北方言語研究 特別号: 47-60（日本北方言語学会 2022）

人文学と情報の過去と現在、そして未来
―人文学におけるデータサイエンスの概念の整理から考える―

小 野　洋 平
（放送大学大学院文化科学研究科文化科学専攻修士選科生）

キーワード：人文学、情報処理、データサイエンス、循環論、数理人文学

1. はじめに

人文学と情報の未来は、短期的には克服すべき困難はあるが、長期的には明るいだろう。インターネットでやり取りできる通信は、容量を飛躍的に増すだけでなく、高速化が進んだことで、情報の規模は大きくなり、種類も文字だけでなく動画や VR のように多彩であり、センサー技術の発展により時間空間による変化に富むようになった。

しかし、「情報」という言葉が指す現象自体は日々変化しており、企業や研究の当事者においても混乱が生じている。「データサイエンス」では、「情報処理技術」と「統計科学」のいずれかを指すか、実際には明瞭でない。情報に関連する概念の混乱は今後も続くだろう。

一方で、人文学は概念の変遷を重視し、歴史的または思想的に変化を整理する。さらに、概念の定義を吟味し、概念間の関係を整理することにも長ける。例えば、言語学はその代表といってもよいだろう。

よって、人文学は、概念が混乱しやすいという情報の欠点を補うことができる。そもそも、情報を扱う分野の一つである統計科学は、哲学・思想の塊である。例えば、不偏性、一致性、有効性、漸近有効性、頑健性など統計科学の教科書で出てくる諸概念は、統計科学の様々な現場で直面した問題を思想として整理したものである。

そのため、統計科学の優れた研究では、序論にて先行研究の概念の整理をおこない、既存の概念に足りなかった新たな視点を持つ概念を提案することが多い。だが、先行研究の概念を理解するにも、新たな概念を提案するにも、数学や統計科学、確率論の深い知識を必要とする。よって、統計科学の哲学や思想は敷居が高い。しかし、実りは多いだろう[1]。

本稿は、人文学とデータサイエンスの関係を例として、人文学におけるデータサイエンスの概念を整理し、先行研究に足りなかった新たな概念を提案し、新たな知見をもたらす一連のプロセスを通じて、人文学と情報の発展の可能性を例示する。

まずは、データサイエンスの現状を簡単に整理してみよう。

[1] もっと広い意味でも、人文学が今の情報やデータサイエンスには必要だと著者は考える。後述のように情報処理技術を用いて、様々な作業の自動化によりコストを削減することを最近は「データサイエンス」という。例えば、レジの会計や注文の自動化が急速に進み、人件費は大幅に削減された。情報処理技術のこの使い方は有用であるが、一定の限界がある。確かに、コストは削減され、機械にでもできる仕事から人間は解放されつつあるといえるかもしれない。しかし、これまでの労働から解放されて、もっと新しい「創造的な」仕事を人間はすぐには習得できない。産業構造の変化によって、労働市場にて需給と供給のミスマッチが生じるのは常であるが、これからの「データサイエンス」は、現状の限界を乗り越えるため、何のために情報処理技術を用いるか更に考え、社会がより良くなるとは何であるか立ち返り、どのように情報処理技術を使うべきか、経験ではなく歴史から学ぶ必要があるのではないかと思う次第である。

2. データサイエンスの昔と今

2.1 言葉の流行りと廃り：モノづくり立国からデータサイエンスへ

著者が子供の頃には、モノづくり立国という言葉が流行りであった。最近は、耳にしなくなった言葉である。科学技術を中心としたモノづくりに関する知識は、統計科学の現場では必須であるので、またモノづくり立国という言葉が流行るのではないかと期待している。

最近は、データサイエンスが流行っている。著者が統計科学専攻の大学院に入学した頃は、データサイエンスという言葉は聞きなれないものであり、今では想像もできないだろうが、統計学や統計科学は日陰の存在であった。大学時代に統計学を教わった先生も「海外にいたころは統計学者というだけで相手が無言になったよ」などと嘆いていたぐらい、実生活との関係のイメージに乏しい学問であった。

ところが、2010 年以降、ニューラルネットの性能が飛躍的に向上し、人間の能力を凌駕する成果を次々と挙げ、モノづくりだけではなくデータが価値を生むという考え方が海外を中心に浸透し始めた。そして、日本においても重要性が認識され、政府が 2025 年までに文理を問わず、すべての大学でデータサイエンス教育の体制を整備することを目指すなど、データサイエンスは時代の申し子となった[2]。

2.2 「データサイエンス」の概念の混乱：情報処理技術と統計科学の違い

しかし、今のところデータサイエンスには 2 つの意味がある。1 つはデータサイエンスの道具である情報処理技術を指す。たとえば、プログラミング、データベース、システム設計、ネットワーク、画像処理、音声認識、ソフトウェアなどである。特に、産業関連において、これらの情報処理技術が様々な作業を自動化し、コストを削減することが期待される。

もう 1 つは、データサイエンスの理論である統計科学、機械学習などを指す。学問は理論がわかっていないと道具を正しく使えない。データサイエンスでも同様に、統計科学や機械学習の正確な知見は情報処理技術の適切な利用に必須である。

著者が大学院生の時は、情報処理技術は各自が必要に応じて自習することが基本であり、大学院では統計科学や機械学習を研究していた。しかし、データサイエンスが情報処理技術を指すことが、最近は多くなった。

具体的には「人文学におけるデータサイエンス」も、画像処理や音声認識を人文学の資料/史料に適用し、データベースの構築を通じて、人文学の研究に貢献する方向性がメインである。人文学では、情報処理技術によるデータベースの構築は長い歴史がある。その一方で、波及効果を考慮したプロジェクトは少なく、作成したデータベースが実際の研究にあまり使われず、データベースの構築を繰り返してきた。

