異なる文献間の数理的な比較研究をふり返る

師　茂樹（花園大学）
s-moro@hanazono.ac.jp

要旨: 数理モデルを用いて異なる内容の文献（特に古典文献）を比較分析する研究について、その方法論を中心に研究史を概観するとともに、問題点を指摘する。
キーワード: 計量文献学，数理文献学，自然言語処理

1. はじめに

文献の研究は、文学・歴史学・哲学をはじめとする人文系諸学の多く分野で行われている。そして 1949 年に Roberto Busa 神父がトマス＝アクィナスの著作の索引の開発をコンピュータ上で開始して以来、文献学は人文学におけるコンピュータ利用の中心的な課題であった[1]。

文献学においては、同一の文献の異本を比較することでそのオリジナルや写本間の系統的な関係などを見出すことが行われているが、一方で異なる文献間での比較を通じて影響関係などを検討することも行われている（例えば、同一作家の異なる作品を比較することで、その間の心境の変化を見たり、異なる思想家の文献を比較することで、両者の師弟関係を推測したりするなど）。文献の比較研究でコンピュータを利用するメリットとしては、研究活動の機械化によるヒューマンエラーの削減や作業の効率化（時間の短縮など）もあげられるが、方法論的には①研究者が気づかない[2]、あるいは抑圧してしまう[3]規則性の発見（知識発見、仮説形成、テキストマイニング）と、②仮説（モデル）の提示とその検証による（しばしば人文学と対比的に論じられる[4]）科学的方法であるという点が特に重要視されているように思われる。

本報告では、異なる文献間の比較研究に焦点を合わせて、そこでコンピュータを利用した研究史を振り返りたいと思う。ただし、筆者の能力と紙幅による制限のため、すべての研究を網羅するものではなく、特に国内の研究、東アジアの文献の研究に偏っているであろうことはあらかじめお断りしておきたい。また、研究者による読解、あるいは複数の文献に対する横断検索や KWIC などの結果を通じて、研究者が文献間の関係を判断し記述するというような方法[5]も、コンピュータを用いた文献比較の方法のひと

つと言えるであろうが、キーワードの選出や結果の判断などにおいて研究者への依存度が大きいこれらの方法は取り扱わない。ここでは、文献を比較するモデルを網羅的に文献に適用するような方法、すなわち研究者の恣意性が低く、研究者への依存度が比較的低い方法に限定する[6]。もちろん、どの文献を比較するのか、どのように文献データベースを構築するのか、どの比較モデルを適用するのか等々において、研究者の恣意性を完全に排除することはできないので、あくまでも程度問題であることは付言しておきたい。

文献学においては、モノとしての写本から推測する「外的証拠」と、書かれた内容から推測する「内的証拠」の二つが用いられるので[7]、本稿でも研究史の分類にこの枠組みを用いる。もちろん両者は互いに関連しあっているものであり、はっきりと分けられるものではない。また、両者のどちらを重視するのかは、学問領域により異なる点も注意しなければならない。

2. 外的証拠による比較

文献の比較においては、成立年代や地域的分布などを検討する場合があるが、現在のところ、時空間情報と結び付けられた文献データベースはほとんど存在しない[8]。

また、筆跡や紙質などが文献の比較研究においては重要な情報を提供してくれる場合がある。前者については、文字の用例データベースがいくつか存在するものの[9]、文献の比較研究を目的としたものではない。文献画像からの文字の自動切り出しなどが実用化されるようになれば、筆跡による比較研究も進むかもしれない。また後者については、文献データベースのメタデータとしてデジタル化されている例もあるが、これを用いた文献の比較研究はまだ見られないようである。

文学作品や哲学書など文献学的に研究している者にとっての「文献」は、そこに書かれた内容の方に重心があり、モノとしての文献の位置づけは内容読解のために必要な前段階であってゴールではない、という意識が強いのではないかと思われる。書誌学や古文書学が文献学や文献史学の「補助学」という位置づけで分類されていることからも、モノとしての文献の位置づけはわかるだろう。したがって、コンピュータを用いた文献研究においても、これらの情報があまり活用されていないのではないかと思われるが、GIS やメタデータの重要性が広く認識されるようになってきた今日、文献学的研究においてももっと積極的に活用されてよいのではないかと思われる。

