京都大学人文科学研究所共同研究プロジェクト:
情報処理技術は漢字文献からどのような情報を
抽出できるか——人文情報学の基礎を築く

文字と非文字のアーカイブズ／モデルを使った文献研究

全国共同利用・共同研究拠点「人文学諸領域の複合的共同研究国際拠点」
2011.2.18

目次

趣旨説明

この冊子は，京都大学人文科学研究所共同研究プロジェクト:「情報処理技術は漢字文献からどのような情報を抽出できるか——人文情報学の基礎を築く」によるシンポジウム「文字と非文字のアーカイブズ／モデルを使った文献研究」の予稿集である。

本プロジェクトの課題名である「情報処理技術は漢字文献からどのような情報を抽出できるか」から連想されるのは，「文献学的研究・言語学的研究」「データマイニング」という分野であろう。

もちろん，われわれの研究には，これらも含まれる。しかし，われわれの研究は，ここにとどまらず，さらにもっと大きなあるいは抽象的な問題を扱うことを目指している。例えば，それは，「非文字資料を扱うアーカイブズはどうあるべきか」という探索であり，あるいは，「モデルを用いた文献研究の可能性とは」という探索である。このような探索の第1歩として，今回のシンポジウムを開催する。

用語について，ひとことお断りしておきたい。本プロジェクトの課題名の副題には「人文情報学」という用語が使ってある。これは“Digital Humanities”の訳語として使用した。しかし，「『人文情報学』は“Digital Humanities”とは異なる」という考えかた，「『人文情報学』は『情報学』の一分野だ」という考えかた，「『人文情報学』は“Digital Humanities”を包含するものだ」という考えかたなど，人によりさまざまな定義が存在する。本プロジェクトが，それらを整理することも，また独自に定義をすることもせずに「人文情報学」という用語を用いて，副題に「人文情報学の基礎を築く」と大見得を切るのは，おかしなことかもしれない。

しかし，上述のように，われわれが目指すのは，文献学的／言語学的研究やデータマイニングのその先である。いずれは，情報学に対して，「文献の構造における情報学的モデルの構築」という寄与をすることも可能になるかもしれない。そう考え，この課題名を使用する。

シンポジウムについて

日時と場所

2011年2月18日（金），13:00-17:30
京都大学人文科学研究所本館101セミナー室

プログラム

0. **趣旨説明**　13:00-13:10

1．**文字資料アーカイブズの現在――特に検索可能性を中心に**／岡本真（ARG）
13:10-13:40

2. **動画のテキスト処理**／安岡孝一（京都大）13:40-14:10

3. **写真の検索可能性について考える**／守岡知彦（京都大）14:10-14:40

＜休憩＞14:40-15:00

4．**ネットワーク分析からみた共観福音書間の比較研究**／三宅真紀（大阪大学）
15:00-15:30

5. **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）15:30-16:00

＜休憩＞16:00-16:20

6. **パネルディスカッション**　16:20-17:20

文字資料アーカイブズの現在

――特に検索可能性を中心に――

岡本真（アカデミック・リソース・ガイド株式会社）
mokamoto@arg-corp.jp

電子書籍ブームを踏まえつつ、改正著作権法の施行や国立国会図書館による大規模デジタル化の進展等、文字資料アーカイブズの現在を考察する上での前提となる近年の動向を概観した上で、特に文字資料の検索可能性の課題を論じる。

文字資料，改正著作権法，国立国会図書館，大規模デジタル化，検索

1. 文字資料アーカイブズの現在

本報告では、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いに対して、主に文字資料アーカイブズを中心に議論を展開したい。その際、これらの文字資料からの情報の発見・抽出にあたって重要となる検索技術の適用可能性を特に考えたい。

1.1 「電子書籍」元年

去る 2010 年は、電子書籍元年と喧伝された。実際、Kindle（アマゾン）や iPad（アップル）に続き、年末には GALAPAGOS（シャープ）や Reader（ソニー）が発売され、大きな話題となった。また、これらの機器・デバイスとして電子書籍だけでなく、ウェブサービスとして電子書籍のプラットフォームが相次いで提供もされている。

文字資料のデジタル化は、国文学研究資料館を中心に各種研究機関や研究者個人の手で行われて来ており、そこには相当程度の蓄積があるが、ここに来て、電子書籍元年の到来によって、画期的といっていい段階に入っている。権利処理を含め、これらのデジタルリソースの利用には、様々な課題はあるものの、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いを立てたとき、利用しうる対象データが爆発的に増加したことは喜ばしい。

1.2 改正著作権法

このように人力では扱い切れないほどの大規模なデジタルデータが文字の世界においても出現してきたが、文字資料アーカイブズの現在を語る上でさらに 2 つ重要な点が

ある。2010 年 1 月 1 日、改正著作権法が施行された。いわゆる検索エンジン合法化等が注目されがちだが、第 47 条 7 として、「情報解析のための複製」が認められたことを忘れてはいけない。議論の正確を期すため、条文を全文引いておこう。

（情報解析のための複製等）

第四十七条の七　著作物は、電子計算機による情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の統計的な解析を行うことをいう。以下この条において同じ。）を行うことを目的とする場合には、必要と認められる限度において、記録媒体への記録又は翻案（これにより創作した二次的著作物の記録を含む。）を行うことができる。ただし、情報解析を行う者の用に供するために作成されたデータベースの著作物については、この限りでない。

これまでは、研究用途であっても文字データを大規模に収集・解析するには、一定の著作権処理が必要であった。たとえば、日本語コーパスの構築を目指して今年度まで 5 ヶ年計画で行われてきた特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」では、著作権処理に多大な労力を要したという[i]。法改正後、1 年を経た現在、まだ法改正の恩恵に関する目立った成果は見られないが、早晩この効果が見られるようになるだろう。

1.3 国立国会図書館による大規模デジタル化

文字資料アーカイブズの現在を語る上で、著作権法改正と並んで重要なのが、2010 年度の補正予算によって国立国会図書館が進める大規模デジタル化だ。同館の館長である長尾真が随所での講演等で語るところによれば、この事業に 100 億円を超える予算を組み、推定通りに進めば、1960 年代後半までに刊行された日本語書籍について、同館に所蔵されている限りデジタル化されるという[ii]。

ただし、一方の利害関係者である出版各社の意向もあり、この大規模デジタル化事業でデジタル化されるのは、あくまで版面の画像データとされている。その意味で、この事業の成果は、厳密には文字資料とは言えないのもまた事実である。しかし、これもあまり認知されていないようだが、国立国会図書館と一部の出版社[iii]の間では、「全文テキスト化実証実験」が昨年中盤から実施されている[iv]。この実験では、今年度中の結果のとりまとめを予定しており、その結果がもたらす影響、特に大規模デジタル化における手法を画像データから文字データへと進める効果をもたらすのか、大いに注目される。

2. 検索可能性という課題

さて、ここまで概観してきた「現在」を踏まえ、「情報処理技術は漢字文献からどのような情報を抽出できるか」という問いを考えていきたい。特に重要な論点と考えるの

は、「検索」である。

2.1 抽出・解析・蓄積の先にある課題

先に述べたように、著作権法の改正により、研究用途での大規模な情報の抽出や解析は極めて実現しやすくなった。この結果、本シンポジウムの開催趣旨にも掲げられている

i) 人手では不可能な大量のデータを扱い

ii) 人手による処理では帰納できない類の情報を抽出し、

iii) 得られた情報を機械可読かつ再加工可能な形式で蓄積する

という人文情報学の目的の最初の 2 点はこの先、飛躍的な進展を遂げる可能性が高い。しかし、3 点目の「得られた情報を機械可読かつ再加工可能な形式で蓄積する」はどうだろうか。蓄積された情報を利用するには、必要とする情報を引き出す仕組みが求められる。ここで出てくるのが、過去 10 年ほどの間に飛躍的に重要性が高まった検索技術であろう。確かに既存の検索技術、たとえば文の類似度を判定し、同一の、あるいは類似度の高い文を検索する技術は相当程度に確立されている。また、類似性とは異なる人間的な連想という思考方法を機械的な検索に適用するいわゆる連想検索の技術も徐々に普及してきている[v]。

しかし、世界の最先端を行く Google の検索技術をしても万人を納得させるには至っていない。なぜか。たとえば、類似度が同一の文と文の間、語と語の間でその優劣を判定する技術が確立されていないからだ。Google がまさに実践しているように、ウェブ上の情報は無数のウェブページ間の関係性をリンクと被リンクの関係性を抽出・計算することで、いわゆる集合知に基づく優劣関係の判定を行えてはいる。だが、要するに人気を指標とするこの方式（ペイジランク）の限界はつとに指摘されているところだ[vi]。とはいえ、一つの確立された方式ではあるが、この技術はこれから予測される文字資料の大規模なデジタル化には必ずしも有効ではない。ウェブ情報と異なり、データ間の関係性を必ずしも有していないためである。ここに文字資料アーカイブズが越えなくてはならない「検索」という課題がある。

