北海道言語文化研究　　北海道言語研究会
No. 18, 19-46, 2020.

北海道アイヌ語方言の分類再考*
－グラフ理論による方言研究の新展開:
人文学データの分析を代数構造から見直す－

小野　洋平

Observations on Lexicostatistical Classifications on Hokkaido Ainu Dialects
－A New Development on Dialectology, Graph-theoretic Approach from Algebraic Structure－

Yohei ONO

要旨：人文学データの分析では当該分野が１）現象のどの側面を重要な点と考え分析対象としているかを捉え、２）どのような記号の体系を規範としているか把握し、３）それらの記号により記述された情報をどのように分析するべきかとしているかを理解し、適切な統計手法を取捨選択することが必須である。今までの人文学データの分析では、上記の作業が十分でないことにより分析結果が当該分野の知見と整合性を示さない場合、示す場合、それぞれについて更なる追究ができない状況であった。本研究は Asai(1974: 92; Table 1)の代数構造にグラフ理論の適用が相応しいことを示し、方言学の分析目的に応じたグラフ理論の新しい応用方法を提案する。分析の結果、北海道アイヌ語方言には、東西型だけでなく ABA 型の情報が２番目以降に強い構造に存在することが、Asai(1974)のデータでは初めて示された。分析目的とデータの性質に応じた適切な統計手法の選択が人文学の更なる発展に寄与することを示した。

キーワード：北海道アイヌ語方言　基礎語彙統計学　ABA 型　グラフ理論　人文学データ

1. はじめに

本研究は Asai(1974)のアイヌ語諸方言に関する基礎語彙統計学による研究の再検討を通じ、人文学データの分析において、当該分野が１）現象のどの側面を重要な点と考えて分析対象としているかを捉え、２）どのような記号の体系を規範としているか把握し、３）それらの記号により記述された情報をどのように分析するべきかとしているかを理解した上で、統計学において１)、２)、３）の全てに対応した分析手法を適切に選択・開発することの重要性を例示し、今後の人文学と統計学の学際的研究の方向性を示すことを目的とする。

具体的には、Asai(1974)のデータの性質を検討することを通じ、先行研究の問題点を示し、基礎語彙統計学データへのグラフ理論の適用に本研究は至る。

分析の結果 Asai(1974)のデータには北海道アイヌ語方言に関して東西型に関する情報だけでなく、ABA 型の情報が存在することを示す。考察では、本研究の分析結果の方法論上及び言語学上の重要性について論じる。

まず、第 1 節では人文学データの特徴、人文学データに統計手法を適用する意義、人文学データの分析における今日までの問題を述べ、本研究の意義を位置付ける。

1.1. 人文学データの特徴

人文学データは以下の点に関して分野ごとの差が大きい。第一に、現象のどの側面を重要な点と考えて分析対象とするか自体、自然科学や社会科学と同様に、人文学においても分野による違いが大きい。第二に、分析対象をどのような記号を用いてどのように記述するべきか、つまり専門分野ごとに規範とされる記号の体系が異なる。第三に、それらの記号により記述された情報をどのように分析するかに関して、専門分野内においても、方向性が大きく異なっている。自然科学や社会科学のデータと比較して上記の三点のより正確な理解が必須であることが人文学データの特徴と考えられる。

1.2. 人文学データに統計手法を適用する意義

人文学データに統計学の諸手法を適用する意義は何か。それは現象の普遍性（ないし類型性）と特殊性を追究する人文学の目的に、記述統計学を中心に発展してきた統計学の諸手法が寄与する可能性があることと著者は考える[1]。

人文学の諸分野は現象の普遍性（ないし類型性）と特殊性を様々な方法を用いて追究している。しかし、現象が普遍性（ないし類型性）を有すると認識するには、現象が特殊性を有することを認識できなければならない。逆に、現象が特殊性を有すると認識するには、現象が普遍性（ないし類型性）を有すると認識できなければならない。普遍性（ないし類型性）の理解と特殊性の理解は表と裏の関係にある。

統計学は、手元にあるデータを生み出した仕組みに関心がある推測統計学と、手元にあるデータを分析目的に応じて如何に記述するかに関心がある記述統計学の 2 つに大別されるが、記述統計学を中心に発展してきた諸手法は人文学データの類型性を把握することに適していると著者は考える。何故ならば、人文学のデータとして記述可能な部分な点ですら様々な特徴を有する多変量のデータであり、人文学データの直感的な把握は一般に人間の認知能力では困難である。例えば、一次元は直線、二次元は平面、三次元は空間として人間は把握することが可能であるが、四次元以上の空間（データ）の把握は一般には困難であり、特殊な訓練を要する。記述統計学を中心に発展してきた諸手法は多変量データの直感的把握を容易

[1] 「統計学」と言っても、確率分布を導入して普遍性を追究する推測統計学と、確率分布の導入を抑制し類型性を記述する記述統計学の区別が大雑把にはあり、その中では様々な思想が発生し、分岐し、融合し、時として対立している。近年、ビックデータとの関連で統計学が注目されているが、それらの中で語られ消費されている「統計学」は、統計学の教科書的な一側面、特に推測統計学の一側面に著者には見える。本研究で示すように、人文学データの分析においては、記録ないしデータに存在する情報を如何に適切な枠組みで扱うかという記述統計学の面で、課題が山積している。近年、データサイエンスが盛んになっているが、本研究の意義は人文学のデータサイエンスのあり方について追究する点にもあるといえよう。

にし、人文学データに存在する類型性（パターン）を明らかにすることを通じて、間接的に特殊性の把握に寄与すると著者は考える。但し、次節で述べる通り人文学の分野でも、推測統計学の考え方を導入し、人文学データの普遍性を追究することが適切な領域も存在する。

1.3. 人文学データの分析における今日までの問題

人文学データを統計学の点から分析する際には、後述する様に、統計学の考え方の理解が欠かせない。推測統計学と記述統計学に関して述べるならば、著者は人文学データの分析においては、分析対象となる現象に応じ推測統計学の考え方と記述統計学の考え方を意識的に使い分ける能力が分析者に要求されると考えており、その実行に伴う困難が今日まで続いている人文学データの分析の混乱を引き起こしている原因の一つであるように思われる。

例えば、著者の研究分野の一つである計量文献学では、作者を特定する際の文体の特徴として「読点が打たれる文字」や「助詞や助動詞」の使用率が用いられる。このような分析の背景には、例えば夏目漱石が「吾輩は猫である」を執筆した時期に全く別の作品 A を書いたとしても、「読点が打たれる文字」や「助詞や助動詞」の使用率は、個々の名詞や動詞など作品の内容に関わる言葉の使用率とは異なり、作品 A と「吾輩は猫である」においてそれほど変わらないという考え方がある。作家が異なる作品を繰り返し書いたとしても「読点が打たれる文字」や「助詞や助動詞」の使用率が同じような傾向を見せるという考え方は、確率分布の導入に繋がるものである。よって、計量文献学では推測統計学の考え方が欠かせない。

一方で、著者の研究分野の一つである統計学による言語類型論データの分析では、分析に確率分布を用いるのが適切かは立場が分かれる。言語変化のような歴史上の一回きりの現象に確率分布のような「繰り返し」を前提とする概念を導入することが適切なのかという問題だけでなく、確率分布が言語類型論データの特徴を適切に捉えることができるか、分析者によって立場が分かれる。Ono (2020a, to appear)では言語類型論データの「欠損値」の観点から確率分布の導入の妥当性について論じている。

推測統計学と記述統計学のどちらの考え方を適用するかだけでも、分析の際の統計手法の取捨選択は異なる。加えて、統計分析の諸手法はさらに細分化され、それぞれ異なった性質のデータの存在を前提として展開されている。一般的な教科書で紹介されている回帰分析、判別分析、因子分析、主成分分析、多次元尺度法などですら明示的には記述されていない様々な前提が存在しており、適用手法の前提と当該分野におけるデータの前提が対応していない場合、それらの分析の実行は可能であるが分析結果が（当該分野にとって）意味のあるものになる保証はない。

この考えは統計学に携わる者には自明のことであろうが実践には常に困難が伴う。人文学データの分析では人文学の当該分野が１）現象のどの側面を重要な点と考えて分析対象としているかを捉え、２）どのような記号の体系を規範としているか把握し、３）それらの記号により記述された情報をどのように分析するべきかとしているか、を分析者は理解した上で統計学において１）、２）、３）の全てに対応した手法を適切に選択することが必須である。人文学データの分析でも場合により対応する方法が存在しないことがある。その場合独自の

統計手法を開発する必要があり、統計学の理論の発展に今後更に寄与するものと考えられる。

統計手法の選択に関する吟味が十分でない場合、以下のような混乱が生じる。まず、分析結果が当該分野の知見と整合性を有していない場合、分析結果は統計手法の問題によるものなのか、データの問題によるものなのか、それとも実は当該分野に関して新しい知見を与え得るものなのか等について更なる追究ができない状態となる。次に、分析結果が当該分野の知見と整合性を有している場合においても、分析結果がなぜ整合性を有しているのか更なる追究ができないため、分析結果を支持する当該分野の知見のみをより合わせた解釈を与えるに留まり、既存の研究の追認に終わることが多い。

特に、数学の使用の有無に関し、人文学と統計学はしばしば文理の両極に位置付けられる。また、現在の教育制度では高等学校の段階から、いわゆる「文系」と「理系」の選択がなされており、数学ないし統計学と人文学の両方に通じた人材を輩出する制度が整っていない。こうした教育制度上の問題から、人文学データにおける統計手法の選択の吟味を行える人材が恒常的に不足し、人文学データの分析における上記の問題は今日まで繰り返し生み出されている。

1.4. 本研究の意義

本研究は、Asai(1974)の、アイヌ語諸方言に関する基礎語彙統計学による研究の再検討を通じ、上記で述べた人文学データの分析における問題が実際の研究でどのような形で現れるかを例示し、基礎語彙統計学のデータの特徴と分析目的に応じた適切な統計手法を選択することが、アイヌ語方言学にどのような知見を与え得るかを示す。

基礎語彙統計学のデータの特徴と分析目的にグラフ理論が対応することが明らかにされ、グラフ理論による方言研究の新展開の可能性が示される。

基礎語彙統計学データという具体例の検討を通じ、人文学データの分析における今日までの問題への取り組み方を示すことによって、人文学と統計学の学際的研究の今後の方向性を提示することに本研究の意義がある。

2. データの性質と方法

2 節では、2.1 節で本研究が取り上げる Asai(1974)のアイヌ語研究における位置付けを述べ、2.2 節で Asai(1974: 92; Table 1)のデータの性質を取り上げ、2.3 節で Asai(1974: 92; Table 1)のデータは負の数を考えることが言語学上意味のない特殊なデータであることを示し、Asai(1974: 92; Table 1)の性質は古典的な統計手法の前提に反することを述べる。2.4 節では、グラフ理論の適用が Asai(1974: 92; Table 1)の分析に相応しいことを述べ、Mcut 法と Ncut 法による分析を提案するとともに、方言学データに統計手法を適用する目的を考慮した 2 つの手法の応用法を提案する。本研究では、アイヌ語方言学の分類でも特に北海道アイヌ語方言の分類に焦点を当て Asai(1974: 92; Table 1)の北海道アイヌ語方言に 2 つの分析を適用する。Asai(1974: 92; Table 1)への分析結果と考察については Ono(2020b, to appear)を参照されたい。

2.1. アイヌ語研究における Asai(1974)の位置付け

金田一京助(1882-1971)以来、日本におけるアイヌ語研究はユーカラを中心とした口承文芸のアイヌ語に注意を向けていたが、服部四郎(1908-1995)はアイヌ語の日常語に関する調査に力を注ぎ、Swadesh(1955)の基礎語彙に関する服部・知里(1960)や調査対象とする方言を絞りより広範囲の語彙を調べたアイヌ語方言辞典(服部編 1964)が公刊された(切替 2010)。

Asai(1974)は、図 1 の 1 から 19 までの地域のアイヌ語諸方言の基礎語彙統計調査の結果をまとめた服部・知里(1960)をもとに、服部編(1964)と帯広、釧路、美幌方言についてそれぞれのインフォーマントに行った独自の調査を基に服部・知里(1960)の資料の修正を行い、千歳方言についてはインフォーマントから新たな調査を行い、更に北千島方言について鳥居(1903)、村山(1971)、そして Pinart の資料(Asai 1974: Appendix)を参照しながらデータを整理している。