著者は「人文学におけるデータサイエンス」を、人文学データに適した統計手法の確立という意味で、一貫して用いてきた。本稿も「人文学におけるデータサイエンス」を特に断りのない限りこの意味で用いる。

---

[2] そんな中で、私は統計科学専攻の大学院にいながら、人文学へのデータサイエンスの応用に興味がわき、研究を進めていくこととなる。今振り返ると、大学時代の神話学の授業で、プロップの「昔話の形態学」を読んだことが一つの契機になっていたように思う。

2.3 「データサイエンス教育」の混乱：プログラミングは道具である

「データサイエンス」の概念の混乱は、「データサイエンス教育」の混乱にも表れているように見える。例えば、大学で行われているデータサイエンス教育も、その実態はかなりの部分がプログラミング教育になっていると聞く。

著者がデータサイエンスに必要と考える要素を以下に列挙する。(1)数学、(2)統計科学や機械学習、(3)情報処理技術、(4)計算機科学、(5)実質科学の知見（言語学、物理学、農学等）、(6)認知科学（自他の考え方の違いを把握し一致点を見出す力/コミュニケーション力）。

著者の経験上、データサイエンスにより優れた活動や研究を実践するには、上記の 6 点の全てに関して、一定以上の能力が必要である。現状の大学教育は、情報処理技術の習得に力点を置きすぎているのではないか。

なぜなら、プログラムを書ける人が何十人集まっても、どんな数理に基づき、どんな統計手法や機械学習をどれくらいの誤差を目指して、どんなアルゴリズムによって実装し、完成したプログラムが当該分野にどんな知見をもたらし、どんな社会貢献ができるかを見通すことはできない。そして、データサイエンティストの役割とは、実質科学の専門家との協働を通じて見通しを立て、上記を実現することである。データサイエンスは茨の道である。

特に、ソフトウェア工学者を目指したり、プログラムの保守性や安全性、安定性に入学時から関心があったりする場合を除けば、自学自習に適した書籍はプログラム教育の成果によりいくらでもあるので、せっかく学費を払い大学に通っているのにわざわざプログラムの授業を受けるのは何とも時間がもったいないような気が、著者にはする。

2.4 人文系の学部学科でのデータサイエンス：学生の関心とカリキュラムの乖離

情報処理技術の習得を中心とした現状のデータサイエンス教育が、果たしてどれほどの効果をあげるかは、今後の検証が必要であろう。著者は、人文系の学部学科で情報処理技術やデータサイエンス教育を実践するには、教育カリキュラムを見直す必要があると考える。なぜなら、高校までの数学教育は、微分積分などの解析学が中心である。このカリキュラムは、自然科学に関心がある生徒には有効である。一方で、人文学におけるデータサイエンスでは解析学よりも、集合と位相や代数の基礎的な理解が重要であると著者は考える。実際に、情報処理技術を理解するには解析学が必要となるが、人文学におけるデータサイエンスで高度な解析学の知識が要求されることは少ない。

現在の人文系の学部学科では、コーパスを用いたソフトウェアによる形態素解析などの自然言語処理がデータサイエンス教育として実施されているが、果たして人文的なものに関心のある多くの学生にとってコーパス言語学は教育の素材として相応しいのだろうか。

2.5 データサイエンス教育のこれから：現場も何をやるべきか混乱している

データサイエンスの現状について、人文学との関連から簡単に整理した。特に、人文系の学部学科ではデータサイエンス教育で何をやるか定まっていない。なぜか。

実は、「人文学におけるデータサイエンス」が、どのような対象についてどのような論理・科学基礎論によって、何を目指して研究していくべきか、規範（ディシプリン）がない。

そのため、「人文学におけるデータサイエンス」を教えようにも、規範が定まっていない

ものを教えることになり、とりあえずプログラミング教育という形になっていると著者は考える。「データサイエンス」だけでなく「データサイエンス教育」も混乱している。

そこで、人文学の視点の出番である。まずは、人文学におけるデータサイエンスについて、歴史的変遷を整理してみよう。

3. 人文学におけるデータサイエンスの歴史

人文学におけるデータサイエンスは、古典期、過渡期、現代の 3 つに大きく区分できる。

3.1 古典期：忘れられた精緻な理論

古典期は、大まかに 1980 年以前を指す。この時期は、計算機の性能が複雑な分析の実行には不十分であったため、データの性質に関する数理的な研究が発展した。言語学者だけでなく、数学者や物理学者などにより学際的な研究がなされた。複雑な計算がコンピュータにより容易に実行可能となった現代において、古典期の研究は極めて示唆に富み、新たな展開を人文学にもたらす可能性が高いが、現代では忘れ去れている。

例えば、アイヌ語方言の統計分析に関し、服部・知里 (1960)と Asai (1974)が古典期の双璧である[3]。特に Asai (1974)は、人文学におけるデータサイエンスの偉業であるが、90 頁以上の大作であり数理的にも難解な部分があるため、今日まで他の言語の研究にて参照されることは少なかった。当時、北海道大学の計算機を用いて統計分析を実行しようとしたが、大学の計算機の利用許可が下りなかった。しかし、計算機を利用できなかったことが、逆に、アイヌ語方言の分類に関して、数理の観点から妥当な基礎づけを行う契機となった (Asai 1974: 59)。まさに、Asai (1974)は執筆や構成の経緯からして、古典期らしい論文である。

3.2 過渡期：計算技術者の軽視が招いた分野の衰退

過渡期は、大まかに 1980 年以降を指す。この時期は、性能の高い計算機が比較的安価になり、複雑な計算を実行できるコンピュータを、人文系研究者が利用できるようになった。古典期の理論を実際の調査データに適用し、計算結果を既存研究と比較・検討することで、計算機に基づく新たな展開が人文学にもたらされた。