2.2 「検索」を実現するための構造化

結局、課題は「検索」へと行き着く。一つの解として考えられるのは、ウェブ創発の初期から指摘され、最近も論議が高まっている文そのものに意味を持たせる手法、いわゆるメタデータ付与という手法であろう。この手法に関しては、たとえば国立国会図書館が 2010 年に発足させたデジタル情報資源ラウンドテーブル[vii]での議論や、同じく 2010 年に日本でも具体的なプロジェクトが始動した Linked-Open Data（LOD）[viii]といった動きがある。では、人文情報学の場合、どのような可能性が考えられるだろうか。

可能性の一つが、まさに本シンポジウムの最大の論点であり、開催趣旨で述べられている人文情報学の上述の 3 つの目的の先にある

> 最終的には、「××という特徴をもつ文献はどのような構造をしているのか」を機械可読な形式でモデルとして提出する、つまり「文献の構造の情報学的モデル」を確立する

だろう。実際にどのような手法が考えられるのかは、これからの議論によるが、ここで一つの提案をしておきたい。なお、これは「文献の構造」に直接的に関わることではなく、そのモデルが確立された暁に、必要とする情報を引き出す際のパラメーターの一つとして有用と思われるデータの整備に関わることである。

本稿の前半でふれたように、文字資料のデジタル化は大いに進展しつつある。特に近年の特徴は、原資料そのもののデジタル化であり、本文のアーカイブズ化である。しかし、いささか本文、データベースの区分で言えば、ファクトデータベースに傾斜しすぎていることを懸念する。言うまでもなく、人文学的な学問領域においては、特に先行研究の参照が強く求められる。いつ誰がどこでどのような形で、ある本文に言及しているのか、という情報を知らずして、文学や歴史学といった分野の研究は成り立たない。しかし、その割には、研究文献のデータベース、いわゆるレファレンスデータベースの整備が滞ってはいないだろうか[ix]。

もちろん、単に書誌情報を集約したレファレンスデータベースでは、「文献の構造の情報学的モデル」の確立に寄与するところは少ない。しかし、どのような論文でどのような本文が論じられているのか、ということを起点に、本文と言うなればその解釈の関係をメタデータとして内包する書誌情報であればどうだろうか。モデル確立に寄与することはもとより、その先での必要とする情報への到達の容易さ、つまり検索可能性にも大きく影響するだろう。本文だけでなく、この方面の研究にも力が注がれるよう期待を表明して、本稿を終えたい。

---

[i] 2010 年 3 月 10 日に開催された情報処理学会創立 50 周年記念（第 72 回）全国大会におけるシンポジウム「改正著作権法と IT」での前川喜久雄の発言等。http://www.ipsj.or.jp/10jigyo/taikai/72kai/event/39.html

[ii] 岡本真・仲俣暁生編著『ブックビジネス 2.0』（実業之日本社、2010 年）所収の長尾論文ほかを参照。

[iii] 「国立国会図書館における全文テキスト化実証実験の出版社等との共同実施について」＜http://www.ndl.go.jp/jp/aboutus/digitization_fulltext.html＞ によれば、2010 年 10 月 12 日時点で以下の各社である。

アーバンプロ出版センター、暁印刷、旭印刷、有田・海南のフリーペーパー Arikaina、岩波書店、イングカワモト、大月書店、快晴堂、紀伊國屋書店、共同印刷、語研、実

業之日本社、実務教育出版、渋沢栄一記念財団、寿限無、小学館、新人物往来社、新潮社、スタイルノート、青弓社、第一法規株式会社、第三書館、大修館書店、大日本印刷、太郎次郎社エディタス、筑摩書房、中央公論新社、東京創元社、東京大学出版会、東京電機大学出版局、読書工房、トランスビュー、日外アソシエーツ、パンローリング、フライの雑誌社、文藝春秋、ポット出版、まむかいブックスギャラリー、ミルグラフ

[iv] 国立国会図書館における全文テキスト化実証実験の出版社等との共同実施について、http://www.ndl.go.jp/jp/aboutus/digitization_fulltext.html

[v] 主に国立情報学研究所（NII）で開発が進められている連想検索エンジン「GETA＜http://geta.ex.nii.ac.jp/＞のほか、東京大学発のベンチャー企業であるプリファードインフラストラクチャーによる reflexa ＜http://labs.preferred.jp/reflexa/＞等がある。

[vi] たとえば、情報通信研究機構（NICT）の委託研究として京都大学の田中克己研究室を中心とする「電気通信サービスにおける情報信憑性検証技術に関する研究開発」の「Web コンテンツ分析技術」＜http://www.dl.kuis.kyoto-u.ac.jp/i-believe/＞で研究・開発が進められている。

[vii] デジタル情報資源ラウンドテーブル、http://www.ndl.go.jp/jp/aboutus/roundtable.html

[viii] たとえば、国立情報学研究所（NII）の武田英明らによる研究グループによって、LODAC Projects＜http://lod.ac/projects/＞が開始されている。

[ix] 詳しくは、岡本真「日本史研究におけるインターネットの学術利用－これまでの成果と、これからの課題」（『日本歴史』740、吉川弘文館、2010 年 1 月）を参照。

動画のテキスト処理

安岡孝一*

はじめに

2008年春、引越作業中の京都大学人文科学研究所で、$7\frac{1}{4}$inch径ブリキ缶に入った16mmフィルム4巻が発見された。ほぼ1年に渡るリストアおよびデジタル化作業、さらに半年に渡る解読作業の結果、これら4巻のフィルムは、1934年と1936年に撮影されたものがそれぞれ1巻ずつ、1938年に撮影されたものが2巻で、撮影地はいずれも中国北部であることが判明した。すなわち、東方文化学院京都研究所および東方文化研究所が撮影したフィルムだったわけである†。

これらのフィルムのうち、筆者は、まず1938年撮影の2本を、デジタルアーカイブとして公開することを考えた。というのも、この2本は、東方文化研究所が1938年4月9日〜6月15日に撮影した『雲岡石窟』調査記録映画の前巻・後巻であり、学術的な価値が非常に高いと考えられるからである。しかしながら、デジタルアーカイブとして公開すると言っても、合計35分の白黒サイレントMPEGをただノンベンダラリと見せるだけでは、あまりに芸がなさすぎるし、見る方も何が映っているのか全く理解できない。すなわち、デジタルアーカイブとしての公開に際し、この『雲岡石窟』という調査記録映画には何が映っているのかを、わかりやすく記述する必要が生じた。端的にいえば、動画をテキストで記述しそれを処理する、ということを考察する必要が生じたのである。

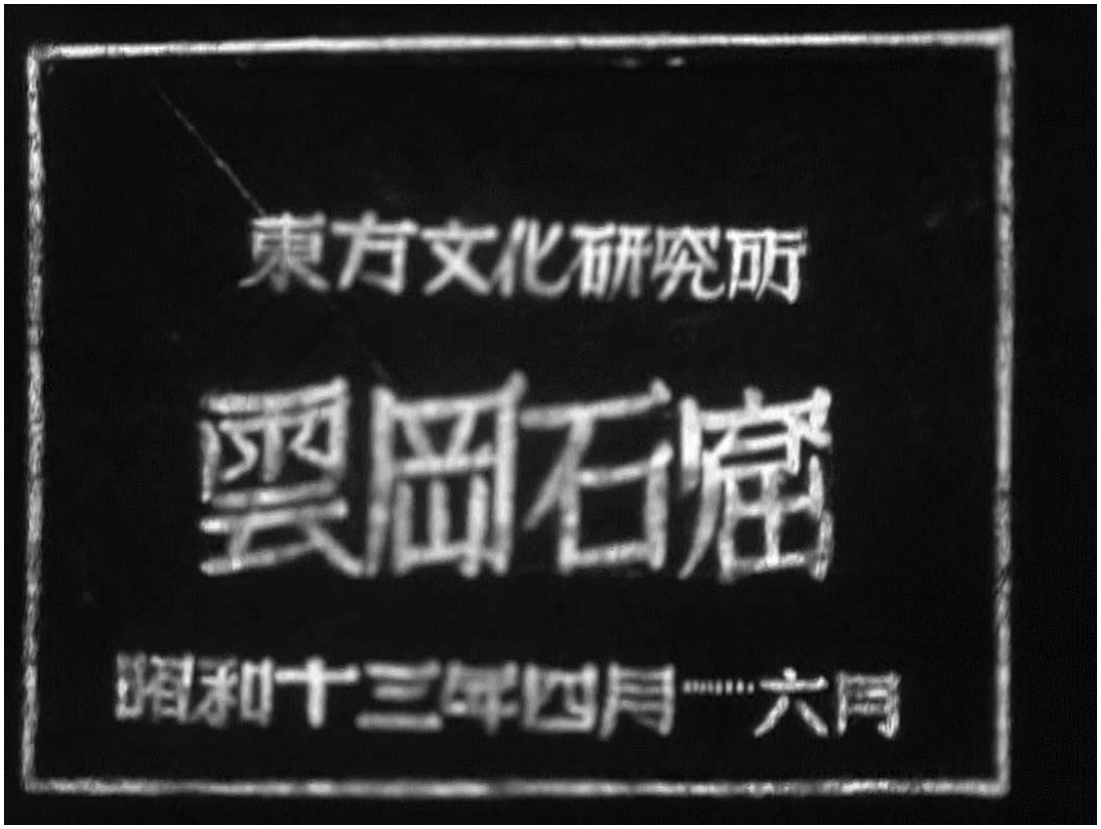
映画『雲岡石窟』の概要

1938年4月9日〜6月19日撮影。白黒16mmサイレント35分。撮影者の水野清一は、東方文化研究所の嘱託員。撮影場所は、北京〜張家口〜天鎮〜大同の鉄道風景、山西省大同の下華厳寺・上華厳寺・市街地・城壁南門・南善化寺、大同郊外の観音閣・雲岡寺など。そして、雲岡石窟およびその周辺の映像が、作品の後半を占める。挿入されている字幕は以下のとおり。