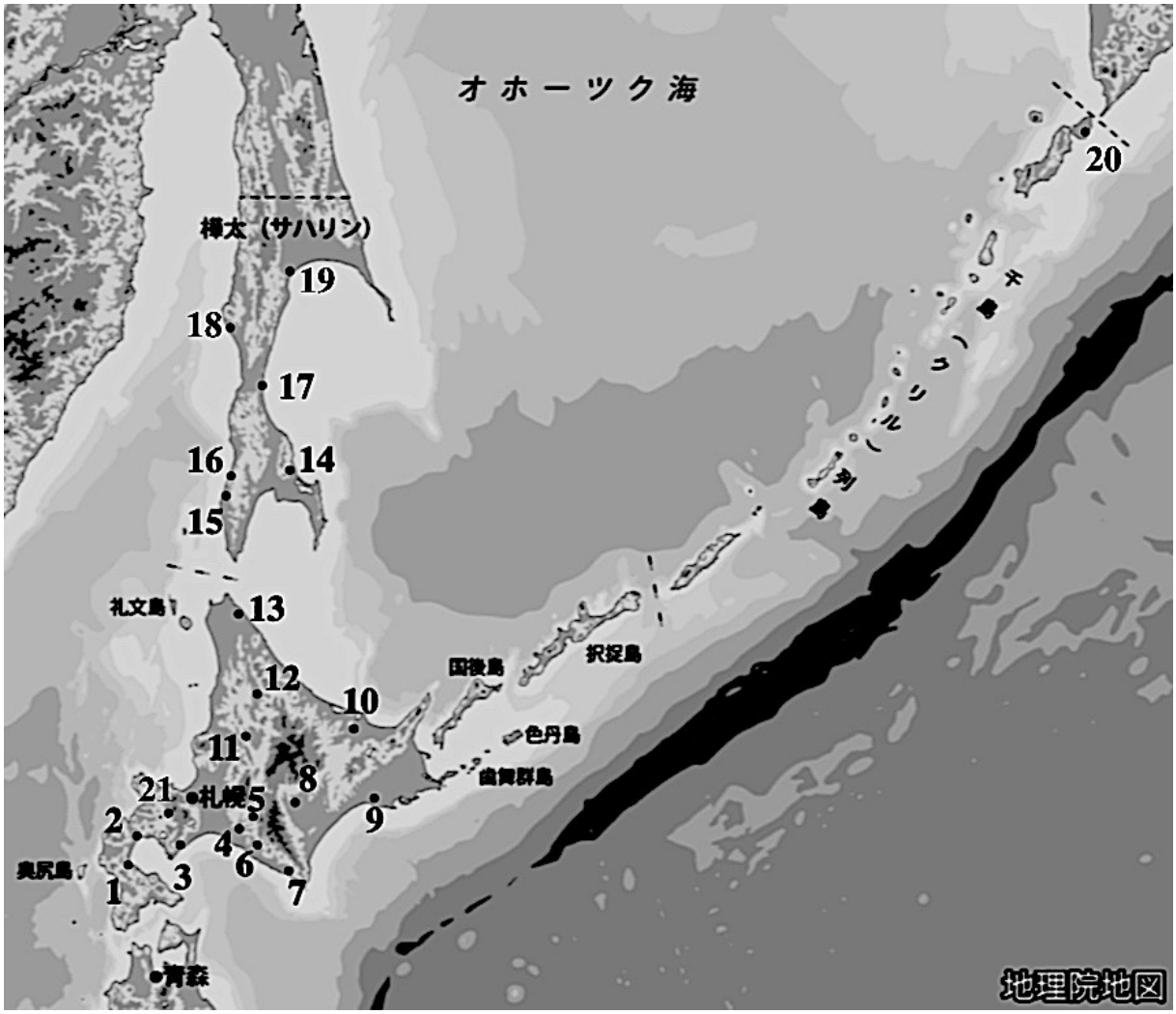

図 1. Asai(1974)の分析したアイヌ語諸方言一覧. 1: 八雲, 2: 長万部, 3: 幌別, 4: 平取, 5: 貫気別, 6: 新冠, 7: 様似, 8: 帯広, 9: 釧路, 10: 美幌, 11: 旭川, 12: 名寄, 13: 宗谷, 14: 落帆, 15: 多蘭泊, 16: 真岡, 17: 白浦, 18: ライチシカ, 19: 内路; 20: 北千島(占守), 21: 千歳. 図は国土交通省国土地理院(2019)を基に著者が編集した.

アイヌ語に関する資料保存の観点からも Asai(1974)は貴重な研究であるが、更に Asai(1974)は Swadesh(1955)の 200 語を参照し、上記の資料の中から 110 語を選び出し各語に関して 21 のアイヌ語方言の間の同根性判断を行い、得られた方言間の類似度にクラスター分析を適用している。

アイヌ語方言の分類という観点からの Asai(1974)の貢献は、アイヌ語の専門家として語の同根性に基づき北海道アイヌ語方言、樺太アイヌ語方言、北千島アイヌ語方言の類似度を示したことと、クラスター分析の結果から、北海道アイヌ語方言(図 1 の 1 から 13 と 21)、樺太アイヌ語方言(図 1 の 14 から 19)、北千島アイヌ語方言(図 1 の 20)のいわゆる「三大分類」を確立し、北海道アイヌ語方言を南西(図 1 の 1 から 6 と 21)と北東(図 1 の 7 から 12)に分ける分類を示したことであろう。

表 1 は Asai(1974: 92; Table 1)の類似度のデータを日本語に著者が直したものである。例えば、八雲と長万部では 110 語の中で 104 語において同根と考えられる語形を少なくとも一つ共有していることを意味する[2]。

表 1. 21 方言の類似度を Asai(1974: 92; Table 1)から日本語に著者が直したもの．列の番号は行と対応する．

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_八雲	110	104	102	97	96	95	83	88	88	83	94	89	77	40	52	43	39	42	43	54	94
2_長万部	104	110	101	95	94	93	80	84	84	78	92	84	77	37	50	39	36	41	42	50	95
3_幌別	102	101	110	104	98	98	83	91	90	85	96	93	76	41	56	45	41	44	46	59	99
4_平取	97	95	104	110	103	105	79	87	86	81	92	89	73	42	58	47	42	45	48	59	104
5_貫気別	96	94	98	103	110	98	75	81	81	74	87	83	72	38	51	41	36	40	41	54	101
6_新冠	95	93	98	105	98	110	73	84	84	78	89	83	70	40	54	43	39	42	45	56	101
7_様似	83	80	83	79	75	73	110	94	92	84	85	81	74	33	49	36	33	33	38	52	80
8_帯広	88	84	91	87	81	84	94	110	106	100	95	93	79	38	54	43	38	31	46	59	84
9_釧路	88	84	90	86	81	84	92	106	110	101	98	98	83	38	54	42	38	41	45	63	85
10_美幌	83	78	85	81	74	78	84	100	101	110	92	91	78	36	52	42	36	39	43	58	79
11_旭川	94	92	96	92	87	89	85	95	98	92	110	102	82	43	58	47	40	45	47	60	93
12_名寄	89	84	93	89	83	83	81	93	98	91	102	110	84	44	59	51	42	48	49	61	87
13_宗谷	77	77	76	73	72	70	74	79	83	78	82	84	110	47	65	56	48	55	52	51	72
14_落帆	40	37	41	42	38	40	33	38	38	36	43	44	47	110	66	90	91	84	60	29	40
15_多蘭泊	52	50	56	58	51	54	49	54	54	52	58	59	65	66	110	78	67	67	74	44	55
16_真岡	43	39	45	47	41	43	36	43	42	42	47	51	56	90	78	110	92	87	68	32	46
17_白浦	39	36	41	42	36	39	33	38	38	36	40	42	48	91	67	92	110	93	68	27	40
18_ライチシカ	42	41	44	45	40	42	33	31	41	39	45	48	55	84	67	87	93	110	62	33	41
19_内路	43	42	46	48	41	45	38	46	45	43	47	49	52	60	74	68	68	62	110	38	39
20_北千島	54	50	59	59	54	56	52	59	63	58	60	61	51	29	44	32	27	33	38	110	54
21_千歳	94	95	99	104	101	101	80	84	85	79	93	87	72	40	55	46	40	41	39	54	110

しかし、Asai(1974: 92; Table 1)のデータには以下の二点に関して問題があった。第一に、Asai(1974: 64-93)の記述からは Asai(1974)は 202 語の語の中から 67 語は分析の対象外としたことを明記しているが、残りの 135 語の中から 110 語をどのように選択したかに関して明確な記述がない。このことにより、管見の限り、今日まで Asai(1974)が分析に用いた 110 語の特定ができていない。第二に、110 語の語形に関する同根性判断に関しても、Asai(1974)が、どのような基準を採用したかについて、管見の限り、今日まで特定できていない。

これらの問題により、服部・知里(1960)のアイヌ語方言の分類と Asai(1974)のアイヌ語方言の分類を、現代のアイヌ語学の観点から比較、検証、統合することが今日まで困難となり、

[2] 表 1 の具体的な作成方法については 2.2 節で述べる。

アイヌ語方言の基礎語彙統計学による分類の研究における足枷となっていた。

Ono(2020c, to appear)は Asai(1974)のデータの特徴を利用することで、Asai(1974)が用いた 110 語の中の一部を特定し、Asai(1974)が採用した同根性判断についても。一部を特定した。服部・知里(1960)と Asai(1974)では同根性判断が大きく異なることも Ono(2020c, to appear)では示している。

よって、Asai(1974)が採用した 110 語と同根性判断の特定については Ono(2020c, to appear)に詳細を譲り、本稿では Asai(1974: 92; Table 1)のデータの性質を詳述することとする。

2.2. Asai(1974: 92; Table 1)のデータの性質

表 2 は Asai(1974: 68)の“many”に関する各地点の語形である。但し Ono(2020c, to appear)と同様に、服部・知里(1960)および服部編(1964)に基づき poronno と poroonno のデータを修正してある。また表 3 は Asai(1974: 68)の“who”に関する各地点の語形である。Asai(1974: 67)は“The following list of basic vocabulary is taken from Hattori’s article. Every entry is headed by the same word guide number and the English word as Hattori used, followed by the corresponding Ainu forms. Each of them is then followed by the place or dialect number in parentheses where the same form is found.”と述べており、例えば poronno(1, 3-13, 15, 16, 18, 19, 21)は“many”という語に関して、図 1 の 1(八雲)、3(幌別)、4(平取)、5(貫気別)、6(新冠)、7(様似)、8(帯広)、9(釧路)、10(美幌)、11(旭川)、12(名寄)、13(宗谷)、15(多蘭泊)、16(真岡)、18(ライチシカ)、19(内路)、21(千歳)が同じ poronno という語形を有していることを意味する。

また、Asai(1974: 67)は“Obviously similar or nearly similar Ainu forms are also put in parentheses, but considerably many forms which may be easily taken as cognate are not placed in parentheses.”と述べており、表 2 の場合 poronno(1, 3-13, 15, 16, 18, 19, 21)と poroonno(2, 14, 17)は括弧の中に表記されており、“Obviously similar or nearly similar Ainu forms”と Asai(1974)が判断したことがわかる。

ただし、Asai(1974: 92; Table 1)を計算する際に Asai(1974: 85)は“identifying only the same or nearly the same forms”としたと述べているが、どのような語形が“the same or nearly the same forms”であるかについてその基準を Asai(1974)は明確に述べていない。Ono(2020c, to appear)では、Asai(1974)が括弧の中に入れた語形を区別している例を発見しているため、以降本稿では括弧の中の語を区別し Asai(1974: 92; Table 1)のデータの性質を説明する。

表 2. “many”の 21 方言に関する語形．Asai(1974: 68)より．服部・知里(1960)および服部編(1964)に基づき poronno と poroonno のデータを修正．

((poronno (1, 3–13, 15, 16, 18, 19, 21), poroonno(2, 14, 17)),
inne(21),
okajno(15, 18, 19),
renkajne(15, 16, 18),
tumanpiki(20).

表 3. “who”の 21 方言に関する語形．Asai(1974: 68)より．

nen(1-3, 7-13),
hunna(4-6, 21),
hunat(20),
naata(14-19).

