日本語の量的テキスト分析

カタリナック，エイミー　・　渡　辺　耕　平

要　旨

本稿は、欧米の政治学者の間で近年注目を集めている量的テキスト分析（quantitative text analysis）と呼ばれる手法の日本語における利用について論ずる。まず、量的テキスト分析が登場した背景を述べたうえで、欧米の政治学においてどのように利用されているかを説明する。次に、読者が量的テキスト分析を研究で利用できるように、日本語の分析において注意すべき点に言及しながら、作業の流れを具体的に説明する。最後に、欧米で利用されている統計分析モデルを紹介した上で、それらが日本語の文書の分析にも利用できることを研究事例を用いて示す。本稿は、近年の技術的および方法論な発展によって、日本語の量的テキスト分析が十分に可能になったことを主張するが、この手法が日本の政治学において広く普及するためには、データの整備など制度的な問題に対処していく必要性があることにも触れる。

Quantitative Text Analysis in Japanese

CATALINAC, Amy and WATANABE, Kohei

Abstract

We introduce quantitative text analysis, an exciting new set of tools for social scientists working on Japan and in Japanese. First, we explain why these methodologies are useful in attempting to answer questions of substantive interest. Second, we outline the workflow of the typical project that employs quantitative text analysis, paying particular attention to how texts written in the Japanese language can be handled at each stage. Third, we provide an overview of the statistical models most commonly used by political scientists working with text and demonstrate that, with the adjustments we identify, they can be readily applied to Japanese-language texts. Finally, we point out challenges, including the relative absence of adequate investment in large-scale, publicly available data collections and a sparsity of methodological training. The discipline of political science in Japan will need to grapple with these issues to enable the use of these tools to become as widespread as they are elsewhere.

はじめに

近年、欧米の政治学者の間で量的テキスト分析（quantitative text analysis）と呼ばれる、自然言語処理技術を用いた文書の統計的な分析が注目を集めている。この手法を用いることで、政治現象を分析する材料として従来の数値データだけではなく、文章データを利用することが可能になる。たとえば、利益団体が支持する政党を明らかにするために政党に対する献金の額ではなく、プレスリリースを分析する、あるいは、上院議員がイラク戦争を支持する程度を明らかにするために、記名投票の結果ではなく、議会での発言を分析するといったことが可能である。さらに、量的テキスト分析を用いることの利点は、データが存在しない場合や入手しにくい場合でも、文書を分析対象とすることで政治現象の研究が可能になることである。たとえば、法案が作成される過程を観察することは難しいが、その基になった法案を文書の類似性から推定することができる。また、政府が秘密裏にメディアに介入していたと考

えられる場合も、その存在をニュース記事に使われる語彙の変化から推測できる。

量的テキスト分析が注目されている主な理由は、政治学的に重要な文書がインターネットを通じて大量かつ容易にダウンロードできるようになったことと、文書を量的に分析するためのソフトウェアが多数開発され、大規模なデータを分析するハードウェアの性能が高まったことにある。さらに、いわゆるビッグデータの利用の観点から、レストラン、映画、航空会社等のレビューやブログ、そしてソーシャルネットワーキングサイトへの投稿を分析しようとする研究者が社会科学の各分野で増えてきたことも、この分析手法が飛躍的に発展している背景である。近年では、量的テキスト分析は、計算社会科学（computational social science）の一分野として認知されている。

量的テキスト分析で用いられる手法は先端的なものが多く、統計学や計算機科学の研究者を巻き込んだ技術的および方法論的な議論も活発である。文書データは事前に変数が定義されることなく生成された非構造化データ（unstructured data）であり、それぞれの語が変数となる多次元データ（high-dimensional data）であるため、そこから社会科学的に意味のある情報を抽出する方法自体も重要な研究の対象である。さらに、文書データから信頼性の高い情報を抽出するためには、多くの文書を効率的に処理する必要があり、量的テキスト分析は情報処理の観点からも興味深い手法と言える。

このように欧米で活用が進む量的テキスト分析であるが、日本の政治学に浸透しているとは言い難い。そこで、本稿は、近年の技術的な発展によってコンピュータによる日本語の処理が大幅に容易になり、量的テキスト分析が日本語の文書に対しても適用できることを示す。前半では、政治学における量的テキスト分析について概説し、この手法を用いた研究の流れを具体的に説明する。後半では、政治学の量的テキスト分析で使われている先端的な統計分析モデルを紹介するとともに、それらを用いて日本語の文書を分析した研究を事例として示す。事例の一つ目は、Catalinac（2016; 2018）による衆議院の選挙制度改革が候補者の政治的なイデオロギーと彼らが訴える政策に与える影響を、選挙公報から分析するもので、二つ目は、Trubowitz & Watanabe（2018）による日本とイギリスのアメリカに対する認識の長期的な変遷を、過去 30 年間の日本とイギリスの新聞記事を通じて分析したものである。なお、本稿は、政治学の事例に的を絞って議論を進めるが、ここで紹介する量的テキスト分析の手法は、社会科学の研究に広く応用できるものである。

政治学における量的テキスト分析

量的テキスト分析はさまざまな研究で用いられているが、その主な目的は（1）文書の内容の検証、（2）文書の影響力の検証、（3）文書を代理とした測定である。文書の内容の検証とは、文書の系統的な分析を通じて、特定の主題に関する記述の傾向を記述することである。文書の影響力の検証では、文書を世論調査などの社会的な反応を示すデータと併せて分析することで、その影響力を推定する。文書を代理とした測定では、文書を通じて直接には観察できない社会現象を計量化する。欧米の政治学では、研究者が直接観察できない潜在変数（latent variable）を文書の系統的な分析から推定することが、量的テキスト分析の主な目的となっている。

