件が欠けたために、ただ二つ(以上)の単語が並べ報告されている場合には、±という記号をつけるという処置をとらざるを得なかった。
○	対応関係の不明な場合、または比較のためにいずれの形態素をとるべきか、不明の場合。
?	調査結果に疑問のある場合
・	一方の方言に、該当の単語のない場合、被調査者が答えられなかった場合
( )	調査漏れ

( )の調査漏れのデータは統計で言えば典型的な欠測値(missing data)の問題であるが、「○, ?, ・」については、欠測値と扱って良いか明らかではない。

この表 1 から、彼らは以下のような手続きで、方言間の類似度を求めている。

1. ＋は 1 とカウントする
2. ±は＋の部分を 0.5 と、−の部分を 0.5 とカウントする
3. ―は 1 とカウントする
4. それ以外の記号は無視する
5. 方言間の語彙に関して、＋, ±, −のカウント総数に対する、同根のカウント(＋は 1、±の＋は 0.5 とカウントする)の割合(%)を「方言の類似度」として用いる。

例えば、仮に 2 つの地点、八雲と長万部について、＋が 60 個、±が 30 個、−が 10 個の場合である場合は、同根と判断されたカウント数は 60×1＋30×

0.5=75 となり、またカウント総数は 60×1＋30×1＋10×1=100 であるので、八雲と長万部との類似度は 75/100=75%となる。

このような計算を各方言間に関して行うと、表 2 が得られる。

表 2: アイヌ語方言間の類似度(%). 服部・知里(1960: 338)より

第 3 表	19. 内路 Na	18. ライチシカ Ra	17. 白浦 Sh	16. 真岡 Ma	15. 多蘭泊 Ta	14. 落帆 Oc	13. 宗谷 Sô	12. 名寄 Ny	11. 旭川 As	10. 美幌 Bi	9. 訓路 Ku	8. 帯広 Ob	7. 様似 Sa	6. 新冠 Ni	5. 貫気別 Nu	4. 平取 Hi	3. 幌別 Ho	2. 長万部 Os
1. 八雲 Yakumo	72.0	70.5	71.9	70.8	70.3	73.3	84.8	86.7	89.5	87.7	88.2	87.5	88.7	89.1	88.8	90.3	91.8	96.1
2. 長万部 Oshamambe	71.3	70.0	71.3	70.0	68.9	74.5	83.2	85.2	88.6	86.7	87.9	87.0	89.8	90.0	90.5	90.2	90.7	
3. 幌別 Horobetsu	75.4	73.8	74.6	73.6	73.2	74.2	85.1	88.2	90.3	87.9	90.6	88.5	87.9	92.0	90.6	93.1		
4. 平取 Hiratori	75.7	73.8	74.3	74.1	73.4	75.3	83.5	87.2	88.7	86.9	88.0	86.7	85.8	96.1	95.6			
5. 貫気別 Nukkibetsu	72.7	70.9	72.5	71.7	70.5	74.9	82.9	85.9	87.7	83.8	85.7	84.6	85.0	95.8				
6. 新冠 Niikappu	75.4	73.3	74.1	73.5	73.1	75.3	82.2	85.2	88.4	85.8	88.0	86.2	87.1					
7. 様似 Samani	71.2	70.3	72.1	70.8	69.7	72.5	84.4	85.2	89.4	89.2	92.1	90.7						
8. 帯広 Obihiro	70.1	68.6	69.6	70.2	69.4	70.5	84.2	88.7	90.0	93.1	94.7							
9. 訓路 Kushiro	72.1	70.6	71.9	71.7	71.7	72.0	86.8	88.7	91.4	94.0								
10. 美幌 Bihoro	70.5	70.1	71.1	70.6	71.1	71.2	86.6	88.4	89.2									
11. 旭川 Asahikawa	75.1	73.4	74.2	73.9	73.6	73.4	85.4	90.8										
12. 名寄 Nayoro	73.2	72.3	73.0	72.8	73.0	73.9	87.6											
13. 宗谷 Sôya	76.8	77.5	78.3	79.4	78.8	79.7												
14. 落帆 Ochiho	92.4	89.6	91.7	91.1	88.8													
15. 多蘭泊 Tarantomari	89.8	89.0	88.5	92.6														
16. 真岡 Maoka	92.3	90.3	91.6															
17. 白浦 Shiraura	93.3	92.1																
18. ライチシカ Raichishka	90.5																	

表 2 の値は類似度であるが、100%から表 2 の値を引いたものを非類似度(方言間の距離)と定義し、試みに、クラスター分析(最長距離法、ウォード法、最短距離法)にかけると、以下の図 4、図 5、図 6 のような結果が得られる。(本章のクラスター分析と対応分析のソフトウェアは R を用いた。)

図 4、図 5、図 6 の結果と、服部･知里(1960)の表 2 の解釈とを比較すると、以下の 3 点が挙げられる。

1. 「北海道方言とカラフト方言との間には大きな断層がある」という点と「北海道方言は北東と南西に大きく分かれる」という服部･知里(1960)の見解は図 4、図 5、図 6 でも確認された
2. しかし、「宗谷方言は、他の北海道方言から比較的遠く、且つカラフト方言に最も近い北海道方言である」という彼らの主張は、図 4、図 5 では、宗谷

方言が名寄と旭川の方言と比較的早い段階でクラスターをなしていることから、宗谷方言は特別に樺太方言と近いというわけではなく、支持されなかった。ただし、最短距離法を用いた図 6 では「宗谷方言は、他の北海道方言から比較的遠く、且つカラフト方言に最も近い北海道方言である」という主張は支持された。

3. さらに、図 4、図 5 では、樺太西海岸にある多蘭泊(15)と真岡(16)がまとまり、樺太東海岸の落帆(14)、白浦(17)と内路(19)と樺太西海岸の北にあるライチシカ(19)がまとまるのは、樺太が南北に山脈が走り、交通も海岸沿いを中心としていることから、地理的に疑問が残る。地理的には, 西海岸の 15,16,18 と東海岸の 14,17,19 がまとまると推察される。さらに、図 6 においても、樺太の東海岸の方言と、西海岸の方言は分かれていない。

以上のように、表 2 を非類似度に変換してクラスター分析した結果、彼らの言語学的知見と必ずしも一致せず、また、地理的条件としても理解し難いものとなるようである。

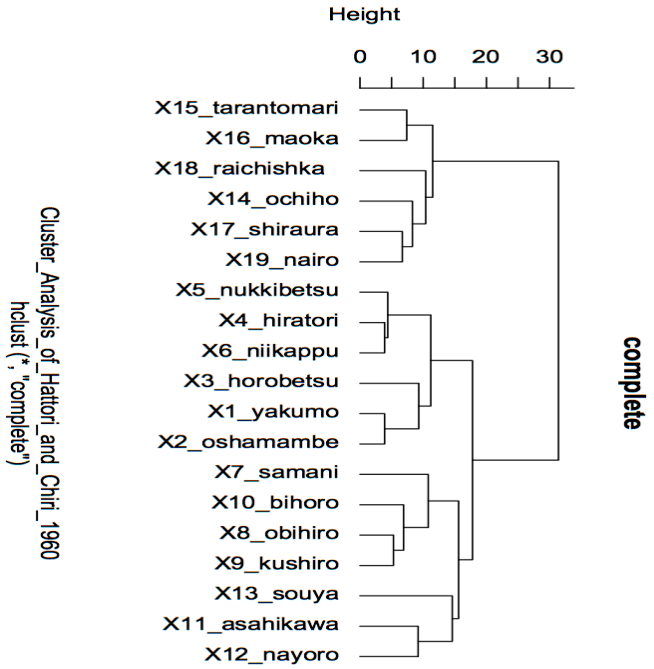


図 4: 表 2 を非類似度に変換し、最長距離法によって、クラスター分析を行った結果。

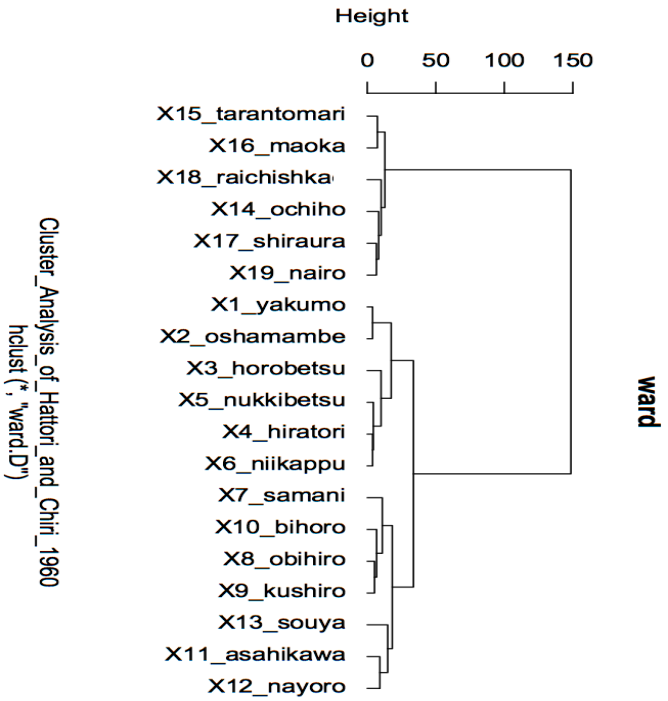


図 5: 表 2 を非類似度に変換し、ウォード法によって、クラスター分析を行った結果。

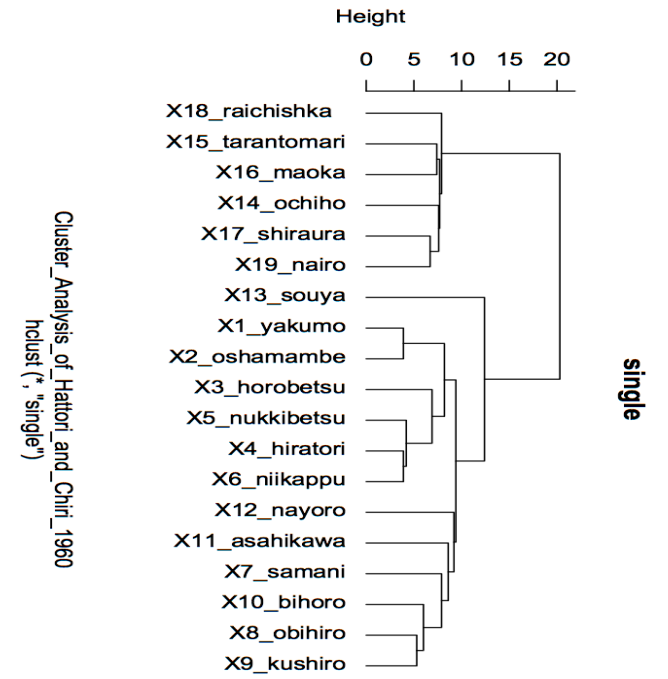


図 6: 表 2 を非類似度に変換し、最短距離法によって、クラスター分析を行った結果。

5.3.1.2 Asai(1974)

Asai(1974)の研究では、前述の服部･知里(1960)のデータに千島(Kuril)と千歳(Chitose)のデータを追加して分析している。ただし、Asai(1974)では、服部･知里(1960)とは方言の類似度の定義が少し異なることに注意する。すなわち、服部・知里(1960)では「±」の「＋」に関しては 0.5 の「＋」については 1.0 の重み付けがなされていたが、Asai(1974)では、どちらも 1.0 の重み付けがなされている。重みづけについては、それ以外の点は服部･知里(1960)と同じである。

Asai(1974)と同じ定義で得られた非類似度データに、クラスター分析（Asai[1974]と同じ最短距離法）を適用した結果が、以下の図 7 である。さらに、最長距離法とウォード法を適用した結果が図 8、図 9 である。Asai(1974)から転用した図 10 を並べてみると以下のようなことがいえる。ここではAsai(1974)の表記に従い、平取は hiratori ではなく piratori としている。

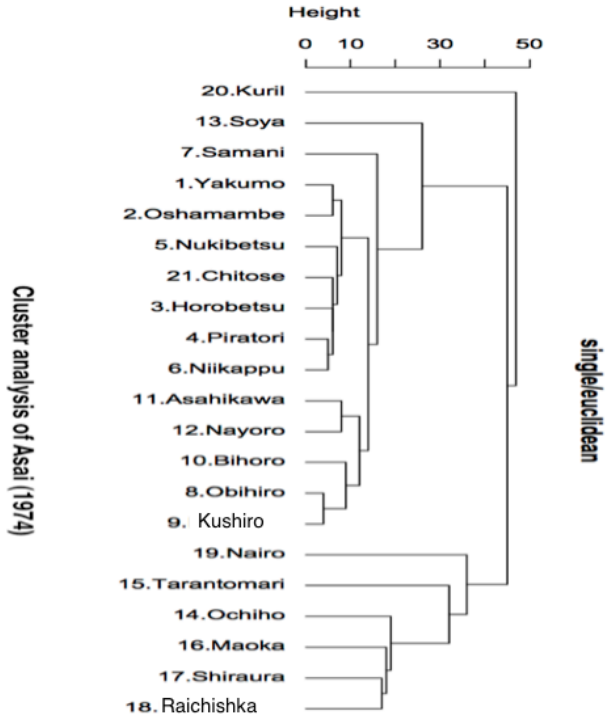


図 7: Asai(1974)の手法で得られた非類似度データにクラスター分析を適用した結果。クラスター分析には Asai(1974)と同じ最短距離法を使った。北海道方言が北東と南西にわかれている。宗谷方言が北海道方言の中ではもっとも樺太方言に近くなっている。

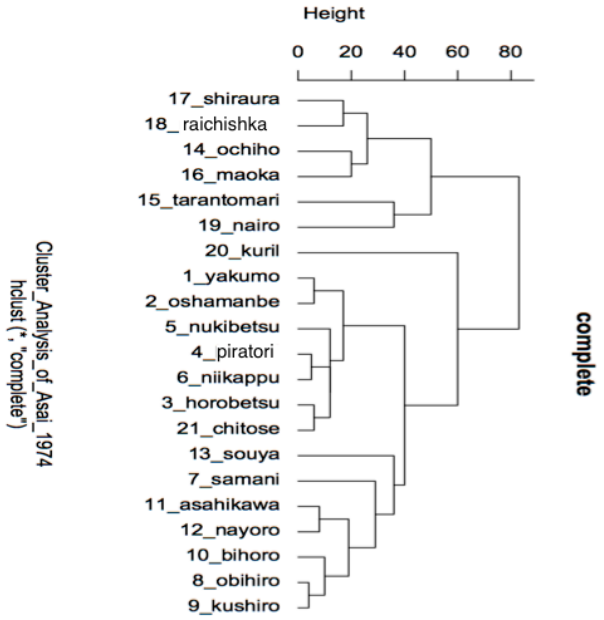


図 8: Asai(1974)の手法で得られた非類似度データにクラスター分析を適用した結果。クラスター分析には最長距離法を使った。北海道方言が北東と南西にわかれている。図 4 とは異なり、宗谷方言が北海道方言の中ではもっとも樺太方言に近くなってはいない。

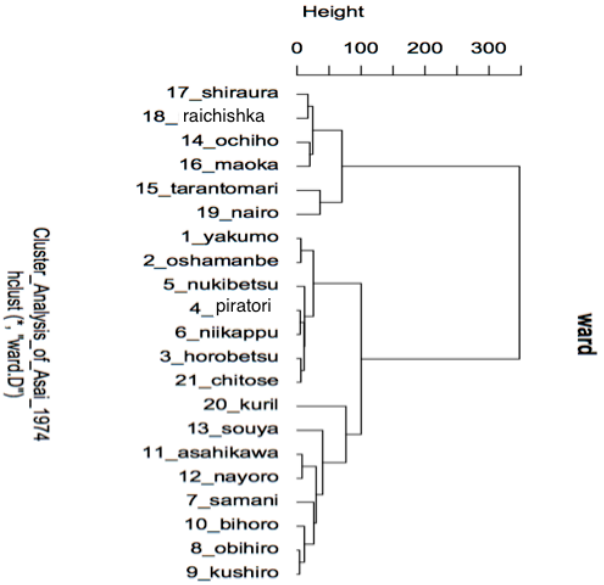


図 9: Asai(1974)の手法で得られた非類似度データをクラスター分析を適用した結果。クラスター分析にはウォード法を使った。北海道方言が北東と南西にわかれている。図 4 とは異なり、宗谷方言が北海道方言の中ではもっとも樺太方言に近くなってはいない。