表 4 は poronno, poroonno, inne, okajno, renkajne, tumanpiki の 6 つの語形が全て同根でないと仮定し、各方言同士が共有する「語形の数」を類似度として数えたものである。表 2 と表 4 から、15 の多蘭泊方言と 18 のライチシカ方言は poronno, okajno, renkajne の 3 つの語形を、15 の多蘭泊方言と 19 の内路方言は poronno, okajno の 2 つの語形を、15 の多蘭泊方言と 16 の真岡方言は poronno, renkajne の 2 つの語形を、16 の真岡方言と 18 のライチシカ方言は poronno, renkajne の 2 つの語形を共有しており、これら 4 つの方言は“many”という一つの語に関して「語形の数」が多いことで他の方言との類似度が大きくなっている。

よって、方言データから方言同士の類似度を計算する場合には上記の各方言の「語形の数」による影響を除く必要がある。

Asai(1974)は分析単位を「語」とし、「各語に関して方言同士が少なくとも一つ同根語形を共有しているか」という同根語形の共有の有無に着目し、“relation index”を導入した。Asai(1974: 61-62)は“relation index”について“we assume that the value of the relation index is 1, if there is at least one similar or the same form in both of any two given dialect (P_i, P_j) and if there is no common form in the two given dialects, the relation index of the two dialects equals a value of 0.”と述べている。この方法は言語年代学の残存語の有無を測る方法と対応している。

Asai(1974: 49-53)では言語年代学における残存語の有無を言語間の距離として用いることを構想しており、Asai(1974)の分析法はその考えを反映していると考えられる。

表 4. 表 2 について各方言が共有している「語形の数」を数えたデータ.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_八雲	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
2_長万部	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
3_幌別	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
4_平取	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
5_貫気別	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
6_新冠	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
7_様似	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
8_帯広	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
9_釧路	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
10_美幌	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
11_旭川	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
12_名寄	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
13_宗谷	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
14_落帆	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
15_多蘭泊	1	0	1	1	1	1	1	1	1	1	1	1	1	0	3	2	0	3	2	0	1
16_真岡	1	0	1	1	1	1	1	1	1	1	1	1	1	0	2	2	0	2	1	0	1
17_白浦	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
18_ライチシカ	1	0	1	1	1	1	1	1	1	1	1	1	1	0	3	2	0	3	2	0	1
19_内路	1	0	1	1	1	1	1	1	1	1	1	1	1	0	2	1	0	2	2	0	1
20_北千島	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21_千歳	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	2

表 5. 表 4 を“relation index”の点から計算し直したデータ.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_八雲	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
2_長万部	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
3_幌別	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
4_平取	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
5_貫気別	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
6_新冠	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
7_様似	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
8_帯広	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
9_釧路	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
10_美幌	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
11_旭川	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
12_名寄	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
13_宗谷	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
14_落帆	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
15_多蘭泊	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
16_真岡	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
17_白浦	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
18_ライチシカ	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
19_内路	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1
20_北千島	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21_千歳	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1

本研究では“relation index”の定義を「一つの語に関して方言同士が少なくとも一つ同根語形を共有していれば“relation index”の値は 1 であり、一つの語に関して方言同士が全く同根語形を共有していない場合は“relation index”の値は 0」と訳し、この日本語の定義を適宜参照する。但し、Asai(1974: 61-62)の“relation index”は、各語に関して各方言が少なくとも 1 つの語形を有していることを“relation index”の前提とし、Asai(1974)はこの前提を満たす語を分析の対象としている。

表 4 を“relation index”の点から再計算した結果が表 5 である。例えば、15 の多蘭泊方言と 18 のライチシカ方言は poronno, okajno, renkajne という 3 つの語形を共有しているが“relation index”の定義より、2 つの方言の類似度は 1 となる。

表 6. 表 3 を“relation index”の点から計算し直したデータ.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1_八雲	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
2_長万部	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
3_幌別	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
4_平取	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
5_貫気別	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6_新冠	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7_様似	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
8_帯広	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
9_釧路	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
10_美幌	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
11_旭川	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
12_名寄	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
13_宗谷	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
14_落帆	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
15_多蘭泊	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
16_真岡	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
17_白浦	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
18_ライチシカ	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
19_内路	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0
20_北千島	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
21_千歳	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

表 3 について“relation index”の点から類似度を計算した結果が表 6 である。このように各語について表 4 と同様の計算を行うことで、類似度に関する表を 110 語に関してそれぞれ得ることができる。それらの 110 個の表を、各方言の組み合わせに関し合計したデータが表 1 である。

Asai(1974)は“relation index”を導入することで、方言の「語形の数」による方言間の類似度への影響を避けることにある程度成功した。しかし、次節で説明するように、“relation index”を導入したことよって、表 1 は特殊な代数構造を有している。

次節では、古典的な統計手法の前提と表 1 の特殊な代数構造が異なることが、表 1 を適切に分析することを困難にしていることを示す。

2.3. Asai(1974: 92; Table 1)の特殊な代数構造

2.3.1. 負の数を Asai(1974: 92; Table 1)からは定義できない

Asai(1974: 92; Table 1)は、“relation index”(Asai 1974: 61-62)によって計算されていた。再度、その定義を以下に記す。

「一つの語に関して方言同士が少なくとも一つ同根語形を共有していれば“relation index”の値は 1 であり、一つの語に関して方言同士が全く同根語形を共有していない場合は“relation index”の値は 0」。

上記の“relation index”(Asai 1974: 61-62)から我々は「-1」という数を言語学上意味のある数として定義することができるだろうか。実はできないのである。

Asai(1974: 92; Table 1)のデータでは、各語に関して各方言が少なくとも 1 つの語形を有すという“relation index”の前提が成立しているため、「1」の定義である「(方言同士が)少なくとも一つ同根語形を共有していること」の論理上の否定は「(方言同士が)全く同根語形を共有していないこと」となり、「0」の定義となる。逆に「0」の定義である「(方言同士が)全く同根語形を共有していないこと」の論理上の否定は「(方言同士が)少なくとも一つ同根語形を共有していること」となり「1」の定義となる。

更に、「1」の定義である「(方言同士が)少なくとも一つ同根語形を共有していること」と、「0」の定義である「(方言同士が)全く同根語形を共有していないこと」は一つの語に関する方言同士の同根語形に関する判断の全ての状態を記述している。

論理学におけるベン図を想像すれば、一つの語に関する方言同士の同根語形に関する判断は「1」の領域とそれ以外の「0」の領域で全て記述できることがわかる。

また、我々が扱う通常の数は具体的な「量」を測っているため単位が同じ「量」に引き算を適用すること(負の数を用いること)が意味のある操作となる場合が多い。

しかし、Asai(1974: 92; Table 1)の「1」と「0」は「状態」によって定義されているため、Asai(1974: 92; Table 1)に引き算を適用すること(負の数を用いること)が意味のある操作となるには、「-1」を「1」と「0」の定義と同様に「状態」として定義する必要がある。

そのため「-1」には、“relation index”(Asai 1974: 61-62)から得られる「1」に「-1」を足すと“relation index”(Asai 1974: 61-62)から得られる「0」になることが要請される。

例えば「1」の定義を「1 語で(方言同士が)少なくとも一つ同根語形を共有していること」として「-1」を「-1 語で(方言同士が)少なくとも一つ同根語形を共有していること」と定義した場合、「0」の定義は「0 語で(方言同士が)少なくとも一つ同根語形を共有していること」となるが、これは「1 語で(方言同士が)全く同根語形を共有していないこと」という「0」の定義と矛盾する。

上記の議論より「1」と「0」の定義と矛盾しない形で、「-1」を「状態」として、定義することは不可能である。表 1 のデータの性質は、代数学における加法に関する可換モノイド(commutative monoid)に対応すると考えられる。表 1 のデータを加法に関する可換モノイドとして捉えたときに有効な操作は「足し算」とそれらの値の「比較」である。次節では表 1 のデータが加法に関する可換モノイドと考えられることを説明する。

2.3.2. Asai(1974: 92; Table 1)は加法に関する可換モノイドと考えられる

Gondran and Minoux (2008: 1-13)によれば可換モノイドは 1)結合則と可換則が成立する 1 つの二項演算(a single associative and commutative binary operation)、2)単位元ないし中立元(neutral element)、3)順序単位(order unit)、4)代数的前順序(an algebraic preorder relation)を 4 つの性質を備えるが、5)逆元が存在しない。

上記の性質のうち最初の 3 つは Asai(1974)が 1) Asai(1974: 92; Table 1)を作成するために表 5 や表 6 のデータの間に加法の適用を認めていること 2) 「0」を「一つの語に関して方言同士が全く同根語形を共有していない場合」と定義していること 3)「1」を「一つの語に関して方言同士が少なくとも一つ同根語形を共有していること」と定義していることに対応する。

また表 1 のデータ同士に加法を適用し、その値を比較することが言語学上意味のある操作と Asai(1974)が考えている(認めている)ことが 4 番目の代数的前順序の性質に対応する。

最後に、2.3.1 節で述べたように、Asai(1974: 61-62)の“relation index”の定義からは Asai(1974: 92; Table 1)において「-1」を言語学上意味のある数として定義できないことが、5 番目の逆元(負の数)が存在しない性質に対応する。次節では、可換モノイドの「逆元が存在しない」という性質によって、表 1 の分析の際にどのような問題が生じるかについて述べる[3]。

2.3.3. Asai(1974: 92; Table 1)には古典的な統計手法が適用できない

2.3.1 節の議論で、Asai(1974: 61-62)の“relation index”から「-1」を「1」と「0」の定義と、矛盾なく定義することが不可能であることを示した。このことは表 1 のデータを負の数として扱う如何なる統計手法も、実行することは可能であるが、言語学上意味のある分析となる保証が全くないことを意味する。

なぜならば、表 1 のデータは“relation index”の定義に基づいた方言同士の「類似度」の情報のみを有しており、「非類似度」の情報を有していない。

上記の性質は古典的な統計手法の前提に反する。まず方言間の類似度を示す表 1 のデータに対して一般的に適用される多次元尺度構成法や対応分析は内積モデル(inner scalar product model)を利用している。内積モデルは表 1 の *n* 番目の方言と *m* 番目の方言の類似度を S*nm* と

[3] 但し、基礎語彙統計の対象となった言語に関する調査状況や記述的研究の状況、特に同根性判断のレベルによっては、加法が成立するとは言い切れないデータも考えられる。この問題については結語で触れる。

表記した場合、*n* 番目の方言と *m* 番目の方言の距離(非類似度)を D*nm* と表記すると、(D*nm*)^2=S*nn*+S*mm*-2S*nm* と計算するモデルである。(D*nm*)^2 は D*nm* の二乗を示す。但し、実際の分析では S*nn*, S*mm*, S*nm* には様々な前処理が施される。

注目すべきは上式の右辺で-2S*nm* という負の数が用いられていることである。上述のように Asai(1974: 61-62)の“relation index”からは「-1」を矛盾なく定義することができず、表 1 のデータを負の数として扱う如何なる統計手法は(実行することは可能であるが)言語学上意味のある分析となる保証が全くない。つまり、内積モデルを前提とした統計手法も表 1 に実行することは可能でも言語学上意味のある分析となる保証はない。

次に多くの古典的な統計手法では類似度の情報が非類似度に関する順序の情報を保存していると仮定し分析を行っている。この仮定は、S*ij*<S*kl*⇔D*ij*>D*kl* と式で表すことができる。

この式は非計量多次元尺度法などが前提としている。非計量多次元尺度法は Kruskal(1964a; 1964b)によって発展し、近年では Dyen, Kruskal, and Black (1992)により印欧語族の語彙データに対して適用されている手法である。

しかしながら、上述のように Asai(1974: 61-62)の“relation index”に基づく表 1 は、方言同士の「類似度」の情報のみを有しており、「非類似度」の順序に関する情報すら保存されている保証はない。そのため、類似度の情報が非類似度に関する順序の情報を保存していると仮定する非計量多次元尺度法も、表 1 に適用することは可能でも言語学上意味のある分析となる保証はない。

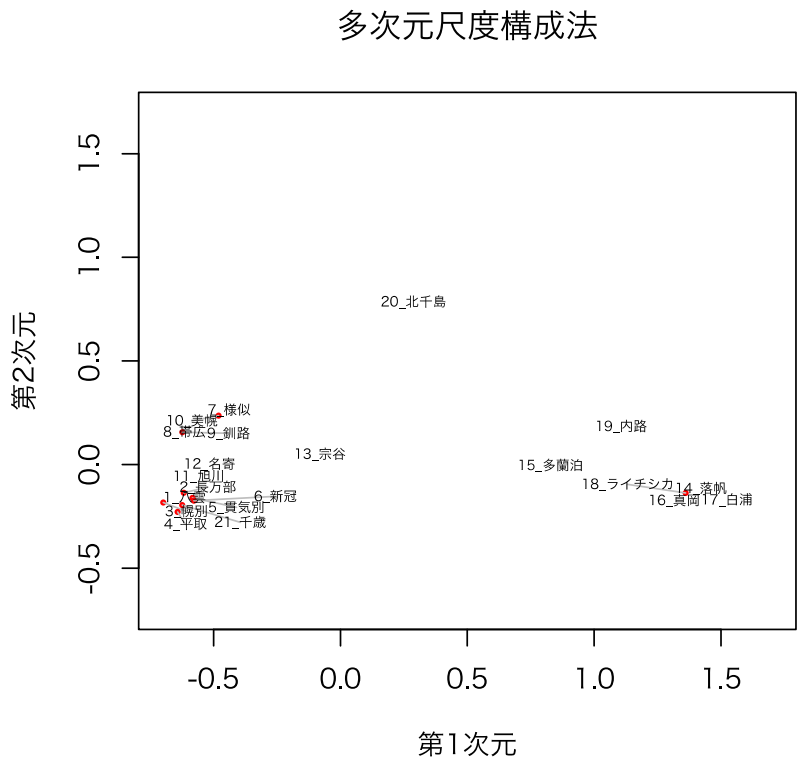

図 2. 表 1 に多次元尺度構成法を適用した結果.