- ●文書の内容の検証
  - ○17 世紀から 19 世紀にわたって米国政府と先住民との間に結ばれた条約の公正性を検証する（Spirling 2012）。
  - ○イスラム教徒に関する英国の新聞記事を分析して報道の公平性を検証する（Baker et al 2012）。
  - ○ロシア政府が国際プロパガンダのために国営通信社の報道に与えるバイアスを測定する（Watanabe 2017）。
- ●文書の影響力の検証
  - ○人種問題に関するニュースを世論調査のデータと併せて分析することで、世論形成にメディアが与える影響を推定する（Kellstedt 2010）。
  - ○米国下院に提出された法案を分析することで、可決済みの法案がまだ可決されていない法案に与える影響力を推定する（Wilkerson et al 2015）。
  - ○米州議会の法案を分析することで、豊かな州が貧しい州の立法過程に与える影響力を推定する（Jansa et al 2018）。
- ●文書を代理とした測定
  - ○選挙公報やプレスリリース、議事録を分析することで、政治家や政党にとっての政策の優先順位を推定する（e.g. Slapin and Proksch 2008;

Catalinac 2018; Grimmer 2010）。

- ○国営メディアの経済に関するニュースを分析することで、ロシアの大統領が政治的支持を獲得する方法を推測する（Rozenas and Stukal 2018）。
- ○ブログを継続的に分析することで、米国の大統領候補に対する世論の好感度の変化を測定する（King and Hopkins 2010）。
- ○米国の新聞の経済ニュースを分析することで、経済政策に関する不安感の変化を30年間に渡って計測する（Baker et al 2016）。
- ○利益団体の政策提言を欧州連合で可決された法案と併せて分析することで、利益団体の政策形成過程への影響力を測定する（Kluver 2009）。

量的テキスト分析の流れ

一般的に、政治学における量的テキスト分析は（1）データの収集、（2）テキストの前処理、（3）文書行列の作成、（4）統計分析の適用、（5）結果の解釈から成り立っている。2から4の作業には、量的テキスト分析用のソフトウェアが不可欠であり、欧米ではPythonのNLTKやGensim、RのTM、Tidytext、Quantedaなどのパッケージが使われているが、日本では、RとJavaを組み合わせたKH Coderが広く使われている。なお、プログラミングが得意な研究者は独自のソフトウェアを開発して研究を行う場合があるが、分析の再現性の観点からは広く用いられているツールを用いた方が良いと言える。

1　データの収集

量的テキスト分析をおこなうための文書データは、さまざまな形で存在する。政治学的に重要な日本語の文書の一部はデジタル化されており、インターネットからWordやPDF、XMLの形式でダウンロードすることができる。ウェブサイトから多数のファイルをダウンロードする場合は、PythonのBeautiful SoupやRのRvest、またはウェブブラウザをSeleniumを通じて操作してスクレ―ピングを行うことが多い。また、全国紙に掲載された過去の記事は、朝日新聞の「聞蔵 IIビジュアル」や読売新聞の「ヨミダス歴史館」等のデータベースで検索できるが、大規模な分析のためにCD-ROMとしても販売されている。しかし、研究者が紙の文書を収集し、文字認識プログラム（OCR）を用いてデジタル化しなくてはならない場合もある(1)。このようにして集められた文書の集合はコーパス（corpus）と呼ばれる。

コーパスの構築においては、収集したデータが標本として適切かどうかを常に考えなければならない。たとえば、Twitterへの投稿は、Web API(2)を通じて容易に収集できるが、Twitter利用者は年齢や性別、職業などにおいて偏りがあるため、収集された文書が国民全体の政治意識を反映していると考えることは適切ではない。標本と母集団との関係に注意してデータを収集する点は、アンケート調査などと同じだと言える。

2　テキストの前処理

文書をコンピューターで効率的に処理するためには、まず、トークン化（tokenization）という処理によって、文を単語や数字、記号などの要素に分割する。英語の文書では、文を空白で分割するだけでトークン化ができるが、日本語は語の境界が明確ではないため、より複雑な処理が必要となる（詳細は次の節を参照）。トークン化の後、データを単純化するために、記号やストップワーズ（stop words）といわれる文法的な語を削除することが多い。英語のストップワーズは「to」「the」「for」などである。日本語の標準化されたストップワーズが見当たらないが、政治学的な研究では平仮名のみで構成されるトークンを削除することで同様の処理ができる。

データをさらに単純化するため、トークンを語幹や原型へと変換することがある。たとえば、「食べる」「食べた」「食べます」「食べなければなりません」「食べたくない」を、「食」という語幹、または「食べる」という原型に変換すると5種類のトークンを1種類にまとめることができる。語尾の削除をステミング（stemming）、原型への変換をレマティゼーション（lemmatization）と呼ぶが、このような処理は、文書の間に共通するトークンが少ない場合には分析の質を改善するために有効だが、語の微妙な意味の違いを捨象するので逆効果の場合もある。

語の文脈を加味しながらトークンのあいまいさを低減したい場合は、隣接する語を結合して、「肉_食べる」「魚_食べる」「野菜_食べる」などのNグラム（N-gram）を生成することもできる。しかし、この処理は、コーパス全体で一回しか現れず、統計

的な分析の役に立たないトークンを多数生み出すため、注意して利用すべきである。トークンの種類を増やすと、データの複雑性が高まり、統計的な処理が難しくなる。なお、N グラムを生成する場合は、共起分析（collocation analysis）の結果に基づいて、統計的に強く関連した語だけを連結することでデータが過度に複雑になることを避けることができる。

3　文書行列の作成

量的テキスト分析を行うためには、トークンを文書ごとに集計し、文書行列（document-feature matrix）を作成する。文書行列は行が文書に列が語に対応し、セルの値はそれぞれの語が文書に登場する頻度を表している。文書行列を作成した後は、以降の統計的な分析で計算量が大きくなりすぎることを避けるため、頻度が低い語（コーパス全体で 5 回以下など）を削除することでデータを単純化することが一般的である。この作業は特徴選択（feature selection）と呼ばれる。特徴選択によって統計モデルを用いた分析の結果が変化するので、どの程度の影響があるかを確認するのが理想的である（Denny and Spirling, 2017）。