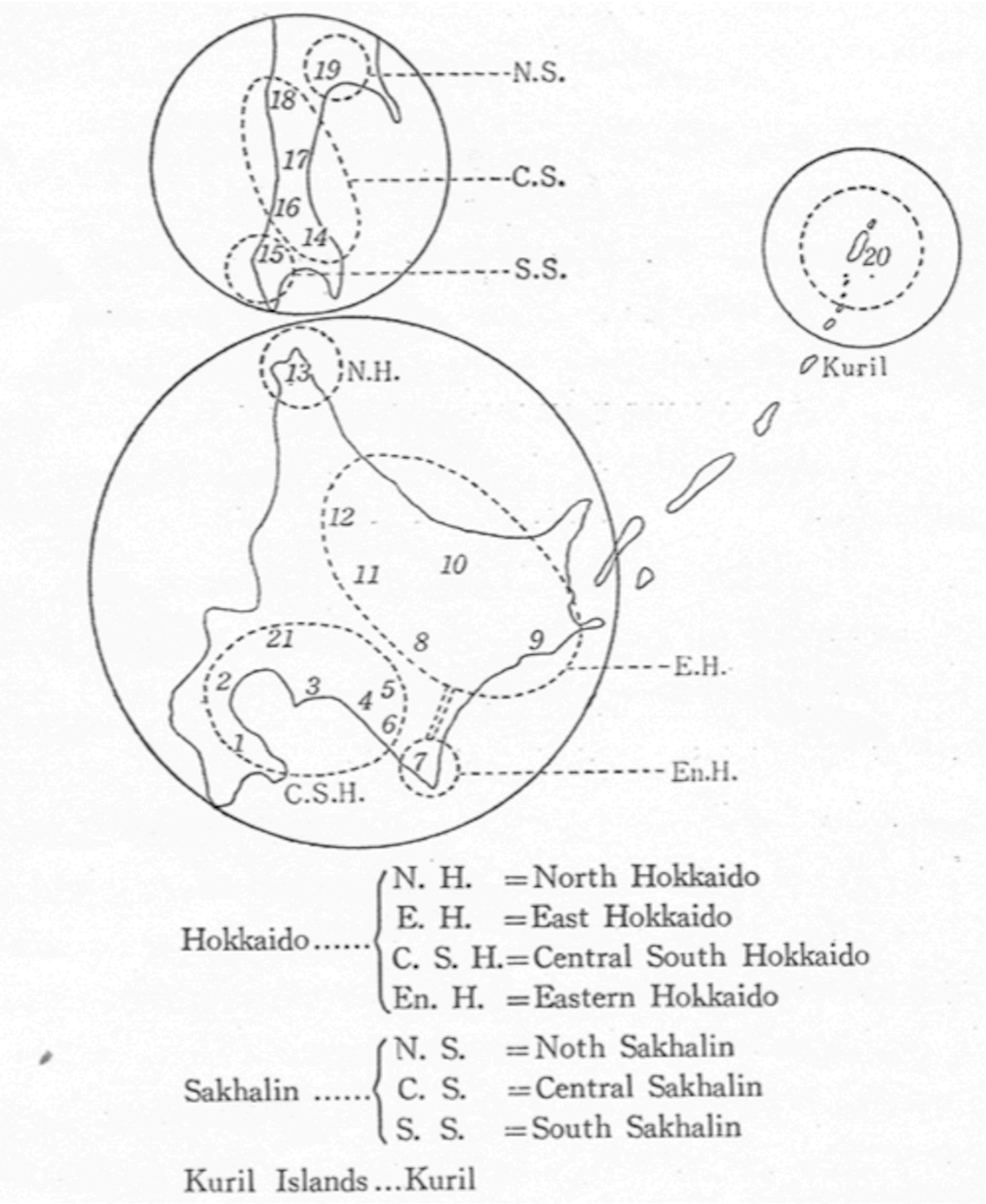


図 10: Asai(1974)のクラスター分析の結果を地図上に示したもの。Asai(1974: 100)

図 7、図 8、図 9 に関して以下のようなことがいえる。

1. 服部·知里(1960)の結果である図 4 と同じように、北海道方言と樺太方言との間には大きな隔たりがある。

2. 服部·知里(1960)の結果と一致している点として、北海道方言は大きく北東方言と南西方言に分かれる。
3. 図 4 からは確認されなかったが、服部·知里(1960)の考察にある「宗谷方言は、他の北海道方言から比較的遠く、且つカラフト方言に最も近い北海道方言である」ということと図 7 は整合している。ただ、図 8、図 9 はその結果を示していない
4. Asai(1974)はこのクラスター分析の結果、多蘭泊(15)を南サハリン方言、落帆(14)、真岡(16)、白浦(17)、ライチシカ(18)を中央サハリン方言、内路(19)を北サハリン方言としている。しかし、これはサハリンには南北に山脈があることなどから地理学的には疑問が残る。

以上のように、Asai(1974)のクラスター分析の結果は、一方で前節の服部·知里(1960)のデータにそのままクラスター分析を適用した結果と整合しているようだが、他方で、やはり地理的条件を考えると一部に疑念の残る分類が残るようである。

5.3.1.3 Lee & Hasegawa (2013)

服部·知里(1960)のデータに基づき、前節の Asai(1974)よりもさらに進んで、ベイズ法による系統樹推定法(Lemmon & Lemmon, 2008 ;Drummond, Suchard, Xie & Raumbaut, 2012)を行ったのが Lee and Hasegawa(2013)である[2]。その結果は、図 11 であるが、Asai (1974)と同様に、北海道方言に関する分類は地理的分布の視点から概ね妥当に見える。ただし、サハリン方言の分類は図 4 の結果と同様に、樺太西海岸にある多蘭泊(15)と真岡(16)が同一のクラスターとしてまとまり、樺太東海岸の落帆(14)、白浦(17)と内路(19)と樺太西海岸の北にあるライチシカ(18)がまとまるのは、樺太が南北に山脈が走り、交通も海岸沿いを中心としている地理的条件からは疑問が残る。彼らの分析結果は、距離的に離れた方言同士の関係が密であるとしている点は興味深いものの、それを説明する言語外要因については何も触れていない。

[2] 系統樹推定法とクラスター分析法は異なった考えに基づいた別の統計手法であり、相互の関係はない。

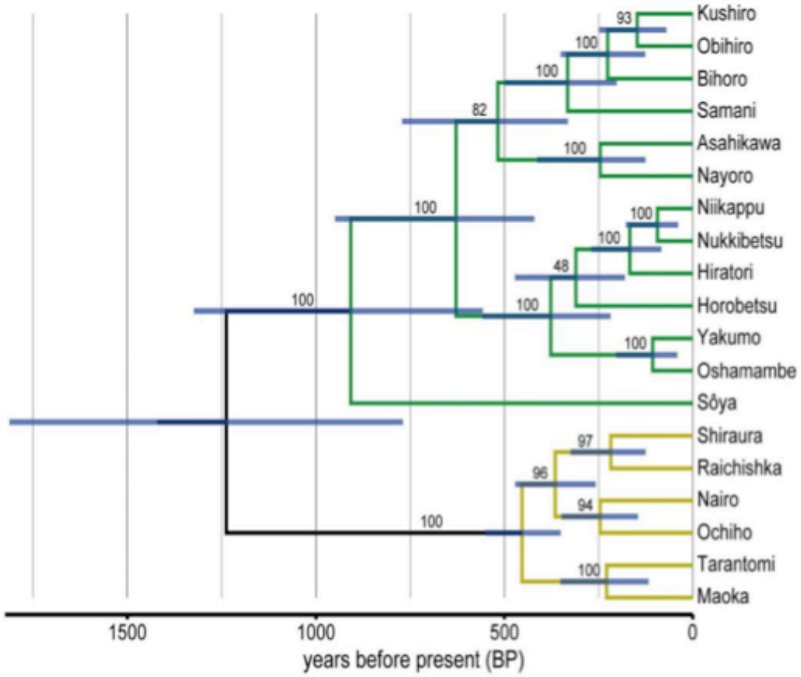


図 11: Lee & Hasegawa(2013)のベイズ法による系統樹推定の結果。Lee & Hasegawa(2013, pp.3)

5.3.2 先行研究の問題点

服部·知里(1960)、Asai (1974)、Lee and Hasegawa (2013)らの分析結果は、サハリン方言に関して、地理的条件の視点から、まだ検討の余地がありそうである。そこで筆者は一つの可能性として、それらの先行研究において、服部·知里(1960)の調査データで(何らかの理由で)「同根性が判断できない」というカテゴリー「〇，?，·」(表 1 参照)を情報として活用していないことに注目した。

つまり、調査時点で、A 方言の語彙 c と B 方言の語彙 c'について「同根性が判断できない」ことはそれだけ A 方言と B 方言とが「遠い」(もしくは「近い」)という情報を表していると考えられないかと、推察してみた。「〇，？，·」を一つのカテゴリーと捉え分析に加えることで、見落とされていた情報が復元される可能性を探る価値がありそうである。

5.4 本研究の目的

服部·知里(1960)における各地方の語彙の同根性の判定データの＋, ±, –だけではなく、先行研究では除外された「同根性が分からなかった」([〇，？，·])

も欠測値ではなく一つの情報とし、4 カテゴリーとして扱い、アイヌ語の方言の分類を行うことを試みる。

統計学的な立場からは、＋，±，−，［○，？，·］を同時に扱う際には「尺度水準」（吉野·千野·山岸，2007，第 2 章）の問題を念頭におくことが重要である。もし＋，±，−だけを扱えばよいのなら、＋>±>− という順序関係は認めることができるだろう（順序尺度）し、さらには、服部·知里（1960）のように＋に 1.0 の重み付け、±の＋に 0.5 の重み付けが間隔尺度として本当に妥当であれば、加減乗除まで仮定することも可能かもしれない。しかし、＋，±，−，［○，？，·］を同時に扱う場合、＋>±>−>［○，?，·］という順序が成立するかも不確かであり、またそもそも服部・知里(1960)や Asai(1974) のような重みづけで＋、±、−が間隔尺度をなすとアプリオリに考える根拠は薄い。このような時には、それぞれのカテゴリーを「名義尺度」として扱うべきであろう。このようなカテゴルカルデータ（名義尺度のデータ）を分析する手段の一つとして、林の数量化Ⅲ類(Hayashi, 1952)（QMⅢ）や多重対応分析（MCA）がある。

5.5 分析方法

5.5.1 データの扱いと距離行列の算出

本研究では、この MCA を服部・知里（1960）の調査データに適用し、その結果得られる各地域の方言のデータの多次元空間の座標をさらにクラスター分析することを考える。

服部・知里(1960)の元のデータは、地域を一つの相、語彙を第二の相とし、19 地域×19 地域×200 語彙の形をもつ 2 相 3 元データである。データの分析に際して、各地域$i(i = 1,2,\ldots,19)$について、他の地域$j(j = 1,2,\ldots,19)$との間で、語彙$k(k = 1,2,\ldots,200)$の同根性判断データから、地域(19)×語彙(200)の 2 相 2 元のアイテムカテゴリー型データを作成した(この 2 相 2 元データ行列は 19 個存在することになる)。それらの地域ごとのデータ行列に個別に MCA を適用し、その各空間座標に基づいて、各地域の対jとj'の距離$d_{jj'(i)}$を求めた。その際、$d_{jj'(i)}$を求める空間座標の次元については、足立·村上（2010）に従い、固有値のプロットの減少が急に緩やかになる前までの次元を採用した。例として 19 地点の一つの旭川のデータの MCA における固有値のプロットを図 12 に示す。この場合、3 次元目までのデータを採用した(累積寄与率は 42.71%)。19 地点それぞれについての MCA の結果も、固有値のプロットから 3 次元目まで採用することが適

切と判定された。19 地点それぞれについて、3 次元の座標から、ユークリッド距離に基づいた距離行列$\{d_{jj'(i)}\}$を得た。またマンハッタン距離に基づいた距離行列$\{d_{jj'(i)}\}$も扱うこととした。

最終的には、19 の距離行列を平均した距離行列$\{a_{jj'}\}$[3]（ユークリッド距離またはマンハッタン距離に基づく）を得た。すなわち、

$$a_{jj'} = \frac{1}{19}\sum_{i=1}^{19} d_{jj'(i)}$$

である。

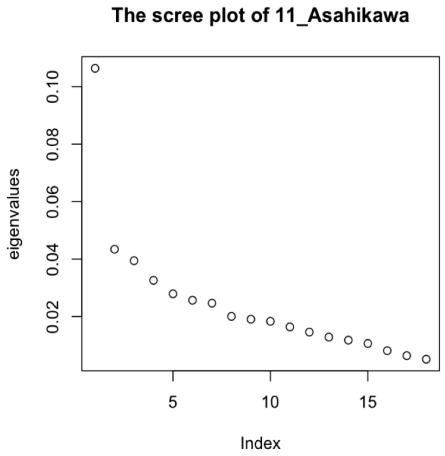


図 12: 19 地点の一つの旭川のデータについての MCA の固有値のプロット。3 次元目までのデータを採用することにした。

5.5.2 数量化Ⅲ類クラスタリングの適用

本節では、まず距離行列$\{a_{jj'}\}$に対してデータの「最も強い」構造を取り出すクラスター分析を適用する。ここでは、もっとも遠いものから分岐していくという最長距離法の考え方が方言の分岐のあり方と近いと考えられるため、最長距離法を用いた。比較のため、最短距離法、ウォード法の結果も示す。

5.5.3 MCA_Neighbor-Net の適用

次に、5.5.1 で計算した距離行列に、Neighbor-Net(NNA と略す)を適用する

[3]距離行列$\{a_{jj'}\}$は、ある地域iから見たときの地域とjとj'の非類似度と別の地点i'から見た時のjとj'の非類似度、さらに別の地域i''から見た時のjとj'の非類似度などをすべての地点に関して平均したときに得られた非類似度と解釈することができる。

ことを考える[4]。今まで、服部･知里(1960)のデータに基づいた研究はいずれも、データの「もっとも強い構造」が取り出されるクラスター分析を用いていた。しかし、NNA を適用することによって、クラスター分析によって隠されていた「2 番目以降に強い」構造が明らかになると期待できる。本稿では、服部･知里(1960)のようなカテゴリカルデータ(名義尺度)に対して、MCA により距離データを得て、それに NNA を適用する方法を MCA_Neighbor-Net と称することとする。対応分析には、R の corresp を、NNA には SplitsTree(Huson & Bryant, 2006)を用いた。2 章で概説したように、NNA は本来、生物進化の系統樹の解析に用いられて開発されてきた（Bryant & Moulton, 2003)。生物の系統進化を解明するには、単純な樹状構造を扱うのでは不十分で、多数の樹状構造を同時に表現するネットワークの方が適切である。特に、データ解析の初期段階では当該の生物進化に関する知見が乏しい場合、推論のための方法ではなく、データの表現の方法として NNA が用いられる。単純な樹状構造などでは、誤ったモデルでの推論となる危惧を避けるためである。本章で取り扱うアイヌ言語の伝搬変容や地理的に隣接する各方言間の解明なども、その生物進化の解明と、方法論としては類似の側面があり、NNA の活用は有益と思え、試行する価値はあろう。

5.6 結果

5.6.1 数量化Ⅲ類クラスタリングによる結果

前節に示した手続きにおいて、ユークリッド距離に基づく距離行列についてのクラスター分析の結果、図 13(最長距離法)、図 14(最短距離法)、図 15(ウォード法)を得た。また、マンハッタン距離に基づく距離行列についてのクラスター分析の結果、図 16(最長距離法)、図 17(最短距離法)、図 18(ウォード法)を得た。

[4] ここでの距離行列は、2 相 2 元データに対して MCA を適用し求めた距離行列とは異なりやや特殊であるが、MCA_Neighbor-Net の一種として扱う。

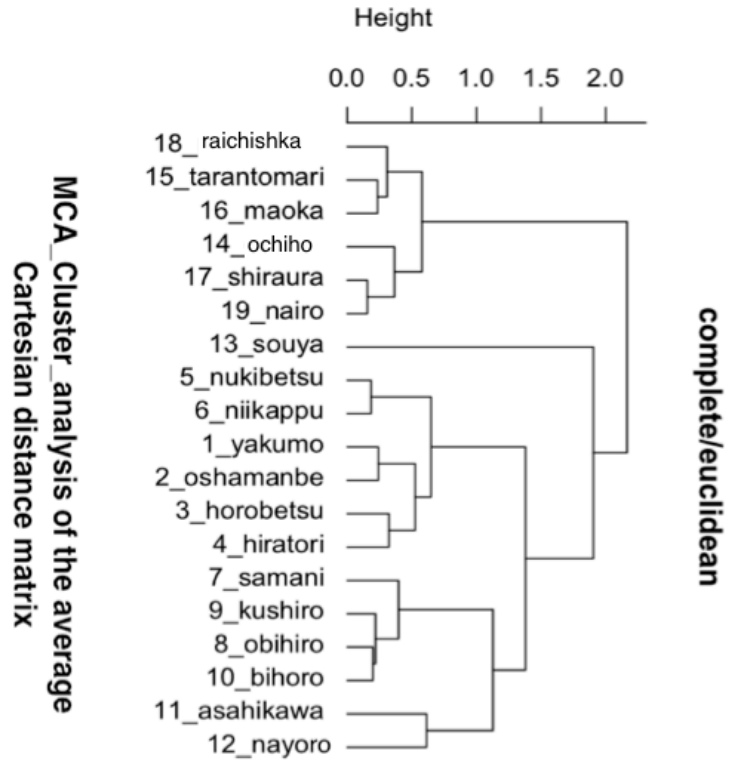


図 13: 19 の MCA3 次元布置からユークリッド距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(最長距離法)を適用した結果。