00:02 「東方文化研究所」「雲岡石窟」「昭和十三年四月—六月」
00:09 「一行」「水野清一」「羽舘易」「小野勝年」「米田太三郎」「徐立信」
00:15 「撮影」「水野清一」
00:18 「四月九日——」「北京出発」「——正陽門車站」
01:27 「南口の谷」「—李花開く」
02:18 「朔北の曠野を行く」「—鶏鳴山—宣化—」
02:39 北京〜青龍橋〜張家口〜天鎮〜陽高〜大同の鉄道略図
02:50 「四月十一日—」「カルガン」「即ち張家口」
03:25 「胸につけた良民票」「昨年の苦々しい戦闘の迹」「を偲ばしめる」「—天鎮」

*京都大学人文科学研究所附属東アジア人文情報学研究センター

†安岡孝一: 人文研所蔵16mmフィルムとそのデジタル化, 東洋学へのコンピュータ利用, 第21回研究セミナー (2010年3月), pp.3-8.



03:40 「車窓瞥見」
04:28 「晋北の高原に」「春は未だし」「雲の山—木の芽は堅い—毛皮」
05:03 「北京より 383 キロ」「大同車站に入る」
05:25 「我々は荷物と一緒に」「大同に到着した」
05:28 「晋北政府」「民生顧問 岩崎継生氏」「警務顧問 森一郎氏」「に迎へられて」
06:32 「四月十二日」「下華嚴寺」
07:01 「上華嚴寺」
08:54 「街頭所見」
12:04 「南門にて」「大同城内を望む」
12:53 「南寺こと」「南善化寺」
13:04 「雲岡途上の」「観音閣」
13:48 「雲岡」「石佛は保護されてゐる」「—雲岡鎮警備隊」
17:07 「我々の」「雲岡生活は始った」「四月十三日から」「六月十六日に至る」
22:41 「雲岡の朝——」「臺上の衛兵」「—我々も護られてゐる」
23:40 「雲岡にも春が来て」「堆肥を運び—」「畑を鋤く」
24:40 「晋北の宝」「大同炭を運ぶ」「—原始的な運搬法で、宝は」「まだ死藏されてゐる」
25:26 「我々の 写真撮影」「屋根から—足場から」「羽舘易」「米田太三郎」
27:55 「雲岡参道の修理に」「工兵隊の活動」「—五月十五日」
28:19 「春深し！」「馬鈴薯を植ゑる」
30:05 「我々の」「拓本作業」「—拓工徐君」
30:17 「討伐隊は行く」「—四月二十五日」

31:50 「夏来る！」「緑陰に集る牧羊」「—第三・第四洞前」
33:16 「我々の」「発掘作業は進む」「小野勝年」
34:41 「芍薬老了」「調査終了」「水野清一」「六月十五日」
35:04 「完了」

サイレント映画のスクリプティング

映画を書写言語の形で記述したものは、一般には「完成台本」と呼ばれる。言い換えると「完成台本」は、映画をテキストとして記述したものの総称だと考えられる。

本来「完成台本」は、スクリプターなど映画制作者側の視点[‡]での記述がおこなわれる。しかし『雲岡石窟』には、そのような「完成台本」など存在していないので、それに向けたスクリプティングをおこなうしかない。ためしに『雲岡石窟』の冒頭 2 分間をスクリプティングしてみよう。

字幕「東方文化研究所」「雲岡石窟」「昭和十三年四月—六月」
字幕「一行」「水野清一」「羽舘易」「小野勝年」「米田太三郎」「徐立信」
字幕「撮影」「水野清一」
字幕「四月九日——」「北京出発」「——正陽門車站」
外套を着た 3 人の男。足元にある多数の鞄を見下ろしている。
箭楼の下、街路に行きかう人々と人力車。
子供連れの女たち。
人力車の周りの男たち。人力車に乗りこむ男。
駅舎に次々と到着する人力車。駅舎の時計塔。
プラットホームで列車を待つ男。奥の線路に貨物車両。遠くに箭楼。
字幕「南口の谷」「—李花開く」
鉄道の車窓風景。谷。
鉄道の車窓風景。谷。
鉄道の車窓風景。万里の長城。

もちろん、筆者は『雲岡石窟』の制作には全く関わっていないので、この記述が映画制作者側の意図に沿ったものかどうかはわからない。ただ、この程度のスクリプトであっても、各カットがどのようなものを映しているかに関して、検索のための手掛かりにはなるように思える。

カットとシーンとシーケンス

映像は、一連のカットを集めた「シーン」と、さらに一連のシーンを集めた「シーケンス」によって、説明されることが多い。通常「完成台本」では、1 カットを 1 行 (あるいは 1 段落) で記述することになっており、さらに各シーンの切れ目に簡単な表題を記すことが多い。ここではあえて、XML 風の構造化記述に挑戦してみよう。

[‡]坂本希代子: スクリプターという仕事, 映像テレビ技術, No.675 (2008 年 11 月), pp.37-39.

<sequence><scene>
<shot>字幕「東方文化研究所」「雲岡石窟」「昭和十三年四月—六月」</shot>
<shot>字幕「一行」「水野清一」「羽舘易」「小野勝年」「米田太三郎」「徐立信」
</shot>
<shot>字幕「撮影」「水野清一」</shot>
</scene></sequence>
<sequence><scene>
<shot>字幕「四月九日—」「北京出発」「——正陽門車站」</shot>
</scene>
<scene>
<shot>外套を着た 3 人の男。足元にある多数の鞄を見下ろしている。</shot>
</scene>
<scene>
<shot>箭楼の下、街路に行きかう人々と人力車。</shot>
<shot>子供連れの女たち。</shot>
<shot>人力車の周りの男たち。人力車に乗りこむ男。</shot>
<shot>駅舎に次々と到着する人力車。駅舎の時計塔。</shot>
</scene>
<scene>
<shot>プラットホームで列車を待つ男。奥の線路に貨物車両。遠くに箭楼。</shot>
</scene></sequence>
<sequence><scene>
<shot>字幕「南口の谷」「—李花開く」</shot>
</scene>
<scene>
<shot>鉄道の車窓風景。谷。</shot>
<shot>鉄道の車窓風景。谷。</shot>
<shot>鉄道の車窓風景。万里の長城。</shot>

さらに<shot>タグで囲まれる部分を、それぞれ HyTime か何かで MPEG の各カットとリンク付けすれば、とりあえず、動画のテキスト処理は可能となるように思える。しかし、それは本当に、動画のテキスト処理として十分なのだろうか？

本稿の記述法の問題点

前節で示した記述法の最大の問題点は、各カットに現れるモノがそれぞれ同一のモノなのかどうなのか、その記述が全くおこなわれていないところにある。たとえば、上記の構造化記述の「正陽門車站」のシーケンス中、「男」という記述が出てくるカットは 3 つある。

<shot>外套を着た 3 人の男。足元にある多数の鞄を見下ろしている。</shot>
<shot>人力車の周りの男たち。人力車に乗りこむ男。</shot>
<shot>プラットホームで列車を待つ男。奥の線路に貨物車両。遠くに箭楼。</shot>

これら 3 つのカットに現れる「男」は、同じ人物なのか、そうでないのか。

映画の基本的な語法においては、同じシーンに現れる同じ人物は、通常は同じ服なり同じ顔なりをしていて、たとえカットが変わっても同じ人物として認識されるよう描かれる。人物以外のモノについても同様である。そもそも、各カットの順序は撮影順とは違うかもしれないし、あるいはスタントマンが演じていたりするかもしれないのだが、それでも、何らかの映像的な語法によって同じモノであると認識されるよう描かれる。その視点が、本稿で示した記述法からは、決定的に欠落している。言い換えるなら、動画における「連続性の表現」という観点[§]が、本稿の記述法には欠けているのだ。

この「連続性」という観点は、本稿の方法を、サイレントからトーキーに拡張する際にも、重要な要素となることが予想される。しかしながら、現時点の筆者は、「連続性」を記述する方法を見いだしていない。というか、そもそも記述可能なのかどうかすら、はっきりしていない。まったく心もとない話ではあるが、今後の研究に期待されたい。