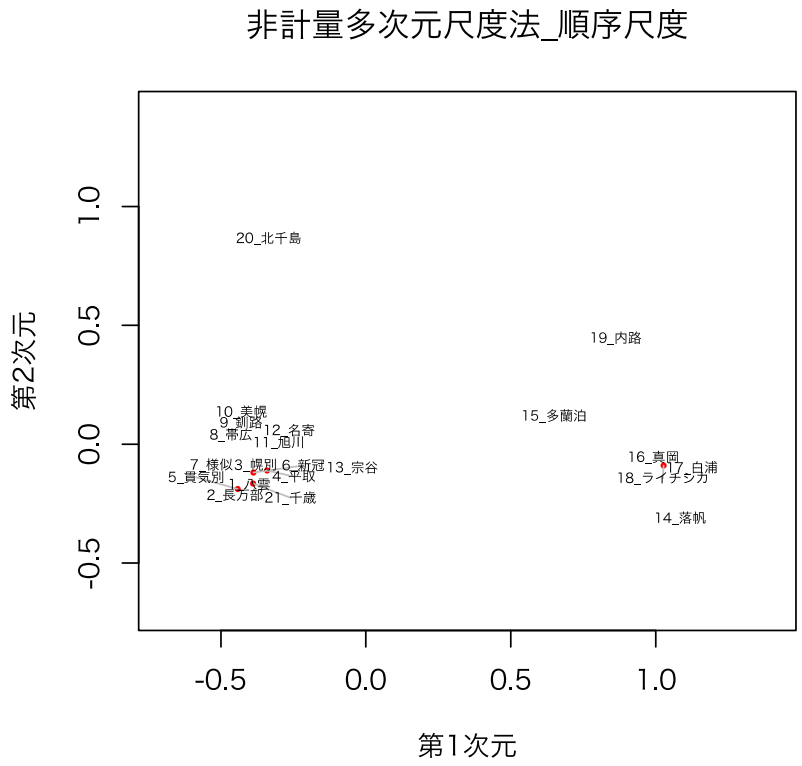

図 3. 表 1 に非計量多次元尺度構成法(順序尺度)を適用した結果.

表 1 に多次元尺度法と非計量多次元尺度法を適用した結果を図 2、図 3 にそれぞれ示す[4]。図 2、図 3 は分析結果の第 1 次元と第 2 次元をプロットしたものである。

注記する点は、図 2、図 3 では北千島アイヌ語方言の北海道アイヌ語方言と樺太アイヌ語方言に対する位置が大きく異なる点である。今までの人文学データの分析では、1 節で述べたように統計手法の選択に関する吟味が十分ではなかったため、図 2、図 3 の結果に関する更なる追究ができなかった。

しかし、既に本節の議論から、図 2、図 3 の分析は何も表 1 の分析に相応しくないことが示されている。次節では表 1 のデータの分析に相応しい手法としてグラフ理論を導入する。北千島方言に関するグラフ理論に基づいた分析は Ono(2020b, to appear)を参照されたい。

2.4. グラフ理論の導入

各方言をグラフの頂点とし、各語において、方言同士に同根語形がある場合を頂点同士が繋がっている(1)とし、方言同士に同根語形が全くない場合を頂点同士が繋がっていない(0)とした場合、出来上がったグラフは、Asai(1974: 61-62)の“relation index”に基づいたデータを

[4] 分析には R 言語(R Core Team 2018)を用い、多次元尺度法の実行には cmdscale コマンドを、非計量多次元尺度法の実行には smacof パッケージ(de Leeuw and Mair 2009)の mds コマンドを、図の作成には wordcloud パッケージ(Fellows 2018)の textplot コマンドをそれぞれ用いた。総語数(110)から表 1 の値を引き、総語数で割ったデータを入力データとして用いた。分析の次元は 3 次元に設定した。

正確に表現している。但し、以降のグラフでは各方言が各方言自身と同根語を共有していること、つまり各方言が各方言自身と繋がっていることは図の上では省略する。

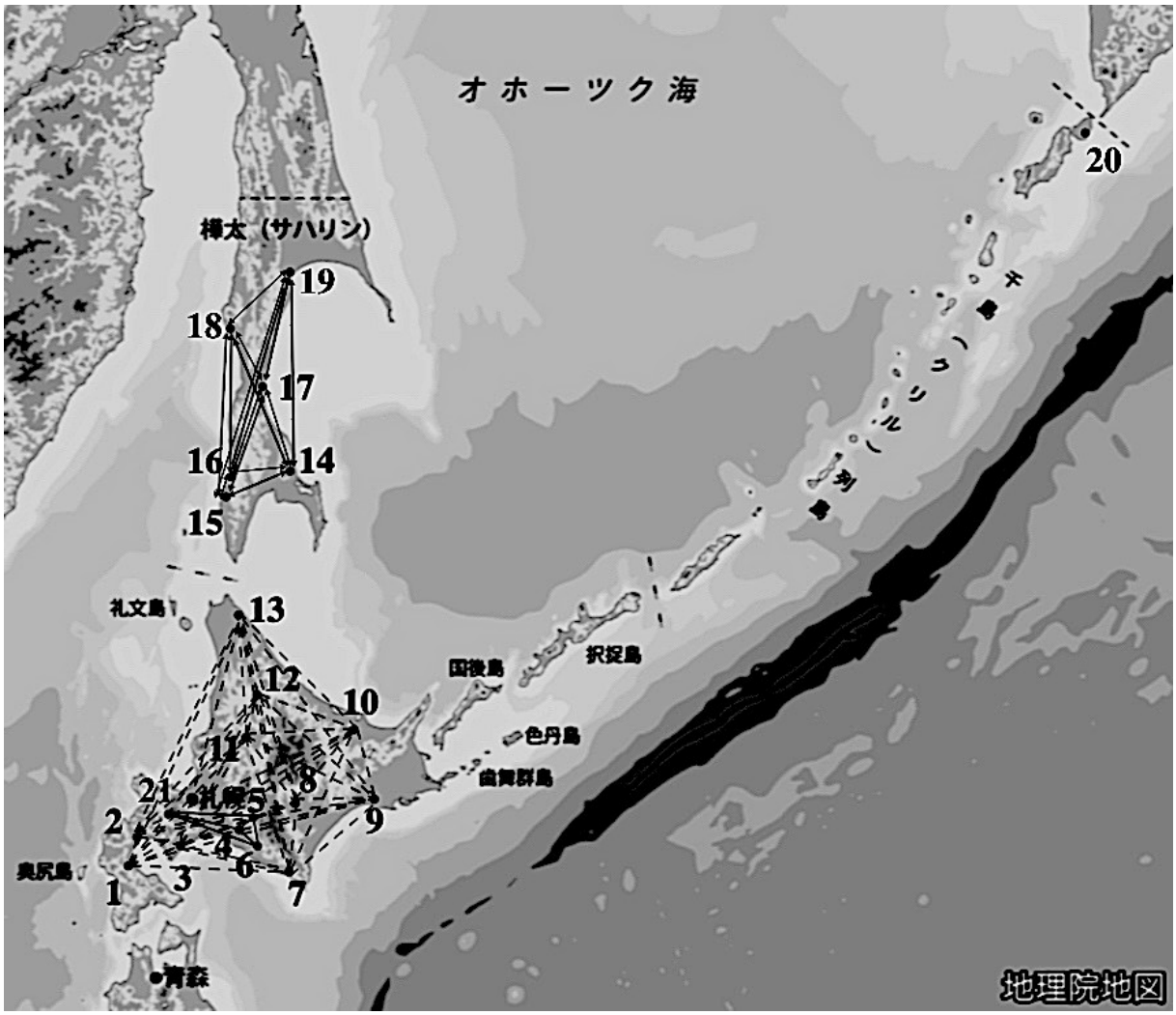

図 4. 表 6 を地図上に表したグラフ．但し、nen(1-3, 7-13)は点線で、hunna(4-6, 21)は実線で、naata(14-19)は両矢印付きの線でそれぞれ示した。図は国土交通省国土地理院(2019)を基に著者が編集した.

例えば、図 4 は表 6 のデータをグラフとして表している。表 6 は“relation index”(Asai 1974: 61-62)に基づき表 3 から計算されている。表 3 では、八雲(1)、長万部(2)、幌別(3)、様似(7)、帯広(8)、釧路(9)、美幌(10)、旭川(11)、名寄(12)、宗谷(13)が“who”に関し、nen という共通の語形を有しているため、図 6 ではこれらの方言の間を点線で繋げている。同様に表 3 では、平取(4)、貫気別(5)、新冠(6)、千歳(21)が“who”に関し、hunna という共通の語形を有しているため、図 6 ではこれらの方言の間を実線で繋げている。また表 3 では、落帆(14)、多蘭泊(15)、真岡(16)、白浦(17)、ライチシカ(18)、内路(19)が“who”に関し、naata という共通の語形を有しているため、図 6 ではこれらの方言の間を両矢印付きの実線で繋げている。北千島(20)は hunat という語形を有しているが、表 3 では nen, hunna, hunat, naata が同根語でないと仮定しているため、北千島と北千島以外の方言は繋がっていない。

表 7. 表 1 の八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13), 千歳(21)の北海道アイヌ語方言に関するデータを抜き出したもの.

	1	2	3	4	5	6	7	8	9	10	11	12	13	21
1_八雲	110	104	102	97	96	95	83	88	88	83	94	89	77	94
2_長万部	104	110	101	95	94	93	80	84	84	78	92	84	77	95
3_幌別	102	101	110	104	98	98	83	91	90	85	96	93	76	99
4_平取	97	95	104	110	103	105	79	87	86	81	92	89	73	104
5_貫気別	96	94	98	103	110	98	75	81	81	74	87	83	72	101
6_新冠	95	93	98	105	98	110	73	84	84	78	89	83	70	101
7_様似	83	80	83	79	75	73	110	94	92	84	85	81	74	80
8_帯広	88	84	91	87	81	84	94	110	106	100	95	93	79	84
9_釧路	88	84	90	86	81	84	92	106	110	101	98	98	83	85
10_美幌	83	78	85	81	74	78	84	100	101	110	92	91	78	79
11_旭川	94	92	96	92	87	89	85	95	98	92	110	102	82	93
12_名寄	89	84	93	89	83	83	81	93	98	91	102	110	84	87
13_宗谷	77	77	76	73	72	70	74	79	83	78	82	84	110	72
21_千歳	94	95	99	104	101	101	80	84	85	79	93	87	72	110