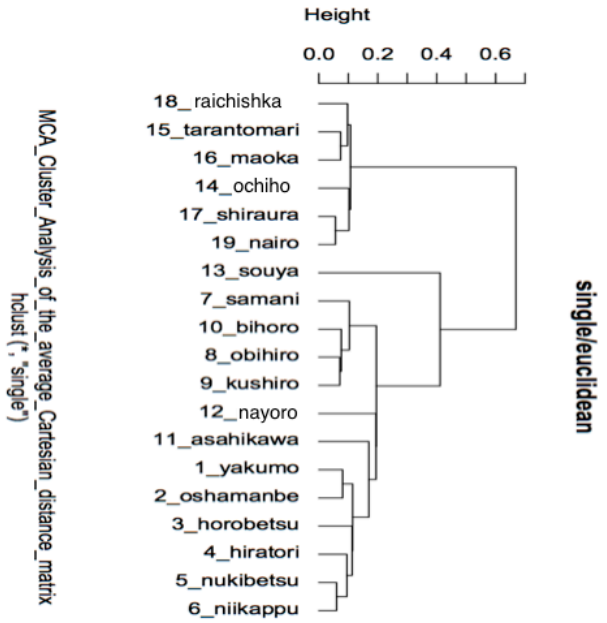


図 14: 19 の MCA3 次元布置からユークリッド距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(最長距離法)を適用した結果。

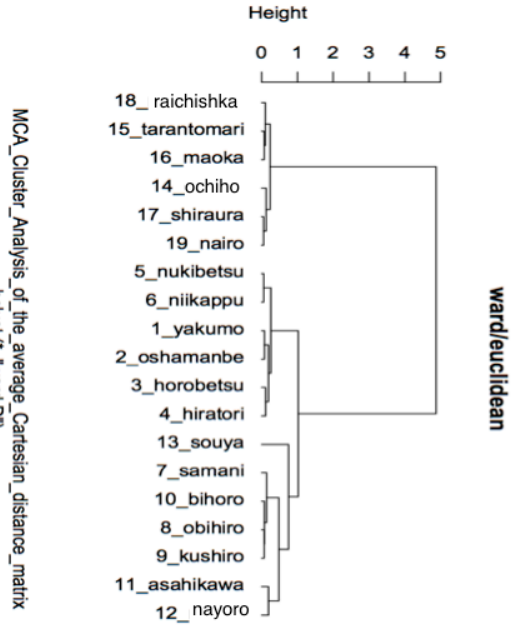


図 15: 19 の MCA3 次元布置からユークリッド距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(ウォード法)を適用した結果。

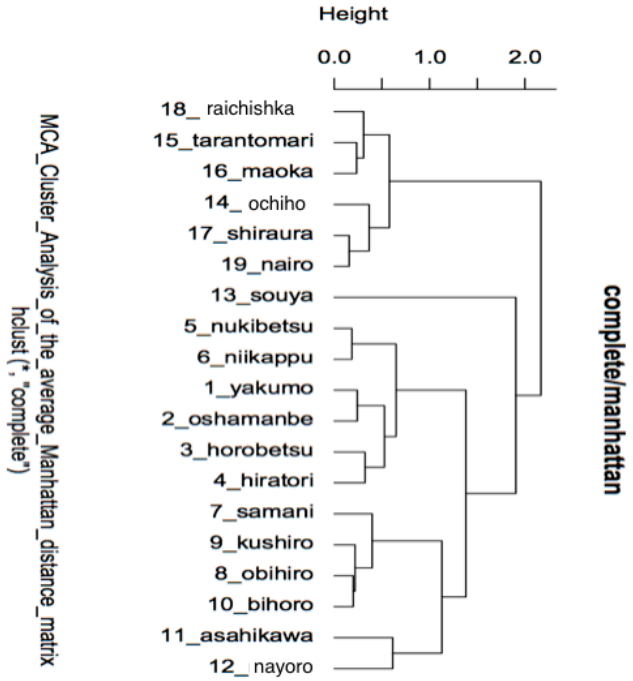


図 16: 19 の MCA3 次元布置からマンハッタン距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(最長距離法)を適用した結果。

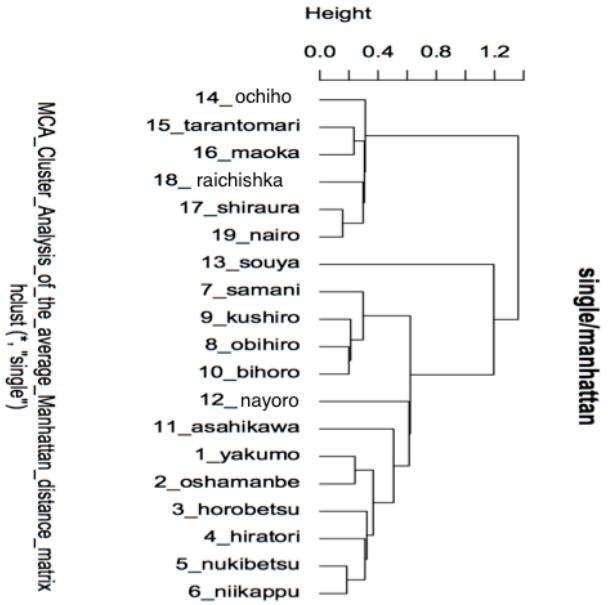


図 17: 19 の MCA3 次元布置からマンハッタン距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(最短距離法)を適用した結果。

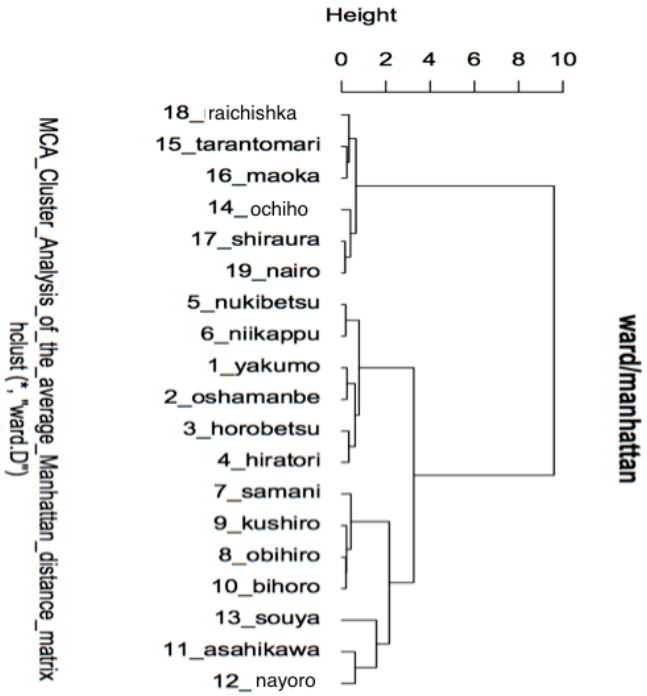


図 18: 19 の MCA3 次元布置からマンハッタン距離に基づいて得られた 19 の距離行列を平均した距離行列にクラスター分析(ウォード法)を適用した結果

図 13、図 14、図 16、図 17 においては北海道方言に関する結果は Asai (1974) と整合して、北海道方言の中では宗谷方言が最も樺太方言に近く、北海道方言は北東方言と南西方言に分かれるという傾向が確認できる。ただし、ウォード法を使った図 15、図 18 では確認できなかった。他方で、Asai (1974) や Lee and Hasegawa (2013)の分析結果では地理的条件とは整合せず疑念の残ったサハリン方言については、図 13、図 14、図 15、図 16、図 17 では西海岸の方言(多蘭泊(15)、真岡(16)、ライチシカ(18)) と東海岸の方言(落帆(14)、白浦(17)、内路(19)) とが分かれ、これは樺太には南北に山脈が走り、交通は海岸沿いを基本としていたという地理学的条件とも整合した結果が示されている。(サハリン方言が、なぜ西海岸と東海岸で分かれることが妥当かということに関しては、文化人類学的、文献学的観点から詳細な議論もなされている[Ono, 2016])。

以上のように、服部・知里 (1960) の調査データについて、尺度水準や「語彙の同根性が不明」というデータを含めた分析の方が、先行する Asai(1974)や Lee and Hasegawa (2013)の結果よりも、既存の言語学的考察や地理的条件と整合性のある結果が得られた。

また、既存の研究と同様の「方言間の単語の同根性がわからない」という情報を用いないデータについて、MCA を全く同じ手順で適用した結果(19 の地点のすべての MCA の結果で採用した次元は 3 次元であった)も見ておくこととする[5]。これを 5 章附録図 1~図 6 とした。この結果は、

1. 5 章附録図 1 から 5 章附録図 6 においては、西海岸の方言(多蘭泊(15)、真岡(16)、ライチシカ(18)) と東海岸の方言(落帆(14)、白浦(17)、内路(19)) とが分かれ、これは樺太には南北に山脈が走り、交通は海岸沿いを基本としていたという地理学的条件とも整合した結果が示されている。この樺太の方言が西海岸と東海岸に分かれるという結果は、図 13、図 14、図 15、図 16、図 18 の結果と一致する。
2. 図 13~図 18 では北東のグループに分類される旭川(11)がいずれの図においても南西のグループに分類されている。

[5] 既存の研究と同様に「方言間の単語の同根性がわからない」という情報を用いないデータについて、MCA を適用した分析を行う理由は、図 13 から図 18 の結果(特に、樺太の方言が東海岸と西海岸に分かれる結果)が、MCA を分析に用いたことによるものなのか、「方言間の単語の同根性がわからない」という情報を用いたことによるものかを明らかにするためである。

3. 最長距離法と、最短距離法を用いた 5 章付録図 1〜図 2、5 章付録図 4~図 5 においては、北海道方言で最も離れた方言が、服部･知里(1960)や Asai(1974)や Lee and Hasegawa(2013)の結果とは異なり、名寄(12)となっている。図 13、図 14、図 16、図 17 の結果や、服部･知里(1960)や Asai(1974)や Lee and Hasegawa(2013)では宗谷方言が樺太方言に最も近い方言と指摘されていた。

5.6.2 MCA_Neighbor-Net による結果

前節では、先行研究に対して、尺度水準や欠測値の扱いの観点などからより妥当な分析の方法が示唆された。本稿では、前節でクラスター分析によって言語地理学的に妥当な分類をあたえたアイヌ語方言の距離行列に、さらに NNA を適用することで、アイヌ語における方言周圏論的構造について統計科学的な考察を進めてみる

5.5.1 において計算した距離行列に、NNA を適用した結果、図 19 を得た。これは、19 点のデータに関する MCA の結果、固有値のプロットから空間座標を全て 3 次元目まで考慮することとし、その空間座標からユークリッド距離に基づき、距離行列$\{d_{jj'(i)}\}$を得て、さらにそれら 19 の距離行列を平均した距離行列$\{a_{jj'}\}$に NNA を適用したものである。また、参考のために、19 点のデータ全てで MCA の 18 次元目（最大次元）まで考慮し、累積寄与率 100%とし、18 次元の空間座標からユークリッド距離に基づき、距離行列$\{d_{jj'(i)}\}$を得て、さらにそれら 19 の距離行列を平均した距離行列$\{a_{jj'}\}$に NNA を適用したものを図 20 に示す。

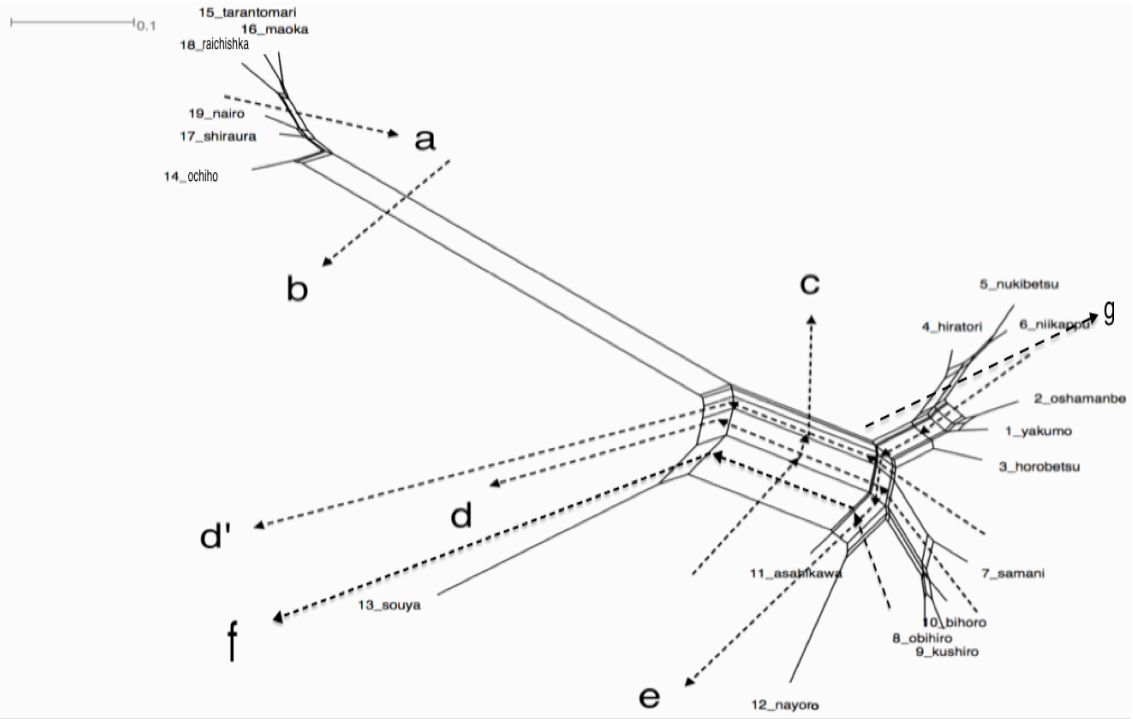


図 19: MCA の結果得られた 3 次元目までの座標に基づいた 19 の距離行列(ユークリッド距離)を平均した距離行列に Neighbor-Net を適用した結果

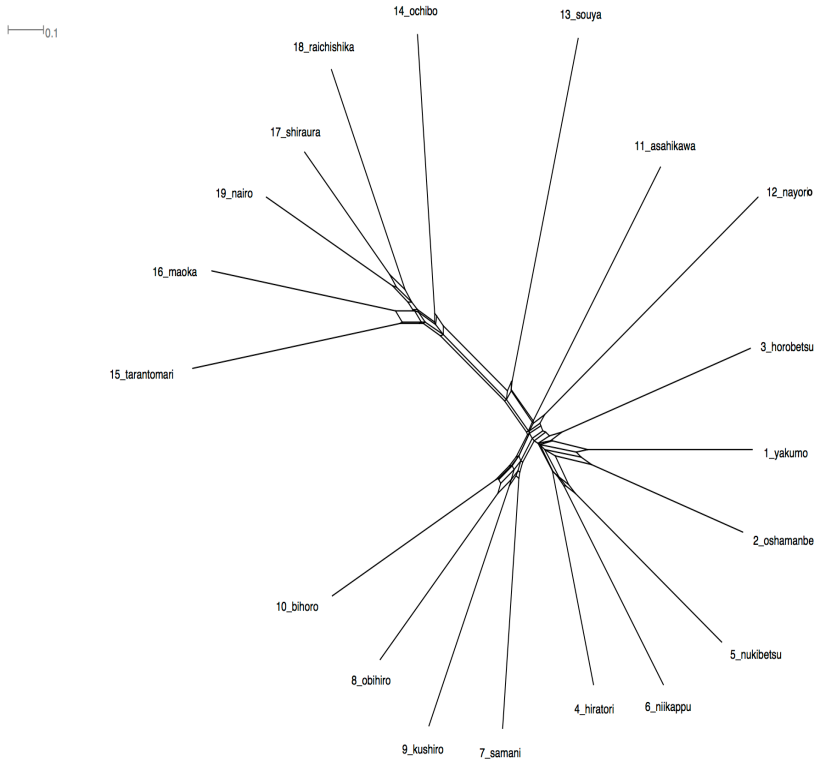


図 20: MCA の結果得られた全次元(18 次元)までの座標に基づいた 19 の距離行列(ユークリッド距離)を平均した距離行列に Neighbor-Net を適用した結果

図 19 と図 20 を比べると分かるように、高次元まで考慮すると、Neighbor-Net の結果における各方言の枝が長くなっている。図 19 では、a, b, c, d, d',e,f,g の線で分割し解釈することが可能であろう。他方、図 20 のような図で「枝が長くなっている」ことは、図 19 に対して、図 20 の方が、距離行列がよりノイズを取り込んでいる可能性を示唆する。さらに、図 20 では、「どこで図を切るべきか」（クラスター分類）という議論がしにくい。そこで、本稿では、一応、3 次元目までの結果に基づいた図 19 の解釈を行うこととし、実際に言語学的にも図 19 の方が図 20 よりも有意味な解釈を得ることができた。このことは、MCA において「累積寄与率」や「説明率」がただ高ければよいというわけではないことを示唆している。