[§]実は、この「正陽門車站」のシーケンスの連続性を支えているのは、筆者の私見では「正陽門」という字幕と「箭楼」だ。その意味では、「男」の存在は、本編全体としては重要であるものの、シーケンスとしては大した意味がなかったりする。

写真の検索可能性について考える

守岡 知彦

1　はじめに

写真を探すという行為は、ある目的を満たす写真を探すということだといえる。例えば、『土偶の写真』とか『明治時代初期の京都市内の写真』とか『1930 年台のファッションの写真』というのはその一例である。こうした写真の検索にはしばしば写真に付けられたアノテーションやメタデータが用いられるが、これはある意味において、写真を属性の束からなる概念のようなものとして捉える行為だと看做すことができるだろう。写真は、本来、『この写真』という風にしか指し得ないものかも知れないが、しばしば、『〜の写真』という風に指されるような『概念上の写真』(のインスタンスのようなもの）として扱われる。これは『写真のセマンティクス』という風に言い換えることができるかも知れない。

属性の束で示されるような『概念上の写真』はそれに属する無数の写真の集合である。iPhoto の『スマートアルバム』のような検索に基づく写真管理手法は有限の写真の集合に対してこれを実現したものと看做すことができる。多数の同様な写真の中から 1 枚（ないしは少数）の写真を選び出すという行為は、この概念上の写真を示すための代表を選ぶ行為だと看做すことができる。これは符号化文字（抽象文字）に対して例示字形を付けることに似ているかも知れない。

『概念上の写真』は指定の詳細さによって、その集合の包含関係を考えることができる。例えば、『土偶の写真』の中には『青森県の土偶の写真』があるし、場所や時代をもっと特定することでより詳細に指定することもできる。こうした属性の束によって指されるものとしての『概念上の写真』の性質は、同様に視覚的に表現される記号である文字（特に、漢字）の場合と共通する部分が多数あるといえ、素性の集合（属性の束）によって文字を表現する Chaon モデル [3] と同様な手法で表現できるといえる。

しかしながら、『写真のセマンティクス』は文字の場合とは異なる部分もある。文字の場合、通常、作者はいない[1] 『概念上の文字』はその文字を解する不特定多数によって共有されており、多少の揺れはあったとしても、『作者の意図』や『受け手の解釈』の介在する部分は少ない。言い換えれば、『文字のセマンティクス』はそれをやりとりする解釈共同体の中に置かれているといえる。しかしながら、写真の場合は写真を撮った人がいるし、写真に撮られた被写体が（少なくともその写真が撮られた時には）存在したはずである。そして、これらは属性の束としての写真、すなわち、『概念上の写真』という枠組に収まり切らないような面がある。

写真資料のデータベース（あるいは、そのための写真のセマンティクス）というものを考えた場合、写真が持つこの２つの側面をうまく扱う必要があると思われる。ここでは、

[1]歴史的・伝説上の作者がいる場合はあるが、通常、誰かの著作物とは看做されないし、そのことが文字のセマンティクスを規定したりしない。

この二面性、特に、『概念上の写真』からはみ出てしまうような部分に焦点を当てて、『検索可能性』という観点から議論するとともに、素性の集合で表されるような枠組に基づきながら写真のセマンティクスをうまく扱うことに成功しているポップな Web サービスについて述べ、写真のセマンティクスを支えるものとしてのソーシャルメディアについて議論したい。

2 『概念上の写真』の外

写真の ID は、文字符号のように内容と関係なしに連番で振ることもできるし、ハッシュ関数等を使ってデータ[2]自身から作ることもでき、これらを使って写真の URI を生成することができる。しかしながら、これらは写真の ID にはなっても、写真を説明するものとはならず、写真に関する知識を与えてはくれない。

この問題は、文字符号は文字の ID にはなっても文字を説明するものとはならず、文字に関する知識は文字符号を定義する文書（規格票）ないしは文字符号の定義者の頭の中にのみあって、計算機の中には存在しなくても良いという『符号化文字モデル』の枠組の持つ性格（問題）と同様であるといえる。

CHISE ではこの問題を、特定の恣意的に作られた ID に依存することなく文字を指示する問題と捉え、文字の知識を使って文字を指示することによって解決しようとした。CHISE では、文字に関するなんらかの特徴を『素性』とし、そうした素性の集合によって文字を表現する『Chaon モデル』を用いて文字を処理している。

これは、『確定記述の束』によって対象物を指示するという立場の一種であるといえ、文字を関係性によって記述されるような（言語で表現可能な）概念の一種として扱っているものといえる。それにより、その概念化された範囲において、文字符号依存性を排除した文字表現と文字の検索可能性を実現しているといえる。

しかしながら、このことは文字が持つある種の要素を捨象しているといえる。石川九楊氏がいう所の『筆蝕』[14] はこうしたもののひとつかも知れない。[3] アートとしての書の持つ幾つかの魅力は（客観的な）情報交換の外にあるものかも知れず、CHISE の枠組ではそうしたものに関する言説や、あるいは、感覚器の出力を符号化した情報のようなものは記述できたとしても、（主観的な）感覚自体は記述できない、あるいは、記述できたとしても交換可能とならないといえる。

この問題は、写真に関する２つの要素として議論されてきた。

例えば、バルトは「明るい部屋」[2] の中で、『ストゥディウム』と『プンクトゥム』という２つの要素を取り上げている。前者はここでいう『概念化されたもの』と同様だと考えられる。そして、『プンクトゥム』はその外にあるものだと考えられる。

検索可能性が写真に関する（交換可能な）知識や概念に依存している以上、『プンクトゥム』的な写真はおそらくは計算可能なものとはならないのではないかと考えられる。

但し、対象範囲を特定の個人に絞れば、特定の個人の心を刺激した『プンクトゥム』的な感覚（好き、嫌い、良い、悪い、フォトジェニック、フェティッシュ、エロス、タナトス、萌え、燃え、たぎる、etc.）は記述可能なものとなるかも知れない。

[2]写真を「画像」[17] と言ったり単なる「データ」として捉えるのが妥当かどうかは微妙な問題を含んでいるように思われるが、ここでは、特に、デジタル写真を対象に、そうした単純化を行った場合のことを考える。

[3]あるいは、デリダの「パレルゴン」論 [1] もこうした問題を考える上で示唆を与えてくれるかも知れない。

例えば、CHISE の枠組を例に考えれば、もし美的感覚や性的感覚や数学的感覚に関する感覚器（ないしはそれらを構成する感覚器群や認識システム等から構成される認知システム）があれば、その感覚の種類を素性名とし、その感覚器（認知システム）の出力を素性値とすることでその主観的な感覚を表現することはできるだろう。多くの写真管理システムは写真のセレクトを支援するために、写真に 1〜5 のような評価値を付けれるようになっているが、これはこうした主観的情報の表現を手動で実現したものといえるかも知れない。写真ではなく音楽を対象としたものであるが、iTunes のレート（★（１つ）〜★★★★（５つ））も同様なものといえる。パターン認識技術の進展によって、こうした主観的な情報の幾つかはある程度自動的に認識できるようになるかも知れない。

今の所、計算コストやコーパスや認識辞書等の制作コストの問題から、完全にパーソナライズされた認識システムを作ることはあまり容易なことではなく、限定された範囲の中でカスタマイズするしかないが、一般用コンピューターのための CPU の並列化（GPU の汎用化）や特定用途向けのコーパスや認識辞書を作るための技術の進展とともに、完全にパーソナライズされた認識システムもいずれ比較的容易に実現できるようになるかも知れない。ただ、問題はある人にとって意味のある情報（その人の感覚）は他の人にとっては必ずしも意味のあるものではないということである。[4]

3　主観的情報の交換可能性

例えば、有名人の場合、その人の主観的な情報は商品価値を持つかも知れない。例えば、ある有名人が好きな食べ物やある有名人が好きなブランドといったもの（素性）はその有名人に対して関心を持っている集団（その有名人の有名度が高い程その集団の人口は大きくなる訳である）の中では交換可能となるといえる。これは、（本来交換可能でなかった）主観的情報を、有名人というハブを介して、集団の中で交換可能にした、という風に捉えることができるかも知れない。

『一般キャラクター論』[16] [15] では、この現象を、有名人の『キャラ化』による解釈共同体の形成として捉える。ここでキーポイントとなるのは、ある一人の人間としてのある有名人そのものではなく、その有名人に関心を持つ人の集合（解釈共同体）で、『この人はこんな人』と理解されている『キャラとしての有名人』である。ここで、その『キャラとしての有名人』の好きなものは、そのキャラ [12] を構成する素性の一種になっており、もはやその（肉体を持った一人の人間としての）有名人の生理的な感覚とは関係ないもの、言い換えれば、（言語で説明可能な）概念の一種となっており、それゆえに交換可能になっているという風に説明できる。

このことを応用すれば、有名人ではないある特定個人の主観的な感覚をある程度交換可能にすることもできるかも知れない。すなわち、ある感覚を共有する多数の人間が繋がっていれば、その集団の中でその感覚は『概念化』され（その感覚を是とすれは、必ずしもその『概念』を定義する必要はない)、交換可能となる訳である。このためには、個人を集団にまとめあげる装置があれば良いといえ、2ch のような大規模 BBS や YouTube やニコニコ動画 [10] 等の動画共有サイト、あるいは、SNS や Twitter のようなソーシャル

