第 24 回図書館総合展　国立国会図書館主催フォーラム
「#NDL 全文使ってみた～「次世代デジタルライブラリー」＆「NDL Ngram Viewer」（2022 年 11 月 1 日）

「第二部　研究者活用編」概要報告

3 人のパネリストから次の内容について発表後、ディスカッションが行われた。

① 全文テキストに関わるご自身の研究
② 各分野における全文テキストを使った研究の状況
③ 研究における国立国会図書館（以下「NDL」）全文テキストの活用可能性

瀧川　裕貴・東京大学大学院人文社会系研究科准教授　発表概要

全文データを使った自身の研究を二つ紹介する。

一つ目は、国会会議録の全文データを使った道徳社会学の研究。特定の道徳概念や感情を表すと考えられる単語をリスト化した辞書を使った頻度分析を行った。分析結果から、政治家の感情や道徳原理の利用の仕方について、立場による差異や時代による変遷などを考察した。

二つ目は、社会学評論の全文データを用いた戦後社会学史の研究。トピックモデルを用いて論文のテーマを抽出・分類し、社会学におけるテーマの変遷を記述した。また、先入観をもたない機械の「読解」により、隠された研究分野も発見した。

社会学における全文データを用いた研究事例として、Google の Ngram コーパスを用いた単語の文化的意味の抽出を紹介する。この研究では、単語埋め込みモデルを用いて、単語の性質をあらわす低次元ベクトルを抽出するという手法を用いている。また、ある単語のベクトルからその対義語のベクトルを引き算することで、特定概念の文化的意味次元を抽出することも可能らしい。たとえば、”rich”から”poor”を引くと、「経済的豊かさ」を表す意味次元が抽出できる。抽出した意味次元におけるほかの単語の位置や、ほかの意味次元と組み合わせたときの関係なども考察できる。

今後の研究計画としては、NDL から提供された全文データを使って、社会的実践・思想・概念の文化的意味づけの構造と歴史的変遷を明らかにしたい。具体的には近代日本における「幸福」概念の持続と受容について調べたい。方法としては、全文データに単語埋め込みモデルを適用して、考察していく予定である。

実験的に数年分のデータをサンプリングして、単語埋め込みモデルを試してみた（スライドには記載なし）。ただし、この実験から実質的な意味を読み取ることはできず、あくまでデモンストレーションである。

まずは、幸福・不幸の類義語をワードクラウドで年代ごとに可視化してみた。結果、可視化した単語や表記に、ある程度の違いがみられた。つぎに、「夢」「宗教」「家族」「結婚」な

どのワードを幸福の次元上に配置することなどを試してみた。

日比　嘉高・名古屋大学大学院人文学研究科教授　発表概要

古典文学研究では、著作権関係の処理の容易さや本文校訂の必要性などを背景として、以前からデジタル化が進んでいる。

デジタルデータを活用した方法論的に興味深い研究事例として、近藤みゆきによる『古代後期和歌文学の研究』(2005)や『王朝和歌研究の方法』(2015)などがある。これらの研究では、Ngram を用いて、詩句における、序言葉・枕詞以外の新たな定型性の発見や、男女間における言葉遣いの差異などが論じられている。

近現代文学については、近年急速に全文データベースが整備され始めた。これらのデータベースは、これまでにも単純なキーワード検索の対象となったり、国内外の高等教育現場で入手・閲覧が容易な作品本文として使われてはいた。だが、研究における方法論の革新には結びつかないでいた。最近になって、研究手法の革新も始まりつつあり、たとえば計量的方法論を用いた Hoyt Long による *The Values in Numbers* (2021)が出版された。近現代文学の分野でのデータを使った研究も、今後に期待できると考える。

全文データを使った活用例として、NDL 全文データを使って、明治から第二次世界大戦前の期間における芭蕉受容の様相を調査した。具体的には、芭蕉の全発句を全文検索にかけ、ヒット数を引用回数とみなして考察を行った。結果、今まで人の作業量の限界から限定的な資料調査によってしか答えられなかったことが、全文データを活用することで一定の根拠とともに回答可能になった。また、句ごとの引用回数の時系列変化をグラフ化することで、時代ごとの引用の波を可視化できた。ただし、年代による刊行点数の違いに注意する必要がある。

全文データを用いた研究をするときに特徴的なのは、人が行う研究と異なり、機械は文脈を無視するという点。この危うさと面白さを理解して、人と機械の強みを組み合わせていくことに期待する。

また、研究では、「全ジャンル」「全時代」が必要な場合は少ない。ジャンルや年代、著者の属性などといった適切な範囲で絞りこんで意味づけをした「意味を持った資料体」を用いることが大事である。そのために、芭蕉の全発句などのような、他のデータ・セットと、全文データを組み合わせることがカギになるだろう。