5.6.2.1 —a, b による解釈

まず、a, b の線で、図 19 を分割することにより、まず第 1 に北海道方言と樺太方言の断絶が非常に大きいという従来の知見が確認される。また、樺太のアイヌ語の方言が樺太西岸方言{15, 16, 18}と樺太東岸方言{14, 17, 19}とに分かれていることは、Ono (2016)の結果を再確認させる。さらに、{14, 17, 19}がより北海道方言に近いことから、樺太のアイヌ語においては西岸方言より東岸方言の方が、北海道方言に近いことが示唆されている。この点については、語彙だけではなく文法など総合的な面から今後研究がなされることが必要であろう。

5.6.2.2 —c による解釈

次に、c の線で、図 19 を分割することにより、宗谷方言は北海道方言の中でも最も樺太方言に近いことが示される。これは、服部·知里(1960)でも示唆されたことであり、その後の Asai(1974)などの研究も支持する結果である。

5.6.2.3 —b と d, d'による解釈

さらに、b と d の線で、図 19 を分割することにより、北海道方言は北東のグループと南西のグループに分かれる。この知見は、服部·知里(1960)の分類や、Asai(1974)のクラスター分析の結果と合致する。先行研究を検討した第 3 節の図 4、図 5、図 6、図 7、図 8、図 9 とも概ね一致しているため、北海道方言を北東のグループと南西のグループにわけることは服部·知里(1960)のデータの「強い」構造の一つと考えられる。ただし、b と d の線で分割することは、様似方言(7)を南西のグループに分類することになる。Asai (1974)などにおいて、様似方言(7)を北東のグループに分けていることは、b と d'の線で分割することに

相当する。様似方言をどのように考えるかについては、今後の研究の余地がある。

5.6.2.4 —e による解釈

また、前述の e の線だけで分割することは、中川(1996)で指摘されている「沙流･千歳･樺太型」に対応することも注目される。ただし、旭川(11)及び宗谷(13)が「沙流･千歳･樺太型」に属するかは、今後の議論を待ちたい。

5.6.2.5 —b,f による解釈

図 19 は b と f の線で分割することもできる。このことによって、図 19 は樺太(北蝦夷)のグループと、旭川(11)、名寄(12)、宗谷(13)の西蝦夷のグループと、それ以外(東蝦夷)のグループに分かれる。これらの西蝦夷型、東蝦夷型の地理的パターンは中川(1996)においても指摘されており、本章の分析においては、服部・知里(1960)の「方言間の単語の同根性がわからない」という情報を用いた MCA_Neighbor-Net による分析によってのみ観察されたものである。西蝦夷型・東蝦夷型のパターンが観察されたことは、本章の「方言間の単語の同根性がわからない」という情報を用いた MCA_Neighbor-Net による分析が有用であることを示すと考えられる。

5.6.2.6 —g による解釈

さらに、図 19 は g の線で分割することができる。これを地理的に見ると平取(4)、貫別(5)、新冠(6)が中心を、それ以外の方言が外輪を構成していることがわかる。これは、柳田国男が「蝸牛考」で述べた「方言周圏論」的考えと合致する(柳田, 1930)。アイヌ語が沙流、平取を中心とした方言周圏論的構造をとることは、中川(1996)においても指摘されている。

アイヌ語における方言周圏論的構造は、今まで文献学的な言及は多くあったが、統計科学的に示されたことはなかった。服部･知里(1960)や Asai(1974)、Lee and Hasegawa(2013)などにおいても、示されたのは「強い」構造の一つである、アイヌ語の方言を北東と南西に分けるものであった。アイヌ語の方言周圏論的構造が、服部･知里(1960)のデータではそれ以外の「2 番目以降に強い」構造であったことがその原因と考えられる。

この結果は、本章の MCA_Neighbor-Net による統計分析により初めて明確になった。これは、直感的にはこれまでにも少なからずの人々に気づかれていたと思われる、沙流方言とそれと関係の深い方言群がアイヌ語諸方言で持つ位置づけの大きさを、本稿では統計的に示したことになる。

なお、5 章付録に、「方言間の単語の同根性がわからない」という情報を用いない以外は、5.6.2 節と同様の手順で MCA_Neighbor-Net を適用した結果を示す(5 章付録図 7[ユークリッド距離]、図 8[マンハッタン距離])。図 19 と同様に、5 章付録図 7 と図 8 は、1)樺太の方言が東海岸と西海岸にわかれ、2)宗谷方言が北海道方言の中では樺太方言には最も近く、3)平取、貫別、新冠を中心とした方言周圏論的構造が確認された。ただし、図 19 とは異なり、中川(1996)が指摘している沙流・千歳・樺太型及び西蝦夷型・東蝦夷型は確認されなかった。

5.7 考察

本研究では、服部･知里(1960)の語彙の同根性判別データにおける「わからない」というカテゴリーを、先行研究のように無視するのではなく、1 つの情報をもった独立したカテゴリーとして扱うことによって、樺太のアイヌ語の方言が、言語地理学的に妥当な分類となることを示した。

さらに、その際に用いられた距離行列を「もっとも強い」構造のみを導く「クラスター分析」ではなく、「2 番目以降に強い」構造も把握することができる Neighbor-Net を適用することによって、今まで文献上は言及されてきたアイヌ語における方言周圏論的構造を示すことができた。

本章で扱った研究の結果の違いを表 3 に整理した。

表 3: 本章で扱った研究の結果の違い。○は「完全に整合」、△は「部分的に整合」、×は「整合しない」ことを意味する。

	服部·知里(1960)のクラスター分析の結果	Asai(1974)のクラスター分析の結果	Lee & Hasegawa (2013)の系統樹	本章(小野, 2015b)の数量化Ⅲ類クラスタリングの結果	本章(小野, 2015b)のMCA_Neighbor-Netの結果	5章付録の、「わからない」という情報を用いず数量化Ⅲ類クラスタリングを適用した結果	5章付録の、「わからない」という情報を用いずMCA_Neighbor-Netを適用した結果
北海道方言と樺太方言の間に大きな隔たりがあるか	○	○	○	○	○	○	○
宗谷方言が北海道方言の中で最も樺太方言に近いか(最も離れた方言か)	△	△	○	△	○	×	○
北海道方言が北東と南西に分かれるか	○	○	○	○	○	○	○
樺太方言が東海岸と西海岸で分かれるか	×	×	×	○	○	○	○
沙流·平取を中心としたアイヌ語の方言周圏論的構造が見られるか	×	×	×	×	○	×	○
沙流·千歳·樺太型(中川, 1996)が確認されるか	×	×	×	×	○	×	×
西蝦夷型·東蝦夷型(中川, 1996)が確認されるか	×	×	×	×	○	×	×

表 3 から、服部·知里(1960)、Asai(1974)、Lee and Hasegawa (2013)とそれ以外の結果を比較すると樺太方言が東海岸と西海岸で分かれているという点でそれ以外の結果が優れている。理由としては、それ以外の結果が数量化Ⅲ類を適用したことが推察される。また、それらの中でも、沙流·平取を中心としたアイヌ語の方言周圏論的構造が観察されるという点で数量化Ⅲ類クラスタリングの結果よりも、MCA_Neighbor-Net の結果は優れている。また、同じ MCA_Neighbor-Net の結果でも、「方言間の単語の同根性がわからない」という情報を考慮した結果の方が、中川(1996)が指摘した沙流·千歳·樺太型及び西蝦夷型・東蝦夷型が確認できるという点で優れている。

5.8 今後の課題

ここでは、本章の実質科学的な課題と方法論的な課題について簡潔に述べる。

まず、実質科学的な課題としては、本章の MCA_Neighbor-Net の結果をアイヌ語学の知見に基づいて、より深く考察することが挙げられる。詳細な内容は本論文の内容を超えるので割愛したが、アイヌ語方言の専門家との共同研究と

して機会を改めて報告したい。

次に、方法論的な課題としては以下のの 2 点が挙げられる。

第一に、本研究においては 2 相 3 元の質的データにおける欠測値の考察が十分に尽くされていない。方法としては、独立なカテゴリーとして扱い MCA を適用したが、個々の分析でカテゴリーに与えられた重みについて、吟味することが必要であろう。ただし、2 相 3 元データを複数の 2 相 2 元データに分割したものに MCA を適用した結果であるので、通常の MCA の分析とは異なる側面もあろう。

第二に、クラスター分析や MCA_Neighbor-Net を中心に、服部・知里(1960)のデータの検証を行ってきたが、そもそも、(非)類似度データに適用される統計手法にはクラスター分析以外にも、多次元尺度構成法(MDS)がある。MDS を適用することによって、得られたアイヌ語方言の多次元の布置を考察することによって、本論文で得られた結果がどの程度頑健な知見であるかを検証することが今後の課題の一つと言えよう。

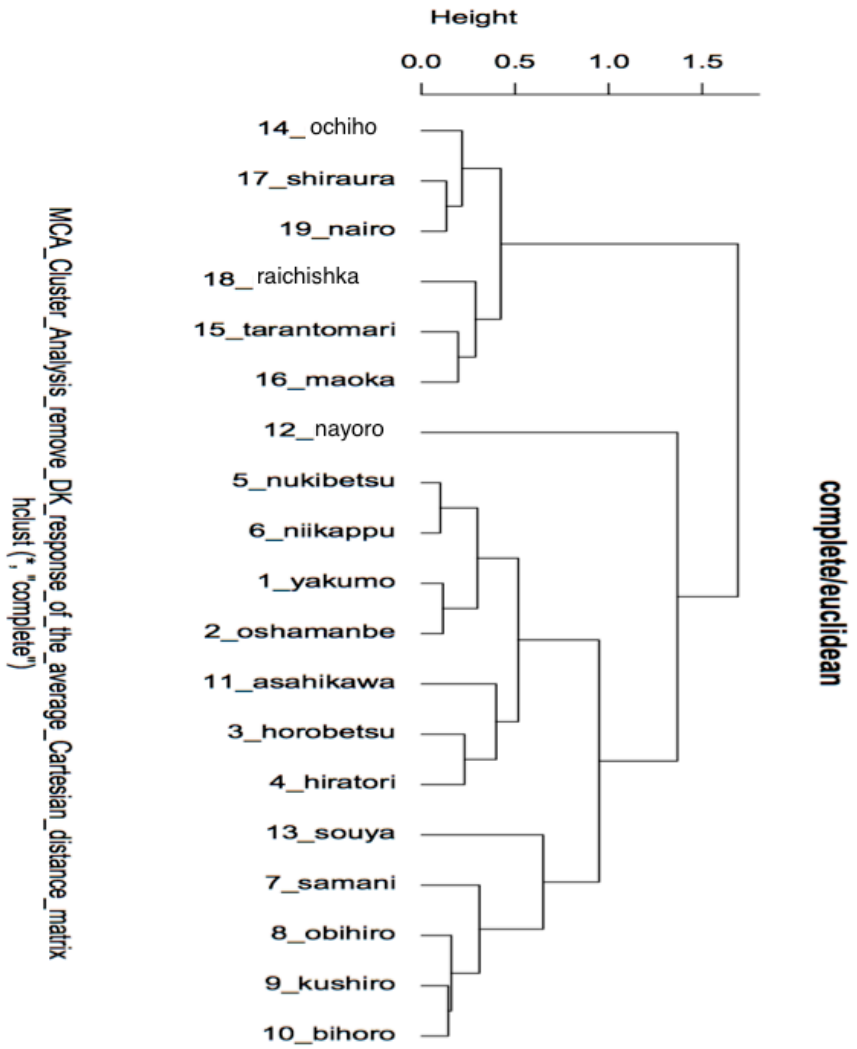
最後に、本章の主題を離れるが、本章において MCA_Neighbor-Net の持つ柔軟な表現力が確認されたので、今後は、日本語などその他の言語のデータにおいても MCA_Neighbor-Net の適用を試行する価値はありそうである。

統計手法は世の中に数多存在しているが、そのデータの性質をよく観察した上で、まず「どのような尺度水準でデータを扱うべきか」ということを十分に考慮することである。その上で、本稿で提案した MCA_Neighbor-Net のように、より柔軟な方法を用いることによって、今までの統計解析では不十分であったことが、明らかになることがあり得る。