例えば、社会調査で発展した林の数量化は、当時の計算機にとって高い計算能力を必要とする固有値問題を解く必要があり、古典期において分析を実行することは容易でなかった。しかし、過渡期には、林の数量化が日本語方言の語彙データに適用され、成果を上げた。

一方で、今日から振り返ると、人文学におけるデータサイエンスは過渡期において徐々に衰退していったと見ることもできる。

なぜ衰退したのか。それは、計算機による統計分析の正確な実行について技術者の貢献が

[3] 浅井亨 (1930-2006)は言語学者、医者であると同時に人文学におけるデータサイエンスの先駆者である。既に 1974 年の時点で、浅井亨は Asai (1974)においてアイヌ語基礎語彙統計学の類似係数を位相の点から考察し、相関係数を用いる因子分析の適用に疑問を呈している。この考え方は、人文学データの数理上の性質に対応した統計手法を吟味、選択するべきという数理人文学の先駆である。浅井亨は先駆的な考えを更に発展させ、同年に発表された浅井 (1974)で展開している。浅井 (1974)は、アイヌ語方言の研究においてもあまり認知されていない論考であるが、独創性に富んでおり精読の価値がある。興味のある読者は是非図書館などで入手されたい。

正しく評価されなかったことによると著者は考える。今でも、計算というと、決まりきった手順をただ実行していくだけで、独創性がなく機械的な作業と思われる方もいらっしゃるだろう。しかしながら実は、コンピュータを正しく使い統計分析を正しく実行するためには、相当な専門知識を要する。

例えば、1000000+0.1=1000000, 1000001-1000002=0, 3*(1/3)=0.9999997 というようなことが昔の計算機では必ず生じた。また、統計学でよく用いられる「分散」を計算する際、データを Xi, {i=1, 2,…,n}とすると、「分散」には以下の 2 つの計算法がある[4]。

「分散」= (1/n)*(X1^2+X2^2+…+Xn^2)-((1/n)*(X1+X2+…+Xn))^2

「分散」= (1/n)*((X1-(1/n)*(X1+X2+…+Xn))^2 + (X2-(1/n)*(X1+X2+…+Xn))^2 +…
+ (Xn-(1/n)*(X1+X2+…+Xn))^2)

式の形からみると、前者の方が簡単で早く計算できそうであるが、前者と後者の計算結果は異なり、前者の方が後者よりも正しい計算結果と食い違いが大きい。

昔の計算機は、今よりも限られた桁数で２進数の計算をする必要があったためである。

このように、コンピュータにより統計分析を正しく実行することは、独創性がない単なる機械的な手続きではない。しかし、過渡期では、「コンピュータは人文学の奴婢である」といった発言や、論文において、実際に計算を担当した研究者に謝辞で感謝を述べるだけで、論文の共著者にしていないといった例も散見される。

今日から振り返ると、計算技術者の貢献の軽視により、計算を正確に実行できる研究者が少なくなり、人文学におけるデータサイエンスは衰退した。日本では、人文学の研究は単独で行うことが基本であり美徳でもあるが、過渡期に生じた衰退の原因から、人文学におけるデータサイエンスの今後のあり方に関して学ぶことは大きい[5]。

3.3 現代：人文学とデータサイエンスの循環論

現代では、計算機の能力が極めて高くなり、価格も個人が気軽に購入できるものとなった。過渡期との違いは、複雑な計算があらかじめソフトウェアに組み込まれ、組み込まれた関数の組み合わせによって、計算機科学の知見がなくとも複雑な分析が可能となった点にある。

例えば、Excel には平均や分散などを AVERAGE()や VAR()により計算できるが、利用者は前述の問題を気にする必要はないだろう。SPSS, SAS, R, Python といったソフトウェアでは、より複雑な関数が組み込まれ詳細な分析が可能である。

しかし、分析が容易になった代わりに、様々な手法の適用により分析結果が無限に生成し得るのが現代の特徴である。自然科学や社会科学などでは、分析の結果を「予測の良さ」、「予測精度」から取捨選択できるため、データサイエンスの規範は保たれた。

一方で、人文学では、ソシュールの事例のように「予測」から分析を評価すること自体は原理的に可能である。ただし、自然科学や社会科学と比較して、遺跡の発掘や新史料の発見

[4] Xi^2 は Xi の 2 乗=Xi*Xi を表す。

[5] 著者の研究は、本稿を含めて、人文学におけるデータサイエンスを、古典期の視点を再考し過渡期や現代の先行研究に代替案を提示することが多い。研究の性質上、再考察や再分析とは先行研究のネガティブな側面を強調しているように見えるかもしれない。しかし、著者の本心においては、人文学におけるデータサイエンスを実践した先駆者の苦労は、計算技術やデータベースの作成など、身に染みてわかる。ただ、論文という形式の文体上の制約により、文章がややきつくなってしまう。その点、ご寛恕いただきたい。

などが必要な場合も多く、分析を予測の点で検証するための時間が比較的長い。

よって、実際には、分析結果を予測の点から評価するのではなく、人文学の知見から評価せざるを得なかった。このことにより、人文学とデータサイエンスが循環論となった。

すなわち、妥当な分析とは、人文学において妥当とされる結果と一致するものとなった。これでは、人文学のデータにデータサイエンスを適用する意義がない。人文学データに統計手法を適用した多くの研究で、人文学の分析の妥当性を示す根拠として、先行研究のデータサイエンスの結果を引用し、データサイエンスの分析の妥当性を示す根拠として、先行研究の人文学の結果を引用する状況である。循環論となっている。これでは、当該分野の研究を深めていった方がよい（図 1)。

更に、循環した議論がネットワークとなり、現在の知見は形成されている。初期の人文学あるいはデータサイエンスの研究が見直され、より妥当な知見が得られると、ネットワークにおいて再検討を要する研究は、幾何級数的に増える。この過程が繰り返されている。