文書行列の単純化は、特徴を削除するのではなく、辞書を用いて特徴をグループ化することによっても実現できる。量的テキスト分析で利用される辞書は、何らかの概念に関連する多数のキーワードで構成されている。たとえば、感情辞書では、数千個の肯定的な語と否定的な語が「ポジティブ」と「ネガティブ」のカテゴリーに含まれている。これを用いた感情分析では、辞書に含まれる語が二個の変数にまとめられるため、大幅なデータの単純化を行える。なお、文書行列に対しては、特徴の頻度を文書の長さで標準化し、割合に変換するなどの算術的な処理を容易に行える。

4　統計分析の適用

コーパスから作成した文書行列の統計的な分析法は多岐にわたるが、代表的なものとして、(1) コーパスの全体もしくは一部からの特徴の抽出、(2) 文書または特徴の類似性の測定、(3) 類似性に基づく文書または特徴のクラスター化、(4) 文書の一次元または多次元空間における位置の推定 (5) 文書の話題による分類などがある。量的テキスト分析ではこれらを組み合わせることで、文書データから研究者が明らかにしたい質問への答えを導き出す。

5　結果の解釈

量的テキスト分析の結果を解釈する際は、文書データが多次元であることに常に注意しなくてはならない。つまり、データが非常に多くの変数によって構成される場合、その中に何らかの外部データと強く相関するものを見つけることは容易だが、それは単なる偶然であることが多い。さらに、政治学における量的テキスト分析の目的は、文書そのものの理解ではなく、文書を通じた社会現象の理解なので、分析結果の解釈はその文書が生成された背景との関連で行う必要がある。このことは、量的テキスト分析を用いる際には、分析手法だけでなく対象とする社会現象それ自体に幅広い背景的な知識を持たなくてはいけないことを意味し、そのために分析対象を含んだ多数の文書を自ら読むことが欠かせない。

日本語のトークン化の詳細

量的テキスト分析の作業の流れは、言語を問わず基本的に同じだが、テキストの前処理では文法や語彙の違いを考慮しなくてはならない。日本語は英語のように語が空白で区切られていないことから、日本語の文書のトークン化では語の境界を検出するために MeCab や Chasen などの形態素解析ツールが使われることが多い。これらのツールは、文中の語の前後関係に基づいて語の境界を検出するだけではなく、語の品詞（名詞、形容詞、動詞など）を判定することもできる。KH Coder の Windows 版は形態素解析ツールを同梱し、すぐに日本語の分析が行えるようになっている。また、R では RMeCab というパッケージを使うことで容易に MeCab を用いた形態素解析を行える（Ishida 2010）[(3)]。

これまでは、日本語の前処理には形態素解析ツールを用いることが一般的だったが、R では Quanteda を用いると外部ツールを用いずに日本語のトークン化を行うことができる。Quanteda はトークン化を、ほぼすべての言語に対応するユニコードの標準ライブラリ（ICU）の境界検出の機能を用いて行っており、その日本語処理の結果も、形態素解析を用いた場合との違いが少なく、表 1 の例では「及ぼし」にだけ違いがみられる。この理由は、ICU による語の境界の判定は、大規模な辞書を用いて機械的に行われており、その辞書は IPA 辞書が元になってい

るからである。IPA 辞書は、情報処理推進機構によって作成され、MeCab でも利用されている。

表 1　MeCab と ICU によるトークン化の違い

MeCab	政治・と・は・社会・に対して・全体・的・な・影響・を・及ぼし・、・社会・で・生きる・ひとりひとり・の・人・の・人生・に・も・様々・な・影響・を・及ぼす・複雑・な・領域・で・ある・。
ICU	政治・と・は・社会・に対して・全体・的・な・影響・を・及・ぼ・し・、・社会・で・生きる・ひとりひとり・の・人・の・人生・に・も・様々・な・影響・を・及ぼす・複雑・な・領域・で・ある・。

MeCab と ICU によるトークン化では、新奇な語の境界を正確に判定できない場合がある。この問題は、機械的に分割を行う後者で深刻であり、たとえば「フェイクニュース」は「フェ・イクニュース」とトークン化されてしまう。このような場合は、NEologd などのより新しい辞書を形態素解析で用いるか、共起分析に基づいてトークンを結合すると良い。

統計分析モデルの種類

量的テキスト分析では、記述的統計からニューラルネットワークまでさまざまな統計モデルが利用されている。以下では、近年の政治学の研究で使われたものの中から、一般に「バッグ・オブ・ワーズ」（bag of words）と呼ばれる、文中の語の順序を考慮しないモデルを手短に紹介する。これらのモデルは、語の順序を無視することで文書データを単純化するため、統計処理にかかる計算量が少なく、特別に高性能なコンピューターを必要としない。

教師あり学習モデル

教師つき学習モデル（supervised learning models）とは、機械が手作業で分類された文書に基づいて学習をすることで、人間の判断を再現する統計モデルのことである。この代表は単純ベイズ（naïve Bayes）であり、このモデルは我々の日常生活の中でも、迷惑メールの自動分類などに用いられている。ここでは、電子メールソフトの利用者による迷惑メールかどうかの判定に基づいて、統計モデルが学習を行い、それ以降に届く電子メールを自動的に分類している。単純ベイズは、利用者によって迷惑メールだと判定されたメールの中から頻度の高い語を選択し、重みづけを行うことで、電子メールの分類モデルを構築する。このモデルは、名前の示す通りベイズの定理に基づいているが、「単純」だとされるのは、文書の中にある語が独立して出現していることを前提としているからだ。この前提はあまり現実的ではないが、このモデルは少ない計算量で高い精度の分類を行えるため、量的テキスト分析において広く使われている。これ以外に政治学で人気のある教師あり学習モデルには、ワードスコア（wordscores）やランダムフォレスト（random forest）、ラッソ回帰（LASSO regression）などがある（c.f. Laver et al 2003; Benoit et al, forthcoming）。