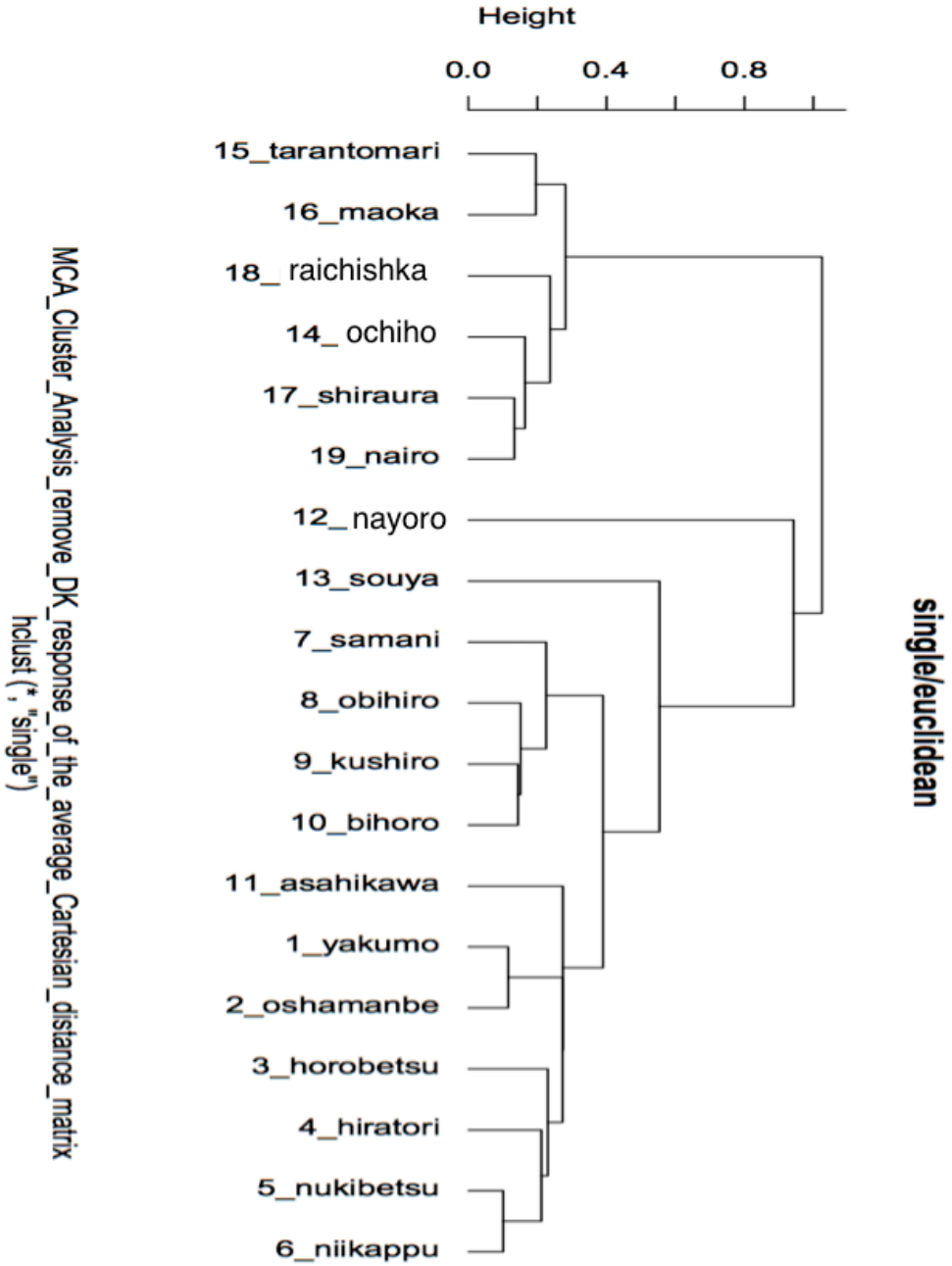
5 章付録

5 章付録 1.「同根性がわからない」情報を用いない数量化Ⅲ類クラスタリング

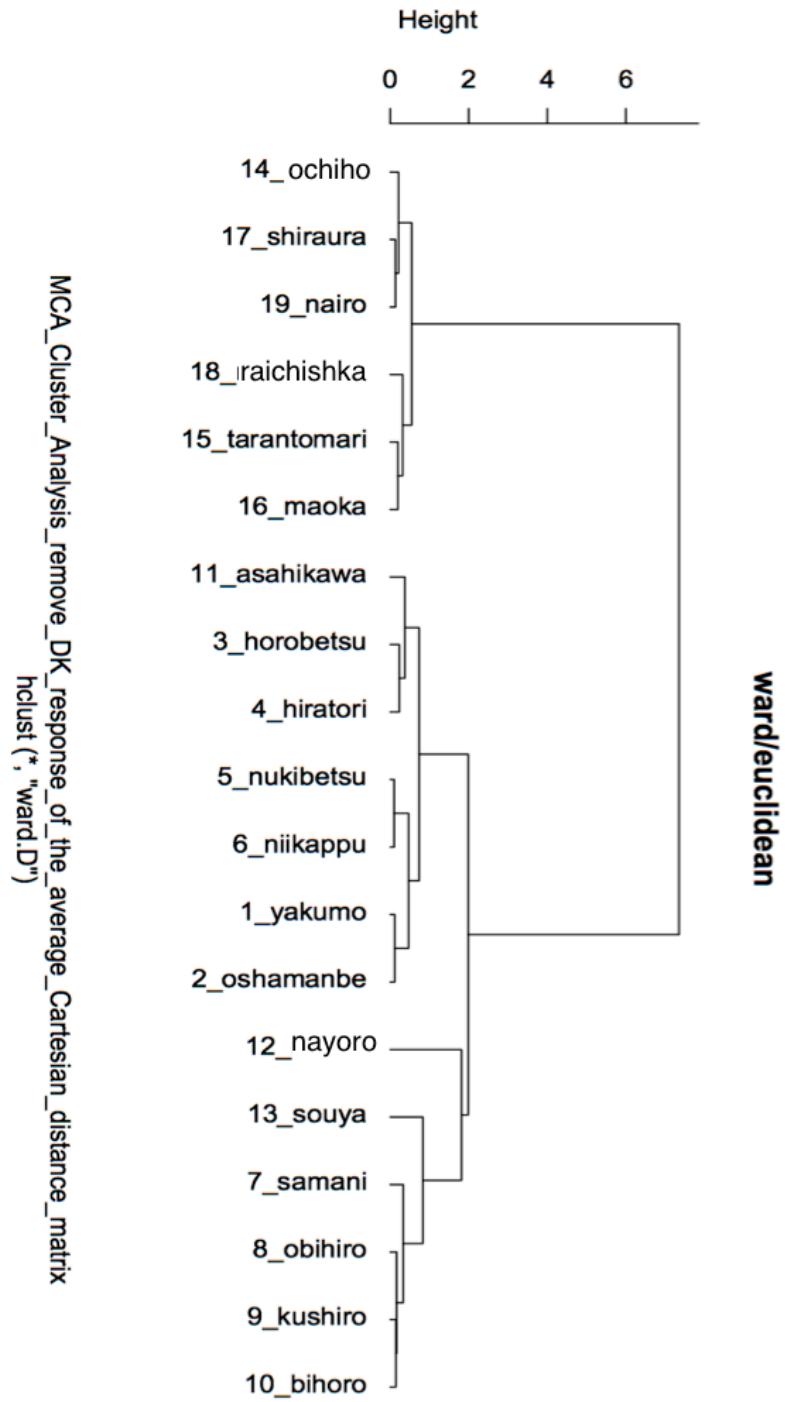
以下、先行研究と同様に「方言間の単語の同根性がわからない」という点以外は 5.6.1 節と同様の手続きで数量化Ⅲ類クラスタリングを適用した結果を示す。距離としてユークリッド距離とマンハッタン距離を用い、最長距離法、最短距離法、ウォード法を組み合わせた分析による 6 つの樹形図を示す。



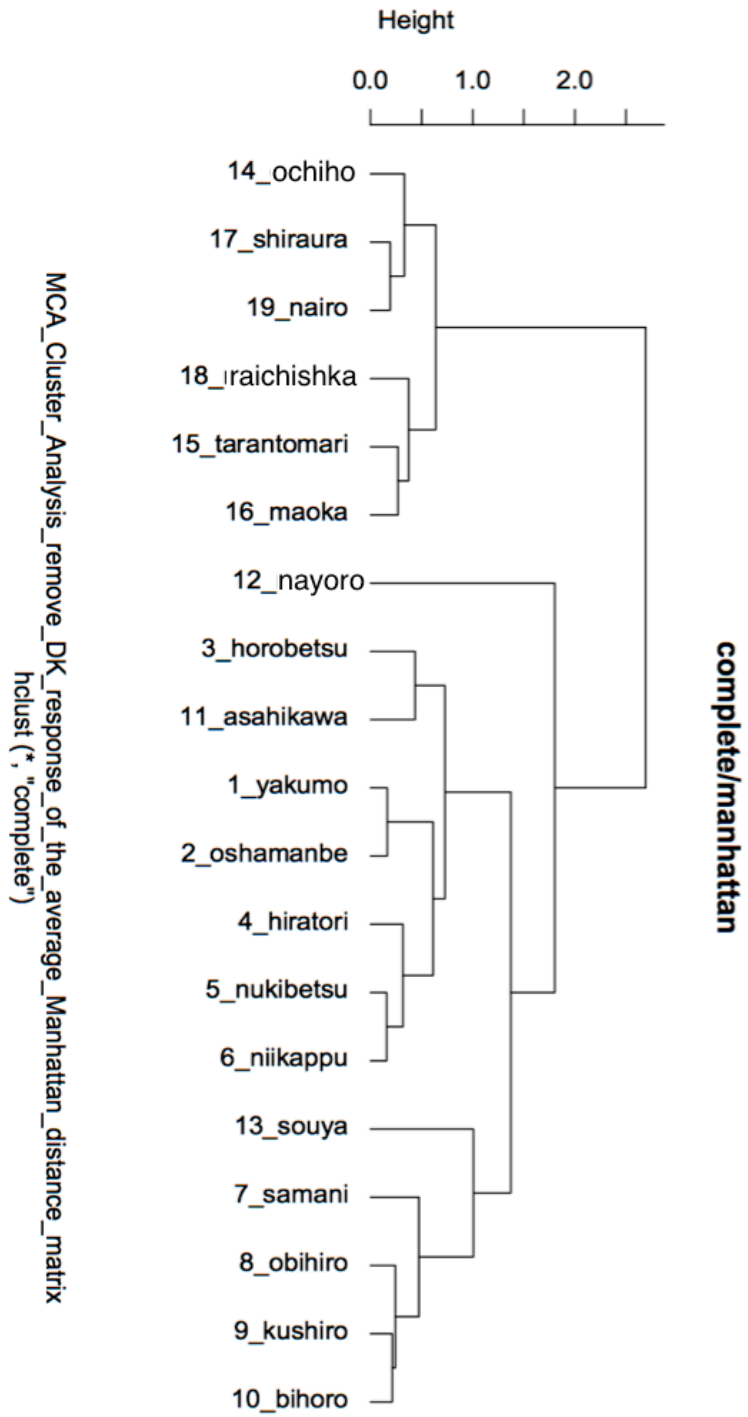
5 章付録図 1: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(ユークリッド距離、最長距離法)



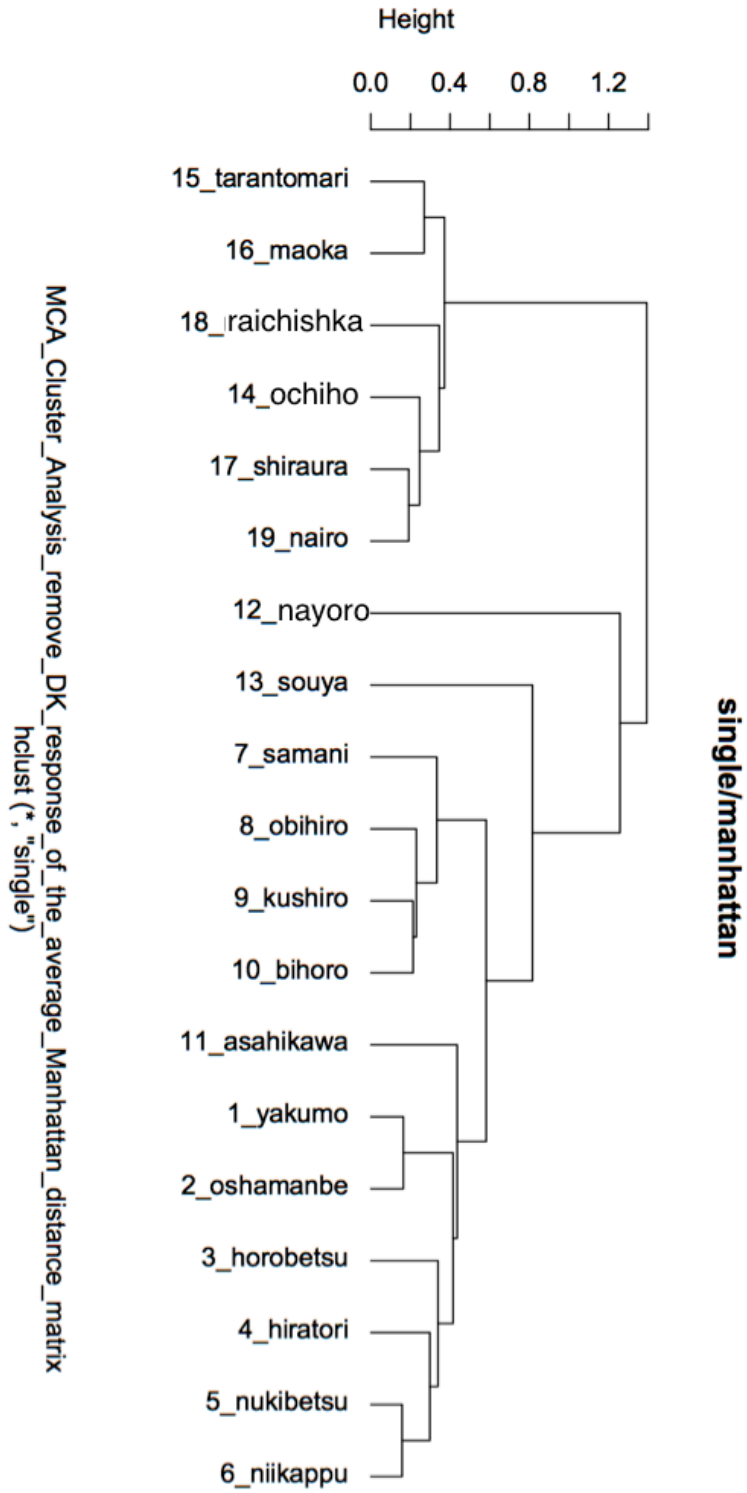
5 章付録図 2: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(ユークリッド距離、最短距離法)



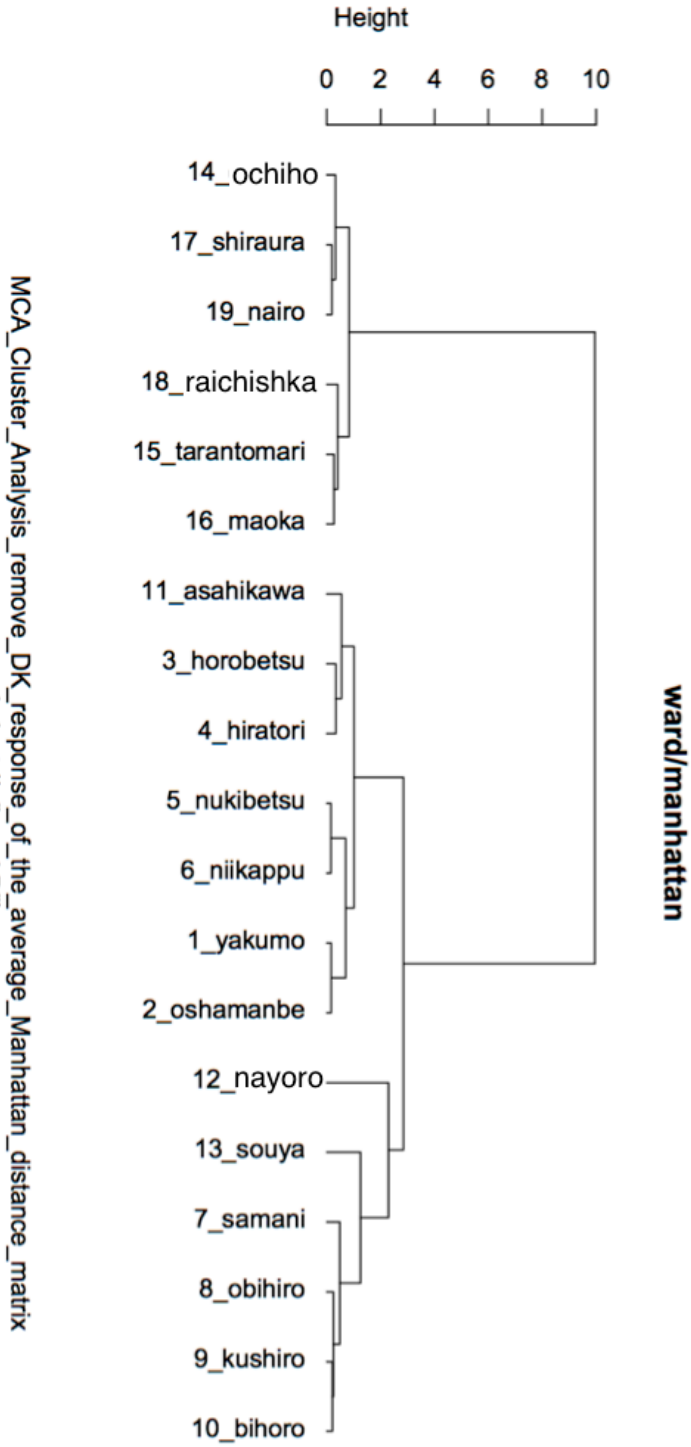
5 章付録図 3: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(ユークリッド距離、ウォード法)



5 章付録図 4: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(マンハッタン距離、最長距離法)



5 章付録図 5: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(マンハッタン距離、最短距離法)



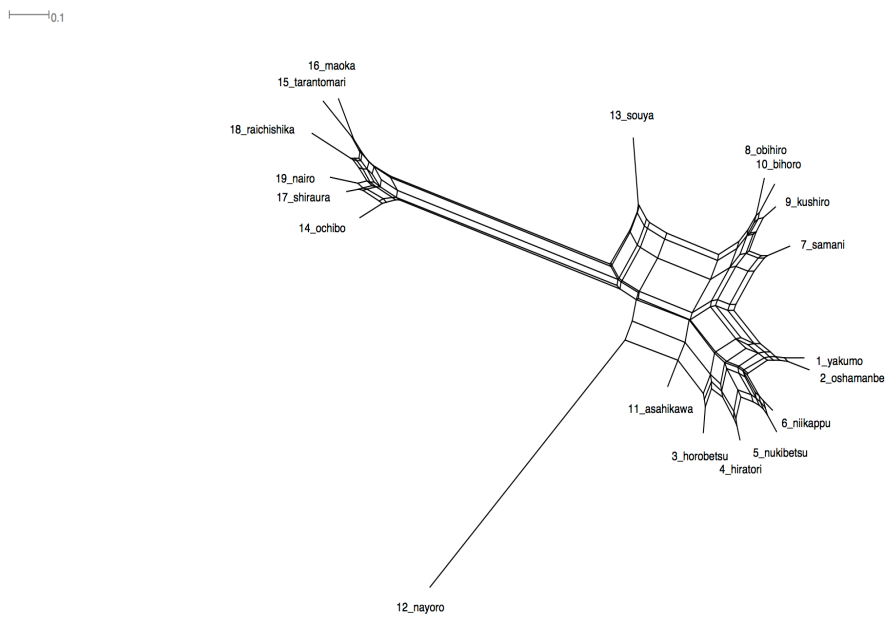
5 章付録図 6: 「方言間の単語の同根性がわからない」という情報を用いずに、数量化Ⅲ類クラスタリングを適用した結果(マンハッタン距離、ウォード法)

5 章付録の図 1 から図 6 から、以下のようなことがわかる。

1. 図 1 から図 6 においては、西海岸の方言（多蘭泊(15)、真岡(16)、ライチシカ(18)）と東海岸の方言（落帆(14)、白浦(17)、内路(19)）とが分かれ、これは樺太には南北に山脈が走り、交通は海岸沿いを基本としていたという地理学的条件とも整合した結果が示されている。
2. 通常は、北東のグループに分類される旭川(11)がいずれの図においても南西のグループに分類されている。
3. 最長距離法と、最短距離法を用いた図 1、図 2、図 4、図 5 においては、北海道方言の中で最も離れた方言が、服部･知里(1960)や Asai(1974)や Lee and Hasegawa(2013)の結果とは異なり、名寄(12)となっている。服部･知里(1960)や Asai(1974)や Lee and Hasegawa(2013)では宗谷方言が樺太方言に最も近い方言と指摘されていた。