著者は、これまでの研究で、人文学におけるデータサイエンスがこの循環論から脱却する方法論を検討してきた。人文学データの数理上の性質を検討し、データの性質と適切に対応する統計手法の取捨選択こそが、循環論から脱却する新たな規範と考える。本稿では、この新たな規範を「数理人文学」と仮に呼ぶ。

次節では、数理人文学に至る道筋を、具体例を通じて詳しく説明する。

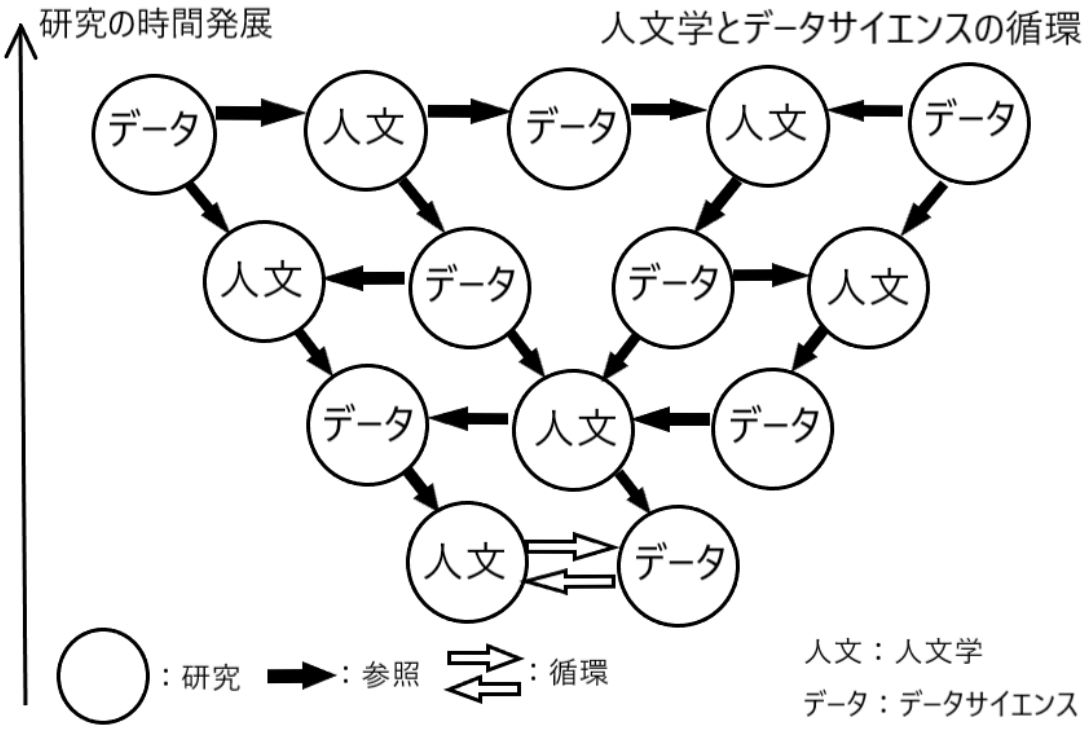

図 1. 人文学とデータサイエンスの先行研究の関係. いずれの先行研究も初期の循環関係を参照している. 初期の人文学またはデータサイエンスの知見が再検討を要する場合、すべての研究を見直す必要がある.

4. データサイエンスの考え方：予測と分類

データサイエンスの考え方を、予測と分類の点から整理してみよう。自然科学では、実験が可能であり分析結果を予測から取捨選択できる。社会科学でも、実験条件をコントロールすることはより難しいが、調査など実験に準ずる方法が実施可能である。

一方、人文学では主に過去の出来事に関心があり、実験がそもそもできないことが多い。よって、人文学におけるデータサイエンスでは、予測ではなく分類に主たる関心がある。

4.1 予測には何が必要か：社会科学の例－マーケティング

マーケティングでは、ある商品 A の売れ行きを、天候や気温、湿度、曜日、時間帯などから予測することが多い。商品 A が惣菜などの保存の効かない食品の場合は、売れ残りは損失に直結する。よって、商品 A の売れ行きを正確に予測し、損失の最小化により利益を最大化することが求められる。

今、データを（商品 A の売り上げ、天候、気温、湿度、曜日、時間帯）の組であらわし、（商品 A の売り上げ、天候、気温、湿度、曜日、時間帯）が 2000 個あったとする。

例えば、（120 個、晴、25℃、30%、水曜日、12 時）、（140 個、雨、12℃、60%、金曜日、18 時）の組が 2000 個あると想像してもらいたい（表 1）。

売れ行きの予測に有力な手法が 6 つあり、商品 A の実際の売り上げと 6 つの手法による売り上げ予測が図 2 となったとき、6 つの手法のどれが優れているだろうか。

表 1. あるスーパーでの商品 A の売り上げと天候条件、曜日、時間帯の仮想データ.