3. 内的証拠による比較[10]

3.1. 表記の特徴

欧米語圏では、単語の長さや文の長さ、単語や品詞の分布などを統計的に分析することで複数文献を比較し、著者などを推定する研究が古くから行われてきている[11]。東アジアの言語についても同様の研究がなされている[12]。

3.2. 文字列・単語列などの共起関係

文字や単語の共起関係に基づいて文献の特徴を見出し、それによって文献を比較する研究は数多く行われている。もっとも、同じ「共起関係」と言っても、どのような関係なのか（隣接しているのか、一定範囲内での登場なのか、など）、共起関係が何を意味するのか等々が研究ごとに異なる点には注意が必要であろう。

また単語の共起関係については、分かち書きになっている言語や形態素解析の結果が比較的安定している言語においては研究が進んでいるが、そもそも文法についての知識が乏しいような古典語の場合、形態素解析についての研究が進んでいないうえ[13]、外部データベースの整備や文献データに対するマークアップなどが必要であるため、比較的研究が進んでいないように思われる。

3.2.1. 数理モデル

文献の比較研究のための数理モデルについては、非常に多くのモデルが提案されている。ここではその中のごく一部をとりあげたい。

多言語コーパスに対して N グラムをはじめとする確率的言語モデルによる分類（クラスタ分析）を行う方法は早くから提案されていたが[14]、2000 年ごろから文字単位の N グラムモデルを用いた東アジアの古典文献の比較研究が行われるようになった[15]。漢字は一文字が単語もしくは形態素と見なすことも可能なので、文字単位の分析を単語（形態素）単位の分析と同列に見なす見解もある[16]。

また、単語の共起関係をネットワークとして表現する分析する手法も用いられている。赤間啓之氏[17]、三宅真紀氏[18]らを中心としたグループは、フランス語やギリシア語で書かれた古典文献の語彙の隣接関係を意味ネットワークとし、それをクラスタリングすることで複数文献間の比較をする方法を検討している。また山元啓史氏は、同一和歌内に登場する名詞を共起関係と見なし、赤間氏らと同様の方法によって、歌集間の歌ことばの変遷などを分析している[19]。

数理モデルの多くは統計的なものであるが、それ以外のモデルもいくつか見られる。矢野環氏は、進化系統樹の推定に用いられるスプリット分解[20]に基づいた Splits Graph や Neighbor-net[21]の方法を古典文献の分析に応用している[22]。この方法は、「歴史言語学や比較文献学では、生物系統学のなかでも（中略）分岐学と事実上同一の方法論が別個

に開発されてきた」[23]という点を踏まえると、文献学的にも興味深い方法であると思われる。

3.2.2. 文字の知識を配慮した分析

上述してきた先行研究では、文字列の比較において文字コードに依存した形となっている。しかし、一般に形・音・義を持つと言われる漢字の場合、「犬」⇔「狗」の違いと「A」⇔「B」の違いを同様に考えるのは不自然である。アルファベットの場合も、「l」⇔「I」のような書き間違えやすさなどを考慮すれば、文字間の違いを同列で扱うのは不自然であろう。そのような点を配慮して、文字どうしの比較をする際に文字知識データベースを用いる方法が提案されている[24]。

これに関連するものとして、漢字文献データと音韻データベースを組み合わせて作られた音韻列に対して数理的分析を行う研究も試みられている[25]。中国古典戯曲文献の場合、メロディーを伴わない台詞と比べて歌詞の音韻は変化しにくいという特徴がある一方、音韻が通じていれば表記する漢字が容易に交替するという側面もあるため、文字（コード）による比較では不十分なのである。

3.2.3. 文法的な情報を用いた分析

形態素解析などによって得られた品詞情報などをもとに、その共起関係（≒統語構造）などを分析する研究がいくつか存在する[26]。

また、これに関連するものとして、近代日本語文献に特有の分析手法ではあるが、読点を用いた分析がある[27]。句読点（あるいは punctuation）が文の構造と強い関係があることは言うまでもないが、一方で日本語の読点などは“息継ぎ”としても用いられたりするなど、複数の構造にまたがっているものだとも言えるので、注意が必要であろう。