たとえば、Nielsen（2017）は、イスラム教の指導者を過激派と穏健派に区別するため、インターネット上で 100 人の指導者によって書かれた 2 万 7 千件の文書を収集し、以前から政治思想が知られている指導者によって書かれた約 3 千件の文書から単純ベイズモデルを構築した。この分類モデルを用いて、まだ政治思想が知られていない指導者によって書かれた残りの 2 万 4 千件を自動的に分類することで、西洋諸国に対する聖戦を訴える可能性が高い者を特定した。

教師あり学習モデルを量的テキスト分析で利用する場合、訓練（training）と検証（test）の両方を手作業で分類した文書を用いて行い、モデルによる予測が十分に正確であることを確認したうえで、それをデータの全体に適用する。訓練と検証のためには、コーパスから一部の文書を無作為に取り出して手作業で分類し、その半分以上を訓練データに割り当てる。訓練データの大きさは、コーパスの 10％程度が理想的だが、データが複雑な場合は 50％以上でなくては十分な精度が得られない場合もある。Nielsen の研究では、モデルを単純化するため前処理として頻度の低い（全文書の 10％以下）もしくは高い（全文書の 40％以上）に出現する語が削除され、残りの語は原型に変換された。

教師なし学習モデル

教師なし学習モデル（unsupervised learning models）とは、あらかじめ定義されたコーパス内の語と文書の関係に基づいて、機械が最適な結果を導き出す統計モデルである。この代表は、LDA（latent Dirichlet allocation）などのトピックモデル（topic

models）であり、これらのモデルを用いると、内容に応じて文書を分析者が指定した数の話題に自動的に分類できる。LDA は、話題と関連する語彙の存在を仮定し、著者が文書の話題を選択してから、それに対応する語を選択するというデータの生成過程をモデル化しており、文書と語の関係に最も適合する形で、話題がそれぞれの文書に割り当てられる（Blei et al 2003）。しかし、LDA によって特定される話題は、分析者によって定義されたものではないので、モデルが提示する語彙を事後的に解釈することで意味が与えられる。量的テキスト分析では LDA よりも単純な LSA（latent semantic analysis）が文書のグループ化に用いられることもある（Landauer et al 1998）。また、政治学的分析に特化した教師なし学習モデルとしては、ワードフィッシュ（wordfish）がある。

Slapin と Proksch（2008）は、東西ドイツが統一した 1990 年から 2000 年代中盤の間に、同国の社会民主党と緑の党が保守化してきた、という仮説を検証するためにワードフィッシュを開発し、選挙マニフェストに適用した。このモデルは、特徴に共起頻度を考慮した重みづけを行うことで、訓練データなしで、文書の政治イデオロギー的な位置を推定できる。ワードフィッシュが作られたのは、ナイーブベイズやワードスコア（Laver et al 2003）のような教師あり学習モデルを用いると、訓練データに含まれる文書の特徴だけがモデルに組み込まれ、残りの文書にしか現れない特徴が無視されてしまうからである。この問題は、文書の語彙が時間と共に変化する場合に特に深刻であった。例えば、緑の党の選挙マニフェストで言及されている政治課題が 1990 年代から 2000 年代にかけて、自然環境の保護から同性愛者の権利へと変化した場合、初期の文書で訓練されたモデルでは、末期の文書に表れる自由主義的なイデオロギーを認識することができない。

教師なし学習モデルは、訓練データの作成を必要としないため非常に便利だが、文書の特徴の選択によって結果が大きく変化することがある。また、教師なし学習モデルでは、事前に分析の尺度や分類が定められていないので、結果の解釈が恣意的にならないように注意して利用するべきである。たとえば、ワードフィッシュを政治イデオロギーの分析に利用した場合、抽出された尺度のどちら側が保守主義（もしくは自由主義）を表しているかの判断は、分析者の解釈にゆだねられる。

準教師あり学習モデル

準教師ありモデル（semi-supervised learning models）とは、人間によって与えられた語彙を手がかりに、機械がコーパスから語の関係を学習し、文書の分類を行う統計モデルである。準教師あり学習モデルは、上述の教師ありモデルと教師なしモデルの弱点を補うために、最近になって使われ始めた手法である（Boiten, Schoonvelde, & Schumacher, 2018, Watanabe 2017, 2018b）。その一例であるニュースマップ（Newsmap）は、文書の地理的分類のためのモデルで、国名と都市名から構成される種語（seed words）を用いた辞書[(4)]で訓練データを作成し、ナイーブベイズと同様の仕組みで学習を行う（Watanabe 2018b）。準教師あり学習の利点は、訓練データの作成に直接人間が関わらないため、コーパス全体からなる大規模な訓練データを利用できること、そして、種語を通じて分類の結果を柔軟に制御できることにある。これ以外の準教師あり学習モデルとしては、以下で説明する LSS がある。

Watanabe（2017）は、ウクライナ危機のロシア政府による国際的なプロパガンダの研究において、大規模なニュースの感情分析を行うために LSS（latent sematic scaling）を開発した。このモデルは、ベクトル空間モデル（vector space model）の応用であり、特異値分解（SVD）によってノイズが取り除かれた文書行列の中で、一般的な感情語と政治的な特徴語の距離を計算し、文書の政治的な感情を予測するモデルを構築する。Watanabe は、このモデルを用いてロシアの国営通信社である TASS によって書かれた 3 万 5 千件の英語の記事を一年半に渡って分析し、政治的な記事に現れるバイアスの強さを測定した。