	商品Aの売り上げ(個)	天候(晴/曇/雨/雪)	気温(℃)	湿度(%)	曜日	時間帯
1	120	晴	25	30	水	12
2	140	雨	12	60	金	18
3	220	雪	−2	50	土	19
4	50	晴	32	60	月	9
（中略）	（中略）	（中略）	（中略）	（中略）	（中略）	（中略）
2000	180	曇	17	20	日	18

方法 1 を基準に比較する．方法 2 は、売り上げを過大評価したり過小評価したりするので、予測の値に基づき、商品 A の売れ残りによる損失を最小化できない。方法 3 は、売り上げが小さいときに過小評価し、売り上げが大きいときに過大評価する．予測の値と過大・過小評価に関連があるので改善が見込める．方法 4 は、予測値よりも実際の売り上げが平均すると 5 個多い。予測値に 5 個加えることで、売れ残りを予測できそうである。方法 5 は、予測した値によらず、実際の売り上げとの差が平均すると 0 に近い。予測値は損失の最小化に使えそうであるが、予測した値によって差のバラつきが違う。損失の程度を事前に知って、経営に反映させたい．方法 6 は、予測した値によらずバラつきが一定であり、バラつき程度が方法 1 よりも小さい。よって、方法 6 が有力な手法である。

予測したい値（商品 A の売り上げ）の情報が予め得られることが、予測には必要である。

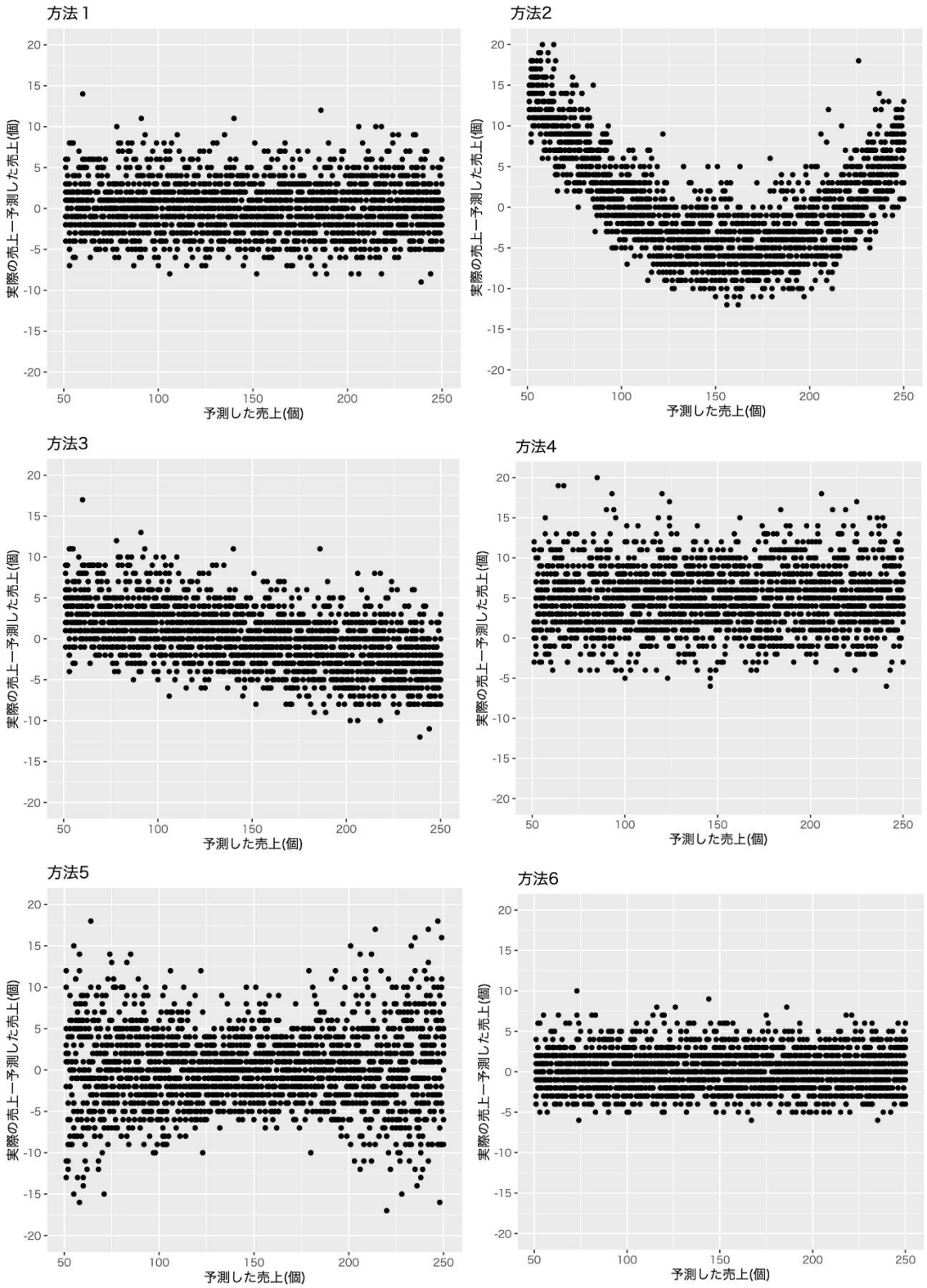

図 2．６つの手法による予測結果．横軸：各手法が予測した売り上げ（個）．縦軸：実際の売り上げから予測した売り上げを引いた予測誤差（個）．上段左：方法１．上段右：方法２．中段左：方法３．中段右：方法４．下段左：方法５．下段右：方法６．縦軸が 0 のとき実際の売り上げと予測が一致する．

4.2 分類の難しさ：人文学の例－方言学

人文学では表 2 のようなデータを扱う。表 2 は、Asai(1974)のアイヌ語方言の類似度を著者が訳した。例えば、八雲(1)と長万部(2)の値は 104 である。よって、110 語の基礎語彙の 104 語で、八雲と長万部は類似した語を少なくとも一つ共有する。

表 2．アイヌ語諸方言の類似度．Asai (1974: 92; Table1)より著者が訳す．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、表 2 から何を予測するか/予測できるか、不明である。

例えば、アイヌ語方言の分類を予測したいとする。予測ができるには、前節の「商品 A の売り上げ」に対応する「予測したい値」が必要である。だが、アイヌ語方言の分類の答えが予め分かっているならば、そもそも予測などする必要はなく、その答えを用いれば良い。

統計学や機械学習の分野では、前節で述べた予測/分類の答えについて「教師」と呼び、「教師あり学習 (supervised learning)」「教師なし学習 (unsupervised learning)」という用語を用いる。表 2 のデータの分析は「教師なし学習」に該当する。