[4]万人にお勧めできるものがあると信じる人や、それを自分が一番好きなものと同一視できる人はいるようであるが…[19]

メディアはこうした装置としての性格を有しているといえる（SNS 等のソーシャルメディアは多分にその種のコミュニケイションを意識した設計がなされていると考えられる)。

しかしながら、一旦、『概念化』された情報は、それが本来個人の主観的な感覚に由来するものであっても、そこから切り離された高度に記号化された情報として流通することとなってしまう。そして、流通力が高いものはやがてその主観的感覚を共有していた集団の外へも拡散していったりもする。

つまり、このような『概念化』が行われた状況では、その『概念』やそのソースとなった『主観的な感覚』の表現たる素性の集合や画像データからだけでは、それが（本来の）主観的なものかその文脈を越えたものなのかは区別が付かず、写真の『プンクトゥム』的な側面を表現するものとしては不十分であろう。

4　ではどうすれば良いか

基本的に『確定記述の束』でしか写真が検索可能にならないのだとしたら、写真の意味が持つ『概念化』され得ない側面は原理的に検索可能とはならないものとして諦めるというのはひとつの立場であろう。しかしながら、問題は、そうした形式的には『概念化』された情報の表現、すなわち、素性の集合として表現されたメタデータの類や写真の画像そのものは、必ずしも良く定義され広く流通している概念だけに基づいて作られないし、また、受け止められもしないということである。よって、たとえ原理的に不可能であり、限定的にしか実現できないとしても、写真に対する主観的な意識に基づくさまざまな意味を扱うために努力する必要はあるだろう。

特に、写真が携帯写真やブログ等によって個人や小さなコミュニティーでのパーソナルなものとして消費される度合が高くなり、[17] 自らが撮影者や被写体として能動的に写真に向き合う機会が増えた今日、『複製芸術としての写真』というメディア論的な視点だけでなく、身近な読み書き可能なテキストとしての写真という視点が重要になってきているといえ、こうした写真との距離感の変化はパーソナルでない写真の見方に対しても影響しているように思われる。

とはいうものの、抽出可能な情報や実行可能な営為は限られていると言わざるを得ない。結局できることは、基本的に、写真を媒介としたコミュニティーのフィールド調査を行い、ある解釈共同体での写真の意味をもたらす文脈や背景等を分析し、それを情報学的に記述するということにならざるを得ないだろう（これとて容易なことではないが)。つまり、一般キャラクター論的な解釈共同体の分析を行うことによって、はじめて、その外にあるものを浮かび上がらすことができるのではないかと考える。

また、こうした調査は比較的コンパクトなコミュニティーにおいて、例えば、構図や色使いなどの目に見える要素を使って、そのコミュニティーにおける言説の中身を記述するというようなことが考えられる。例えば、「フォトジェニック」だとかコス写真（第 5 節）における「雰囲気（のある）写真」といったものがどういうものを指すかをその傾向性だけでも明らかにすることができれば、検索機能の強化という点では有益かも知れない。

5　ソーシャルメディアとエコシステム

『確定記述の束』で表現されるようなものが強く意識される写真ジャンルとして、アニメや漫画、ゲーム等のコスプレ写真（以下、『コス写真』とする）がある。コス写真は、外形的にはポートレート写真やファッション写真、グラビア写真、スナップ写真等と同じような人物を写した写真であるが、写真に写っている人物が、通常、物語の登場人物のようなキャラクターとして写っていることに特徴がある。つまり、写ってるものの背景として別の物語や設定があり、そうした元の作品やキャラクター、エピソード等の知識を持って写真を撮ったり鑑賞したりすることが少なくなく、視覚的記号性やテキスト性の高い写真のひとつだといえる。

コス写真に関するコミュニケーション・サイト（WWW サービス）は、現在、“Cure” [11] と「コスプレイヤーズアーカイブ」（以下、『アーカイブ』と略す）[8] の２つにほぼ収斂している。前者はコスプレイヤー[5]向けの写真投稿機能とコミュニケーション・サービスを中心に発展してきたものであり、近年は、SNS 機能を採り入れている。一方、アーカイブの方は初期の mixi に似た SNS をそのままコス専用にしたようなもので、構造的には特筆すべき点はない。[6]

このうち、写真の検索サービスとして見た時、Cure はなかなか興味深いシステムである。Cure ではコス写真に対して階層的なキャラクター分類を付与するようになっており、コス写真を写ってる人によってだけでなく、作品やキャラクターによって検索することができる。また、キャラクターの下位分類として、特定エピソードでの衣装や２次創作等のバリエーションを書くこともできる。また、作品のジャンル等の上位分類も行われており、類縁関係にある作品を探すことも可能である。[7]

コス対象となるキャラクターはしばしば平面上で表現されたものであり、画面の外にあるもの、例えば、背面がどうなってるかは良く判らないことがある。また、マンガなどでは白黒で表現されていて色が判らない場合もある。また、形や色が判っていたとしても、物理的に実現が難しかったり、実現に非常に手間がかかりそうなものの場合、どうするかが問題となる。こうした場合、他の人がどういう風にコス化しているかサーベイしたくなる訳であるが、キャラクター（とそのバリエーション）による写真の検索機能はこうした場合に大変有益であるといえる。あるいは、検索結果の件数を見ることで、どういうキャラクター（のどういうバリエーション）が良くコスされているか（≒人気があるか）を調べることもできる。[8] ボーカロイド（以下、『ボカロ』と略す）[13] [5] [6] [7] や UTAU [9] [4] [18] 等の歌声合成技術を用いた曲のように、衣装が特定されていないような作品の場合でも、しばしば、幾つかのパターンに収束する現象が観測される。これはコスの視覚的記号としての性質を反映していると考えられる。すなわち、そのコス（や写真）を見てそれが何のコスであるかを理解できるかどうかということを鑑みた場合、ある程度広く共有可能な代表的なコス表現（文字における『例示字形』のような）があることが望ましいということである。Cure の機能はこうした要請に応えるとともに、こうしたコス写真の解

[5]コスプレをする人

[6]『一般人』との分離を望む人が少なくないことと、さまざまな『改悪』のために mixi を離れた人の受け皿になっている面があると考えられる。

[7]但し、単一継承関係の純粋な階層構造になっているため、上位階層に関する実用性は乏しいかも知れない。

[8]無論、ここに上がっているものが全てではないし、タイムラグもあるので、あくまで傾向性を見るためにしか使えないが。

釈共同体の傾向性を強化する働きを持っている面もあるかも知れない。

ボカロ・UTAU 系作品の場合、ニコニコ動画 [10] や pixiv [20] というイラスト共有サービス等の影響も少なくない。近年のテレビ離れの傾向性とともに、テレビアニメのキャラの比率が減り、ゲームやボカロ系作品の比率が増えてきたように思われる。ボカロ系作品に対するニコニコ動画の影響はいうまでもないが、ゲームの場合も、実際にゲームをした人は必ずしも多いとはいえず、ニコニコ動画上の二次創作を通じて好きになった人が少なくないといえる（同じことは、テレビアニメに関してもいえるだろう）。こうしたことを考えれば、ニコニコ動画が果たしている役割は少なくなく、そのアーキテクチャーの影響力は無視できないと考えられる。

ニコニコ動画は画面内に表示されるコメントが特徴的であるが、検索という観点で見た場合、（人手によって付けられる）タグと「ニコニコ大百科」の集合によって構成されるシステムと捉えることができる。タグは分類項目であり、CHISE における素性に相当するものと看做すことができる。ニコニコ大百科はタグの説明文であり、CHISE におけるメタデータ素性の一種（`*note` 等）と看做すことができる。つまり、ニコニコ動画（のメタデータ）は素性の集合からなるという意味で CHISE と似たものとなっている。しかしながら、基本的に人間が読むためのものとなっており、タグを使った遊びなどが行われ、タグは必ずしも検索だけを意識したものとはなっていない。この人手によって行われ、（ネタも含めて）人が読むことを意識したという部分が、YouTube 等の Google 的な機械処理的アプローチと大きく異なるポイントであろう。すなわち、ニコニコ動画の場合、コミュニティーにおける合意形成やそれを意識した作業というものに意識的にならざるを得ないアーキテクチャーになっている訳である。

Cure にせよニコニコ動画にせよ、その形式自体は CHISE と共通するような、確定記述の束に基づくものであるが、それがコミュニティーのありようやそこでの振るまい方を可視化したり、[9] それを意識したようなアーキテクチャーを実現することで、コミュニティーとシステムが相互作用するような場を提供しているという風に考えることができるだろう。そして、ボカロ系コスの例のように、こうしたユーザー参加型のシステムは、少なくともそのコミュニティーのありようとしては、互いに繋がったものとしてエコシステムが形成されるようになったといえる。