準教師あり学習モデルでは種語の選択が決定的に重要なので、分析者が専門知識に基づいて慎重に行う必要がある。また、準教師あり学習モデルは、十分に大きなデータで訓練された教師あり学習モデルと比較すると、分類の精度が低いので、結果を集約して誤差を相殺するなどの処理が必要になる場合がある。

研究事例

以下では、本稿の著者が日本語の文章を対象にし

て量的テキスト分析を用いて行った研究（Catalinac 2016; 2017, Watanabe 2018）を事例として紹介する。Catalinac は、教師なし学習モデル（ワードフィッシュとトピックモデル）を利用して日本の選挙公報を分析し、Watanabe は準教師あり学習モデル（ニュースマップと LSS）を用いて日本とイギリスの新聞記事を分析した。なお、これらの研究で用いられているモデルは全て、R のパッケージとして配布されているので、読者も容易に同様な分析を行える。

事例 1：選挙制度改革による候補者のイデオロギー的な位置と主張する政策の変化

政治学においては、候補者のイデオロギー的な位置が選挙制度に左右されると以前から考えられている。しかし、候補者の発言や政党の出版物の分析を通じてこの仮説を検証した研究はほぼ皆無であった。そのため、Catalinac（2018）は、選挙候補者が政治的なイデオロギーが選挙制度に応じてどのように変化するかを明らかにするために、1994 年の衆議院選挙制度改革（中選挙区制から小選挙区・比例代表並立制へ）を事例として、選挙公報の量的テキスト分析を行った。この研究で、選挙公報を分析対象としたのは、公職選挙法の規定で候補者が利用できる宣伝媒体が非常に限られているため、選挙公報が候補者の戦略を十分に反映していると考えられるからである。また、各候補者の選挙公報は、選挙管理委員会が選挙区の全家庭に配らなければならないと法的に定められている。

Catalinac は、1986 年から 2009 年の間に発行された約 8 千件の選挙公報を自ら収集し、神戸大学の品田裕教授の支援を得ながらデジタル化した。候補者の政治的なイデオロギーの推定では、文書を MeCab によってトークン化し、ワードフィッシュを適用した。図 1 は、1986 年の総選挙の際に配布された 800 個の選挙公報から、ワードフィッシュが推定した語のパラメータを表している。技術的な詳細は省くが、縦軸は語の頻度、横軸は語のイデオロギー的な位置を示していると解釈できる。たとえば、「政治」「教育」「平和」「円」「国民」は、頻度は高いがイデオロギー的に中立な語である。「間違う」「採算」「圧倒的」「集団」は保守的な候補者がよく使う語であり、反対に「切りすて」「ツケ」「スジ」「ニセ」は革新的な候補者がよく使う語である。ワードフィッシュは、推定された語のパラメーターに基づいて、文書のイデオロギー的な位置も計算する。

分析の第一の結果は、選挙制度改革によって小選挙区で競合する候補者間のイデオロギー的な距離が縮まったことを示していた。この理由は、中選挙区では一部の有権者からの支持を獲得するだけで当選できるため、候補者がイデオロギー的に過激な訴えをすることが多かったが、小選挙区では大半の有権者からの支持を得なくては当選できないので、候補者が穏健な訴えをするようになったからである。分析の第二の結果は、選挙制度改革によって同じ政党から出馬している候補者のイデオロギー的な距離も縮まったことを示していた。この理由は、小選挙区・比例代表並立制では議席の過半数を獲得しようとしている政党が同じ選挙区で複数の候補者を出馬させる必要がないため、党内競争が和らいだからである。

Catalinac（2016）は、党内競争が激しい場合、候補者が政党よりも自分自身の人柄や実績、公約を強調するという説を検証するため、1994 年に行われた選挙制度改革による党内競争の緩和が、候補者が訴える政策に与える影響を分析した。本研究では、トピックモデルを用いることで、選挙公報の主な話題が教育、郵政民営化、安全保障、憲法改正などの政策課題や、公共工事の約束、過去の実績の訴えなどであることが明らかになった。とりわけ、地方の候補者の選挙公報では農業と漁業の話題が多い一方、大都市圏の候補者は保育所の不足など子育て支援の話題が多いことも明らかになった。選挙制度改革の前後を比較すると、自民党の候補者の選挙公報で利益誘導に関連する話題が減り、党の政策に関連する話題が増えた。

上記の研究で、Catalinac は、選挙制度改革が候補者のイデオロギー的な立場と主張する政策の変化をもたらしたと論じているが、経済や国際情勢などの選挙制度以外の要因が変化をもたらしたのではないことを、中選挙区内と小選挙区内における党内競争の激しさの違いを利用して確認した。もし、選挙制度改革ではなく 1990 年代中項の経済や国際情勢が候補者の立場と主張に変化をもたらしたのであれば、1986 年と 1990 年の総選挙で、党内競争が激しい選挙区の候補者間のイデオロギー的な距離が遠いことを説明できない。