人文学におけるデータサイエンスの難しさとは、「教師なし学習」における困難である。具体的には、「教師なし」のため前節のように「予測の良さ」、「予測精度」の観点から分析結果を取捨選択できず、手法の数だけ分析結果が生じる。

総語数 110 から表 2 を引いたデータをアイヌ語方言の違いとみなし、諸手法を適用した例を、図 3 に示した。

図 3 には同じ結果も一部あるが、手法によって分析結果が異なる。このように分類では、無数の方法から無数の異なる分析結果が生じ、妥当な結果を選択することが難しい。

人文学の知見とデータサイエンスの分析が循環しないように、分析だけから妥当な結果を取捨選択することは一見難しそうである。「議論が循環している」という指摘は、人文学におけるデータサイエンスへの殺し文句であった。著者自身も、以前ある機会に、他分野の研究者の方から同様のご指摘をいただき、問題の重要性を認識し、解決に向けて取り組み始めた。その答えが「数理人文学」であった。

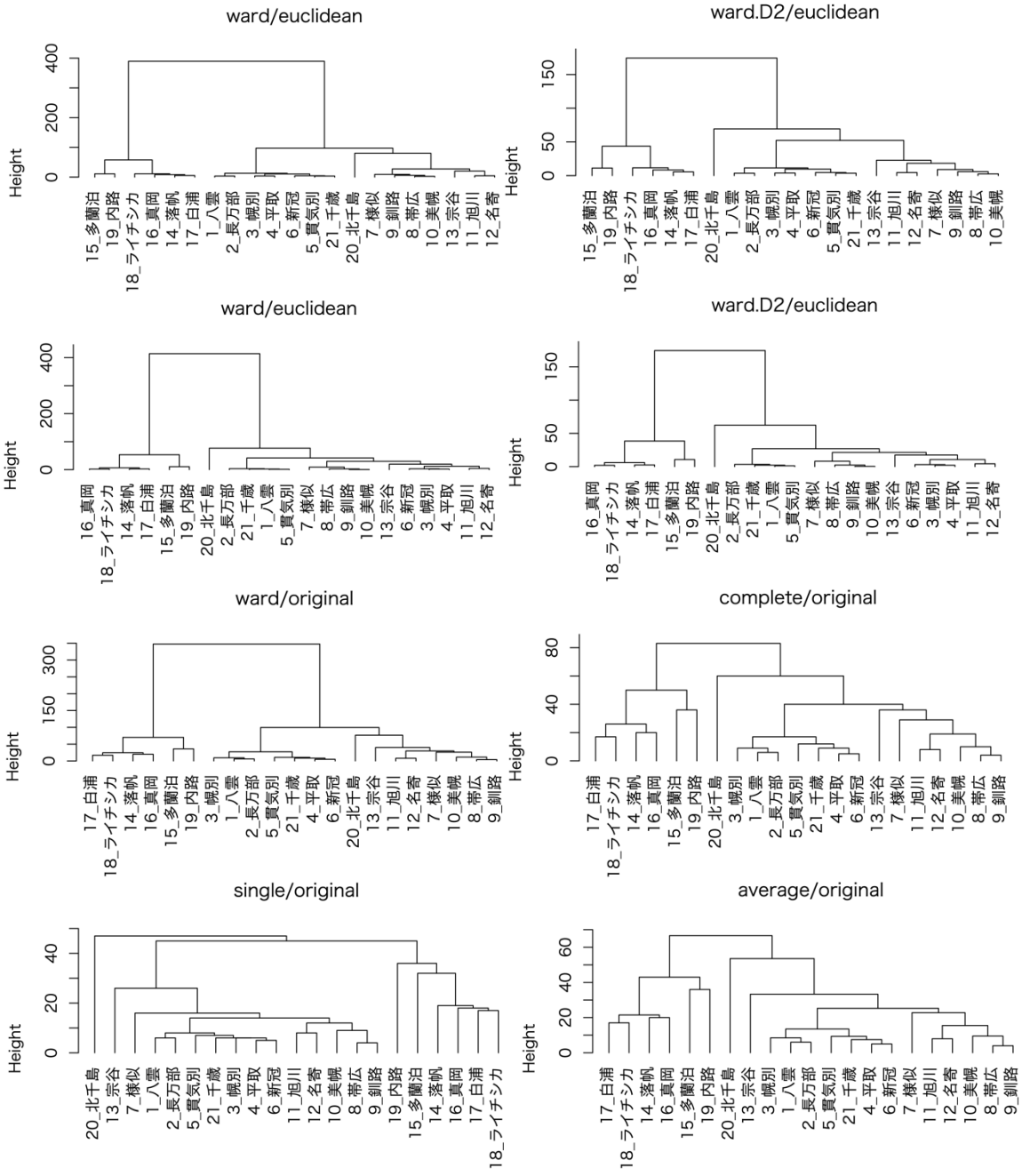

図 3. 表 2 をアイヌ語方言の非類似度とみなし、クラスター分析の諸手法を適用し得られた分類結果.
最上段左: 多次元尺度構成法（次元数 3)、距離行列は Ward 法 (Ward 1963). 最上段右: 多次元尺度構成法（次元数 3)、距離行列を 2 乗し Ward 法. 2 番目左: 多次元尺度構成法（次元数 2)、距離行列は Ward 法. 2 番目右: 多次元尺度構成法（次元数 2)、距離行列を二乗し Ward 法. 3 番目左: Ward 法. 3 番目右: 最長距離法 (Sørensen 1948). 最下段左: 最短距離法. 最下段右: 群平均法.