こうしたことを鑑みれば、仮に写真を確定記述の束からなるようなキャラクター的なものに限定したとしても、写真に限定して考えるのは不十分であり、その背景となるような語彙や概念と繋がるような形で形式化・データ化する必要があるという風にいえるだろう。また、こうしたことがコミュニティーの意識とうまくマッチする形で実現できた場合、そのダイナミクスの情報化の実現に繋げることができるかも知れない。いずれにせよ、こうした点では、趣味的に使われているようなポップなサービスに学ぶ点は多々あるように思われる。

[9] 但し、可視化は必ずしも利用者（参加者・生産者）にとって望まれるものではないかも知れない。実際、Cure よりもアーカイブが好まれるようになってきたり、最近の mixi の改変が個人のプライバシーの流出問題として捉えられるという現象を見ても判るように、情報の可視化（や、それを可能とする API の公開等）は微妙な問題を孕んでいるといえる。

6　おわりに

コンピューター上で写真をアーカイブし Web サービス等によってそれらを検索し活用することを考えた場合、写真に対する人文情報学的に妥当なモデル化が不可欠であると考えられる。このためには、写真の持つ多面性と写真を媒介とする多様なコミュニケイションあり方を分析するとともに、『画像データ』や『作品』としての写真そのものをテキストとして読み解き、そのセマンティクスを明らかにする必要があるように思われる。パーソナルコンピューターの高速化によるパターン認識技術の普及によって、従来は困難だったような複雑な解析が可能になりつつあるが、そうした技術を実際に利用するためには写真に関わるさまざまな要素を適切に切り分けていく必要があるだろう。また、デジタルカメラや携帯写真、ブログ等が普及して写真が大量に撮影・消費されるようになってからの写真は、それ以前の写真とはその意味付けが変わっているかも知れない。写真のセマンティクスを考える場合、写真との関わりや写真に対する意識というものを明らかにする必要があるといえるだろう。

いずれにせよ、コンピューターは基本的に素性の集合で表されるような、『概念化された写真』しか理解することはできないので、その枠組の中でいかにして個人的・感覚的・感情的な要素を扱うかということが問題となるが、このためには、『概念としての写真』を明らかにし、それを拡張していくことによって、そこからはみ出る部分を浮かび上がらせるような手法が必要となるのではないかと思われる。プラクティカルにはソーシャルメディアの分析は重要であろう。また、『概念としての写真』からはみ出る部分を分析したり、写真のセマンティクスを形式的に研究するためには、『写真のシンタックス』を見つける必要があるのではないかと思われる。特に、ジャンル固有の傾向性や価値観に関係するようなものを『シンタックス』として取り出していくことが重要ではないかと思われる。このためには、伝統的（教科書的）な写真における美学や『お約束』などとともに、コス写真のような、一見、極めて一般キャラクター論的に見える対象における傾向性や価値観、『お約束』等を抽出して、それらのオントロジーを書くといった手法が有用なのではないかと思われる。

いずれにせよ、写真のセマンティクスを扱うためには写真や写真に直接関わるようなメタデータだけでは不十分であり、対象領域や隣接領域に関わるような情報と連携可能なサービスを設計・構築し、それらが相互作用したシステム全体がうまくエコシステムとして回るような工夫が必要となるだろう。こうしたことを鑑みた場合、学術系データベースはポップな Web サービスに学ぶ所が少なくないと思われる。

参考文献

[1] Jacques Derrida. 絵画における真理 (上). 法政大学出版局, 1997 年 12 月. 高橋允昭、阿部宏慈 訳.

[2] ロラン バルト (Roland Barthes). 明るい部屋—写真についての覚書 （原題：La Chambre claire : Note sur la photographie）. みすず書房, 1997（原書：1980）年. 花輪 光 (訳).

[3] MORIOKA Tomohiko. CHISE: Character Processing Based on Character Ontology. In Takenobu Tokunaga and Antonio Ortega, editors, *Large-Scale Knowledge Resources*, Vol. 4938 of *LNAI*, pp. 148–162. Springer, 2008 年.

[4] 重音テト. http://kasaneteto.jp/, 2008 年 4 月.

[5] MEIKO. http://www.crypton.co.jp/mp/do/prod?id=25220, 2004 年 11 月.

[6] KAITO. http://www.crypton.co.jp/mp/do/prod?id=27720, 2006 年 2 月.

[7] VOCALOID$_2$ — キャラクター・ボーカル・シリーズ. http://www.crypton.co.jp/mp/pages/prod/vocaloid/, 2007 年 8 月.

[8] コスプレーヤーズアーカイブ. http://www.cosp.jp/, 2007 年.

[9] 飴屋／菖蒲. 歌声合成ツール UTAU. http://utau2008.web.fc2.com/, 2008 年 3 月.

[10] 株式会社ニワンゴ. ニコニコ動画. http://www.nicovideo.jp/, 2006 年 12 月.

[11] Cure. http://ja.curecos.com/, 2001 年.

[12] 伊藤剛. テヅカ・イズ・デッド — ひらかれたマンガ表現論へ. NTT 出版, 2005 年 9 月.

[13] 剣持秀紀, 大下隼人. 歌声合成システム VOCALOID. 情処研報, Vol. 2007, No. 102, pp. 25–28, 2007 年 11 月. 2007-MUS-72 (5).

[14] 石川九楊. 筆蝕の構造 — 書くことの現象学. ちくま学芸文庫. 筑摩書房, 2003 年 2 月.

[15] 守岡知彦. キャラクターを考える. 守岡知彦（編）, 人文情報学シンポジウム—キャラクター・データベース・共同行為— 報告書, pp. 55–64. 京都大学 21 世紀 COE プログラム 「東アジア世界の人文情報学研究教育拠点」, 2007 年 12 月.

[16] 守岡知彦. 序文： 『ディープな人文情報学』としての一般キャラクター論への誘い. 守岡知彦（編）, 人文情報学シンポジウム—キャラクター・データベース・共同行為— 報告書. 京都大学 21 世紀 COE プログラム「東アジア世界の人文情報学研究教育拠点」, 2007 年 12 月.

[17] 小林美香. 写真を〈読む〉視点. 写真叢書. 青弓社, 2005 年 7 月.

[18] 藤本萌々子ほか. 桃音モモ. http://www36.atwiki.jp/momonemomo/, 2008 年 5 月.

[19] 千野帽子. あなたが文学を必要としているかどうかは、iTunes マイレートの星のつけかたでわかる。毎日が日直。「働く大人」の文学ガイド, 第 25 章. 日経ビジネスオンライン, 2009 年 4 月. http://business.nikkeibp.co.jp/article/life/20090403/191007/.

[20] 上谷隆宏. pixiv. http://www.pixiv.net/, 2007 年 9 月.

ネットワーク分析からみた共観福音書間の比較研究

――共観表のネットワーク描画――

三宅 真紀（大阪大学大学院言語文化研究科）
mmiyake@lang.osaka-u.ac.jp

要旨：本研究では、新約聖書における共観福音書を分析対象として、単語の共起情報からネットワークを構築し、ネットワーク分析を適用する。ネットワーク構造や視覚化されたグラフ図をもとに、文書間の類似性について考察する。

キーワード：新約聖書, 共観福音書問題, ネットワーク分析

1. はじめに

グラフ理論に基づいたネットワーク分析は、WWW やソーシャルネットワークのような実世界の複雑なネットワーク体系を直感的に把握しやすいデータ解析手法として有効である。近年、自然言語データにおいても、ネットワーク分析の観点から語彙や文書間の関係性を捉える応用研究が報告されている。Gfeller et al. (2005)や Dorow, B. et al.(2005)は対象とするコーパスは違うが、曖昧性をもつ単語に対してマルコフクラスタリングを適用し、クラスタリング精度を上げる手法を提案している。Steyvers & Tenenbaum (2005)は、英語コーパスによる意味ネットワークと複雑系ネットワークの共通性を示して、グラフ理論の有用性を提唱した。

新約聖書学においては、村井・徃住（2007）が、内容が重複するテキストに対して階層的クラスタリング手法を適用し、編集的中心メッセージを抽出することに成功した。三宅（2006）は、共起情報にもとづく福音書のネットワークを構築し、テキストデータの整形処理のためのネットワーク基本特性の活用に関して報告している。さらに、三宅（2008）では、登場人物や活動場所の固有名詞に限定したソーシャルネットワークを構築し、インタラクティブに登場人物の関係性が追跡可能な Web アプリーケションとして実装した。

本稿では、共通する部分が多いとされる共観福音書に対して、ネットワーク分析の適用を試みる。ネットワークグラフ図として視覚化された単語の関係性をもとに、文書間の類似性について考察する。

2. 共観福音書の類似性

2.1. 福音書

新約聖書には、福音書、書簡、歴史書（行伝）、黙示録といった４つの文学類型に分けられた 27 文書が、正典としておさめられている。福音書は、キリスト教会において新しく作り出された概念であり、宣教的意味を持つ。元来、福音書を表すギリシャ語 ευαγγελιον は、特有な文学類型を表すのではなく、救済のよい知らせそのものを意味するものであったが、キリスト教会により、一つの文学類型を意味するようになった (Conzelmann & Lindermann, 1998)。