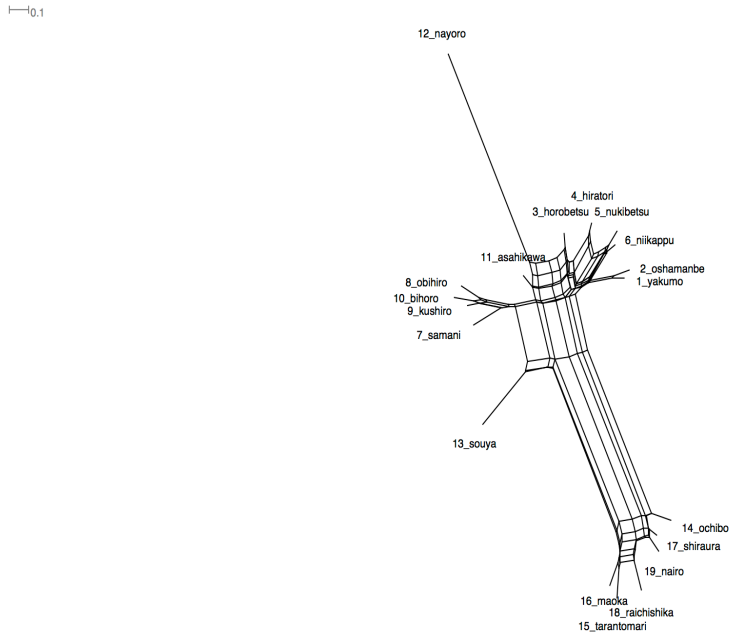
5 章付録 2.「同根性がわからない」情報を用いない MCA_Neighbor-Net

さらに、「方言間の単語の同根性がわからない」という情報を用いない点以外は 5.6.2 節と同様の手続きで MCA_Neighbor-Net を適用した結果を示す。



5 章付録図 7:「方言間の単語の同根性がわからない」という情報を用いずに、MCA_Neighbor-Net を適用した結果(ユークリッド距離)

樺太の 6 方言が、西海岸と東海岸にわかれ、言語地理学的な知見と一致している。



5 章付録―図 8:「方言間の単語の同根性がわからない」という情報を用いずに、MCA_Neighbor-Net を適用した結果(マンハッタン距離)

樺太の 6 方言が、西海岸と東海岸にわかれ、言語地理学的な知見と一致している。

第 6 章 結語

6.1 本研究の言語研究における方法論での成果

6.1.1 本研究の分析手法の流れ

本節では、本研究の各章における試行錯誤の結果、既存の(個体と変数の行列形の)データに対して推奨される分析手順を以下のフローチャートの形で図 1、図 2、図 3、図 4 に示す。

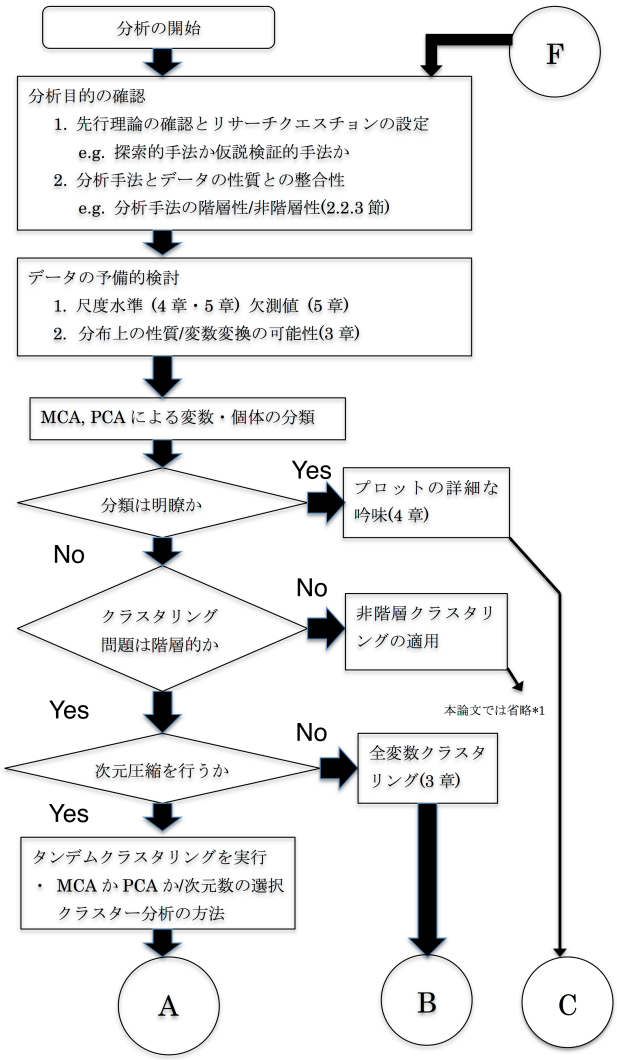


図 1. 本研究の結果得られた分析方針(フローチャート)

*1 非階層クラスタリングについてはタンデム/同時最適化のどちらも扱わなかった

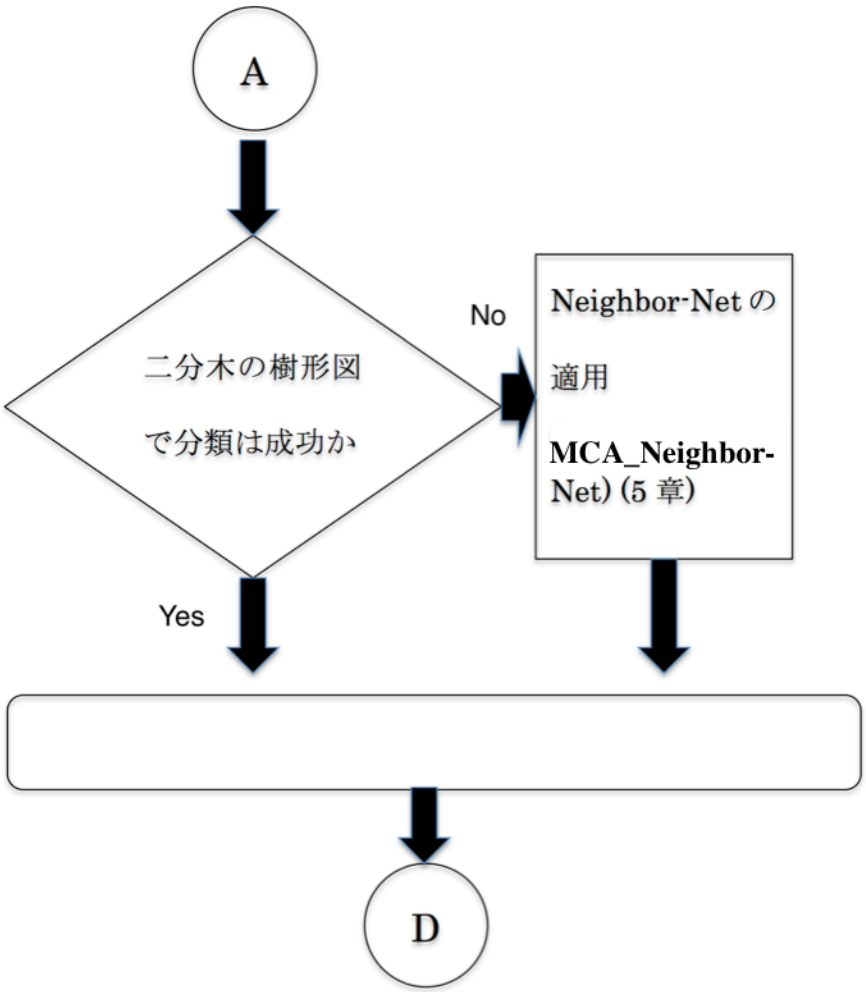

A
二分木の樹形図
で分類は成功か
No
Neighbor-Net の
適用
MCA_Neighbor-
Net) (5 章)
Yes
D

図 2. 図 1 の続き

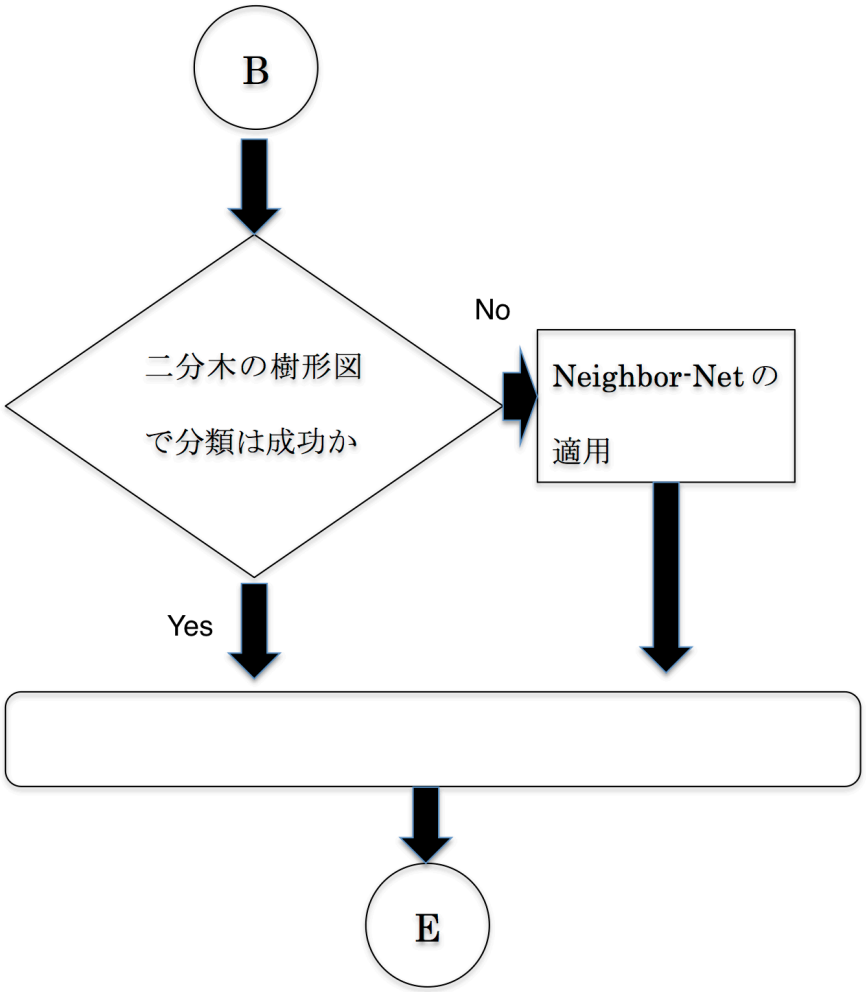


図 3. 図 1 の続き

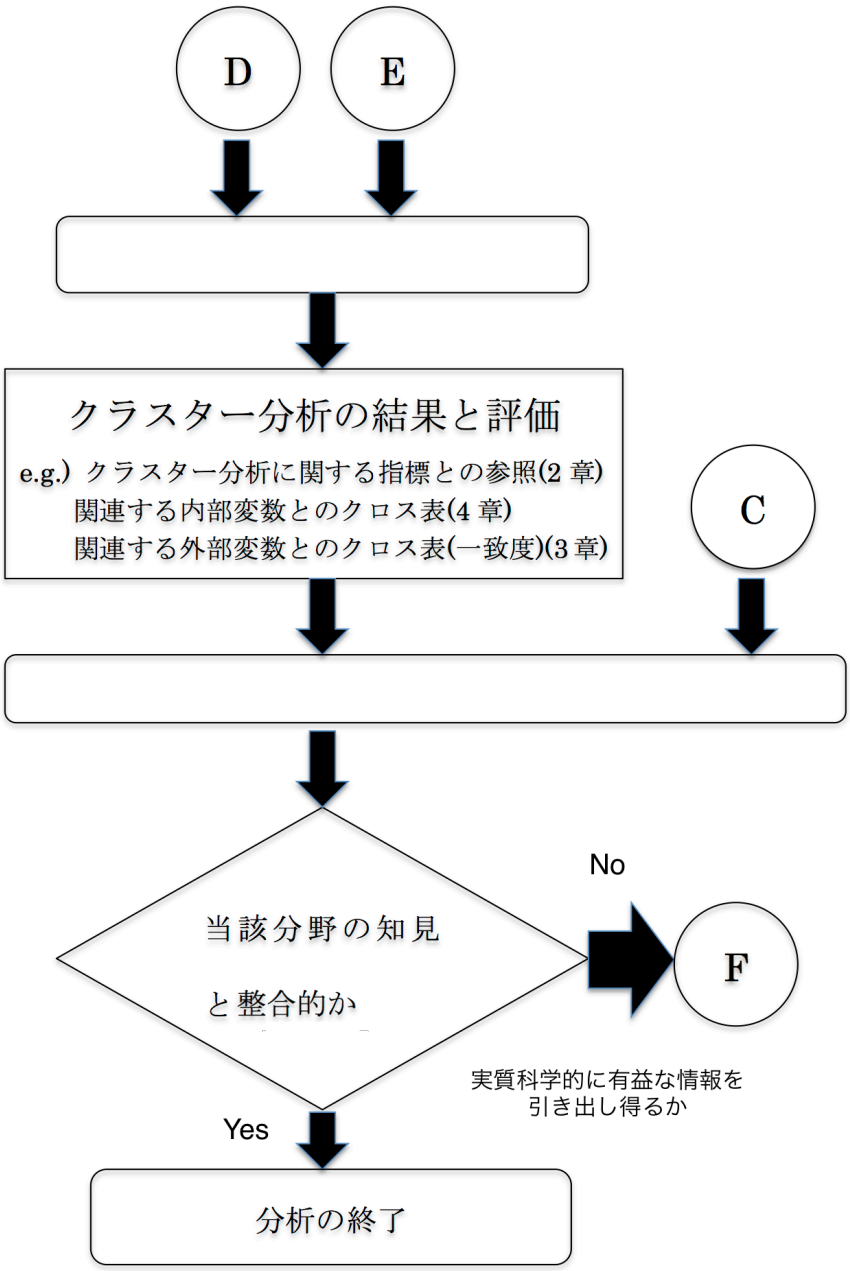

D
E
クラスター分析の結果と評価
e.g.) クラスター分析に関する指標との参照(2 章)
関連する内部変数とのクロス表(4 章)
関連する外部変数とのクロス表(一致度)(3 章)
C
当該分野の知見
と整合的か
No
F
実質科学的に有益な情報を
引き出し得るか
Yes
分析の終了

図 4. 図 2、図 3 の続き

6.1.2 フローチャートに沿った本論文の提言

ここでは、6.1.1 で示したフローチャートに沿って、本論文の検討結果に基づく提言をまとめる。なお、この提言はデータの対象を言語データに限らないが、言語データへの適用を意識して説明する。

まず既存のデータの分析を始めるに際して、以下の 2 点に関して検討を行うことが前提となる。第一に、先行理論の確認とリサーチクエスチョンが何であるかによって、探索的方法と仮説検証的方法とのどちらを分析手法として用いるか検討する必要がある。本研究では、MCA を探索的な手法として主に用いたが、仮説検証的な枠組みで MCA に関連した方法を適用していくことも十分に考えられよう。第二に、分析手法とデータの性質との整合性について考える必要がある。分類の手法は、問題の性質によって、階層的クラスタリングか非階層的クラスタリングかを選択する必要がある。この点について分析にあたって、十分に検討しておくことが求められる。具体的には、本研究では、5 章でのアイヌ語の方言における基礎語彙の分析においては、言語の伝搬は人間の移動と密接な関連があること及びアイヌ語においては言語接触が比較的少ないと考えられることから、非階層クラスター分析ではなく階層クラスター分析を選択した。さらに言うと、階層的手法の中でも二分木的なデンドログラムを表現する従来のクラスター分析を適用に止めるか、より柔軟な構造を取り出す Neighbor-Net を適用に踏み込むかという選択もあるが、この点については後に述べる。いずれにせよ、このようなこと(分析目的と手法の整合性)を事前に綿密に検討しておくことは、本節で提案したフローチャートの各所における選択に影響するため、極めて重要である。

次に、リサーチクエスチョンに沿ってデータを本格的に分析する前に、データの予備的検討が不可欠である。まずデータの尺度水準を考慮し、欠測値の扱い方について注意する必要がある。4 章では、Tsunoda et al. (1995b)のデータが尺度水準の観点からは、順序尺度もしくは名義尺度であることに注目し、より適切な多重対応分析を適用することによって、Tsunoda et al. (1995a)の知見の再考を示唆するとともに、言語類型論の観点から新たな興味深い知見を得た。5 章では、欠測値にあたる「方言の単語の同根性がわからない」という情報を無視できる欠測値と捉えるのではなく、なんらかの情報を持ったものとして独立したカテゴリーで扱うという指針を提示した。

さらに、データをプロットし分布の歪みなどを把握する必要がある。データ

が量的変数の場合には、変数変換などによって、データのバラツキを均一化することによって結果的にデータの特徴をよく捉えた「望ましい」距離行列が得られることがある。本論文では 3 章において、データに歪みがある場合には Box-Cox 変換などを施すことにより、改善を図るという指針を得た。

その上で、MCA や PCA を実際にデータに適用する段階に移る。実際に、本研究の 4 章では MCA の適用のみで、個体と変数の明瞭なプロットが得られ、言語学的に有益な情報をもたらした。MCA による個体と変数の同時分類が成功しているか否かについては、変数のプロットと個体のプロットの両方の詳細な吟味が必要である。その際、注目する変数に関する各言語のカテゴリーについて、視覚的に分かりやすい表示を行うことで解釈を容易にした。