4.3 循環論の克服：数理人文学への道

人文学データの数理上の性質を吟味し、適切な統計手法を取捨選択するとは、どのような

ことだろうか。表 2 のデータを具体例として説明する。

まずは、類似度の定義を検討しよう。Asai (1974: 61-62)の定義を簡単に訳せば、「1 語において、方言同士が類似した語を少なくとも一つ共有するとき 1」、「1 語において、方言同士が類似した語を一つも共有しないとき 0」である。1 の定義において「少なくとも一つ」という条件がある。この条件が類似度の代数構造に制約を与える。

実は、表 2 の類似度へは小学校で習う「足し算/引き算」だけしか適用できない。小学校と中学校の「足し算/引き算」は代数構造が異なる。小学校の算数では、「5-2」の答えを「2 を足したら 5 になる数」と習い、「2-5」を計算しない構成である。

小学校までの「引き算」は「足し算」なのである。

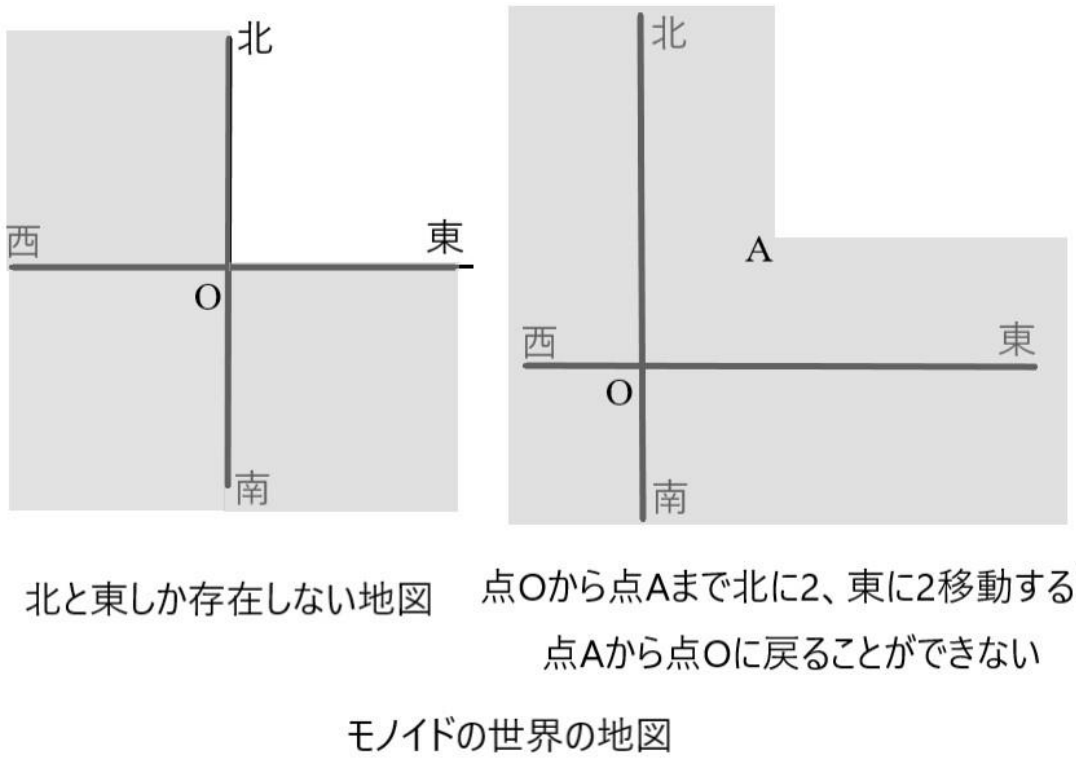

図 4. モノイドの特徴を表した地図．灰色の領域には、負の数がなく辿り着くことができない．

ところが、中学校の数学では最初に「-1」や「-2」といった負の数が定義され、「2-5」の答えは「-3」となる。そして「(-3)*(-3)=9」のように、負の数同士に掛け算が定義される。これは中学生にとって一大事である。「2-5」は「5 を足したら 2 になる数」であり、小学校の算数にはなかった数である。また、「0」が最も小さい数であったのに、「0 より小さい数」が無限に存在する。小学校の算数と中学校の数学では、数の性質や計算規則が違うのである。

ところで、小学校のとき、「足し算/引き算」を使って色々と計算をしたと思う。お小遣いで欲しいものが買えるかなどは、なかなか難問であった。しかし、その「足し算/引き算」で実は事足りていた。また、小学校の「足し算/引き算」で計算の最中に「2=3」といった矛盾が生じないことは代数学が保証してくれる。もっとも、著者は計算間違えをよくやったが。

小学校の「足し算/引き算」をモノイドという。一方で、中学校の「足し算/引き算」は群という。Asai (1974: 61-62)の 1 の定義から、-1 を「-1 語において、方言同士が類似した語を少なくとも一つ共有するとき-1」と定義すると、1+(-1)=0 より「0 語において、方言同士が類似した語を少なくとも一つ共有するとき 0」が 0 の定義となる。これは本来の 0 の定義「1 語において、方言同士が類似した語を一つも共有しないとき 0」と異なる。

よって、表 2 は負の数がないモノイドであり、小学校の「足し算/引き算」が実行できる計算規則となる。地図の世界でモノイドを考えてみよう（図 4)。この世界では、北と東しかなく、北にも東にも 0, 1 ,2,…,N しか移動できない。

そのため、この世界では「北に 2、東に 2」動くと、もとに位置に戻れない。もとの位置に戻るには「北に-2、東に-2」動く必要がある。しかし、負の数がない。

多くの統計手法では、方言を地図上の点として捉える方法を採用している。この地図では負の数の存在を前提する。よって、表 2 を既存の統計手法で分析した結果は、アイヌ語方言の地図には「負の数があること」になってしまう。