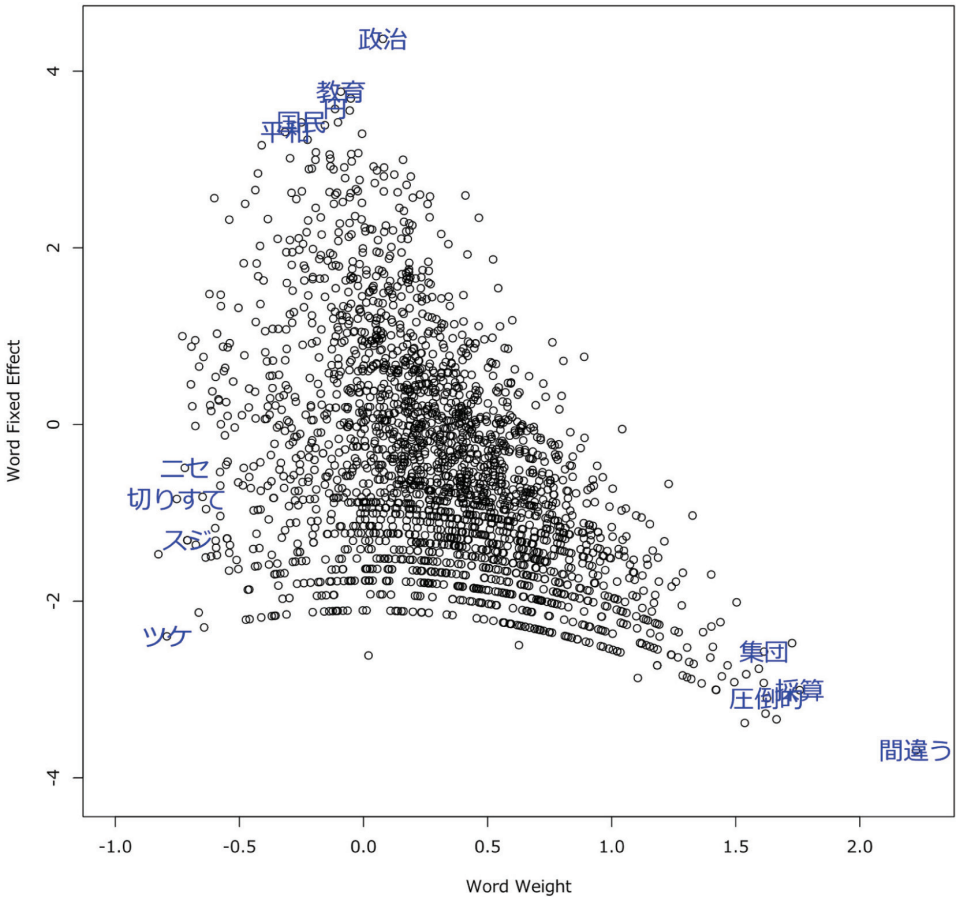

図 1　ワードフィッシュによって推定された言葉のイデオロギー的な位置

事例 2：日本とイギリスのアメリカに対する認識の長期的な変遷

国際政治学では、国家が互いをどのように認識しているか理解することが外交政策を説明する上で重要とされる。しかし、従来、国際関係の研究に利用できるデータは、紛争に関するイベントデータや経済や軍備に関する統計データなど限られてきた。この問題を解決するため、Watanabe は、日本とイギリスが主要な同盟国であるアメリカをどのように認識しているかを、過去約 30 年間分の新聞記事を系統的に分析することで量的に把握しよう試みた（Trubowitz & Watanabe 2018）。この研究では、同盟国としてのアメリカの信頼性に関わる重要な国内外の出来事（湾岸戦争、9・11 同時多発テロ、リーマンショックなど）および、日本やイギリスの安全保障に関わる出来事（ソ連の崩壊、台湾海峡危機、イラク戦争、ウクライナ危機、北朝鮮の核実験など）が、どのように日本人とイギリス人のアメリカに対する認識を変化させたかを、朝日新聞とガーディアンに掲載されたアメリカに関する新聞記事を分析することで明らかにすることを目標としている。

日本とイギリスの新聞は、アメリカに関する記事を多数掲載するため、1980 年代から今日までの全記事を収集すると、一紙当たりの記事の数は 10 万件を超える。これらの記事を手作業で分類するとなると、大変な労力がかかるだけではなく、現代の解釈が入ってしまう恐れがあるため量的テキスト分析を用いた。また、新聞記事の大規模な分析では、辞書分析が用いられることが多いが、現代的な語に偏るのを避けるため、準教師あり学習モデルを選択した。

本研究では、まず、日本とイギリスの新聞から収集した記事をニュースマップを用いてアメリカ、日本、イギリスについての記事だけを選択した。次に、収集したすべての記事から LSS のベクトル空間モデルを作成し、｛懸念, 危惧, 疑惧, 憂慮, 不安, 心配｝、日本に対する脅威を ｛脅威, 危険, 恐怖｝ などの種語を用いて 30 年間に使われた数千の語を重みづけし、アメリカに対して日本人が抱く懸念を測定した。イギリスの新聞の分析のためには、日本語の種語に対応する ｛concern*, worr*, anxi*｝ および ｛threat*, danger*, fear*, risk*｝ などを用いた。下の

二つの図は、この分析の結果を表しているが、図2の赤い線は1990年代と比べて2000年代は日本人がアメリカに対する懸念を強く抱いていたこと、緑の線はイギリス人がアメリカに対する懸念を2010年代から強く抱き始めたことを示している。懸念の高まりは、その時期の重要な出来事に対応しており、日本人にとっては2001年の同時多発テロと2003年のイラク戦争と2008年のリーマンショックが重要であったことがわかる。イギリス人も同時多発テロ以降に懸念を強めたが、アメリカの政治が両極化し、公的機関が財政難から閉鎖された2013年頃に向けて急速に高まったことがわかる。

図3によると、イギリス人は、湾岸戦争と同時多発テロの時期を除けば、1990年代から2000年代まで政情不安を感じてはいなかったが、中東が「アラブの春」によって不安定化した2010年から政情不安を強く感じてきた。その一方で、日本人は、イラク戦争の頃から政情不安を感じてきたが、中国が南シナ海に進出し、北朝鮮がミサイル開発を加速させた2010年後半から今日にかけて政情不安を非常に強く感じるようになってきた。