3.3. 構造分析

説話や神話などの分析、あるいは哲学・思想・宗教史などの研究においては、時代的・地域的関連性や単語の一致などの表層的な近親性よりも、物語や思想が持っている構造の類似性が重要視される場合がある。そのような研究でのコンピュータの応用については、物語の構成要素（モティーフなど）をゲノム情報学の方法論によって記述、分析する方法などが提案されたりもしているが[28]、現時点では民話研究におけるモティーフ・データベースなどの整備の段階にとどまっているようである。

今後は、ハイパーテキストやコンピュータゲームのように、小説のような直線的な構造ではなく、アルゴリズム的な構造を持つ物語形式についても、分析方法の検討が必要になってくるのではないかと思われる[29]。

4. 人文学における研究結果との関係

コンピュータを用いた文献の比較研究においては、通常、人文学において確立されたジャンルや慣習を前提として文献が選ばれたりすることが多い[30]。また、数理的な文献の分析におけるモデルの妥当性については、しばしば人文学における研究成果との比較を通じて検証される。これによって人文学における先行研究とのあいだに小さな齟齬が見出された場合には、先行研究に対する批判的再検討も含めて人文学にフィードバックされることがあるが、両者の結果が大幅にずれる場合には数理モデルがそもそも妥当ではないと判断されることが多いように思われる[31]。

数理的な分析結果が人間の読解による分析結果と一致する場合に、それがたまたま一致したのか、それとも人間の読解を数理的なモデルで説明することができているのかについては、今後様々な角度から議論される必要があると思われる[32]。そして、この議論を通じてこそ、文献学的研究の方法論のうち、人文学独自の部分、コンピュータによって代替可能な部分、コンピュータでしかできない部分などを仕分けすることができるのではないだろうか。

5. まとめ

以上、非常に雑駁ながら、コンピュータを用いた文献比較の研究史について、管見の範囲で概観した。筆者の理解不足はもとより、無理に短くまとめようとしたために誤解が生じたり、あるいは重要な研究を見落としたりしているかと思われるので、諸賢のご教示をいただければ幸いである。

謝辞

本稿は科学研究費補助金による研究（課題番号 20520338、20720013、22300087）による成果の一部である。

---

[1] Michael Fraser. “A Hypertextual History of Humanities Computing: Introduction.” (http://users.ox.ac.uk/~ctitext2/history/intro.html、2011 年 1 月 23 日最終確認)、Susan Hockey. “The History of Humanities Computing.” *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004.など参照。

[2] 近藤泰弘氏は、数理的なテクスト分析のメリットとして「徹底的に網羅的な研究」をあげ、「それによって現代人には通常認知できないデータの構造性や規則性を探り出す。それは、現代人の古典語に対する「内省」(introspection)（語感）の欠如を補うことができ、文学研究に貢献する。なぜなら、古典文学の正しい読みにとって、「内省」（文法的直観と言語外知識など）の欠如は大きな障害のひとつだからである」と述べる（「コンピュータによる文学語学研究にできること　―古典語の「内省」を求めて―」全国大学国語国文学会夏季大会シンポジウム「情報技術は文学研究をいかに変えるか」要旨、2001 年)。なお、この種の研究すべてが網羅的であるわけではなく、頻出語の上位にのみ分析を限定するなどの（統計的、あ

るいは恣意的な）操作を行っているものも多い。

[3] アンソニー・ケニー氏は、「文体に指紋があるとすれば、それはどのようなものだろうか？それはおそらく、ある著者の文体的な特徴—例えば‘such as’の生起度数といった、まったく取るに足りないと言ってもよいような特徴を組み合わせたもの—であって、指紋と同様にその人に特有のものであろう。文体上些細で取るに足りぬ特徴だからといって、文体分析に利用しない理由にはならない。指先にある渦巻や輪が我々の容姿においては大切でも目につくわけでもないが、指紋が一生涯変わらないように、そういったものこそが著者の叙述において変化することのない特徴となるはずであり、他の書き手には見られないその人だけのものとなるはずであろう」と述べ、研究者が「些細で取るに足りぬ」と判断し無視してしまうような規則性を見出す方法のひとつとしてコンピュータの利用を評価している（吉岡健一訳『文章の計量　文学研究のための計量文体学入門』南雲堂、1996 年、24 ページ）。