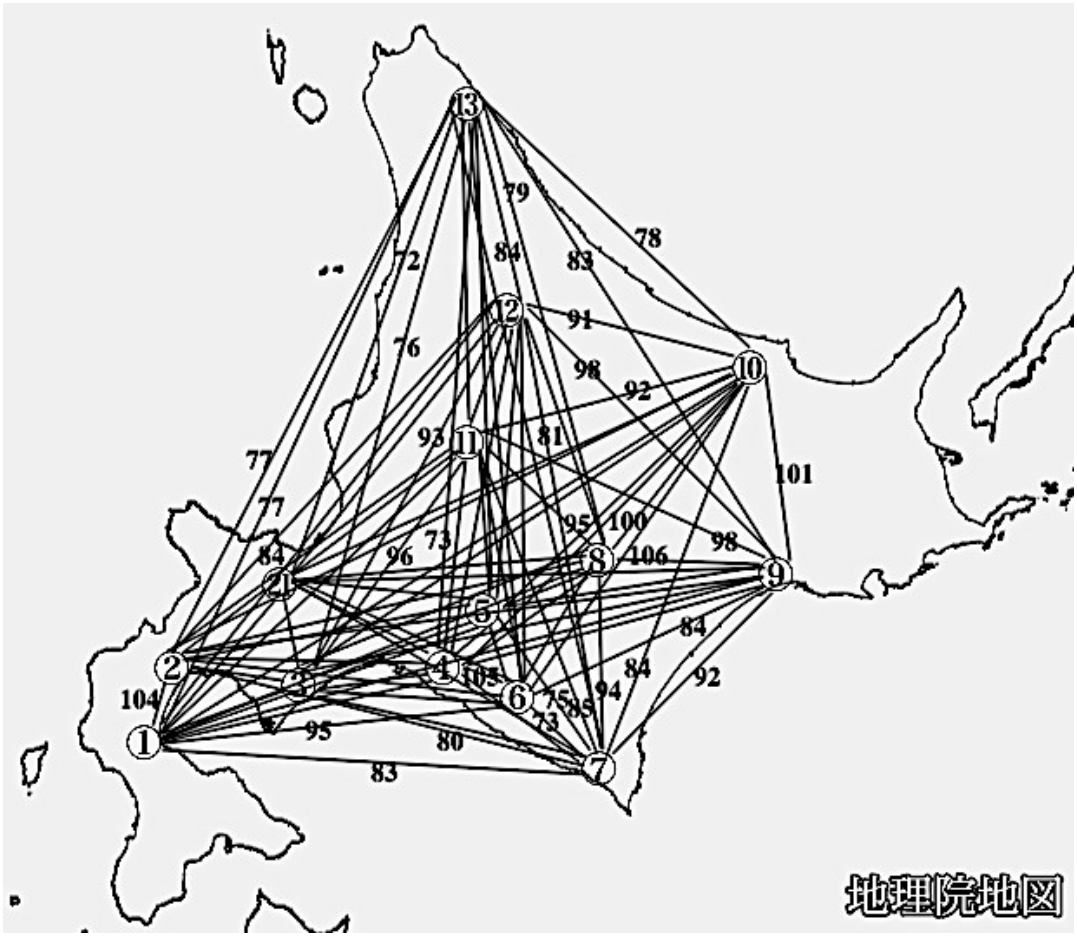

図 5. 表 7 をグラフとして地図上に可視化した図．地図上の各番号は表 7 の各方言の番号に対応している．例えば、八雲(1)と様似(7)は 83 となっている．これは、八雲(1)と様似(7)は表 7 では 110 語中 83 語において少なくとも一つの同根語形を有していることを表している．可視化の都合上、表 7 の一部の値のみ記載．図は国土交通省国土地理院(2020)を基に著者が編集した．

本研究は表 1 の八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13), 千歳(21)のいわゆる北海道アイヌ語方言のデータを分析する。表 7 は表 1 から北海道アイヌ語方言に関するデータを抜き出したものである。図 5 は表 7 をグラフとして地図上に視覚化している[5]。

数学ではグラフ理論という分野で、グラフの一般的な性質が研究されている。服部・知里(1960)や Asai(1974)のように基礎語彙統計学データから方言分類を考える場合、グラフ理論における知見を利用することが今後有効になると著者は考える。

本節では、グラフ理論の知見を用いることで、基礎語彙統計学データから方言分類を行う目的にふさわしい分析手法を導入する[6]。

2.4.1. 直感的な方法とその問題点

2.3 節の議論より、表 7 のデータに対して有効な操作は、「足し算」とそれらの値の「比較」の 2 つの操作である。この 2 つの操作から表 7 の方言を分類するにはどのような方法が考えられるだろうか。以下では von Luxburg (2007)の記法と説明を参照しこの問題について考える。

W(A, B)をグループ A とグループ B の間の類似度の和としグループ A とグループ B の両方に属する要素(方言)はないものとする。また、表 7 のデータを$\{\omega_{ij}\}$の行列として表記する。但し i=1, 2,…14, j=1, 2, …14 である。W(A, B)の定義を以下に記す(von Luxburg 2007: 396)。

$W(A, B)=\Sigma_{i\in A, j\in B}\ \omega_{ij}$

例えば、A={八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 千歳(14)}として B={様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13)}とすると、W(A, B)は

$W(A, B)=\Sigma_{i\in A, j\in B}\ \omega_{ij}$=4076 となる。但し i=1, 2, 3, 4, 5, 6, 14, j=7, 8, 9, 10, 11, 12, 13 である。

今、$A_1, A_2, A_3,\ldots,A_k$ の k 個のグループについて、$cut(A_1, A_2, A_3,\ldots,A_k)=W(A_1, A_1^c)+ W(A_2, A_2^c)+ W(A_3, A_3^c)+\ldots+ W(A_k, A_k^c)$という関数を定め、$cut(A_1, A_2, A_3,\ldots,A_k)$を最小化する分割 $A_1, A_2, A_3,\ldots,A_k$ を求めることが、直感的には考えられる。但し A_i^c はグループ A_i 以外の集合(補集合)を意味する。

例えば 2 分割の場合には、グループ A_1 とグループ A_2 の内部の類似度の和が最も大きく、グループ A_1 とグループ A_2 の間の類似度の和が最も小さいグループ A_1 とグループ A_2 を求めることが、直感的には望ましいと考えられる。ここで $W(A_1, A_1)+W(A_2, A_2)+W(A_1, A_2)$は常に定数である。$W(A_1, A_2)$を最小にする分割 A_1, A_2 は全ての組み合わせを「足し算」により計算しそれらの値を「比較」することで求まる。そのため、$W(A_1, A_2)$を最小にする分割 A_1, A_2 は表 7 のデータに対して有効な 2 つの操作のみで求まることがわかる。

しかしながら、von Luxburg(2007: 401)が指摘するように、上記の方法は実用性に乏しい。例えば 2 分割の場合、グループ A_1 に属する方言が少なければ少ないほど $W(A_1, A_2)$は小さく

[5] 例えば、図 3 の①と②を間にある 104 という数字は表 7 の八雲(1)と長万部(2)の値に対応している。
[6] 方言データをグラフとして捉える考えは、代表的な論文としては柴田・熊谷(1985; 1987)まで遡ることができる。しかし、柴田・熊谷(1985)の「ネットワーク法」は項目の有無を数えたデータを扱っている。項目の有無を数えたデータは負の数が言語学上意味のある代数構造である。そのため柴田・熊谷(1985)の考え方を引き継いだ諸手法と本研究の手法は異なる。扱うデータの代数構造が異なっているためである。

なる傾向がある。そのため $W(A_1, A_2)$を最小にする組み合わせは、ある方言とそれ以外の方言全てという分割になる。また、他の点の集まりから外れて一点だけ非常に遠くに孤立した点 O は、他の点との類似度が小さい。そのようにデータ同士の類似度が小さい場合、$W(A_1, A_2)$を最小にする組み合わせを計算すると、$W(O, O^c)$が最も小さくなり、O と O 以外の全ての点(O^c)という自明なグループ分けが生じ、残った O^c においても同様のグループ分けが生じる。上記の $W(A_1, A_2)$の性質は方言分類を行う目的に相応しくない。よって、新たな指標を考える必要がある。次節では、上記の方法に代わるものとして Mcut 法、Ncut 法を導入する。

2.4.2. Mcut 法、Ncut 法の導入

本研究ではグラフ理論で用いられる MinmaxCut 法(Ding et al. 2001)、NormalizedCut 法(Shi and Malik 2000)のそれぞれの指標を参照することを提案する。以下では MinmaxCut 法は Mcut 法と、NormalizedCut 法は Ncut 法とそれぞれ略す。

2 分割の場合、Mcut 法の指標 $Mcut(A_1, A_2)$、Ncut 法の指標 $Ncut(A_1, A_2)$はそれぞれ以下のように記述できる。

$$Mcut(A_1, A_2)=W(A_1, A_2)/W(A_1, A_1)+W(A_1, A_2)/W(A_2, A_2)$$

$$Ncut(A_1, A_2)= W(A_1, A_2)/vol(A_1)+ W(A_1, A_2)/vol(A_2)$$

但し、$vol(A_1)$はグループ A_1 に属する点(方言)に関して、各点(自身を含める)とそれ以外の点の間の類似度を合計した値を, $vol(A_2)$はグループ A_2 に属する点(方言)に関して、各点(自身を含める)とそれ以外の点の間の類似度を合計した値を、それぞれ意味する。各手法はそれぞれの指標を最小化する分割 A_1, A_2 を求める[7,8]。

2.4.3. 方言データに統計手法を適用する目的

方言学のデータに統計手法を適用する目的の一つはデータに存在する方言関係についての様々な情報を可視化することにある。しかし、多くの方言データでは地理的パターンが矛盾する情報がデータの中に同時に存在している。例えば、日本語方言学における東西型と中央周圏型では、東北方言は東日本と高い類似性を示すと同時に、西日本に属する九州とも高い

---

[7] $Mcut(A_1, A_2)$, $Ncut(A_1, A_2)$の指標は、2.3 節で述べた可換モノイドに適用可能な「足し算」と、それらの値の「比較」以外に、「割り算」と「割り算を適用した値同士の足し算」の 2 つの操作を導入している。本注はその妥当性について補足する。今、$Mcut(A_1, A_2)$, $Ncut(A_1, A_2)$は以下の形に式変形することが可能である。

$Mcut(A_1, A_2)=W(A_1, A_1^c)/W(A_1, A_1)+W(A_2, A_2^c)/W(A_2, A_2)$

$Ncut(A_1, A_2)=W(A_1, A_1^c)/vol(A_1)+W(A_2, A_2^c)/vol(A_2)$

但し、A_1^c は A_1 を除いた集合を A_2^c は A_2 を除いた集合をそれぞれ意味する。

上式の右辺の第一項と第二項はそれぞれ A_1 と A_2 の関数であり、グループ A, グループ B のグループとしての望ましさを測っている。例えば、$W(A_1, A_1^c)/W(A_1, A_1)$と $W(A_2, A_2^c)/W(A_2, A_2)$の値が小さい分割 A_1, A_2 は、グループの間の類似度が小さくグループの内部の類似度が大きいという条件を満たすため 2.4.1 節で述べた直感的な望ましさを備えている、と考えることができる。$Mcut(A_1, A_2)$を導入することはグループ A_1 の良さを測る指標として $W(A_1, A_1^c)/W(A_1, A_1)$を、グループ A_2 の良さを測る指標として $W(A_2, A_2^c)/W(A_2, A_2)$を我々がそれぞれ新たに定義し、その指標の間に「足し算」と比較を導入することを意味する。$W(A_1, A_1^c)/W(A_1, A_1)$と $W(A_2, A_2^c)/W(A_2, A_2)$の分子と分母はいずれも可換モノイドに適用可能な「足し算」により計算できる。両者とも 2.3.2 節で述べた“relation index”(Asai 1974: 61-62)の「1」を、単位として共有するため、割り算を適用し $W(A_1, A_1^c)/W(A_1, A_1)$をグループ A の良さとすることは意味のある操作と考えられる。しかしながら、$W(A_1, A_1^c)/W(A_1, A_1)$, $W(A_2, A_2^c)/W(A_2, A_2)$には単位がないため、$W(A_1, A_1^c)/W(A_1, A_1)$と $W(A_2, A_2^c)/W(A_2, A_2)$は「足し算」や「比較」ができない。そのため 2 つの指標の間に「足し算」と「比較」を導入する必要がある。$Ncut(A_1, A_2)$の導入に関しても同様に議論できる。指標間に「足し算」と「比較」を導入することの言語学上の意味と妥当性については今後議論すべき課題と考える。本研究は Asai(1974: 92; Table 1)では意味のない負の数を Mcut 法と Ncut 法が用いない点に鑑み、指標間に「足し算」と「比較」を導入することを妥当と考え、以下の議論を進める。

[8] 3 グループ以上の場合の Mcut 法、Ncut 法の目的関数の定義は von Luxburg(2007: 401, 412)を参照されたい。

類似性を示すという矛盾した情報を示す。

データに存在する(矛盾した)複数の分類情報を認識することは、人間の認知能力が得意とすることではない。例えば表 7 を眺め、北海道アイヌ語方言でも南西部(図 1 の 1 から 6 と 21)と北東部(図 1 の 7 から 13)それぞれの内部の類似度が高く、南西部と北東部の間の類似度が低いことも認識することは可能であろう。