本研究はまだ初期の段階であるが、予備的な分析

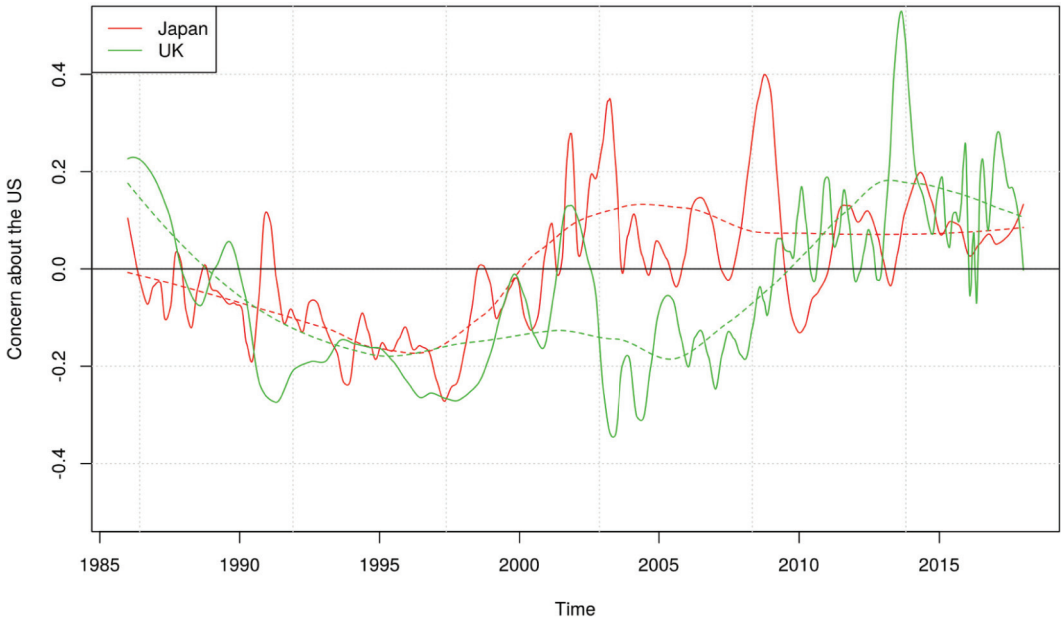

図2　日本とイギリスの新聞で見るアメリカに対する懸念

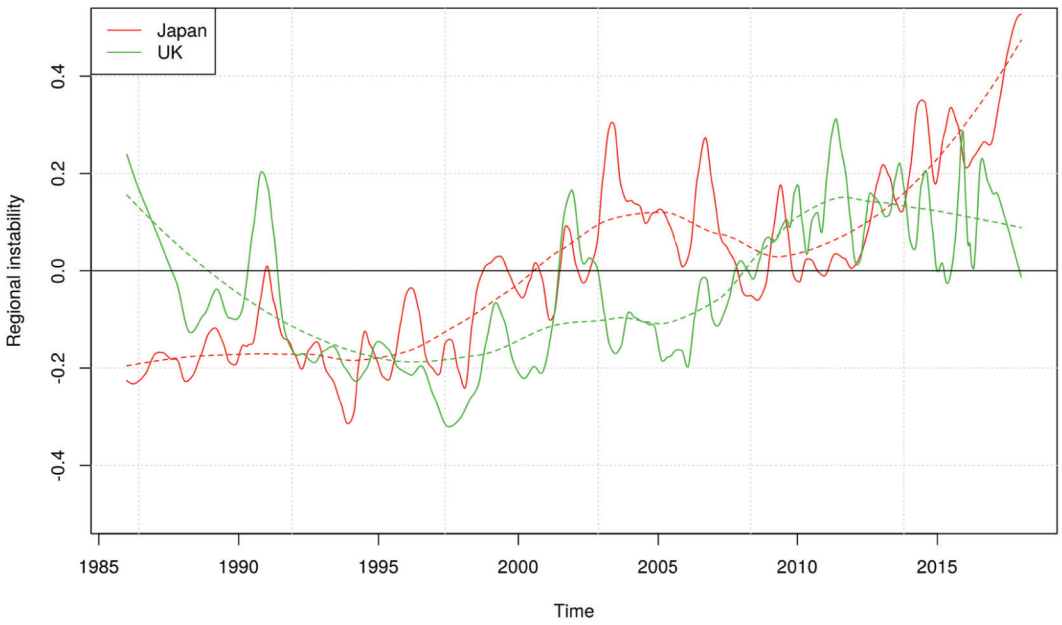

図3　日本とイギリスの新聞で見る周辺地域の情勢不安の認識

結果も筆者の認識と一致する部分が多く、準教師あり学習の有用性をはっきりと示しているように思われる。新聞記事は多様な話題や意見を含み、非常にノイズが多い文書であるが、LSS のような準教師あり学習モデルを用いると、分析者は種語によって測定の尺度を理論的な関心に従って指定できる。このような大規模な新聞記事の分析を行うためには、教師あり学習モデルであれば、数千件の記事を手作業で分析して訓練データを作成する必要があるだろう。

結論

インターネットによって文書の収集が容易になるとともに、ハードウェアが高性能化し、効率的ソフトウェアが登場したことで、欧米の政治学者によって量的テキスト分析の手法を用いた研究が広く行われるようになったが、日本の政治学者による研究はまだ多くない。その理由のひとつは、欧米で使われている量的分析のソフトウェアがアジア言語の処理を苦手としており、日本語の分析のためには日本独自のソフトウェアを使わざるを得なかったことにある。しかしながら、近年、開発された量的テキスト分析のためのソフトウェアは高い水準でアジア言語処理に対応しているため、日本語の文書の分析においても独自のツールを利用する必要がなくなり、量的テキスト分析を標準的な手続きに基づきながら、先端的な手法を用いて行えるようになっている。

本稿では、政治学者によって広く使われている教師ありおよび教師なし学習モデルと、これから注目を集めると予想される準教師あり学習モデルを紹介した。これらのモデルは語の順序を考慮せず、頻度だけで統計的な分析を行う比較的単純なものであるが、統計分析の際の計算量が少ないため特別なコンピュータを必要せず、さらに文法の違いを無視できるため言語の壁を越えて、同じモデルをすべての言語に適用できるという点で優れている。これらのモデルが、日本語の量的テキスト分析に利用できることは、二つの事例を通じて示した通りである。