他方で、3 章のように MCA の適用の結果得られたプロットが、必ずしも個体や変数の分類に関して明瞭な結果をもたらさない場合もある。このような場合においては、手続きとしてクラスター分析を適用する段階に入る。その際、初期の検討においてもたらされた、分析手法の階層性/非階層性に応じて、チャートが分岐している。本論文では階層的問題のみを扱った。

さて、階層的な問題においては、次元圧縮を行うか/行わないかの選択肢がある。本論文では素データに対して何らかの次元圧縮を行い、得られた少数の次元に対して階層的クラスター分析を適用する「タンデムクラスタリング」と素データに対して階層的クラスター分析を適用する「全変数クラスタリング」の比較を行なった。

本章で提案した、MCA を適用し得られた多次元の座標の布置に対してクラスター分析を実行する「数量化Ⅲ類クラスタリング」は、いわゆる「タンデムクラスタリング」とされ、次元圧縮によるクラスタリングがうまく行かないケースが例証されていた(Arabie & Hubert, 1994; DeSarbo, Jedidi, Cool & Schendel, 1990; Timmerman, Ceulemans, Kiers & Vichi, 2010; Vichi & Kiers, 2001; Yamamoto, 2012; Yamamoto & Hwang, 2014)。実際に、3 章においては「タンデムクラスタリング」はうまく行かない例が、5 章では逆に「タンデムクラスタリング」がうまく行く例がそれぞれ示された。この点について本研究は、クラスター分析とタンデムクラスタリングの両方を実行し、両者の結果を何らかの基準から比較しより適切なものを選ぶことを提案せざるを得ない。統計的な観点からクラスタリングの「良さ」の基準に加えて得られた知見について外部の理論などとの整合性を参照することを提案した。クラスタリング結果の評

価については後述する。

タンデムクラスタリングを含めてクラスタリングの適用の際には、距離の選択やグルーピングの選択が重要である。タンデムクラスタリングなどの次元圧縮を伴う手法においては、解の次元数の選択や、素データを使うクラスター分析とタンデムクラスタリングとの比較が重要である。次元圧縮とクラスタリングの同時最適化の手法も適用可能な文脈においては、それも含めた 3 手法の比較が望ましいであろう。例えば、本研究では、実際には、グルーピングの選択において、最短距離法は均一性の観点から有効なクラスタリングの結果をもたらさなかった。最短距離法のいわゆる「チェーン現象」はすでに多くの文献で指摘されているが、言語データの分析においても最短距離法は同様の問題を抱えがちであることが示唆された。

解の次元数の選択については、スクリープロットを用いた方法や累積寄与率に基づいた方法を併用し、複数の次元数を選び、それらを比較した。結果、第一に、スクリープロットから固有値の減少が緩やかになる一つ前の次元を選択する手法を採用しても、必ずしも外的基準と整合性のある樹形図が得られる訳ではなかった。第二に、単に固有値の大きい少数の次元を選択しても、逆に単純に累積寄与率を上げても、必ずしも外的基準と整合性のある樹形図が得られるわけではなかった。よって、タンデムクタスタリングの次元数の選択においては、万能な方法はなく、本論文の 3 章で例示したように以下の 3 水準を目安に、外的基準との整合性を考慮しつつ、結果の比較を行う方針を推奨しておくこととする。第一に、寄与率の大きい上位少数(高々3 程度)の次元を選択する方法、第二に、固有値のスクリープロットから固有値の減少が緩やかになる一つ前の次元を選択する方法、第三に、累積寄与率をある程度まで大きくとった(70%～80%まで)次元を選択する方法である。第一の手法の利点としては、データの重要な情報が寄与率の大きい上位少数の次元で表現される場合においては、データを特徴付ける個体や変数の解釈が容易である点が挙げられる。第二の手法の利点としては、寄与率の比較的小さい上位の次元まで選択することによって、データの解釈可能な相関・連関構造を余すことなく取り込んだ成分によるクラスタリングができることが挙げられる。第三の手法の利点としては、固有値の小さい下位次元まで選択することによって、データが複雑かつ高次元で表現されている場合には、データの構造をより正確に捉えることができる点が挙げられる。ただし、第三の手法の欠点としては、高次元までの情報を取り込んでい

ることによってデータを特徴付ける個体や変数の解釈が容易ではない点がある。

しかし、クラスター分析のように二分木的な樹形図でデータの「最も強い」構造を取り出す手法には、限界がある可能性もある。こうした場合においては、データの「2 番目以降に強い」構造を取り出すことができる、より柔軟な分類手法である Neighbor-Net の適用が有効である。実際に、本論文の 5 章においては、アイヌ語の方言データに対して MCA を適用した結果得られた距離行列に対して Neighbor-Net を適用する手法(筆者は MCA_Neighbor-Net と称した)ことによって、今までは文献学上は言及されていたものの、統計学的には確認されなかったアイヌ語における様々な言語地理学的パターンを確認することができた。

最後に、統計手法の適用によって得られた結果をどのように評価するか、という点に関して、本論文では 3 つの観点から提案を行った。

具体的には、第一に、統計学的な観点からは、均等性(石田·西尾·椿, 2011)などにより統計学的に受け入れ可能なものを選んだ上で、実質科学的な知見からクラスタリングの「良さ」を検討することが可能であろう。第二に、4 章のように、分析に用いた変数(内部変数)とのクロス表を作成し、分類の結果を最もよく説明できる変数を探索する方法があろう。第三に、3 章のように先行研究の蓄積によってある程度参照するに足る区分(外的変数)が存在している場合には、分類の結果と外的変数とのクロス表を作成し、分類結果を評価する方法がある。すでに確立した外的基準としての分類変数が存在している場合には、クラスター分析の必要性は薄い。外的基準ではなく外的変数と表記した理由もここにある。この場合の評価の指標も「精度」という言葉ではなく、「一致度」という語を用いた。

最後に、統計手法の適用によって得られた結果は、当該分野の実質科学的知見との整合性によって、評価されるべきである。実際に、3 章 5 章で述べたように、クラスタリングの「良さ」というものは、統計学的な知見と統計学以外の文献学や言語学、人類学などの知見を総合し、多面的に評価する必要があり、当該分野の知見との間に著しい乖離が見られた場合には、フローチャートの始めの段階に立ち返り、分析目的の確認から再検討を行う必要があろう。

ただし、分析の開始前に、データの質の検討や「どのようにデータが獲得されたか」、「どのようにデータの値が与えられたか」ということを、統計側と実質科学的側の対話により十分に吟味することが、後述する「林のデータの科学」においては、重要であり、本研究ではデータを取る段階で、これらの試行錯誤

がなされていない点に留意する必要がある。

6.2 本研究の言語研究における実質科学的貢献

「数量化Ⅲ類クラスタリング」の実際の言語学データへの試行によって、本研究では 4 つの点が示唆された。

第一に、村上·今西(1999)の源氏物語の成立論に関する「A→C→B→D」という推測は、MCA の 2 次元の布置の「解釈」に基づいており、実際にはやや踏み込み過ぎと考えられた。統計学的には MCA と平方根変換を組み合わせることで「(A,C)→(B,D)」という成立順序が示唆された。また、松風、初音、夕霧の帖に関して、既存の分類とは異なるグループに属している可能性を指摘することができた。

第二に、Tsunoda et al. (1995b)のデータに対しては、MCA を適用した結果だけで言語類型論の観点から十分興味深い結果が得られた。「側置詞」の観点からは、「前置詞言語」「後置詞言語」「無側置言語」の 3 つのタイプを想定することが妥当であることが示唆された。変数の分析から、2 つの分類が重要であることがわかってきた。一つは、名詞句内で「主要部前置型」であるか「主要部後置型」であるかということと、もう一つは節レベルで「主要部前置型」であるか「主要部後置型」であるかということである。さらに、興味深い点として、名詞句内であっても「所有格と目的語の順序」だけは、節レベルの変数と同じ振る舞いをすることがわかった。この現象については、Whitman and Ono (2017, to appear)において、言語変化の観点から説明が試みられている。

第三に、樺太のアイヌ語の方言に関して、従来の先行研究の結果と比較して、より言語地理学的に妥当な結論を得た。南樺太には南北に山脈が走っており、また交通も沿岸の海上交通を主としているため、樺太のアイヌ語の方言が東海岸と西海岸に分かれることは、今後のアイヌ語の研究に一石を投じる可能性がある。

第四に、MCA_Neighbor-Net を利用することによって、今まで文献学的には指摘されていたものの統計学的には示されることのなかった沙流·平取地域を中心としたアイヌ語に置ける方言周圏論的構造を明らかにした。

第三、第四の知見が得られた背景としては、今までの基礎語彙統計学における研究方法は、「明示的に得られた方言間の単語の同根性に付いてのみ研究する方法」であったが、本研究では、「方言の同根性がわからなかった」という

データを無視せず、1 つの貴重な情報として扱うという工夫があった。このことは、既存の基礎語彙統計学研究に、新たな視座を提供し、既存のデータから新たな事実を統計的に「再発見」することにつながる可能性がある。

我々は各々の文化に生きているが、その文化において言語が重要であることは言うまでもない。現在、特にアイヌ語のように話者が僅かしかおらず、既存のデータの見直しが必要となっている言語の研究においては、このように統計学の知見によって、新たな進展が期待できる。

ただし、本研究によって示唆された知見については以下の点を留意しなければならない。

第一に、3 章で扱った村上･今西(1999)のデータは、源氏物語の助動詞に関する情報のみであった。実際には、名詞や、助詞、形容詞など特徴量となるデータは様々なものがある。また、源氏物語のデータのみを扱っている限り、文体の違いが、作者の違いによるものなのか、経年効果によるものなのか、区別がつかない。よって、紫式部が書いたとされる「紫式部日記」や、同時代に書かれた「うつほ物語」などと比較する必要がある。さらに、源氏物語に関する文献学的知見など、量的データとして扱われていない知見も考慮する必要がある。

第二に、4 章で扱った Tsunoda et al. (1995b)のデータは、地域的にも系統的にも偏ったデータベースであった。地域的には、ヨーロッパの言語に偏り、アフリカの言語が極端に少なく、系統的には、インド･ヨーロッパ語族の言語が多かった。ただし、Whitman and Ono (2017, to appear)では地域的にも系統的にもよりバランスの取れたデータベースで 4 章の知見を検証している。

第三に、5 章で扱ったアイヌ語のデータは、量的な分析がしやすい語彙に関するデータのみである。この他にも音韻論、形態論、統語論など言語学には様々な分野があり、考慮するべき情報は多い。さらに、アイヌ語に関する文献学、考古学、歴史学など、今までの知見は、さまざまな学問の知識を総合して導かれたものであり、語彙の情報のみをもとに導かれた本研究の結果が、既存の知見より勝るとは、即断することはできない。序で述べたように、アイヌ語学の中において、本研究の結果に対して、既存の知見をどのように再検討するかを待つ必要がある。

6.3 本研究の示唆する分析上の哲学

本稿のすべての章に共通していることは、先述した通り、言語学の活動では、

言語学のデータのみではなく、文献学、考古学、歴史学、人類学など、様々な知見を総合し、より妥当な結論を導くべきであるので、統計的に分析できる一部の情報から得られた結論をもとに、既存の知見を直ちに覆すことは難しいということである。同時に、これは分析の結果得られた知見について、統計学以外からの評価が、統計学的な評価に対して全面的に優れていることまでは意味しない。統計学の手法の適用が新たな知見をもたらす事例は 4 章や 5 章の一部に例示されており、発見的な手法としての統計学の有用性が改めて示されている。それでもなお我々は、結果に重要な影響を与えるデータが分析の対象外であったため、既存の知見とは異なった知見が得られただけの可能性もあることに留意すべきである。

第 1 章で述べたように、言語データに対して統計手法を適用することは、統計手法によって計量的な視点を言語学に提示し、これに対して言語学において既存の知見に対する再検討が行われ、今度は言語学から統計学に対して問題提起がなされるという、言語学と統計学の対話によって、統計学と言語学の営みが生産的になる。言語データに対して統計手法を適用した結果は、言語学との対話なしでは、実質科学的な意味に乏しいものとなろう。

学際的な分野においては、統計学者は、データ収集の現場に寄り添い、その中から生まれた問題に対して、その学問の知識と統計学の知識を総合し、新たな手法や分析の切り口を提案し、新たなデータ解析の結果を、データ収集の現場に提示する。そして、その結果に対するデータの現場の見解をもとに、分析の妥当性を考え、「対話」が行われる。こうした実践的な営みは林(2001)が「データの科学」と呼んだ哲学であり、本研究でとった分析の方針もこの哲学に従っている。本研究で得られた結論を端的に示せば、各種言語データの解析もこの理念のもとに推進されるべきである、ということである。

参考文献

足立浩平，村上隆（2010）．非計量多変量解析法-- 主成分分析から多重対応分析へ．朝倉書店.

阿部秋生（1939）．源氏物語の執筆順序．*国語と国文学*, *16*(8~9).

Albu, M. (2006). *Quantitative Analyses of Typological Data*. Ph.D. dissertation, University of Leipzing.

Arabie, P., & Hubert, L. (1994). Cluster analysis in marketing research (pp.160-189). In Bagozzi, R. P., editor, *Advanced methods in marketing research.* Blackwell, Oxford.

新井皓士（1997）．源氏物語・宇治十帖の作者問題:一つの計量言語学的アプローチ．*一橋論叢*, *117*(3), 397-413.

浅井治平（1933）．地勢．地理講座日本編第 1 巻．改造社

Asai, T. (1974) Classfication of dialects: Cluster analysis of Ainu dialects. *Bulltein of the Institute for the Study of North Eurasian Culture. 8*, 45-136.

Austin, W. B. (1966). The Posthumous Greene Pamphlets: A Computerised Study. *Shakespeare Newsletter*, *16*, 45.

Baayen, H., Van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, *11*(3), 121-132.

Baillie, W. M. (1974). Authorship Attribution in Jacobean Dramatic Texts.*Computers in the Humanities*, 73-81.

Benzécri, J.-P. (1973). *L'Analyse des Données. Volume II. L'Analyse des Correspondances*. Paris: Dunod.

Bickel, B. (2008). A general method for the statistical evaluation of typological distributions. *Manuscript, https://core.ac.uk/download/files/542/14515057.* Accessed on 2016-08-04.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statstical Society. Series B (Methodological)*, 211-252.

Bryant, D., & Moulton, V. (2003). Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution, 21*(2), 255-265.

Bryant, D., Moulton, V., & Spillner, A. (2007). Consistency of the

neighbor-net algorithm. *Algorithms for Molecular Biology, 2*(1), 1.

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference.* Springer-Verlag: New York.

Chen, Z., & Van Ness, J. W. (1996). Space-conserving agglomerative algorithms, *Journal of Classfication, 13,* 157-168.

地理院地図 (2016). 国土地理院. URL https://maps.gsi.go.jp/. (Accessed on 2016-12-27)

Chrétien, C. D. (1943). The quantitative method for determining linguistic relationship: Interpretation of results and tests of significance. *UCPL* 1:2, 11-20.

Cichocki, W. (1989). An application of dual scaling in dialectometry. *Journal of English Linguistics, 22*(1), 91-95.

Cichocki, W. (2006). Geographic variation in Acadian French /r: What can correspondence analysis contribute toward explanation?. *Literary and Linguistic Computing, 21*(4), 529-541.

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In *New approaches in classification and data analysis* (pp. 212-219). Springer Berlin Heidelberg.

Desarbo, W., Jedidi, K., Cool, K., & Schendel, D. (1991). Simultaneous multidimensional unfolding and cluster analysis: An investigation of strategic groups. *Marketing Letters*, *2*(2), 129-146.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, *29*(8), 1969-1973.

Dyer, M. S. (1989). Large Linguistic Areas and Language Sampling. *Studies in Language, 13,* 257-292.

Dryer, M. S. & Haspelmath, M. (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2014-07-28.)

Ellegård, A. (1962). A Statistical Method for Determining Authorship: 1769–72.*Gothenburg: Acta Universitatis Gothoburgensis*.