しかし、例えば表 7 の北海道のアイヌ語方言に関し、南西部と北東部という分割の次に、グループ A'とグループ B'の内部の類似度の和が大きく、グループ A'のグループ B'の間の類似度の和が小さいグループ A'とグループ B'を表 7 から正確に認識することは、一般に人間の認知能力では困難である。

服部・知里(1960)を用いた先行研究(Ono 2019a, 2019b)では、Neighbor-Net(Bryant and Moulton 2004)を適用し、アイヌ語方言における複数の(矛盾した)情報を可視化することに成功しているが、Neighbor-Net は距離(非類似度)に基づいた手法である。

そのため、表 7 のような類似度の情報のみを有するデータから、どのようにデータの中に存在する方言関係の様々な情報を可視化するかは、今後の方言学の発展に大きく寄与し得る重要な問題である。よって、この問題は今後も統計学者が追究するべき課題である。

次節では、Mcut 法と Ncut 法の指標を用いることで上記の問題に取り組む。

2.4.4. Mcut 法と Ncut 法の方言データへの応用

Mcut(A, B)の計算には、個体数(言語数)が増加すると組合せ論的に計算量が増大する。だが、Mcut(A, B)を最小にする分割は、ある種の固有値問題を解くことに帰着することが示されている。そのため Mcut 法は固有値分解を用いて表 7 のようなデータの分類を非常に短時間で得ることを特徴としている。このことで、個体数(言語数)が非常に大きい場合の計算時間の問題を解決している。但し、固有値分解により得られる分類は、Mcut 法の指標を最小にする 1 つの分類のみである。

しかしながら、方言データの分析では個体数は多くても 50 程度である。そのため Mcut 法の指標 Mcut(A, B)を可能な全ての分割について計算することがコンピュータの力を使うことで可能である。

この方法により Mcut 法の指標を 2 番目に最小にする分割、3 番目に最小にする分割等々を得ることができる。Mcut 法の指標 Mcut(A, B)のこの性質は類似度のみが存在するデータの様々な情報の可視化を可能にするものであり、距離データを分析する際に用いられる Neighbor-Net の役割と担うことが期待される。以上の議論は Ncut 法に関しても成り立つ。

よって、Mcut 法の指標 Mcut(A, B)と Ncut 法の指標 Ncut(A, B)を全ての分割に関して直接計算する方法を本研究は採用し、北海道アイヌ語方言に関する表 7 のデータに、採用手法を適用する。

2.4.5. 本研究の分析目的

服部・知里(1960)や Asai(1974)の研究以来、北海道アイヌ語方言では南西と北東を区別する分類が支配的であった。しかし、Kirikae(1994: 109-110)が指摘する pa と ca の ABA 型の分布、中川(1996)や深澤(2017)の言語地理学に基づいたアイヌ語分析における ABA 型の重要性など、

服部・知里(1960)や Asai(1974)の北海道アイヌ語方言を南西と北東を分ける分類とは矛盾する ABA 型の情報の重要性がアイヌ語の専門家の中で度々指摘されている。

服部・知里(1960)に関しては Ono(2019a)などで ABA 型の情報の存在が既に示されているが、Asai(1974: 92; Table 1)の北海道アイヌ語方言に関するデータを用いて ABA 型の存在を統計学の立場から追究した研究は、管見の限り、存在しない。

よって本研究では前節までの議論から、Asai(1974: 92; Table 1)のデータの性質を検討した結果、分析に相応しいと考えられた Mcut 法と Ncut 法を表 1 の北海道アイヌ語方言に適用する。更に方言データに統計手法を適用する目的に鑑みて、上記の結果得られた分類の中で Mcut(A, B)と Ncut(A, B)の値が小さい幾つかの北海道アイヌ語方言の分類を可視化することで先行研究の知見を検討することを本研究の分析目的とする。

以下の分析では、北海道アイヌ語方言(表 7)の Mcut 法の指標 Mcut(A, B)と Ncut 法の指標 Ncut(A, B)の全ての組み合わせを計算する。

3. 分析結果

表 8. Mcut(A, B)の値を昇順にまとめた分析結果.

	分割A	分割B	Mcut(A, B)の値
1	八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 千歳(21)	様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13)	2.962862
2	八雲(1), 様似(7), 帯広(8), 釧路(9), 美幌(10), 名寄(12), 宗谷(13)	長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 旭川(11), 千歳(21)	3.072434
3	八雲(1), 長万部(2), 平取(4), 貫気別(5), 新冠(6), 旭川(11), 千歳(21)	幌別(3), 様似(7), 帯広(8), 釧路(9), 美幌(10), 名寄(12), 宗谷(13)	3.079303
4	八雲(1), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 旭川(11), 千歳(21)	長万部(2), 様似(7), 帯広(8), 釧路(9), 美幌(10), 名寄(12), 宗谷(13)	3.090722
5	八雲(1), 様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12)	長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 宗谷(13), 千歳(21)	3.092127
(中略)	(中略)	(中略)	(中略)
8188	八雲(1)	八雲(1)以外	10.961810
8189	平取(4)	平取(4)以外	11.007960
8190	旭川(11)	旭川(11)以外	11.026420
8191	幌別(3)	幌別(3)以外	11.201780

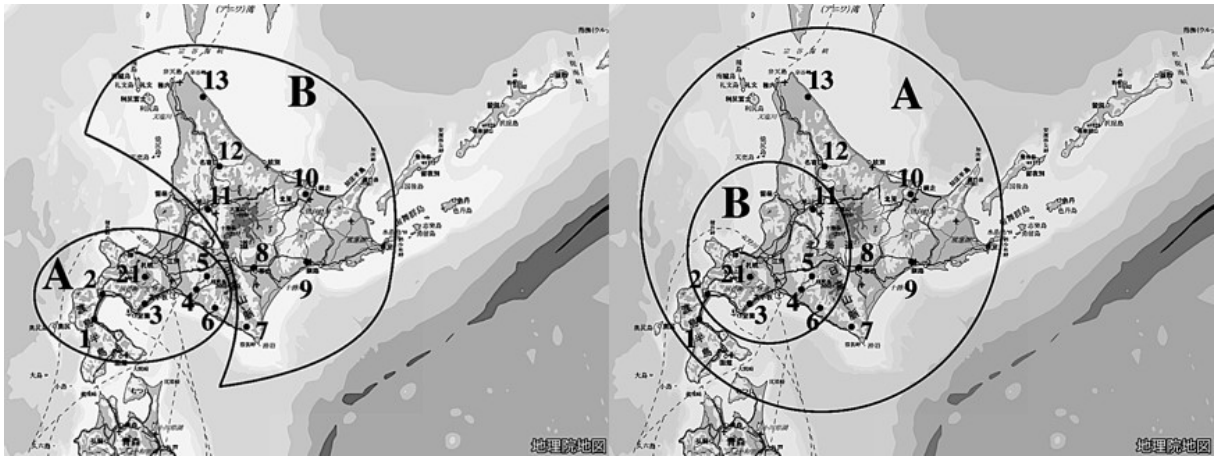

図 6. (左)表 8 で Mcut(A, B)の値が最も小さい分割を地図上に示した．東西型のパターンを示している．
(右)表 8 で Mcut(A, B)の値が 2 番目に小さい分割を地図上に示した．ABA 型のパターンを示している．
図は国土交通省国土地理院(2019)を基に著者が編集した．

表 9. Ncut(A, B)の値を昇順にまとめた分析結果.

	分割A	分割B	Ncut(A, B)の値
1	八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 千歳(21)	様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13)	0.925195
2	八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 旭川(11), 千歳(21)	様似(7), 帯広(8), 釧路(9), 美幌(10), 名寄(12), 宗谷(13)	0.931274
3	八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 宗谷(13), 千歳(21)	様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12)	0.936511
4	八雲(1), 様似(7), 帯広(8), 釧路(9), 美幌(10), 旭川(11), 名寄(12), 宗谷(13)	長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 千歳(21)	0.936596
5	八雲(1), 長万部(2), 幌別(3), 平取(4), 貫気別(5), 新冠(6), 旭川(11), 名寄(12), 千歳(21)	様似(7), 帯広(8), 釧路(9), 美幌(10), 宗谷(13)	0.936683
(中略)	(中略)	(中略)	(中略)
8188	八雲(1), 平取(4), 釧路(9), 旭川(11)	八雲(1), 平取(4), 釧路(9), 旭川(11)以外	0.995245
8189	長万部(2), 平取(4), 釧路(9)	長万部(2), 平取(4), 釧路(9)以外	0.995359
8190	八雲(1), 幌別(3), 平取(4), 帯広(8), 釧路(9), 旭川(11), 宗谷(13), 千歳(21)	長万部(2), 貫気別(5), 新冠(6), 様似(7), 美幌(10), 名寄(12)	0.995557
8191	八雲(1), 平取(4), 釧路(9)	八雲(1), 平取(4), 釧路(9)以外	0.995633

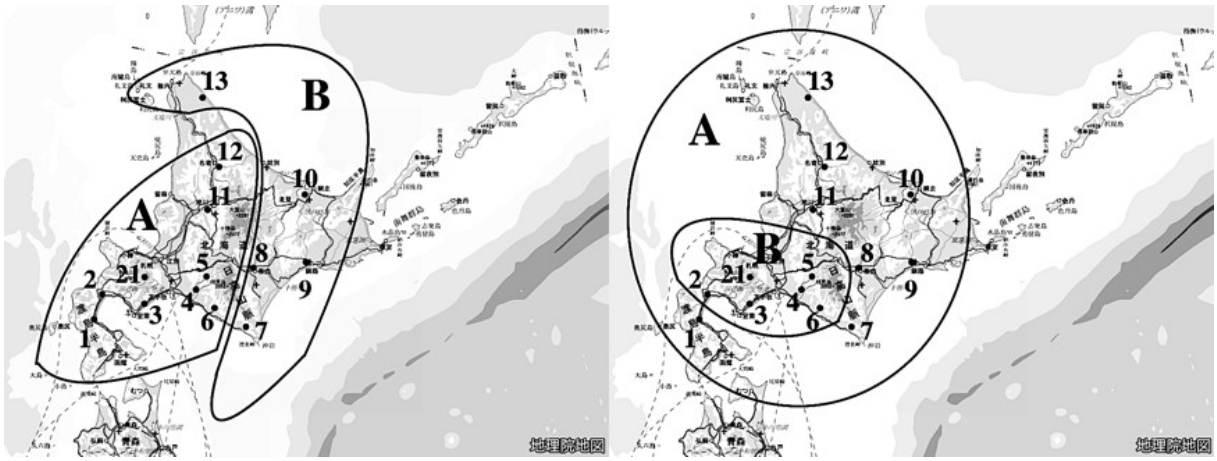

図 7. (左)表 9 で Ncut(A, B)の値が 5 番目に小さい分割を地図上に示した．東西型のパターンを示している．
(右)表 9 で Ncut(A, B)の値が 4 番目に小さい分割を地図上に示した．ABA 型のパターンを示している．
図は国土交通省国土地理院(2019)を基に著者が編集した．

表 8 に、Mcut(A, B)の値を昇順にした Mcut 法の分析結果をまとめた。表 8 の中で、Mcut(A, B)の値が最も小さい分割を図 6 の左に、Mcut(A, B)の値が 2 番目に小さい分割を図 6 の右に、それぞれ地図上に示した。次に、表 9 に Ncut(A, B)の値を昇順にした Ncut 法の分析結果をまとめた。表 9 の中で、Ncut(A, B)の値が 5 番目に小さいものを図 7 の左に、Ncut(A, B)の値が 4 番目に小さいものを図 7 の右に、それぞれ地図上に示した。

Mcut(A, B)の値が最も小さい図 6 の左の分布は Asai(1974: 100)と同じ北海道アイヌ語方言を南西と北東に分けるものである。一方、表 6 において Mcut(A, B)の値が 2 番目以降に小さい分割の中には言語地理学における ABA 型の分布を示すものが多い。

例えば、Mcut(A, B)の値が最も小さい図 6 の左の分布では、八雲、長万部、幌別、平取、貫気別、新冠、千歳が北海道南西部のグループを構成し、様似、帯広、釧路、美幌、旭川、名寄、宗谷が北海道北東部のグループを構成しており、典型的な東西型の分布を示している。