[4] たとえば村上征勝氏は「近年のコンピュータをはじめとする情報分析機器の進歩・普及と、データ分析、感性情報処理、シミュレーションなどの情報分析手法の発展は、自然科学の研究のみならず、文化現象に関する研究にも多大な影響を与えつつある。これまで哲学的、主観的、感性的な方法が中心であった文化現象に係わる研究に、自然科学の領域で用いられている実証的、客観的、数量的な研究手法や種々の分析機器が積極的に導入されるようになってきたのである」と述べている（村上征勝「文化情報学とは」〔『文化情報学入門』、勉誠出版、2006 年 3 月〕）。もっとも後に見るように、文献学における方法の一部は自然科学の方法と「事実上同一」であるものもあり、必ずしも両者は対立するものではないと思われる。

[5] テキストデータベースにおいて文献間の関係を記述する方法としては、永崎研宣氏の諸研究が参考になる（「要素間の関連情報を基盤とする仏教文献デジタル・アーカイブの可能性」〔『情報処理学会研究報告』2007-CH-75、2007 年 7 月〕など）。また、文献画像をベースとした文字列検索や解読、文書間の関係を記述するシステムとして、林晋氏が中心となって開発している SMART-GS（http://www.shayashi.jp/HCP/SMART-GS/）が注目される。

[6] このほかにも、数理的分析のためにはどのようなデータベース（テキストデータベースだけでなく、メタデータ、オントロジなども含む）の設計が必要なのか、という問題や、分析結果の視覚化方法をはじめとする研究者支援システムの開発など、興味深い論点はいくつかあるが、ここではとりあげない。

[7] バート・D・アーマン（松田和也訳）『捏造された聖書』（柏書房、2006 年 6 月）、165〜170 ページ。

[8] かつて国際敦煌プロジェクト（http://idp.bl.uk ほか）に Map Search という地図ベースの写本検索システムが存在したが、現在は稼動していないようである。敦煌文書にしかない文献の研究においては、写本＝文献と混同してしまうので注意が必要。

[9] 「漢字字体規範データベース」（http://joao-roiz.jp/HNG/）、「拓本文字データベース」（http://coe21.zinbun.kyoto-u.ac.jp/djvuchar）など。

[10] 計量文献学については、村上征勝氏の諸著作（『真贋の科学—計量文献学入門』朝倉書店、1994 年など）や金明哲『テキストデータの統計科学入門』（岩波書店、2009 年）などで、様々な方法が紹介されているので参照されたい。また、日本語文献については、入門者向けではあるが伊藤雅光『計量言語学入門』（大修館書店、2002 年）も参考になる。

[11] 村上征勝前掲書『真贋の科学』や John Burrows. “Textual Analysis.”（前掲 *A Companion to Digital Humanities*）など参照。

[12] 上田望「『三国演義』の言語と文体—中国古典小説への計量的アプローチ—」（『金沢大学文学部論集 言語・文学篇』25、2005 年 3 月）に中国古典文献の研究史について紹介されている。現時点では調査が行えていないが、中国にはこの種の研究が多数あると思われる。

[13] 所謂漢文（古典中国語）の形態素解析については、守岡知彦「MeCab を用いた古典中国語の形態素解析の試み」（『情報処理学会研究報告』2008-CH-73、2008 年）が注目される。

[14] 北研二「確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築」（『自然言語

処理』Vol. 4, No. 3、1997 年）

[15] 『漢字文献情報処理研究』を中心とした研究史については、師茂樹「仏教学における自然言語処理」(『漢字文献情報処理研究』第 6 号、2005 年 10 月)、同「N グラム特集、その後」(『漢字文献情報処理研究』第 10 号、2009 年 10 月）などを参照されたい。

[16] 沖本克己「MENSURA ZOILI　禅文献の計量語彙的研究の試み」(『禅文化研究所紀要』19、1993 年)

[17] 近年の研究成果として、赤間啓之・三宅真紀・鄭在玲「近代ストア主義とメスメール主義の思想的類似性に関するグラフ言語学的分析」(『情報処理学会研究報告』2007-CH-74、2007 年 5 月）など。

[18] 本予稿集の三宅真紀氏の論文を参照されたい。

[19] 最近の研究成果として、山元啓史「ブーリアン演算による歌ことばモデルの解析」(『第 16 回公開シンポジウム「人文科学とデータベース」論文集』、2010 年 11 月）など。

[20] H.-J. Bandelt and A. W. M. Dress. “A canonical decomposition theory for metrics on a finite set.” *Advances in Mathematics*, Vol. 92, 1992.