福音書は、マルコ(Mk)、マタイ(Mt)、ルカ(Lk)、ヨハネ(Joh)の４つの文書から構成され、それぞれ別の著者によって書かれたものとされる。様々な口伝伝承、文献資料を用いて叙述されており、イエスの登場・活動を描き、受難と復活で終わっている。

原文である古典ギリシャ語『ネストレ＝アーラントの新約聖書 (Novum Testamentum Graece) 第 27 版 (Nestle-Aland, 1993)』を使用して、テキストの延べ語数や異なり語数などの基本統計量に関して、4 福音書をそれぞれ個別に集計した情報を表 1 に示す。

各文書の長さを述べ語数から判断すると、マルコ福音書が一番短く、ルカ福音書が一番長いことがわかる。ここで、異なり語数は、形態素の処理を行わない生データのままで処理を行っており、同意語の名詞においても形態が異なれば別の語としてカウントしている。

語の延べ数や異なり数に注目すると、4 福音書の中でもルカ福音書が一番長く、また多くの語を使用していることが分かる。一方で、文の区切りの単位である節の長さの平均をみると、4 福音書とも大きな差はみられず、平均 17 語程度で節が作られていることがわかる。また、4 福音書の異なり語数は、8361 語であった。

表 1：基本統計量

	マタイ	マルコ	ルカ	ヨハネ
章数	28	16	24	21
節数	1068	673	1149	878
1節あたりの平均語数	17.4	17.0	17.1	17.8
述べ語数	18541	11427	19696	15635
異なり語数	3944	2859	4579	2572

2.2. 共観福音書問題

福音書のうち最初の３福音書（Mt, Mk, Lk）は、ヨハネ福音書と比較して、全体構成枠や内容が似通っているほか、言い回しが一致している箇所も多い。共観福音書の共通部分の比率をみると、マルコ福音書全体の約 95 パーセントが、マタイ・ルカ福音書のいずれかと共通している。その共通部分は、マタイの約 58 パーセント、ルカの約 41 パーセントに相当している。また、マタイ、ルカ福音書において、マタイ・ルカにのみ共通している部分については、それぞれ約 20 パーセントの割合である（小林, 1996）。

このように共通性が著しいことから、18 世紀以降「共観福音書（Synoptic Gospels）」と一般に呼ばれている。 18 世紀の終わりには、文書間の文学上の関係性、共通部分の要因を探り、文書間の成立上の相互関係を整合的に説明しようとする、「共観福音書問題」の提起にいたった。これまで、原福音書説、断片説、伝承説など様々な仮説が立てられ、長い間議論されている。その中で、「二資料説」は、最も説得力のある説とされている。

「二資料説」は、マルコ福音書が最も古く、マタイとルカが個別にマルコを資料として用いたと考える「マルコ優先説」を前提とし、マタイ・ルカ福音書のみに共通して現れる箇所が頻出することから、マタイとルカは、マルコ福音書とは別の資料（Q 資料）を用いていたと想定した。つまり、マタイ・ルカ福音書は、共通の資料としてマルコ福音書と「Q 資料」の二資料をそれぞれ用いたと考える説である。

2.3. 共観表

共観表(Synopsis)は、福音書の共通箇所を並べて表にしたものであり、「共観福音書問題」を扱うにあたって欠かせないツールである。古いものでは、Griesbach (1976) の”Synopsis Evangeliorum Matthaei, Marci et Lucae（共観福音書対観表）”がある。近代聖書学研究では、Aland（1983）の Synopsis がよく参照されているが、最近では、『岩波版新約聖書』を使用した、佐藤研（2005）の原文の基づき共通語を彩色した、『福音書共観表』が秀でている。

共観表の一例として、「２資料説」で説明されうる典型的な２つの共通部分（３福音書、マタイ=ルカ共通）を含んだ並行箇所を、表２と表３に示す。いずれも、佐藤（2005, 2006）の福音書共観表から転載した。平行箇所の小見出しは「来るべき者の告知」（マタイ 3, 11-12、マルコ 1, 7-8、ルカ 3,15-18）の部分である。表２は岩波版の日本語訳共観表であり、それに対応するギリシャ語原文が表３にあたる。紙面の都合上、共通部分が表れない節は省略した。共通部分の色に関しては、佐藤氏の彩色原則に従い、原文の単語が一致している場合は「網掛け」にし、形態が異なる語は「２重下線」で色をつけ

た。３福音書共通部分(Mt=Mk=Lk)は水色、マタイ・マルコ共通(Mt=Mk)は明るい緑色、マタイ・ルカ共通(Mt=Lk)は黄色、マルコ・ルカ共通(Mk=Mk)はピンク色で表している。

表 2：平行箇所「来るべき者の告知」（佐藤研訳）

マタイ 3: 11-12	マルコ 1: 7-8	ルカ 3:16-18
この私はお前たちに、 回心に向け、 水によって浸礼を施している。 しかし私の後から来たるべき者は 私よりも強い。 〔私は〕皮ぞうりを 脱がす値打ちもない。 彼こそは、 お前たちに聖霊と火とによって 浸礼を施すだろう。 彼はその箕を 手に持ち、 その脱穀場を隅から隅まで掃き清め、 その麦を倉に集めるだろう。 しかしもみ殻は、消えない火で 焼き尽くすだろう。	そして彼は宣教して 言った、 私よりも強い者が私の後から来る。 〔私は〕その者の皮ぞうりの紐を かがんで解く値打ちもない。 この私はお前たちに、 水で浸礼を施した。 しかし彼こそは、 お前たちに聖霊　によって 浸礼を施すだろう。	〔そこで〕ヨハネは皆に答えて 言うのであった、 この私はお前たちに 水で浸礼を施している。 しかし 私よりも強い者が来る。 〔私は〕その者の皮ぞうりの紐を 解く値打ちもない。 彼こそは、 お前たちに聖霊と火とによって 浸礼を施すだろう。 彼はその箕を 手に持ち、 その脱穀場を隅から隅まで掃き清め、 その麦を倉に集めるだろう。 しかしもみ殻は、消えない火で 焼き尽くすだろう。

表 3：平行箇所「来るべき者の告知」（原文）

Matthew 3, 11-12	Mark 1,7-8	Luke 3,16-18
ἐγὼ μὲν ὑμᾶς βαπτίζω ἐν ὕδατι εἰς μετάνοιαν ὁ δὲ ὀπίσω μου ἐρχόμενος ἰσχυρότερός μού ἐστιν οὗ οὐκ εἰμὶ ἱκανὸς τὰ ὑποδήματα βαστάσαι αὐτὸς ὑμᾶς βαπτίσει ἐν πνεύματι ἁγίῳ καὶ πυρί οὗ τὸ πτύον ἐν τῇ χειρὶ αὐτοῦ, καὶ διακαθαριεῖ τὴν ἅλωνα αὐτοῦ, καὶ συνάξει τὸν σῖτον εἰς τὴν ἀποθήκην τὸ δὲ ἄχυρον κατακαύσει πυρὶ ἀσβέστῳ	καὶ ἐκήρυσσεν λέγων Ἔρχεται ὁ ἰσχυρότερός μου ὀπίσω μου, οὗ οὐκ εἰμὶ ἱκανὸς κύψας λῦσαι τὸν ἱμάντα τῶν ἐγὼ ἐβάπτισα ὑμᾶς ὕδατι αὐτὸς δὲ βαπτίσει ὑμᾶς ἐν πνεύματι ἁγίῳ	ἀπεκρίνατο λέγων πᾶσιν ὁ Ἰωάννης Ἐγὼ μὲν ὕδατι βαπτίζω ὑμᾶς: ἔρχεται δὲ ὁ ἰσχυρότερός μου οὗ οὐκ εἰμὶ ἱκανὸς λῦσαι τὸν ἱμάντα τῶν ὑποδημάτων αὐτὸς ὑμᾶς βαπτίσει ἐν πνεύματι ἁγίῳ καὶ πυρί: οὗ τὸ πτύον ἐν τῇ χειρὶ αὐτοῦ διακαθᾶραι τὴν ἅλωνα αὐτοῦ καὶ συναγαγεῖν τὸν σῖτον εἰς τὴν ἀποθήκην αὐτοῦ τὸ δὲ ἄχυρον κατακαύσει πυρὶ ἀσβέστῳ

表 4：平行箇所「来るべき者の告知」のテキスト情報

	マタイ 3:11-12	マルコ 1:7-8	ルカ 3:1-5
述べ語数	58	30	58
異なり語数	43	28	49

表２,３から、平行箇所「来るべき者の告知」は、３文書共通部分が内容の中心部分であり、マタイ・ルカ共通箇所が続いており、マルコ文書には表れない”πυρὶ（火）”に関する内容となっている。前半部分には、“ὀπίσω μου（私の後から）”といったマタイ・マルコだけに出現する単語もみられ、マルコ・ルカに共通する文章が３文書共通文章に入り込む形をしているところもある。表４に、ギリシャ語の並行テキストに対して、各文書の出現単語の情報をそれぞれ示す。