一方、Mcut(A, B)の値が 2 番目に小さい図 6 の右の分布では渡島半島の八雲が様似、帯広、釧路、美幌、名寄、宗谷の北海道北東部のグループに属し、石狩川が流れる旭川が長万部、幌別、平取、貫気別、新冠、千歳の北海道南西部のグループに属し典型的な ABA 型の分布を示している。また、Ncut(A, B)の値が最も小さい分布は Mcut(A, B)の場合と同様に、北海道アイヌ語方言を南西と北東に分けるものである。

Ncut(A, B)の値が 5 番目に小さい図 7 の左の分布は、図 6 の左の分布と同様に、東西型の分布を示している。但し、図 6 の左では旭川と名寄が東西型の東に属すのに対し、図 7 の左の図では旭川と名寄が東西型の西に属している点で分類が異なっている。

一方、Ncut(A, B)の値が 4 番目に小さい図 7 の右の分布は、図 6 の右の分布と同様に、ABA 型の分布を示している。但し、図 6 の右の図では旭川が ABA 型の B に属すのに対し、図 7 の右の図では旭川が ABA 型の A に属している点で分類が異なる。

次節では、本節の分析結果について方法論上及び言語学上の観点から考察を行う。

4. 考察

本節では本研究の分析結果が示す方法論上及び言語学上の意義について考察を行う。まず、方法論上の観点からは、本研究は Asai(1974: 92; Table 1)が有する加法に関する可換モノイドという特殊な代数構造に適用可能と考えられる統計手法を採用したことで、より妥当な分析を行ったと考えることができる。更に Asai(1974: 92; Table 1)のデータの妥当性を認めれば、既存の研究と違い本研究の分析結果の意味を追究することが可能になった。本研究は序論で述べた人文学データの分析の今日までの問題を克服する一つのアプローチとして、今後参照され得るものと考えられる。

次に、本研究の提案手法は距離構造を有さず正の類似度情報のみを有するデータについて、データの代数構造に適切な方法を用いることで、距離構造を有するデータにおける Neighbor-Net の役割と同様に、2 番目以降に強い構造について追究することを可能にした。人文学の当該分野のデータの有する代数構造の再検討を通じ、適切な分類手法を構築するという本研究のアプローチはデータの 2 番目以降に強い構造に関心がある言語学以外の人文学の分野においても、今後の追究の際に参考となり得ると考えられる。

また、本研究の提案手法によって、Asai(1974: 92; Table 1)の北海道アイヌ語方言のデータには、Asai(1974: 100)が示した南西と北東を分ける分類だけでなく、ABA 型のパターンが、分析手法の値が 2 番目以降に小さい分割において存在していることが初めて示された。

服部・知里(1960)のデータの統計分析では、Ono(2019a)などの先行研究では既に ABA 型の情報が存在していることが指摘されている。しかし、Ono(2020c, to appear)は、服部・知里(1960)と Asai(1974)では基礎語彙に関して分析に採用した語と語形間の同根性判断の基準が大きく異なっていることを述べている。それにも関わらず、服部・知里(1960)と Asai(1974)において ABA 型のパターンの存在が示されたことは、ABA 型のパターンを今後更に追究することの重要性を統計分析の立場から示すものと考えられる。

最後に、Kirikae(1994), 中川(1996)などの言語学・文献学の先行研究でアイヌ語方言の分類

における ABA 型のパターンの重要性はすでに指摘されており、本研究は Asai(1974)を適切に分析することにより ABA 型のパターンを統計学の点から再確認している。しかしながら、実際の言語研究においては先行研究が存在しない状況にしばしば直面する。そのような状況において有力な分類を考え、言語調査の対象を絞る、もしくは広げる必要がある。例えば、Asai(1974)のような基礎語彙統計学データだけが手元にあった時、上記の言語研究に資する適切な分析手法を追究することは、今後の重要な課題である。

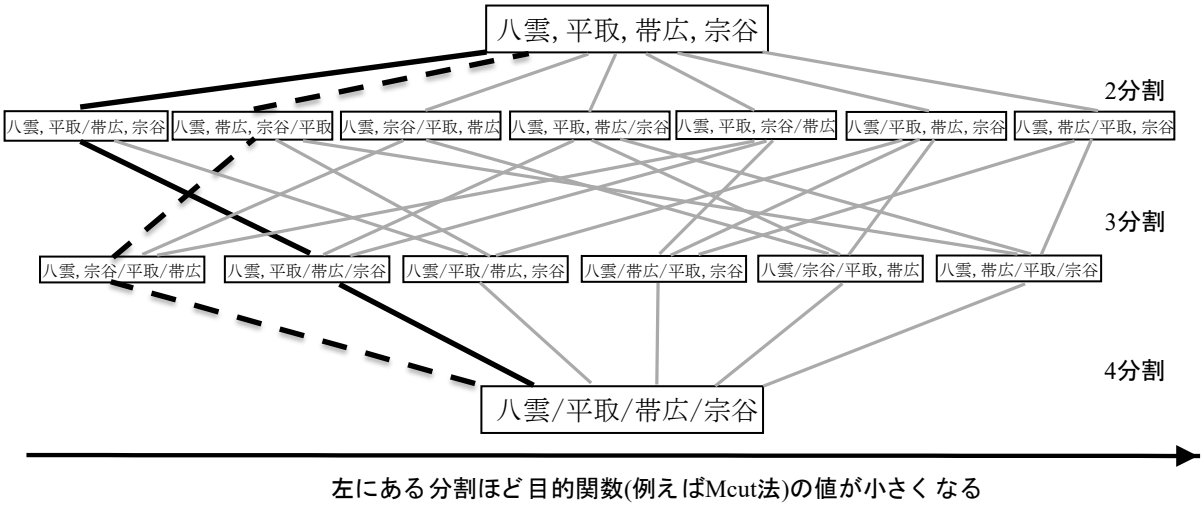

図 8. 八雲、平取、帯広、宗谷に関する仮想の分析結果を可視化したハッセ図.
八雲,平取/帯広,宗谷は 4 つの方言を八雲,平取と帯広,宗谷の 2 グループに分割したことを意味する。
ハッセ図では、ある分割が別の分割の細分になっているとき、2 つの分割は線で繋がる。

この問題について、Mcut 法や Ncut 法の分析結果のハッセ図による可視化が有効であると著者は考える。図 8 は、八雲、平取、帯広、宗谷の 4 方言に関する仮想の分析結果について、ハッセ図を用いて可視化したものである[9]。本研究、は図 8 のハッセ図では表 7 のデータに Mcut 法または Ncut 法の目的関数を適用し、2 分割の場合に目的関数が小さい幾つかの分割を比較したことに対応する。

ハッセ図による可視化の利点は距離データ(非類似度)のおける Neighbor-Net の役割と同様に、階層クラスター分析による可視化の問題点を克服できることが挙げられる。Ono(2020b, to appear)のように、Mcut 法の目的関数に基づき方言を分割し階層的な方言の分類を求める手法は図 8 の太い黒の実線に対応する。一方、Asai(1974)の古典的な階層的クラスター分析は、ある目的関数に基づき方言を集めていき階層的な方言の分類を求める手法であり、図 8 では太い黒の点線に対応する。

図 8 はあくまでも仮想の分析結果を可視化したハッセ図であるが、前者の手法においては、

[9] ハッセ図は数学では束論(lattice theory)という分野で扱われる。本研究では方言の「分類」という点から、束(lattice)による可視化を行った。ただし、方言の系統関係を考える場合には半束(semilattice)による分析が有効である。詳しくは、半束を用い方言の系統関係を追究した濱田(2019)の優れた研究を参照されたい。

2 分割の場合に目的関数を最小化する分割が「八雲,平取/帯広,宗谷」であったため、3 分割の場合に目的関数を最小化する分割である「八雲,宗谷/平取/帯広」が「八雲,平取/帯広,宗谷」と繋がらず(「八雲,宗谷/平取/帯広」が「八雲,平取/帯広,宗谷」の細分でなく)、分析において「八雲,宗谷/平取/帯広」という 3 分割の場合に目的関数を最小化する有力な分割が見落とされるという問題がある。

同様に、後者の手法においても、3 分割の場合に目的関数を最小化する分割「八雲,宗谷/平取/帯広」が 2 分割の場合に目的関数を最小化する分割「八雲,平取/帯広,宗谷」と繋がらず(「八雲,宗谷/平取/帯広」が「八雲,平取/帯広,宗谷」の細分でなく)、「八雲,平取/帯広,宗谷」という 2 分割の場合に目的関数を最小化する有力な分割が見落とされるという問題がある。

しかし、目的関数の値に基づいた分割の結果を、図 8 のようにハッセ図により可視化することで上記の問題を回避することができる。ハッセ図による可視化法は先行研究がなく Asai(1974)のような基礎語彙統計学データのみが手元にあった場合に、有力な分類を考え、言語調査の対象を絞るもしくは広げるための探索的手法として有力であると著者は考える。

一方、ハッセ図による可視化の欠点は、計算する組み合わせが方言の数の増加とともに、組み合わせ論的に増大しに計算量が膨大になること、膨大な組み合わせの全てを人間の認知能力で把握しやすい形で可視化する有力な方法が今のところないという点である。

しかしながら、著者は上記の問題は今後の技術の発展とともに解決されると考えている。上記の問題は現在の論文が紙の上ないし電子版の論文ではディスプレイの上で表示されるという制約によるものである。すでに、ビックデータの分野では、バーチャル・リアリティ、すなわち仮想疑似空間(virtual reality)を使ったデータの可視化手法が研究されている(例えば Olshannikova, Ometov, Koucheryavy, and Olsson 2015)。よってハッセ図による基礎語彙統計学データの可視化は今後の言語研究にとり、実用可能な手段になっていくと考えられる。

次に言語学上の観点からは、本研究が示した ABA 型の分布は、言語地理学においては、当該言語に関して A が古い語形、B が新しい語形と一般的に考えられるため、ABA 型の分布を示す語ないし語形を通時的ないし共時的な観点から追究することは、当該言語がどのように形成されたかを解明することに寄与する可能性がある。

北海道アイヌ語方言に Asai(1974: 92; Table 1)が示したとされる南西と北東を分ける分布に次いで、2 番目以降の構造に ABA 型の分布の存在を本研究が示したことで、アイヌ語の形成や起源に関する研究において ABA 型の分布を示す語ないし語形について、アイヌ語学の点から追究することの重要性がより一層明確になったと考えられる。

また、Ono(2020e)は服部・知里(1960)の基礎語彙統計学データに対して、本研究と同様の方法を適用し、服部・知里(1960)においても ABA 型の分布の存在を確認している。アイヌ語方言に関する服部・知里(1960)と Asai(1974)の信頼性が高く同根性判断が異なる 2 つのデータに本研究の手法を適用した結果、ABA 型の分布の存在が確認されたことにより、ABA 型の分布を示す語ないし語形に関する言語学・文献学上の追究の重要性はより一層裏付けられた。

また、旭川や名寄が Mcut 法や Ncut 法の分析結果において Asai(1974: 100)と異なる分類となったことは Ono(2019b)の知見とも一致する。

次節では、本節で述べた考察を踏まえ、本研究で明らかになった課題について述べ、結語とする。

5. 結語

本節では本研究の考察から得られた方法論上及び言語学上の課題について述べる。方法論上の課題として、本研究で分析した Asai(1974: 92; Table 1)のデータは 110 語に関する各語の同根性判断のデータを合計したものであり、各語における詳細な同根性に関する情報が損失している。それにも関わらず、本研究の提案手法により北海道アイヌ語方言の分類に関する重要な情報を得ることができた。

今後は、同根性判断を合計したデータの分析が、どのような言語の場合に有効であるかを検討するとともに、分析手法と分析結果の意味について、言語学および統計学の両面からの追究が必要である[10]。

著者は既に Ono(2020d; 2020e)において、服部・知里(1960)のアイヌ語諸方言の分析での、同根性判断を合計した分析手法の意味を論じているため、詳細は Ono(2020d; 2020e)に譲る。一般論としては、以下の論点に注意を払う必要があると著者は考える。

Starostin(2010: 84-86)は基礎語彙統計学を classic lexicostatistics と preliminary lexicostatistics の 2 つに分類し、同根性判断に関して classic lexicostatistics は信頼できる音韻対応(reliance on phonetic correspondence)に基づくが、preliminary lexicostatistics は音の類似性や音の変化に関する言語学における一般的な知識、他の言語における知見などの音声上の類似性(phonetic compatibility)に基づくとしている。