増田　知子・名古屋大学大学院法学研究科教授　発表概要

私どもが行った全文データに関する研究として、デジタル歴史情報基盤の構築を目的とするデータベースの作成と公開がある。『人事興信録』については、全文データ化・構造化して項目ごとだけでなく全文検索ができるようにした。それを使ったデータ分析を紹介する。『人事興信録』には、採録者の親戚関係を示す「参照」項目があり、この項目を分析したネットワーク図（球形）を作成した。参照者の多い中心部の極めて少数の集団は、血縁関

係による富裕層だと推測できた。一方で、参照者がゼロまたは極めて少ない外縁部の人は、圧倒的多数を占めていたが、『人事興信録』にどのような理由で掲載されたのかにつき、当時の社会的評価を得られるような情報を得たいと考えていた。

そこで、此度 NDL の全文データの活用という機会を得たので、次世代デジタルライブラリーの全文検索機能で、4 分の 1 の人名を検索してみたところ、興味深い結果が出た。まず、血縁関係の参照次数と全文検索のヒット数には相関がないことがわかったが、このことは、参照次数がゼロであっても、全文検索のヒット数によって、社会的に認知・評価されている度合いを推測できることがわかったことである。ただし、人名が地名などの他の単語と混ざる場合があったため、次世代デジタルライブラリーの全文データの形態素解析が必要ではないかと考える。

次に、NDL Ngram Viewer を使った試験的な分析結果を紹介する（スライドには記載なし）。私どもが現在取り組んでいる研究テーマは、なぜ富裕層が大恐慌を経て個人主義・自由主義から軍国主義・超国家主義へと「転向」したのかを解明することである。「自由」「統制」などの国家体制原理を含意する政治思想の単語の出現頻度を年代ごとに可視化することにより、単語グラフの交差が社会の思想の変化を示すのではないか、という作業仮説をたてて分析した。

その結果、異なる単語間の出現頻度が交差した時期が、既知の政治史上の重要な変化と重なっていることがわかった。また、実際にどのような言論や思想が流通していたのかにつき、長期的変化を可視化することができた。明治以来の検閲制度により出版は常時統制下にあったが、NDL Ngram Viewer は、社会に流通した思想動向を知る方法になり得ることが分かった。さらに、それらの単語がどのような文脈を持って流通していたかについては、次世代デジタルライブラリーを活用できると考えている。

ディスカッション（まとめ）

最初に、モデレーターである永崎研宣・一般財団法人人文情報学研究所主席研究員から、北米が中心となって運営する電子図書館の取り組みである HathiTrust について紹介があった。

永崎氏は、パネリストの発表に共通していた論点として「分析対象とするデータの範囲」の偏りの可能性を指摘した。その一例として、NDL の所蔵する戦前期資料（納本制度開始前）の網羅性には限界があることに言及があった。また、地方や大学のアーカイブ機関がデジタル化及びテキスト化に取り組むことの重要性が述べられた。

また、パネリストから、データが多すぎる場合の絞り込みという観点でメタデータ整備の重要性についての指摘があった。逆に、データがまだデジタル化されていない場合や廃棄・検閲によって現存していない場合には、分析データの範囲外の資料を知っている専門家の存在が重要になる点について指摘がなされた。

NDL への期待については、それぞれパネリストから次のコメントがあった。

瀧川氏からは、デジタル化・テキスト化される資料の拡充に関して、1970 年代以降の資料に拡がることへ期待感が示された。また、書誌情報に限らない、著者の生年などの資料をとりまく文脈を含めたメタデータの整備などの必要性を指摘し、メタデータの整備に関しては、どちらかといえば研究者がやるべきであり、協力したいとの意見があった。

日比氏からは、全文データの「国立国会図書館デジタルコレクション」への投入が楽しみであること、特に、雑誌の全文検索についての期待が述べられた。また、「意味をもった資料体」にするために、絞り込み機能の充実への要望も示された。その一例として、既存の十進分類法などを用いた絞り込みだけでなく、外部の協力によりタグ付けを行うソーシャルタギングなどの試みについての提案があった。資料を使う上で、それがどれだけ信頼できる本文かは重要な観点であるため、間違いデータの訂正方法の検討の必要性への指摘もあった。

増田氏からは、文字自体を読むことにも訓練が必要な歴史研究においては、AI による分析が進むはずだが、AI はそこに存在していないデータを知ることはできないことについて指摘があった。存在していないデータには、例えば、占領期に GHQ によって廃棄された文書や、戦時期の検閲によりそもそも出版できなかった文書などがある。このようなデータの不存在に気付けるのが専門家であり、専門家は AI の行うことを見守っていく必要がある。一方で、特定の領域に縛られないという AI の特徴は利点にもなり、専門家がデジタルデータの分析から気づきを得られるということもあるため、うまく協同していきたいという考えが示された。