Everitt, B., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis*

Wiley.
Fisher, L., & Van Ness, J. (1971). Admissible clustering procedures, *Biometrika, 58,* 91-104.
Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics,* **10**, 422-429.
藤井潔 (1966) 源氏物語の構造. 桜楓社.
Gattone, S. A., & Rocci, R. (2012). Clustering curves on a reduced subspace. *Journal of Computational and Graphical Statistics*, *21*(2), 361-379.
Gifi, A. (1990). *Nonlinear multivariate analysis.* John Wiley and Sons.
Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1-26.
Greenberg, J. H. (1966). "Some Universals of Language with Particular Reference to the Order of Meaningful Elements." Greenberg 73-113.
Greenacre, M., & Blasius, J. eds. *Multiple correspondence analysis and related methods.* CRC Press, 2006.
Gries, S. T. (2012). Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: towards more and more fruitful exchanges. *Language and Computers*, *75*(1), 41-63.
Gries, S. T., & Stefanowitsch, A. (2007). *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis* (Vol. 1). Walter de Gruyter.
服部四郎, 知里真志保 (1960). アイヌ語諸方言の基礎語彙統計学的研究. *季刊民族學研究, 24*(4), 307-342.
Hawkins, J. A. (1983). *Word Order Universals.* New York: Academic Press.
Hayashi, C. (1952). On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics*, *3*(1), 69-98.
林知己夫 (1993). 数量化: 理論と方法. 朝倉書店.
林知己夫 (2001). データの科学. 朝倉書店

Hocking, R. R. (2013). *Methods & applications of linear models: regression and the analysis of variance*. John Wiley & Sons.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**, 417–441, and 498–520.

Huang, Z. (1997) A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In *KDD: Techniques and Applications* (H. Lu, H. Motoda and H. Luu, Eds.), pp. 21-34, World Scientific, Singapore.

Huson, D. (1998). SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics, 14*(1), 68-73.

Huson, D., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution, 23*(1), 254-267. Oxford.

池田亀鑑 (1951). 新講源氏物語. 至文堂.

池田亀鑑 (1985). 源氏物語大成. 中央公論社.

今西祐一郎, 室伏信助 (2010). 紫上系と玉鬘系—成立論のゆくえ. 勉誠出版.

Inoue, F. (1988) Dialect Image and New Dialect Forms. *Area and Culture Studies*, Tokyo University of Foreign Studies*, 38,* 13-23.

井上史雄. (1990). 標準語形の計量的性格と地理的分布パターン. *言語研究, 97,* 44-72.

Inoue, F. (1992). Dialect distribution pattern along the Tokaido line. *Gengo Kenkyu, 101,* 35-63.

Inoue, F. (2009). Year of first attestation of Standard Japanese Forms and Gravity Centre by Railway Distance. *Dialectologia et Geolinguistica, 17*(1), 118-133.

石田実, 西尾チヅル, 椿広計 (2011). 2 値変量に基づく教師無し分類における類似係数の選択. *行動計量学, 38*(1), 65-81.

岩坪秀一. (1987). 数量化法の基礎. 朝倉書店.

Jin, M., & Murakami, M. (1993). AUTHORS"CHARACTERISTIC WRITING STYLES AS SEEN THROUGH THEIR USE OF COMMAS. *Behaviormetrika,20*(1), 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*(8), 382-391.

Kaufman, L., & Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the L1–Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, *7*, 1-41.

Kroeber, A. L., & Chrétien, C. D. (1937). Quantitative classification of Indo-European languages. *Language*, *13*(2), 83-103.

Kroeber, A. L., & Chrétien, C. D. (1939). The statistical technique and Hittite. *Language*, *15*(2), 69-71.

Labov, W. (1963). The social motivation of a sound change. *Word*, *19*(3), 273-309.

Labov, W. (1966). *The Social Stratification of English in New York City.* Washington DC: Washington DC: Center for Applied Linguistics.

Lance, G. N. & Williams, W. T. (1967). A general theory of classificatory sorting strategies I. Hierarchical Systems, *Computer Journal, 9*, 373-380.

Lebart, Ludovic (1994) Complementary use of correspondence analysis and cluster analysis. *Correspondence analysis in the social sciences*, 162-178.

Lee, A., & Willcox, B. (2014). Minkowski Generalizations of Ward's Method in Hierarchical Clustering. *Journal of Classification*, *31*(2), 194-218.

Lee, S., & Hasegawa, T. (2013). Evolution of the Ainu Language in Space and Time. *PloS one, 8*(4), e62243.

Lemmon, A. R., & Lemmon, E. M. (2008). A likelihood framework for estimating phylogeographic history on a continuous landscape. *Systematic Biology*, *57*(4), 544-561.

Lewis, M. P, Gary F. S, & Charles D. F. (eds.). 2013. Ethnologue: Languages of the world, (17th ed.) Dallas, Texas: SIL International. (Online version: http://www.ethnologue.com, Accessed on 2014-07-28.)

Linmans, A. J. M. (1998). Correspondence analysis of the Synoptic Gospels. *Literary and Linguistic Computing, 13*(1), 1-13.

リンゼイ·J·ウェイリー. (1997). 大堀壽夫·古賀裕章·山泉実訳. (2006). 『言語類型論入門–言語の普遍性と多様性–』 岩波書店.

MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press. pp. 281–297.

Maslova, E. (2000). A dynamic approach to the verification of distributional universals. *Linguistic Typology, 4*(3), 307-333.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Methods. *Theory and Practice*.

Mealand, D. L. (1995). Correspondence analysis of Luke. *Literary and Linguistic Computing, 10*(3), 171-182.

Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 237-249.

Miller, G. (1971). Empirical methods in the study of semantics. In: *Semantics: An Interdisciplinary Reader.* Ed. By D. Steinberg and L. Jakobovits. Cambridge: Cambridge University Press, 569-585.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measure for cluster analysis. *Psychometrika, 46,* 187-199.

Mirkin, B, (1996). *Mathematical Classification and Clustering,* Kluwer Academic Publishsers, Dobrecht.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics* (Vol. 66). Walter de Gruyter GmbH & Co KG.

Morton, A. Q. (1978). *Literacy Detection*, East Grinstead: Bowker Publishing Company.

Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*,*58*(302), 275-309.

村上征勝 (1994). 真贋の科学: 計量文献学入門. 朝倉書店.

村上征勝, 今西祐一郎 (1999). 源氏物語の助動詞の計量分析. *情報処理学会論文誌*, *40*(3), 774-782.

村上征勝 (2002). 文化を計る: 文化計量序説. 朝倉書店.
村田年. (2000). 多変量解析による文章の所属ジャンルの判別―論理展開を支える接続語句・助詞相当句を指標として―. *統計数理*, *48*(2), 311-326.
永田靖 (2005). 統計学のための数学入門 30 講. 朝倉書店.
永田靖, 棟近雅彦 (2014). 多変量解析法入門. サイエンス社
中川裕 (1996). 言語地理学によるアイヌ語の史的研究. *北海道立アイヌ民族文化研究センター研究紀要*, *2*, 1-17.
Nishisato, S. (2006). *Multidimensional nonlinear descriptive analysis*. Chapman and Hall/CRC.
Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh University Press.
小野洋平 (2015a). 源氏物語成立論の統計科学的再考察: 村上・今西 (1999) を中心に. *計量国語学*, *29*(8), 296–312.
小野洋平 (2015b). 服部・知里 (1960) の統計科学的再考察: アイヌ語方言周圏論の実証. *北方人文研究*, *8*, 25–41.
Ono, Y. (2016). How to Handle "Don't know" Judgments in Lexicostatistical Survey: An Exercise in Ainu Dialectology. Unpublished manuscript. (Available on request)
大野晋 (1984). 源氏物語. 岩波書店.
Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1-9.
Pearson, K. (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space" *Philosophical Magazine 2*(11), 559–572.
Reed, D. W., & Spicer, J. L. (1952). Correlation methods of comparing idiolects in a transition area. *Language*, 348-359.
Rousseeuw, P. J. (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics 20*, 53–65.
R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

齋藤堯幸, 宿久洋 (2006). 関連データの解析法. 共立出版.
佐藤義治 (2009). 多変量データの分類-判別分析・クラスター分析.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.
Székely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of classification*, *22*(2), 151-183.
Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, *5*, 1-34.
Stefanowitsch, A. (2010). Empirical cognitive semantics: some thoughts.*Quantitative methods in cognitive semantics: corpus-driven approaches*, 355-380.
Steinhaus, H. (1957). "Sur la division des corps matériels en parties". *Bull. Acad. Polon. Sci.* (in French) *4*(12), 801–804.
菅民郎 (2001). 『多変量解析の実践（下)』, 現代数学社.
Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics. 21*, 121-137.
武田宗俊 (1954). 源氏物語の研究. 岩波書店.
竹内啓編 (1989). 統計学辞典. 東洋経済新報社.
玉上琢彌 (1940). 源氏物語成立攷. *国語・国文*, *10*(4).
鄭躍軍, 金明哲. (2011). R で学ぶデータサイエンス 17―社会調査 データ解析, 共立出版.
Timmerman, M. E., Ceulemans, E., Kiers, H. A., & Vichi, M. (2010). Factorial and reduced K-means reconsidered. *Computational Statistics & Data Analysis*, *54*(7), 1858-1871.
Tsuchiyama, G., & Murakami, M. (2013). Authorship Identification of Classical Japanese Literature Using Quantitative Analysis, *Journal of Mathematics and System Science. 3*, 631-640.
角田太作 (1991). 『世界の言語と日本語』. くろしお出版.
Tsunoda, T., Ueda, S., & Itoh, Y. (1995a). Adposition in word-order typology. *Linguistcs, 33*(4), 741-762.
Tsunoda, T., Ueda, S., & Itoh, Y. (1995b). Statistical Data Analysis for Word

Ordering Rule. *The Institute of Statistical Mathematics Cooperative Research Report 70,* Tokyo: The Institute of Statistical Mathematics.

Ueda, H. (1993). División dialectal de Andalucía: Análisis computacional, *Actas del Tercer Congreso de Hispanistas de Asia*, Asociación Asiática de Hispanistas, Tokio, 407-419.

上田博人 (2013). 広域スペイン語語彙バリエーション研究における新しい数量化の試み: 日本語計量方言学の方法に学ぶ. *日本語・日本学研究*/東京外国語大学国際日本研究センター編, *3*, 59-90.

Ueda, H., & Perea, M. P. (2014). The degree of union resulting from reaction points expressed in a diatopic table. An application to a Catalan verb morphology database. *Dialectologia et Geolinguistica, 22*(1), 16-38.

上田澄江・伊藤栄明 (1995). 語順規則による言語分類と 2 パラメータモデル. *統計数理*, *43*(2), 341-365.

Venables, W. N., & Ripley, B. D. (2002) *Modern Applied Statistics with S.* Fourth edition. Springer.

Vichi, M., & Kiers, H. A. (2001). Factorial k-means analysis for two-way data.*Computational Statistics & Data Analysis*, *37*(1), 49-64.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, *58*(301), 236-244.

和辻哲郎 (1926). 日本精神史研究. 岩波書店

Wen Chieh-Hua Wen & Wei-Ying Chen (2011). Using multiple correspondence cluster analysis to map the competitive position of airlines. *Journal of Air Transport Management*, 17, 302-304.

Wetzel, R. (1995). Zur Visualisierung abstrakter Ähnlichkeitsbeziehungen. PhD Thesis, University of Bielefeld.

Whitman, J., & Ono, Y. (2017, to appear). Diachronic interpretations of word order parameter cohesion. In Truswell, Robert and Matthieu, Eric (eds). *Micro-change and Macro-change in Diachronic Syntax.* Oxford: Oxford University Press.

Williams, C. B. (1970). *Style and Vocabulary*, London: Griffin

山田文康, 西里静彦 (1993). 双対尺度法に関するいくつかの特性. *行動計量学*, *20*(1), 56-63.

Yamamoto, M. (2012). Clustering of functional data in a low-dimensional subspace. *Advances in Data Analysis and Classification*, *6*(3), 219-247.

Yamamoto, M., & Hwang, H. (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika*, *41*(1), 115-129.

柳井晴夫. (1994). 多変量データ解析法–理論と応用. 行動計量学シリーズ 朝倉書店.

柳田国男 (1930). 蝸牛考. 刀江書院.

安本美典 (1957). 宇治十帖の作者―文章心理学による作者推定. *文学・語学*, 第 4 号 三省堂.

安本美典 (1958). 文体統計による筆者推定–源氏物語・宇治十帖の作者について. *心理学評論*, *2*(1), 147-156.

安本美典, 野崎昭弘 (1976). 言語の数理. 筑摩書房.

安本美典, 本多正久 (1981). 因子分析法. 培風館.

吉野諒三, 千野直仁, 山岸侯彦. (2007). 数理心理学. 培風館.

Yule, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, *30*(3-4), 363-390.

Yule, G. U. (1944). *A statistical study of vocabulary*. Cambridge University Press.

Zwiener, I., Frisch, B., & Binder, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PloS one*, *9*(1), e85150.

謝辞

筆者が、この博士論文を擱筆するに至るまでには、多くの先生方のご指導、ご鞭撻を賜りました。先生方は、人柄はもとより学問においても非常に誠実な方々であり、これから御礼申し上げる方々なしに本博士論文は完成に至ることはなかったと心より感謝します。

まず、最初に筆者の指導教官である総合研究大学院大学及び統計数理研究所の吉野諒三教授に厚く御礼申し上げます。吉野教授は筆者が研究テーマを発見できず大変苦しんでいた時期に、筆者が研究者として活動するために大変心強い後押しをしてくださりました。吉野教授は人の悩みや苦しみに深く共感することができる大変人間味のある方であり、吉野教授を師と仰げたことは筆者にとり、大変名誉なことでした。本博士論文においても草稿の段階から、多大なお時間をご指導に割いていただき、博士論文の内容は勿論、筆者の拙い日本語や誤字脱字等を直していただき、感謝の言葉もありません。

次に、私の博士論文の主査である総合研究大学院大学及び統計数理研究所の前田忠彦准教授に深謝いたします。前田准教授はこの博士論文を草稿から完成に至るまでに、多くの時間を割いて、統計学的及び計量文献学的視点から、適切且つ丁寧なご指導をしていただき、心より感謝申し上げます。前田先生のご指導により、筆者自身自らの研究をより客観的に説得力のあるものとし自らの研究の枠組みを確固たるものとすることができました。

次に、東洋英和女学院大学名誉教授及び統計数理研究所客員教授の林文教授に御礼申し上げます。林文教授は数量化理論の創成期を知り、筆者の研究の初期から、吉野教授と共に相談に乗っていただき私の研究を後押ししていただきました。本博士論文では特に数量化Ⅲ類の適用に関して、その深い見識から適切なアドバイスを数多くいただき、筆者の数量化Ⅲ類の理解にあたって堅固な土台を与えてくださりました。また、本論文の草稿から完成に至るまで、私の拙文に丁寧なコメントと共にご指導いただき、ここに深く謝します。

次に、統計数理研究所名誉教授の村上征勝教授に心より感謝申し上げます。村上教授は、日本語における計量文献学的研究の先駆者であり、また源氏物語研究のパイオニアとして、本博士論文の特に 3 章について貴重なコメントをいただきました。村上教授から本論文が「既に発表された論文が用いたデータ」の再検討に終始している点について、「オリジナルなデータを集めていない」という観点から厳しいご指摘をいただきました。日本の計量文献学という学問

の開闢の時代から、常に先陣を切ってきた村上教授のご指摘は大変重いものがありました。本学で身につけた統計リテラシーと共に、人文科学への統計学の適用を、より豊穣な海とすべく、これから邁進していく所存であります。

次に、総合研究大学院大学及び統計数理研究所の足立淳准教授に、衷心より感謝申し上げます。足立准教授には階層及び非階層クラスター分析のみならず、Neighbor-Net の適用に関して、その背景にある分析の思想について、ご教示いただきました。足立准教授のご指摘によって、筆者の中で、今までのクラスター分析を中心に行っていた研究が確固たる思想をもとに有機的に結びつき研究の展望が一挙に広がりました。改めて御礼申し上げます。

次に、国立国語研究所名誉所員の米田正人先生と総合研究大学院大学及び統計数理研究所の朴堯星助教に感謝申し上げます。両先生は、本論文の草稿段階において言語学及び統計学的な視点からご指摘をいただき、両先生のご指摘により本稿は大幅に改善されました。

次に、国立国語研究所名誉教授の角田太作先生に御礼申し上げます。角田先生には、私の言語類型論に関する初期の研究に対して大変好意的なコメントをいただきました。角田先生の力強い後押しがなければ、本論文は完成に至りませんでした。

最後に、コーネル大学教授のホイットマン・ジョン教授に深謝いたします。歴史言語学のみならず、様々な言語学の分野における大家であるホイットマン教授と共同研究することができたことは、私にとって大変名誉なことでした。ホイットマン教授と議論する中で発展してきた「統計的言語類型論」や「方言データへの統計手法の応用」は、長い時間をかけ漸く実を結び、従来の「言語類型論」の常識を覆すような結果が得られようとしています。言語学と統計学との学際的な研究に携わる中で、筆者は、ホイットマン教授の持つ鋭い洞察力と慧眼、すなわち実質科学的な分野の達人の持つ深い考察に、ある種の畏怖の念を禁じざるを得ませんでした。改めて深謝すると共に、これからも何卒よろしくお願いいたします。

以上の諸先生のみならず、筆者は総合研究大学院大学在学中に、多くの方にお世話になりました。統計数理研究所の皆様、総合研究大学院大学複合科学研究科統計科学専攻の皆様にも大変お世話になりました。ここに感謝の意を記します。

最後に、私を支えてくれた父と母にこの論文を捧げます。