Starostin(2010)は言語間の分類に基礎語彙統計学のデータを用いることを念頭にしている。だが、上記の視点は言語内の方言の分類においても有効である。基礎語彙統計学を方言分類に利用する場合、同根性判断が「信頼できる音韻対応」に基づくならば同根性判断に関するデータは「信頼できる音韻対応」という一つの物差し(尺度)で測られていると考えられる。その場合、表 5 や表 6 そして表 1 のデータへ加法を適用することが言語学上意味のある操作になると考えることができる。本研究の Asai(1974: 92; Table 1)は同根性判断が「信頼できる音韻対応」に基づいた一例と考えることができる。

一方、当該の方言に関する調査環境や調査状況、記述的研究の状況によっては同根性判断が「信頼できる音韻対応」に基づくとは言えず、音声学上の類似性に依らざるを得ない状況も考えられる。この場合、同根性判断に関するデータは「信頼できる音韻対応」という一つの物差し(尺度)で測られているとは言えず、表 5 や表 6 のデータへ加法を適用することが、言語学上妥当であるか議論の余地がある。そのようなデータに対してどのような統計手法を適用することが妥当であるか追究することが今後の方法論上の課題と言えよう。

さらに言語学上の課題として、2 節で述べたように、Asai(1974: 92; Table 1)は 110 語に関し、

[10] 大変興味深いことにネットワーク法成立の初期の段階において、伊藤(1987)が同じ問題を指摘している。残念ながら、伊藤(1987)に対する返信である柴田・熊谷(1988)は、伊藤(1987)が相関係数の使用を提案したことに焦点を当てすぎているように思う。この問題については他稿で論じたい。

「方言同士に少なくとも一つ同根語形があるかどうか」を数えたデータであり、どのような同根性判断を Asai(1974)が行ったかを明らかにすることが課題である。この問題について、Ono(2020c, to appear)はデータ分析の手法を用いることで一部の同根性判断を特定しているため、詳細は Ono(2020c, to appear)に譲る。

各語における同根性判断を明らかにすることは、ABA 型を示す分布と語ないし語形の関連が判明しアイヌ語の形成や起源の解明に寄与するだけでない。一般的に言語により身体語彙、自然語彙、生活語彙などが異なった類似関係を示す場合があり、このことの追究が言語形成の過程の解明に寄与する可能性がある。既に、浅井(1974)はアイヌ語諸方言に関して同様の分析を行い、アイヌ語の成立に関して示唆的な知見を得ている。

本研究は Asai(1974)の北海道アイヌ語諸方言に関する研究の再検討を通じ、人文学データの分析において、当該分野が１）現象のどの側面を重要な点と考えて分析対象としているかを捉え、２）どのような記号の体系を規範としているか把握し３）それらの記号により記述された情報をどのように分析するべきかとしているかを理解した上で、統計学において１)、２)、３）の全てに対応した分析手法を適切に選択・開発することの重要性を例示し、今後の人文学と統計学の学際的研究の方向性を示すことを目的としていた。

具体的には Asai(1974: 92; Table 1)の特殊な代数構造の分析を通じ、方言データに統計手法を適用する分析目的とデータの性質に適した提案手法を、Asai(1974: 92; Table 1)に適用した結果、Asai(1974: 92; Table 1)の北海道アイヌ語方言においても ABA 型の情報が存在することを示し、北海道アイヌ語方言における ABA 型の情報の重要性を統計学の点から再確認した。提案手法による分析結果とその考察から、本節で述べた方法論上および言語学上の課題が、明らかになった。

また、基礎語彙統計学データの代数構造が加法に関する可換モノイドであるという本研究の結果の意義は大きいと著者は考える。安本(1995)は基礎語彙統計学に基づいた方言や言語の系統関係や分類に関する先行研究についてまとめている。それらの先行研究が用いている指標(例えば標本相関係数)などは方言間ないし言語間の同根性データに対して、加法だけでなく減法、乗法、除法などが「言語学上意味のある操作であること」を前提としている。

しかしながら、本研究で示したように Asai(1974)の“relation index”に基づく方言間の同根性データは逆元が言語学的に意味のない加法に関する可換モノイドと考えられる。加法、減法、乗法、除法が成り立つ代数構造と加法に関する可換モノイドは全く異なる(Gondran and Minoux 2008: 13-14)。Asai(1974: 61-62)の“relation index”の定義は、基礎語彙統計学データの類似度の定義を一般的な形に定式化したものである。Asai(1974)の“relation index”に基づく、方言間の同根性に関するデータが加法に関する可換モノイドと考えられるという本研究の結果は、基礎語彙統計学データを用い方言ないし言語の分類を行った既存の先行研究が代数構造の点から再考を要することを示している。

但し、この結果は単に先行研究の結果を否定するものではない。むしろ、基礎語彙統計学データの代数構造からの見直しは、既存の方言研究や言語研究が新たな展開を見せる可能性を示唆している。特に、我が国では日本語方言等に関する詳細な記述的研究が盛んであり、

本研究の成果はそれらの分野で蓄積された知見に貢献する可能性がある。また海外においては印欧語族などに関する基礎語彙統計学による研究が盛んである。よって、本研究の成果は国内にとどまらず、海外の方言研究、言語研究に今後寄与するものと考えられる。

最後に、基礎語彙統計学データを代数構造から見直すという本研究の議論は、言語学以外の人文学にも貢献する可能性が高い。本研究の 1 節で述べたように、一般に人文学データは、１）現象のどの側面を重要な点と考えて分析対象としているかを捉え、２）どのような記号の体系を規範としているか把握し、３）それらの記号により記述された情報を、どのように分析するべきかとしているかについて、他の自然科学や社会科学とは異なり、考慮するべき点が多い。このため、例えば本研究で示したように、データの代数構造が古典的な統計手法の前提と異なることがある。管見の限り、人文学データの分析は、代数構造の問題に限らず、記述統計を含めた広く数学上の点から検討を要する課題が山積している。これらの問題を、人文学と統計学の学際的研究において適切に取り組むことで、既存の人文学の研究が新たな展開を見せる可能性がある。

逆に、本研究で示したように、人文学と統計学の学際的研究はこれらの課題に適切に取り組まなければ、1.3 節で述べたように不毛な議論をこれからも繰り返すことになるだろう。

本研究で示した一連の議論が、今後の人文学と統計学の学際的研究の方向性を示すことに寄与することを期待して結語とする。

謝辞

* 本研究の第 2 節の内容は、著者が計量国語学会第 63 回大会（2019 年 9 月 21 日於国立国語研究所）で、オーストラリア諸語の基礎語彙統計学データを統計学の観点から分析した Dobson and Black(1979)の研究についてグラフ理論の観点から発表した内容を、Asai(1974)のアイヌ語諸方言の基礎語彙統計学データの性質の点から新たに執筆したものである。計量国語学会第 63 回大会関係者の皆様、研究発表会の座長であった荻野綱男先生、およびコメントを頂いた参加者の方々に深く感謝いたします。また著者がアイヌ語諸方言の基礎語彙統計学データによる分類の諸問題に取り組み始めた際に、浅井(1974)の論考の存在をご教示してくださった高橋靖以先生にも感謝いたします。最後に、編集委員会の皆様ならびに貴重なコメントを下さった匿名の査読者の先生方に心より感謝を申し上げます。

参考文献

【日本語文献】

浅井亨 (1974)「言語から見た地域集団」新野直吉・山田秀三(編)『北方の古代文化』119-142. 東京: 毎日出版社

伊藤隆 (1987)「類似度・共有度・距離－柴田・熊谷のネットワーク法との関連で－」『国語学』151 集. 41-42.

切替英雄 (2010)「日常アイヌ語と口承文芸のアイヌ語」『国文学 解釈と鑑賞』75 巻 1 号. 61-65.

国土交通省国土地理院 (2019) 地理院地図. URL: https://maps.gsi.go.jp (2019 年 7 月 6 日閲覧)

国土交通省国土地理院 (2020) 地理院地図. URL: https://maps.gsi.go.jp (2020 年 2 月 3 日閲覧)

柴田武・熊谷康雄 (1985) 「言語的特徴による地域分割のための「ネットワーク法」－特に NT-1(r)につい

てー」『国語学』140 集. 45-60.

柴田武・熊谷康雄 (1987)「ネットワーク法における地点間の言語的類似の新しいとらえかたと処理のしかたー言語的特徴による地域分割のためのネットワーク法Ⅱー」『国語学』150 集. 1-14.

柴田武・熊谷康雄 (1988)「ネットワーク法における「共有度」と「距離」の定義と計算についてー伊藤隆氏の《短信》に答えるー」『国語学』152 集. 66-68.

鳥居龍蔵 (1903) 『千島アイヌ』 東京: 吉川弘文館

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 巻. 1-17.

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 巻 4 号. 307-342.

服部四郎編 (1964) 『アイヌ語方言辞典』 東京: 岩波書店

濱田武志 (2019) 『中国方言系統論: 漢語系諸語の分岐と粤語の成立』 東京: 東京大学出版

深澤美香 (2017) 『加賀家文書におけるアイヌ語の文献学的研究』 博士学位論文. 千葉: 千葉大学

村山七郎 (1971) 『北千島アイヌ語: 文献学的研究』 東京: 吉川弘文館

安本美典 (1995) 『言語の科学』 東京: 朝倉書店

【英語文献】

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

Bryant, D., and Moulton, V. (2004) Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution, 21*(2), 255-265.

de Leeuw, J., and Mair, P. (2009) Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software, 31*(3), 1-30.

Ding, C., He, X., Zha, H., Gu, M., and Simon, H. (2001) A min-max cut algorithm for graph partitioning and data clustering. In Cercore, Nick, Lin Tsauyoung, and Wu Xindong (eds.), *Proceedings of the first IEEE International Conference on Data Mining (ICDM), 1,* 107–114. Washington: IEEE Computer Society, USA.

Dobson, A. J., and Black, P. (1979) Multidimensional Scaling of some Lexicostatistical Data. *Mathematical Scientist, 4,* 55-61.

Dyen, I., Kruskal, J. B., and Black, P. (1992) An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society, 82*(5), 1-132.

Fellows, I. (2018) wordcloud: Word Clouds. R package version 2.6.
URL https://CRAN.R-project.org/package=wordcloud

Gondran, M., and Minoux, M. (2008) *Graphs, Dioids and Semirings: New Models and Algorithms.* New York: Springer Science+Business Media.

Kirikae, H. (1994) Pa/ca correspondence between Ainu Dialects: A linguistic-geographical study. *The proceedings of the 8th international Abashiri symposium: Peoples and cultures of the boreal forest, 8*, 99-113, Abashiri: Hokkaido Museum of Northern People, Japan.

Kruskal, J. B. (1964a) Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis.

Psychometrika, 29(1), 1-27.

Kruskal, J. B. (1964b) Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*(2), 115-129.

Olshannikova, E., Ometov, A., Koucheryavy, Y., and Olsson, T. (2015) Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data, 2*(1), 1-27.

Ono, Y. (2019a) The Ordinal Scale on Lexicostatistical Data in Ainu Dialects: Towards a New Interdisciplinary Research among the Humanities and Statistics. *Journal of the Center for Northern Humanities, 12*, 89-110.

Ono, Y. (2019b) Observations on “Northeastern” Hokkaido Ainu Dialects: A Statistical Perspective. *Northern Language Studies, 9*, 95-122.

Ono, Y. (2020a, to appear) How to Handle “Missing Values” in Linguistic Typology: A Pitfall in the Statistical Modelling Approach. *Northern Language Studies, 10.*

Ono, Y. (2020b, to appear) Reconsideration of “Major Division” of Ainu Dialects: A Statistical Reanalysis of Asai (1974). *Northern Language Studies, 10.*

Ono, Y. (2020c, to appear) Some Remarks on Cognacy Judgments of Ainu Dialects: On Asai (1974). *Journal of the Center for Northern Humanities, 13.*

Ono, Y. (2020d) On the Relationships among Hokkaido Ainu Dialects and Sakhalin Ainu Dialects: A Statistical Observation. Manuscript in preparation.

Ono, Y. (2020e) Some Remarks on Dialect Classification in Lexicostatistical Research: An Exercise on Ainu Dialects. Manuscript in preparation.

Shi, J., and Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888-905.

Starostin, G. (2010) Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship, 3,* 79-116.

Swadesh, M. (1955) Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics, 24,* 121-137.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing, 17*(4), 395-416.

執筆者紹介

氏名：小野　洋平（おの　ようへい）
所属：放送大学大学院文化科学研究科文化科学専攻