筆者は、日本語の量的テキスト分析の普及を妨げる技術的な問題はほぼ解決されたと考えているが、最後に制度的な問題を指摘しておきたい。まず、量的テキスト分析に必要な日本語データの収集に関する問題である。政治学的な研究や教育に利用できる文書データは存在するが、利用者が使いやすい形で整備されているとは言い難い。この問題を解決するためには、研究者が収集してきた重要な文書（選挙マニフェストや選挙公約など）集約的に管理するデータベースを構築することが必要だろう。また、国会会議録検索システム(5)などの有用な情報源については、研究者による利用を促していくべきである。次に、量的テキスト分析では、データ収集や整理のために初歩的なプログラミングを行わなければならない場合が多いが、その能力が不足している研究者が多い。プログラミングの能力が備わった政治学の研究者を育成するためには、大学で社会データ分析のための統計の授業を増やすと同時に、授業では SPSS や Stata ではなく、R や Python を用いることが有効であろう。そして、日本には量的テキスト分析を行っている政治学者が少ないため、国内で学生や若手の研究者が先駆的な研究に触れる機会が少ない。日本における量的テキスト分析の普及を推し進めるためには、欧米の研究者を招へいし、セミナーを開催するなどの国際交流が必要であろう。

謝辞

本稿の執筆において有益な助言を与えて頂いた福元健太郎、川田恵介、前田耕、山岸光、松村尚子、ケイ・シミズ（Kay Shimizu）、アーサー・スパーリング（Arthur Spirling）に感謝の意を表する。

注

(1) 日本語の文書のデジタル化には苦労が付きまとうが、日本語向けの OCR プログラム（Panasonic の読取革命や EPSON の読んde!! ココなど）がある。また、多言語に対応する Adobe Acrobat や ABBYY Fine Reader も日本語を正確に認識できる。

(2) Web API（application programing interface）とは、ソフトウェア同士がインターネット上で情報を交換する仕組みのこと。

(3) より最近では、RcppMeCab という日本語の他に韓国語、中国語にも対応するツールの開発も進んでいる。

(4) ニュースマップには日本語および英語、ドイツ語、スペイン語、ロシア語の種辞書が用意されている。また、種辞書を新たに作成することで、文書の県や州レベルでの分類、文書の話題での分類なども行えることが知られている（Watanabe 2018a）。

(5) 国会会議録検索システムからは、R の kaigiroku パッケージ（https://github.com/amatsuo/kaigiroku）を利用すると容易にデータをダウンロードできる。

参考文献

Baker, P., Gabrielatos, C., & McEnery, T. (2012). Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word “Muslim” in the British Press 1998-2009. *Applied Linguistics*, 34(3), 255-278.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593-1636.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. (forthcoming). "Measuring and Explaining Political Sophistication Through Textual Complexity, *American Journal of Political Science*.

Blei, David M and Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation", *Journal of Machine Learning Research,* 3, pp. 993-102.

Catalinac, Amy, 2018. "Positioning under Alternative Electoral Systems: Evidence from Japanese Candidate Election Manifestos", *American Political Science Review*, 112, 1, pp. 31-48.

Catalinac, Amy, 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections", *Journal of Politics*, 78, 1, pp. 1-18.

Denny, Matthew J. and Arthur Spirling, 2018, "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It", *Political Analysis*, Vol. 26, Issue 2, pp. 168-189.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases", *Political Analysis*, Vol. 18, No. 1, pp. 1-35.

Ishida, Motohiro. 2010. *R* によるテキストマイニング入門. Tokyo, Japan: Morikita.

Jansa, Joshua M. and Eric R. Hansen and Virginia H. Gray. forthcoming, "Copy and Paste Lawmaking: Legislative Professionalism and Policy Reinvention in the States", forthcoming, *American Politics Research,* published online May 31, 2018.

Kellstedt, Paul. 2000. "Media framing and the dynamics of racial policy preferences". *American Journal of Political Science,* 44, 2, pp. 245-60.

King, Gary and Daniel Hopkins, 2010. "A Method of Automated Nonparametric Content Analysis for Social Science", *American Journal of Political Science*, Vol. 54, No. 1, January 2010, pp. 229-247.

Kluver, Heike. 2009. "Measuring interest group influence using quantitative text analysis", *European Union Politics*, Volume 10 (4): 535-549.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data", *American Political Science Review* 97(2): 311-32.

Watanabe, Kohei. (2018a). Conspiracist propaganda: How Russia promotes anti-establishment sentiment online? Presented at the ECPR General Conference, Hamburg.

Landauer, T. K. and RP. W. Foltz and D. Laham. 1998. "Introduction to Latent Semantic Analysis", *Discourse Processes*, 25, pp. 259-284.

Laver, Michael and Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data", *American Political Science Review,* 97, 2, pp. 311-331.

Nielsen, Richard. 2017. *Deadly Clerics: Blocked Ambition and the Paths to Jihad*. Cambridge University Press: New York, NY.

Rozenas, Arturas and Denis Stukal, 2018. "How Autocrats Manipulate Economic News: Evidence from Russia's State-Controlled Television", forthcoming, *Journal of Politics*.

Slapin, Jonathan and Sven-Oliver Proksch, 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts", *American Journal of Political Science*, 52, 3, pp. 705-722.

Spirling, Arthur. 2012. "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784-1911", *American Journal of Political Science*, Vol. 56, No. 1, pp. 84-97.

Trubowitz, Peter, Kohei Watanabe. (2018) International perceptions of US security commitments. Working paper.

Watanabe, Kohei. (2018b). Newsmap: A semi-supervised approach to geographical news classification. *Digital Journalism*, 6(3), 294-309.

Wilkerson, John, David Smith, Nicholas Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach", *American Journal of Political Science*, Vol. 59, No. 4, pp. 943-956.