数理人文学とは、簡単に言えば、「負の数がない」表 2 のデータに「負の数を前提とする」統計手法を適用することは、存在しない負の数を存在するかのように分析するため、分析を採用せず、「負の数を前提としない」分析結果の中から数理上の点でより妥当なものを追究する新たな学問の規範である。「負の数を前提としない」分析手法を用い、表 2 を分析する詳細は和文では小野 (2020)に、英文では Ono (2020)に述べた。

上記の議論では、アイヌ語学の方言分類に関する知見を参照しなかった。よって、議論は循環していない。このように、人文学データの数理上の性質を検討し、対応する統計手法を取捨選択し、より妥当な結果の提示が論理的に循環せずに数理人文学にはできる。人文学とデータサイエンスの循環論からの脱却を可能とする学術領域であると、著者は考える。

5. 人文学におけるデータサイエンスの今後：数理人文学の継続は可能か

数理人文学において、習得すべき内容は多い。数学と人文学は、どちらも積み重ねが重要な学問である。自然科学では、最先端の知見が日々更新される。よって、基礎の習得により先端研究への参入が可能となる。一方、人文学は、膨大な資料や先行研究を読みこなして、やっと先端が見えてくる。数学もまた然り。どちらも長い積み重ねを必要とする学問である。

だからこそ、数理人文学は長期的には豊穣な領域である。一方で、上記の理由により、浅井亨のような研究者の特異性が頼りであった。数学も人文学も半端な著者がこのように提言するのは誠に恐縮であるが、研究者の特異性にたよらず、数理人文学が領域として継続に発展するには、教育カリキュラムの作成から始める必要があるだろう。

そもそも、教科書がないのである。データサイエンスの分野では、最初に統計学の教科書を読むのが通例である。しかし、本稿で扱ったモノイドは統計学の教科書で紹介されることすら、まずない。数学の教科書でも、「群」「環」「体」といった主流に隠れている。

教科書がない限り、やはり後進が育ちにくい。著者などは、統計学の教科書を読んでいき、統計学の教科書の内容が人文学におけるデータサイエンスに不向きである理由がわかって、初めて数理人文学の重要性に至った。教科書通りに研究がうまくいかないことは、人文学に限らず、どんな分野でもよくあることだが、まずは読書案内などを整備できればと思う。

6. まとめ：人文学と情報の未来

本稿では、人文学におけるデータサイエンスを例に、人文学と情報の未来の例示を試みた。人文学が、概念の変遷を重視し、歴史的または思想的に変化を整理する点や、概念の定義を吟味し概念間の関係を俯瞰することに長ける点は、今後も、概念が混乱しやすい情報の欠点を補うだろう。実際に、人文学におけるデータサイエンスの歴史的変遷を振りかえることで、論理的な循環が明らかとなり、数理人文学に至る過程を本稿では示した。

本稿は、著者が大学院入学以降、人文学へのデータサイエンスの応用に興味を持つ中で、直面した課題を人文学の視点を用い、どのように現在の数理人文学という考えに至ったかを描いた個人史の記録でもある。研究で壁にぶつかったとき、歴史や思想といった人文学の視点を取り入れることが、どなたかの何かの参考になれば幸いである。

今日、情報学や統計科学、データサイエンスといった理数系の学問が盛んであり、人文学はあまり元気がない。そんな流れになんだかついていけない、文理の狭間をただよう人間のささやかな抵抗が伝われば、またこれも幸いである。

7. 終わりに：切替先生との思い出

2020 年の冬、そして翌年の年初に津曲敏郎先生と切替英雄先生が亡くなられた。謹んで、津曲先生と切替先生のご冥福をお祈り申し上げます。津曲先生とはお話する機会も少なく、また切替先生とお会いできたのは一回だけであるが、その思い出を記す。切替先生に初めてお会いしたのは、2015 年の北海道大学での国際シンポジウムであった。発表が終わった後、突然、「君の研究はかくかくしかじかだったから是非これから飲もう」と仰り、他の先生に「こいつとこれから飲んでくるから」といった感じで断りをいれ、15 時ごろから 24 時近くまで札幌の街で飲みに連れてくださったのが切替先生であった。見ず知らずの大学院生に本当に実直に色んな話をしてくださり、こんな先生がいらっしゃるのか、と心から驚いた。

研究の話も迂遠ではなく、明快であった。今日は誰に会ったかと聞かれ、津曲先生の名前を挙げたところ「津曲は立派な学者だ」と即答だった。その頃は、まだ何も知らなかったが、津曲先生と切替先生の深い絆が感じられた。

「俺は言語学を知らない人間でも深く考えればその意味がわかる、そんな論文を目指している」と仰っていたことも印象的であった。特に切替 (2017)は、内容だけでなく文章の形式美も備えた名論文である。たった 13 頁の短い世界に、これだけの内容を、様式美まで伴い収めるのは、藝の域である。

帰り際、切替先生がぼそっと一言仰った。「死ぬなよ」と。驚いて何故かと尋ねたところ「だってお前全然笑わないんだもん」と。たった一日で、当時の私の状況を見抜いていたのである。切替先生のこの一言の重みで、なんとか今日まで研究を続け、この論文を書くことができた。灰色の青春というが私の 20 代の記憶はほとんど色がない。しかし、切替先生と飲み歩き話した思い出には、今でも鮮やかな色がある。

再びお会いすること叶わず、痛恨の極みである。

参照文献

【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

小野洋平 (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―」『北海道言語文化研究』 18: 19-46.

切替英雄 (2017)「アイヌ語の作意の見えない言語形式」『北方人文研究』 10: 105-117.

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.

【英語文献】

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

Ono, Yōhei (2020) Reconsideration of “major division” of Ainu dialects: A statistical reanalysis of Asai (1974). *Northern Language Studies*, 10: 231-254.

Sørensen, Thorvald (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. *Biologiske Skrifter*, 5: 1-34.

Ward Jr, Joe H. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301): 236-244.