[21] D. Bryant and V. Moulton. “Neighbor-net: An agglomerative method for the construction of planar phylogenetic networks.” *Algorithms in Bioinformatics WABI 2002*, Vol. LNCS 2452, 2002.

[22] 矢野環「芸道伝書の発展経過の数理文献学的考察 —Split decomposition, Spectronet—」(『情報処理学会研究報告』2005-CH-65、2005 年 1 月)、矢野環・福田智子「茶道伝書の文化系統学的処理」(『日本計算機統計学会大会論文集』20、2006 年 5 月)、矢野環「古典籍からの情報発掘 ：再生としての生命誌、ネットワーク」(『情報知識学会誌』17-4、2007 年 12 月）など。

[23] 三中信宏『生物系統学』(東京大学出版会、1997 年 12 月)、92 ページ。なお、2011 年刊行予定というアナウンスが出ている中尾央・三中信宏編『文化系統学への招待：文化の進化パターンを探る［仮］』(勁草書房）は、この種の方法論に関連するものとして注目される。

[24] 師茂樹「文字オントロジに基づく文字オブジェクト列間の編集距離」(『CHISE Conference 2005 報告書 & CodeFest 京都 2005 資料集』、2007 年 1 月)。ただしこの方法を用いた具体的な文献の比較研究は行われていないようである。

[25] 師茂樹・千田大介・二階堂善弘・山下一夫・川浩二「中国古典戯曲文献の韻律の数理的分析に向けて」(『東洋学へのコンピュータ利用 第 19 回研究セミナー』、2008 年 3 月）

[26] 村上征勝前掲書『真贋の科学』等参照。

[27] 村上征勝前掲書『真贋の科学』等には、読点の直前の文字から著者の特徴を見出す研究などが紹介されている。

[28] 小田淳一「情報生物学モデルによる民話研究について」(『認知科学』8-4、2001 年 12 月）

[29] ハイパーテキストやゲームなどが持つ文学的な構造については、森田均・小方孝「デジタル文学理論の構想と試み」(『情報処理学会研究報告』2000-CH-48、2000 年 10 月)、Marie-Laure Ryan. “Multivariant Narrative.” （前掲 *A Companion to Digital Humanities*）など参照。

[30] Google が人文学などで形成された「媒体とジャンルと慣習」を無視して、独自のモデルによるテクスト群の再構成を行っていることに対しては、ロジェ・シャルチエ氏による批判がある（ロジェ・シャルチエ「デジタル化と書物の未来」〔『みすず』2009 年 12 月号〕)。

[31] 人文学における「定説」と対立した例としては、伊藤瑞叡氏・村上征勝氏らによる日蓮の文献に関する共同研究 をあげなければならないだろう（藤本熙・村上征勝・伊藤瑞叡・春日正三『統計的決定理論の立場からの文献学的判別問題に対する研究—日蓮の三大秘法禀承事の真偽判別解析—』〔文部省科研費一般研究研究報告、1981 年〕、村上征勝・伊藤瑞叡「日蓮遺文の数理研究」〔『東洋の思想と宗教』8、1991 年〕、伊藤瑞叡・村上征勝「三大秘法禀承事の計量文献学的新研究」〔『大崎学報』148、1992 年〕等)。この研究では、従来偽作の疑いが強かった『三大秘法禀承事』の真贋を判定するために計量文献学が用いられ、真作の可能性

が高いと結論する一方、従来真作と考えられていた一部の文献については偽作である可能性を示唆している。この研究に対しては、冠賢一「文部省統計数理研究所の「三大秘法禀承事」真作説に対する疑義」(『大崎学報』148、1992 年)、伊藤瑞叡「三大秘法禀承事の計量文献学的新研究　クラスター分析による真偽判定—本研究に対する批判疑義をも消通する」(『大崎学報』148、1992 年）などで論争が展開された。

[32] その際障害となるのは、所謂「文系」の研究者の数学アレルギーではないかと思われる。今後、この分野が実りある発展をするためにも、教育や啓蒙をはじめとする活動が必要であると思われる。