3. ネットワーク分析

3.1. テキストのネットワーク表示

ネットワークは、ノードとエッジによって表すことができる。言語データをネットワーク分析に応用する場合、ノードの単位は、出現単語や章などの区分がとして考えられる。さらに文書間を比較するときは、文書全体を１つのノードをして表すことができる。エッジは、２つのノード間になんらかの関係があるときに、それらのノードを線で結んで表す。

3.2. 平行箇所の共起情報

小規模なネットワークの作成例として、２節で取りあげた並行箇所「来るべき者の告知」をもとにして、共観表ネットワークを構築する。ここで、ノードは文書の出現単語とし、出現単語に共起する単語をエッジで結ぶ。表２，３からもわかるように、共観表は、文書間で一致する記事の部分をなるべく並べてみられるように単語の配置が巧妙に工夫されている。この配置に従い、共観表の行を単位として共起情報の範囲を設定し、各行に同時に出現する単語を共起単語としてみなした。出現単語には、共通部分のカテゴリー情報を付随情報として加え（便宜上、Mt=Mk=Lk, Mt=Mk, Mt=Lk, Mk=Lk, Mt, Mk, Lk の順に、1-7 までの番号を単語の後に付加）、同じ単語でも出現する共通カテゴリーが異なる場合は、別の単語として認識している。このようにして、共起単語ペアの頻度リストを作成し、その一部を表５に示す。それぞれの文書におけるペア延べ数は、マタイが 103、マルコが 50、ルカが 107 ペアであり、共観表としての全体の延べ数は、107

ペアであった。

表 5：共起単語ペアリスト（抜粋）

単語1	単語2	ペア頻度数
ἁγίῳ_1	πυρί_3	2
ἀπεκρίνατο_7	Ἰωάννης_7	1
ἀπεκρίνατο_7	λέγων_4	1
ἀπεκρίνατο_7	ὁ_7	1
ἀπεκρίνατο_7	πᾶσιν_7	1
ἀποθήκην_3	αὐτοῦ_3	1
αὐτὸς_1	βαπτίσει_1	2
αὐτοῦ_3	ἄλωνα_3	2

各文書のネットワークに関して、ネットワーク基本特徴量のノード数，平均次数，クラスタリング係数の平均値を順に表６に示す。ここで、次数は単語の繋がりの度合いを表す指標であり、Watts&Strogatz (1998)のクラスタリング係数は、ノード間の繋がりの度合いを表す重要な指標である。テキストのネットワーク分析では、単語の曖昧性を抽出するための指標として、クラスタリング係数を活用することもできる。

表６から、マタイとルカの２つのネットワークは、ほぼ同じ大きさであり、平均次数やクラスタリング係数からみても類似した構造をしていることがわかる。マルコは、最も小さなネットワークを形成し、そのクラスター性は極めて高いことが、平均クラスタリング係数からわかる。

表 6：基本ネットワーク特徴量

	マタイ 3:11-12	マルコ 1:7-8	ルカ 3:1-5
ノード数 (カテゴリー別異なり数)	50	30	53
平均次数	4.0	3.3	3.9
平均クラスタリング係数	0.74	0.93	0.77

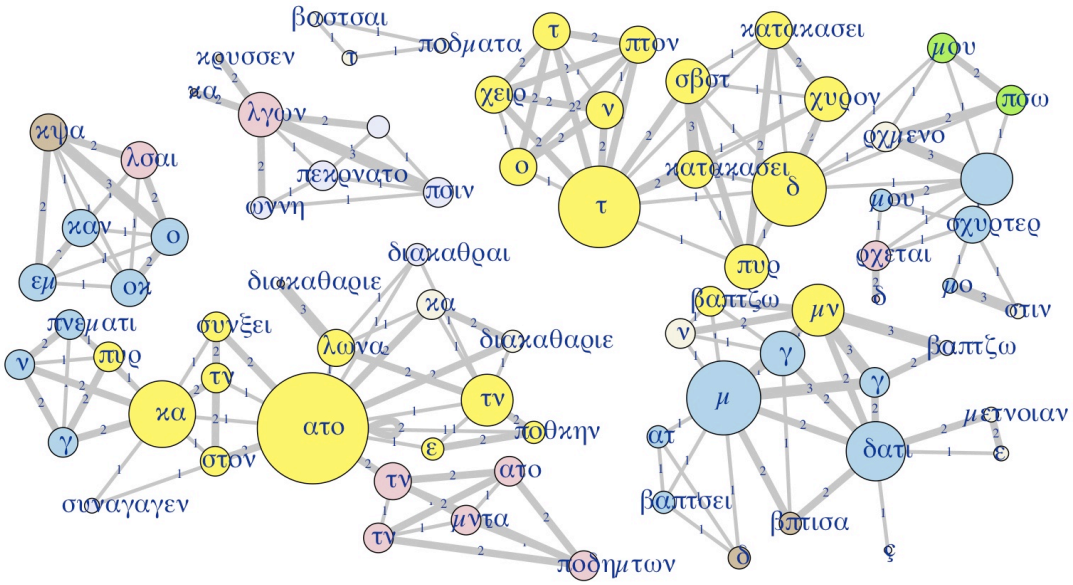


図 1：共観表ネットワーク図

3.3. ネットワーク描画

文書構造の体系をグラフ図として視覚化することにより、単語や文書の関係性を直感的に捉えることができる。ネットワーク視覚化には、統計解析ソフトウェア R の collonet パッケージを使用した（三宅, 2010）。このパッケージは、igraph パッケージを言語データ用にカスタマイズすることを目的として作成したものである。図 1 に共観表ネットワーク図を示す。共観表のカテゴリー分けに従って、ノード色を区別した。４つの共通カテゴリーに加えて、各福音書に独立して表れる単語についても文書別に識別し、文書特有の使用単語を確認できるようになっている。次数の大きさに比例して、ノードの大きさを変化している。単語と単語の結びつけているエッジの太さは、共起単語ペアの頻度数に比例している。図 1 から、３福音書共通部分とマタイ・ルカ共通部分のクラスターが分かれていることが読み取れる。そのクラスターに派生する形で、マルコ・ルカ共通部分のノードの小クラスターが繋がっている部分もあり、ネットワーク構造をもとにして、共観表の並行箇所をさらに細分化することも可能である。

4. おわりに

本稿では、共観福音書の３福音書に対して、共観表の並行箇所から単語の共起データを抽出しネットワークを構築する手法を提案した。共通カテゴリーの情報が付加された単語を使用したグラフはクラスター性を高め、文書の共通部分が把握しやすいネットワーク図として描画された。内容一致を重視したクラスター抽出のための単語の形態素

処理や、クラスタリング係数を閾値としたクラスター形成については、今後の課題としたい。

参考文献

Aland, K. (1985), *Synopsis Quattuor Evangeliorum 15th edition*, Deutsche Bibelgesellschaft.

Conzelmann, H. & Lindermann, A. (1998), *Interpreting The New Testament*, Hendrickson Publishers.

Dorow, B. et al.(2005), Using Curvature and Markov Clustering in Graphs for Lexical Acquisition and Word Sense Discrimination, *MEANING*.

Gfeller, D., et al. (2005), Synonym Dictionary Improvement through Markov Clustering and Clustering Stability, *ASMDA*, 106-113.

Greisbach (1776), *Synopis Evangeliorum Matthaei, Marci et Lucae*, Helle.

Nestle-Aland (1993), Novum Testamentum Graece 27th edition, German Bible Society Stuttgart.

Steyvers, M., Tenenbaum, J. (2005), The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, *Cognitive Science*, 29 (1) pp.41-78.

Watts, D. and Strogatz, S. (1998), Collective dynamics of 'small-world' networks, *Nature*, 393:440-442.

小林稔 (1996),「福音書問題」,『現代聖書講座』, 日本基督教団出版社, 1996.

佐藤研 (2005), 『福音書共観表』, 岩波書店.

佐藤研 (2006),『ギリシャ語彩色共観表』, http://www.rikkyo.ne.jp/web/msato/GrSynAll.pdf.

村井 源, 徃住 彰文 (2007), 正典テキスト群から編集的中心メッセージを抽出するネットワーク解析法,『情報知識学会誌』 17(3), pp.149-163.

三宅真紀 (2006), グラフクラスタリングに基づく共観福音書意味ネットワークの実装,『人文科学とコンピュータシンポジウム論文集 じんもんこん-2006』, pp.161-165.

三宅真紀 (2008), 福音書ソーシャルネットワークにおけるコミュニティ構造の考察,『人文科学とコンピュータシンポジウム論文集 じんもんこん-2008』, pp.319-324.

三宅真紀 (2010), 「ネットワーク分析のための R パッケージの開発―テキストマイニングへの応用に向けて―」,『電子化言語資料分析研究 2009-2010』, 大阪大学大学院言語文化研究科, pp.41-63.

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。

文字と非文字のアーカイブズ／モデルを使った文献研究

発行日: 2011年2月18日
発行者: 全国共同利用・共同研究拠点「人文学諸領域の複合的共同研究国際拠点」
住所: 〒606-8501 京都大学人文科学研究所
印刷: