量的テキスト分析による現代中国研究[*]

―「古くて新しい方法」の展開と課題―

2024 年 11 月
ISS Discussion Paper Series
J-253

伊藤亜聖[†]・于海春[‡]・御器谷裕樹[§]・林載桓[**]

[*] 本稿は東京大学社会科学研究所・中国学イニシアティブの支援を受けています。

[†] 東京大学社会科学研究所・准教授: asei@iss.u-tokyo.ac.jp
[‡] 北海道大学メディア・コミュニケーション研究院・助教: u_kaisyun@imc.hokudai.ac.jp
[§] 慶應義塾大学大学院法学研究科・博士課程生: yukimikiya@keio.jp
[**] 青山学院大学国際政治経済学部・教授: t13025@aoyamagakuin.jp

量的テキスト分析による現代中国研究
—「古くて新しい方法」の展開と課題—

伊藤亜聖・于海春・御器谷裕樹・林載桓

要旨

2010 年代以降、量的テキスト分析による研究が社会科学の諸分野で増加してきたが、中国研究において同手法は「古くて新しい」方法である。本稿では同手法に着目し、(1)冷戦期の研究の存在とその背景、そして後の縮小、(2)近年の増加と主要な応用例、(3)更なる可能性および課題を検討する。まず冷戦期にはデータの制約下で同手法が積極的に採用されたが、1980 年代以降には新たなデータ環境の下で同手法が下火となったことを確認する。次に 2010 年代以降の増加を論文データから示す。そして近年の代表的研究を取り上げ、検閲と情報操作をはじめとする権威主義体制の統治メカニズムの解明に寄与してきたことを確認する。最後に同手法の可能性として機械学習の更なる応用と研究課題の多様化がある一方で、データ規制への対応や質的知見との接合といった課題に加え、地域研究との間には一定の緊張関係があることを論じる。

目次

I はじめに

2010 年代以降、テキストをデータとして扱うことで定量的な分析を加える手法、いわゆる量的テキスト分析を用いた社会科学研究が増加してきた[Gentzkow, Kelly, and Taddy 2019; Frankenreiter and Livermore 2020; Grimmer, Roberts, and Stewart 2022; Ash and Hansen 2023]。この背景にはデジタル化されたテキストの増大、計算機の性能向上、分析手法の発達と普及があり、政治学、経済学、社会学、地域研究の主要学術雑誌に同手法を採用した論文が掲載されている[King, Pan, and Roberts 2013; Baker, Bloom, and Davis 2016; Kozlowski, Taddy, and Evans 2019; Jaros and Pan 2017]。しかし共産圏や中国を対象とした研究分野では同手法は決して新しいものではない。半世紀以上前から新聞や雑誌の記事をコーディングすることで政治家の認識や対外イメージを推定する手法が採用されていた[Leites, Bernaut, and Garthoff 1951; 岡部 1964]。

ではこの「古くて新しい」手法は、現代中国研究の系譜のなかにどのように位置づけられるのか。かつての量的テキスト分析はどのような研究環境下で、如何なる研究課題に取り組んだのか。その知的取り組みは改革開放という環境変化のなかで、どのように継承されたのか、あるいはされなかったのか。2010 年代以降に台頭しつつある「新しい」量的テキスト分析は、過去との対比においてどのような同質性と異質性を持つのか。国家安全保障が強調される 2020 年代に、中国研究者は中国を外部からしか見ることができない、冷戦期のような研究状況に「原点回帰」しつつあるとの見方がある[Shambaugh 2024]。このなかで「新しい」量的テキスト分析は、中国研究にいかなる可能性を提供し、どのような課題があるのか。そして地域研究とはどのような関係性があるのだろうか。

本稿の主な議論は次のとおりである。第一に、初期の研究としての量的テキスト分析は、研究資料の制約を主因として 1960 年代から 1980 年代初頭に日本にもあった一方で、改革開放期に資料環境が激変するなかで研究全体に占める位置づけは相対的に縮小した。第二に、新旧の量的テキスト分析を比較すると新聞が重要なデータであることとテキストから何らかの指標を作成するという基礎的な手法において同質性がある。その一方で、データの大規模化、データ種類の多様化と研究分野の拡大、因果推論の強調、機械学習の応用の面では、今日の量的テキスト分析は明らかに進化を遂げている。第三に、2010 年代以降の量的テキスト分析は、検閲や情報操作、中央と地方関係といった論点で成果を上げている。今後、幅広い応用が期待されると同時に、技術的進歩は必ずしも研究上の進歩を約束しないため、研究者に求められる要件も少なくない。

議論に先立って、本稿で言うところの「量的テキスト分析」と「現代中国研究」の定義や範囲を明らかにしておく必要があるだろう。本稿では過去から現在に至る量的テキスト分析の系譜をなるべく包括的に捉えるために、量的テキスト分析を「テキストの特徴を抽出するために、テキストに何らかの処理を加えて数量化されたデータに転換し、そのデータを使って定量的に分析する手法」と定義する[Benoit 2020, 468]。こうした「データとしてのテキスト・アプローチ(Text-as-Data Approach)」を中心に据えれば、従来の内容分析

(content analysis)に属する研究の少なくとも一部は本稿の検討範囲に含まれる[1]。それに対して、中間処理(データの構造化)を経ずテキストから直接意味を抽出しようとする言説分析(discourse analysis)は本稿の射程から外れることになる。

次に本稿における現代中国研究は、おおむね近現代以降(便宜的にアヘン戦争以後)の中国の政治、経済、社会を分析対象とする研究を念頭に置いている。そして中国は主に中国大陸を念頭に置くが、文脈によっては台湾研究を排除するものではない。また以下でレビューの対象とするのは、主に国際的な研究動向を反映した英語での研究および日本において発表される日本語の論文あるいは書籍である[2]。

本稿の構成は以下のとおりである。第II節において量的テキスト分析を用いた現代中国研究の初期の研究とその縮小を確認する。第III節で同手法の採用例を量的な観点から概観した上で、新旧の手法の同質性と異質性を整理し、主要な応用例を紹介する。量的把握では英語論文のデータを用いて、その量的推移を確認する。第 IV 節では同手法の更なる可能性とともに、限界や課題を指摘する。第 V 節では地域研究への含意を述べる。

II 初期の研究と縮小

１．冷戦期の共産主義研究と内容分析

量的テキスト分析の初期の研究は戦間期のマスメディアの普及を背景として、1930 年代以降にアメリカを中心に確立されてきた内容分析に遡る。内容分析は当初、「コミュニケーションの明示的内容について客観的、体系的、量的記述を行うための研究手法」と定義され、とりわけテキストを特定のカテゴリーにコーディングする手続きが強調された[Berelson 1952, 18]。この手法は第二次世界大戦中に枢軸国の宣伝分析に用いられ[George 1959]、冷戦期にはソビエト連邦や中国などの共産主義諸国の研究に使われるようになった。初期の研究である Leites, Bernaut, and Garthoff [1951]は、スターリンの 70 歳の誕生日を記念してソ連共産党機関紙『プラウダ』に掲載された 13 名(うち 11 名が政治局員)の文章をもとに、それぞれの政治家のスターリンへの態度をコーディングすることでスターリン認識を推論した。しかしその後、内容分析は第一にマニュアルコーディングのため一つの研究で扱える文書量が限られていたこと、第二に特定単語の有無に基づいた印象論的な解釈への批判があったこと、そして第三に研究費獲得が困難化し、一時期停滞した。再び盛んに使われるようになったのは、機械によるデータの処理が可能になった 1960 年代に入ってのことであった[Holsti 1969]。

---

[1] 本稿の対象とする方法は「データとしてのテキスト・アプローチ」と呼ぶこともできるが、カタリナック・渡辺[2019]に基づいて本稿では量的テキスト分析と呼ぶ。

[2] 量的テキスト分析による中国研究の長期的展開(冷戦期を含む)を考えるうえでは、英語圏および日本の研究に目を向けることに価値がある。中国大陸では近年関連研究の発展が著しいが、その点については別の研究課題としたい。

1960 年代から 1970 年代にかけて内容分析は、中国共産党やソ連共産党の対外認識やイデオロギー、国内の政策論争などを分析するための有効な手法とされた。これらの研究の共通点は、第一にコーディングを通じて文書のデータ化を行っていること、第二に資料的制約のため、少数の文書を用いていることである。例えば Wong [1967]は、内容分析が全体主義や閉鎖的な社会の分析に適していると指摘した上で、『人民日報』記事を使って中国共産党のイデオロギーや政策の変化を測定する指標を作成している。また Holsti [1966]は中国共産党 38 件とソ連共産党 44 件の文書を用いて対外関係(主に対米関係)が両国の相互認識に与える影響を検証した。同様に Kringen [1975]は中国大陸雑誌セレクションの翻訳版 81 件を対象に、「紅い専門家(中国語：又红又专)」概念の多次元性とその時系列な変化を解明しようとした。

日本でも 1960 年代から中国および共産圏研究に内容分析による研究群があった[岡部 1964; 衛藤・岡部 1965; 岡部 1971; 猪口 1970; 天児 1982; 高木 1982; 天児 1983; 高木 1983; 田中 1983]。なかでも内容分析を中国研究に応用したパイオニアは岡部達味だった。岡部はインタビューで衛藤瀋吉から内容分析の中国への応用を示唆されたと述べており[岡部 2011, 147-148]、特に嚆矢となった岡部[1964]から岡部[1971] にかけて集中的に『人民日報』の量的分析に取り組んでいる[3]。

ここで岡部[1964]を少し詳しく取り上げておこう。同論文は中国共産党機関紙『人民日報』に現れる対外認識を分析したものである。まず著者は対外政策決定過程の研究上の困難として資料上の制約を挙げながら[4]、『人民日報』に着目する意義について説明する。『人民日報』には、「必然的に自己の立場、主張の正当化という傾向、またいわゆる「都合の悪いものはのせない」という傾向が出てくる」。一方で、公式見解という権威性ゆえに「政策決定者の情勢判断、態度等をかなりくわしく知ることができる」とする[岡部 1964, 30]。そして 1949 年 10 月から 1963 年 10 月までの間に掲載された外交に関する社説合計 27 件を対象とし、社説に占める対外問題への言及スペースの増減、対外関係の緊張と緩和への言及頻度、そしてキーワード(帝国主義、平和、ソビエト、アメリカ、台湾解放等)への言及頻度から中国共産党の対外認識を分析した。

本論文の特筆すべき点は、量的分析の結果の解釈に、多面的観察による質的知見を積極的に活用している点である。対外問題への言及スペースは、朝鮮戦争介入直後の 1951 年 1 月、そしてスターリン死後のソ連共産党 20 回党大会におけるフルシチョフによるスター

---

[3] 岡部は『人民日報』しか資料がないことを指摘したうえでアレクサンダー・ジョージ、ネーザン・ライテスの研究から示唆を受けたと述べている[岡部 2011, 147]。

[4] 同論文によると同時代に西側の研究者が利用できる資料は①中国共産党の公式声明、新聞雑誌等、②訪問者・旅行者の手記と会見記録、③アメリカ、国民党政府、香港ルート等の機密情報。難民の報告、この三種類に限られた[岡部 1964, 28-29)。単著も参照[岡部 1971, 2)。

リン批判後の 1956 年 10 月が二つのピークとなっている。一方、1957 年 10 月と 1963 年 1 月に対外問題を扱うスペースが最低水準となっているのは、「次の新路線が確立されるまでの政策転換期における故意の沈黙と考えて良い」と位置づけている[岡部 1964, 33]。量的計測結果を質的知見で補正しているわけである。

その後、1980 年代の初頭までに天児[1982]、天児[1983]、高木[1982]、高木[1983]、田中[1983]のように、量的テキスト分析は若手研究者の一部にも採用され広がりを見せた。特筆に値するのはこの段階で分析データの多様化も見られたことである。高木[1983]は国連総会における中国代表による演説を対外認識の「定点観測」を可能ならしめるユニークな資料と位置づけて分析している[高木 1983, 30]。また天児[1983]は党中央機関紙である『人民日報』のみならず、山東省党機関紙である地方新聞『大衆日報』を取り上げることで、「重層的なコミュニケーションのネットワーク」に光を当てようとしている[天児 1983, 83-88]。また岡部[1971, 307-314]はシンガポールの『星州日報』と『南洋商報』に適用することで華僑・華人研究に拡張させている。なお、1970 年代まで、中国経済研究においても『人民日報』に代表されるテキストは重要な情報源であったものの、本格的な量的テキスト分析は見られなかった[5]。

２．改革開放と量的テキスト分析の相対的縮小

1972 年の日中国交正常化と 1980 年代以降の市場経済化改革と対外開放の進展により、研究者が得られる情報量は格段に増加していった。公刊される各種資料(雑誌、論文、回顧録、『中国統計年鑑』などを含む)の増加に加えて、研究者が定期的に現地を訪問し、長期滞在することも可能となった。档案館、現地聞き取り調査、アンケート調査、さらにはインターネット上の情報まで幅広く情報アクセスが広がったことは中国研究全般に大きな影響を与えた[Carlson, et al eds. 2010]。例えば Lieberthal and Oksenberg [1988]は、党と政府内の政策担当者への直接のインタビューに基づき、政策形成と実施の複雑な過程を克明に解明した。日本でも、政府関係者へのインタビューと内部文書をもとに共産党・政府関係の変化を実証的に分析した唐[1997]の研究が注目を集めた。

情報アクセスの改善は量的テキスト分析が積極的に採用されてきた対外政策研究にも新たな潮流をもたらした。早くも岡部編[1983]は「中国の内部事情についてのインフォメーションは質量ともに飛躍的に増大」しており、「内部の意見集団、利益集団の存在が、よりはっきりした輪郭をもって浮かび上がってきた」と述べている[岡部編 1983, iii]。さらに 1990 年代後半、「中国において、学問分野に限られてはいるが、言論の自由が大幅に拡大し」、中国人学者の手による研究書・論文が大量に現れ、利用できるようになった[岡部 2002, ii]。そうした状況の中、例えばヤーコブソン＆ノックス[2010]は、政府と軍の政策担

[5] 中兼和津次は「1976 年以前，われわれが中国経済研究で何をやるかというと、まずは『人民日報』をすみからすみまで読むのです」と述べている[中兼 2003, 54]。

当者や学者、企業家への広範な聞き取り調査をもとに対外政策形成過程に関わるアクターの多様化を示した。また量的テキスト分析を用いた分析でも、質的観察や調査データと組み合わせた研究が現れた[Johnston and Stockmann 2007; Stockmann 2010; Johnston 2013]。

経済研究では、改革開放期に入ると公式統計、現地調査、アンケート調査を利用することが一般化していった[中兼 1998]。五か年計画に代表される政府文書の質的な整理は継続的に行われたものの、そこから量的な指標を構築するような取り組みは引き続き限定的であった。珍しい取り組みとして丸川[2008]があり、第 8 次から第 11 次の全国レベルの五か年計画および省レベルの第 11 次五か年計画に現れる産業振興の表現(「重点発展」、「大力発展」等)と省・産業レベルの競争力(顕示的比較優位指数)の関係を検討している。

III　近年の増加と応用例

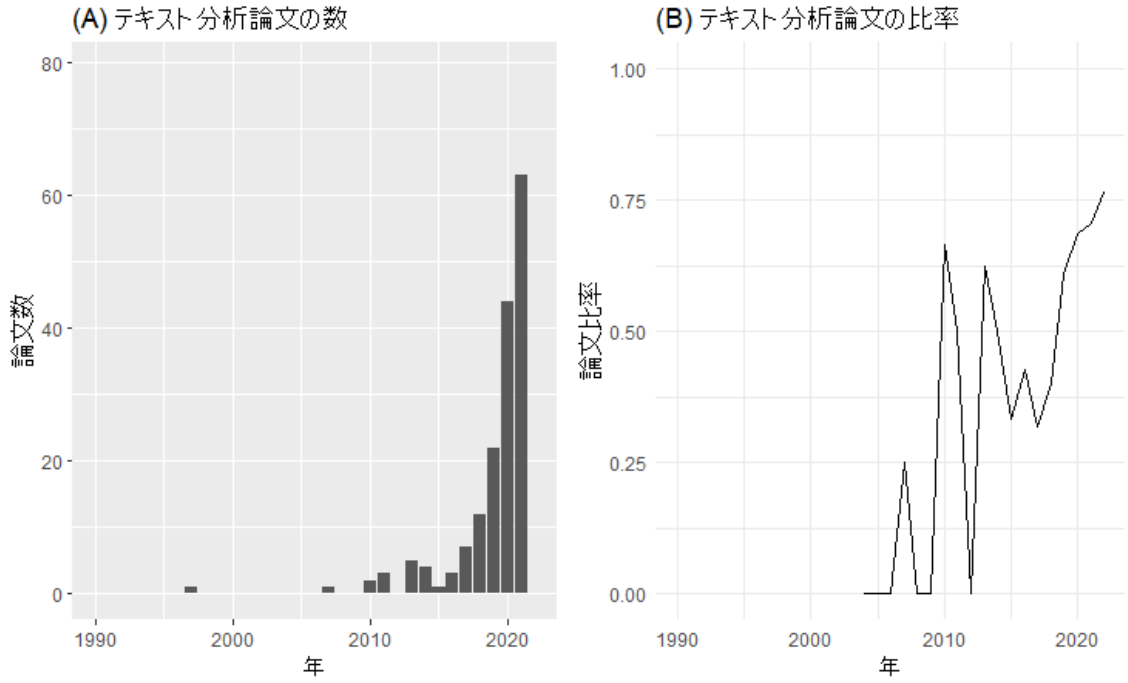
１．量的把握

2010 年代後半以降、量的テキスト分析を用いた現代中国研究の論文の数は増加している。具体的な論文は次節で取り上げることとして、本項では Web of Science データを用いてその趨勢を量的に把握する。

1990 年から 2022 年までを対象として、Web of Science をもとに中国を研究対象とする社会科学分野の論文を抽出し、そのなかから①経済学、政治学、社会学、メディア・コミュニケーション、地域研究の論文 71,889 本を抽出し、②さらに量的テキスト分析に関わるキーワードを論文タイトル・要旨・キーワードに含む論文 400 本を特定した(抽出手順の詳細は補足資料 1 を参照)。この時点での単純集計に基づくと、期間平均で当該分野の論文の 0.556%が(量的)テキスト分析を用いているものと推計される(400/71,889)。さらに 400 本のうち PDF ファイルを入手できた 374 本を対象に内容をマニュアルに確認し、量的テキスト分析論文 247 本を特定した。これは総論文の 0.34%に当たる(247/71,889)。論文要旨から特定できていない論文(偽陰性)もあると思われるが、総じて量的テキスト分析を採用した論文の総数は中国研究のなかで少数に留まっていることがわかる。

ただしその数は近年増加傾向にある。中国研究全体に占める比率は図 1 のとおりである。(A)には論文の実数を示し、該当論文は 1990 年代にはわずか 1 本で、2000 年代にも 1 本にとどまっていた。この状況は 2010 年代に変化し、2020 年には合計 44 本、そして直近の 2022 年には合計 79 本が刊行されている。年ごとに経済学、政治学、社会学、メディア・コミュニケーション、地域研究の総論文数で割って、該当論文比率を示したものが(B)である。ブレはあるものの 2022 年に 0.76%となっており、近年増加傾向にある。

図 1 現代中国を対象としてテキスト分析を用いた論文の数と比率(%)



(注)(A)は該当論文数、(B)は中国研究の論文数を分母とした際の比率(%)である。収集学術分野は政治学、経済学、社会学、地域研究、コミュニケーション・メディア研究である。収集論文の学術分野やテキスト分析を特定した方法の詳細は補足資料 1 を参照。
(出所)Web of Science データより作成。

２．同質性と異質性

それでは冷戦期の分析と近年の分析の間には、どのような異同があるのだろうか。

まず指摘できるのは両者の同質性である。抽出した論文のうち少なくない研究が新聞記事をデータとして利用している。またテキストから何らかの指標を作成する、スケーリングが主たる目的である面でも同質性がある。

一方で、近年の研究には明確な進化もみられる。第一の変化は、前提としてのコンピューターの発達である。冷戦期と比較すると、格段の計算能力を持つツールが今日の研究者には利用可能である。量的テキスト分析では有力な方法としてまず文書単語行列を作成する。この行列は文書ごとにどの単語が何回登場するかを頻度として集計する。文書単語行列は文章数に対してデータ全体に登場する単語数は膨大になるため、一般に横長のデータとなる。なおかつ特定の文書には一部の単語のみしか登場しないため、行列のうちの殆どがゼロを占める疎行列となる。今日の技術環境では文書単語行列を個人の所有する機器で構築・統計処理できる。さらに R、Python といった高水準言語とテキスト分析に特化したツールを用いて、多様な言語処理ができるようになっている[Benoit et al. 2018]。

コンピューターの性能向上と分析ツールの普及を前提として、分析データが大規模化し

た。1950 年から 1960 年代の分析が、少数のサンプルから指標を作成していたのに対して、近年の分析では新聞データであれば数万から数百万のサンプルを扱うものが一般化し、さらに SNS データの場合、数億件の投稿を扱ったものもある。データの大規模化は、有意義な研究結果を保証するものではない。しかし機械を活用することで、技術的な要因によりサンプル数を減らさざるを得ないといった障害は著しく減少した。

第二の変化は、データと研究分野の多様化である。デジタル化されたテキストデータがより豊富になっており、例えば各種ウェブサイト、有価証券報告書、論文データ、特許情報、判例等々が利用されている。研究分野も広がっており、例えば経済学分野でテキストデータが正面から利用されるようになった[Gentzkow, Kelly, and Taddy 2019]。

第三は機械学習などを用いた独自のモデル構築である。従来の辞書方式(キーワード方式)は、言語や分野によっては適切で十分なキーワードを特定できず、適切な分類や評価に限界があった。このような背景から分析領域を絞った特定タスクに対応したモデルを独自に訓練・構築する手法が開発されている。なかでも機械学習はこの課題に柔軟に対応できる手法である。機械学習には大別して教師あり学習と教師なし学習がある。例えば教師なし学習の代表的な手法はいわゆるトピックモデルである。文書に対して一定数の内在的な話題(トピック)があると想定し、そのトピックを推定する潜在的ディリクレ配分法(Latent Dirichlet allocation, LDA)が多くの研究で採用されている[Blei, Ng, and Jordan 2003]。

第四は因果推論の強調である。近年、社会科学全般において因果効果の推定が重視されるようになっており、量的テキスト分析を用いた論文においても、他の統計データ・調査データと結合させて、テキストから作成した指標が他の指標とどのような関係があるかを検証する論文が増加している。冷戦期の分析でも因果関係は検討されていたが、テキストからの指標構築に加えて他のデータを入手して数量分析することが必要となるため、その実行は困難であった[6]。今日では多面的なデータ分析の一部としてテキストデータを用いることが容易になっている。

3．応用例

前節で示した通り、量的テキスト分析を用いた中国研究は近年増加しているものの、論文全体に占める比率は捕捉できていない論文を含めたとしても決して高くない。論文数としては限定的であるが、我々は近年のテキスト分析論文はいくつかの論点において重要な成果を挙げていると考えている。以下では 2010 年代以降の研究を取り上げ、中国地域研究にとっての意義を重視してその成果を整理する。

(1) ソーシャルメディア投稿を用いた検閲と情報操作の研究

第一の研究群は検閲と情報戦略の実証研究である。中国ではウェブ上で発信される情報

[6] 因果関係の推論例として例えば高木[1982]を参照。

を当局が検閲し、ウェブサイトや投稿を閲覧できないように措置を取っていることは2000年代から知られていた。しかしどの程度の規模で、そしてどのような情報が削除されるのか、そして逆にどの程度敏感な投稿でも許容されているのかを明らかにした系統的な研究は存在していなかった。

こうした秘匿された検閲政策の実態を明らかにする研究をけん引したのは政治学者のゲイリー・キングと当時大学院生であった2名の研究者である。King, Pan, and Roberts [2013]はテキストデータから検閲のメカニズムを、遡及的に解明(リバースエンジニアリング)した。著者らは2011年前半に中国の1,382のウェブサイトから収集した 3,674,698件の投稿データをもとに、ここから127,283件をランダムサンプリングした。そしてこれらの記事が一定期間後にアクセス不能になっているか否かを確かめ、教師あり学習の一種であるReadMEを用いて、どのようなトピックにおいて検閲の確率が高いかを検証した。その結果、①全投稿の約13%が検閲されていること、②短期間に投稿量が急増するトピックにおいて検閲が顕著であること、③具体的には集団的行動に関する投稿で検閲が顕著である一方で、政府や政策に対する批判は一般に想定されるほど検閲比率は高くないことを発見した。中国共産党と政府はインターネット情報を厳しく管理すると同時に、ネット世論からフィードバックを得るために、ネット上での政府と政策に対する批判的意見を戦略的に許容している可能性が指摘されている[7]。

King, Pan, and Roberts [2017]では、「五毛党」などと呼ばれる政府に雇用された人員による投稿に着目した。キングらは、末端地方政府から情報漏洩した「五毛党」を含む電子メールのデータを利用することで、「五毛党」の新浪微博、Baidu Tieba、中央・地方レベルの政府フォーラムなどのアカウントを特定し、その投稿を収集した。そして特定した「五毛党」による投稿43,757件をもとに、ReadMeなどを用いてその投稿内容を特定・分類した。その結果、「五毛党」による投稿の約半数が政府への支持を表明する内容であり、外国の嘲笑や論争的な賞賛・批判の割合が10%未満であることが明らかになった[King, Pan, and Roberts 2017, 492]。「五毛党」は党と政府を支持する声を水増しすることで、批判的なコメントへの大衆の注目や影響力を打ち消す介入手段となっている。さらに、キングらは全国レベルでの「五毛党」活動の普及程度を推定し、合計800億のソーシャルメディア投稿のうち、約2.12億が「五毛党」によるものだとした。Roberts [2018]ではこれらの実証分析に基づいて、中国における検閲の一つのパターンとして「情報の洪水」を挙げている。キングらの実証分析を通じて、地方の公務員が実際に世論誘導を担当していることがわかり、中国当局の情報操作は実に大規模で、体系的かつ戦略的に行われていることが示された[8]。

---

[7] Qin, Strömberg, and Wu [2017]も参照。

[8] 関連研究として Gallagher and Miller [2021]、Hobbs and Roberts [2018]も参照。

(2) 全国紙を用いたプロパガンダの研究

第二の研究群は全国紙を対象にした中国共産党のプロパガンダの分析である。プロパガンダは検閲とともに大衆の情報環境を統制し、誘導する主要な手段である。体制親和的なメッセージを一方的かつ持続的に発信することが主な実施形態となる。これまで中国共産党のプロパガンダについては、例えば文化大革命期上海の宣伝・思想工作のように、特定の時期や地域での政策展開に注目したものが多く、長期間にわたる内容の変化や大衆認識上の効果を取り上げるものは少なかった。

工藤・中山[2022]は、中国共産党が正統性を確保するためにいかにして報道内容を変容させてきたかを分析している。著者らは 1950 年から 2020 年までに発行された『人民日報』の記事から「共産党」という単語が含まれる記事を約 2 万 5 千件収集した。そのうえで Newsmap と LSS を用いて中国共産党が自身をどのようにフレーミングしているかを特定した。その結果、改革開放政策の成功と世論監督を強調する基調が 2000 年代に形成されたことを指摘した。著者らによれば 1989 年の天安門事件が画期であったと主張する先行研究とは異なり、2000 年代の構造変化に着目する必要がある。

関連して Hu [2020]は、同じく『人民日報』を対象に、民主主義に対するフレーミングがどのように変わってきたのかを分析している。分析手法としては教師なし機械学習の一種である構造トピックモデルを使い、1946 年から 2003 年までの民主主義(とその類義語)に触れている約 15 万件の記事から主要なトピック(概念)を抽出し、トピック間の相関関係を分析した。その結果、民主主義に対する共産党のフレーミングには①普遍的価値の追求、②現体制の正当化、③政策課題の実現があり、特に③の側面が最も顕著に表れているとした。同論文は共産党の民主主義言説には、「民主主義とは何か」を正面から議論するより、「民主主義のために何をしているか」に焦点を当て直す特徴があると指摘している。

一方、Carter and Carter [2022]は、共産党のプロパガンダに含められたもう一つの意図を明らかにしている。それは威嚇(暗示的な威嚇を含む)を通じて大衆の集団行動への参加を抑制することである。著者らは、都市労働者を主な読者層とする全国紙の『工人日報』の記事約 16 万件から、当局による抑圧を暗示する特定のトピックを含む記事を抽出し、加えて特定のキーワードの使用頻度を数えることで記事ごとに抑圧のトーンを計測している。計測された日次の抑圧指数と実際に発生した集団行動との関係を分析することで、当局のプロパガンダには政権の正統性や統治能力をアピールする目的の他に、大衆の集団行動を威嚇する目的と効果を有していることを指摘した。

(3) 地方新聞を用いた中央・地方関係の検討

第三の研究群は地方レベルの新聞記事をデータとすることで、中央と地方の政治関係のダイナミックスや構造変化を検討するものである。

Jaros and Pan [2017]は、2012 年の習近平政権発足後に生じた変化を定量化するために、2011 年から 2014 年まで 21 の地方党報からの 1,949,409 記事を対象に分析を行った。2012

年 11 月習近平政権の発足を区切りに、その前後のそれぞれ 18 カ月の記事における個人や団体の言及される頻度を比較分析した結果、2012 年以来、省・直轄市レベルの党報において、党の最高指導者や中央レベルの政府機関の言及頻度が高くなった。一方で、外国の政治的人物に対する言及頻度の減少傾向が分かった。また政治的人物に対する報道において顕著な地域的差異が存在していることを実証的に提示した。

Chen and Hong [2021]は、省レベルの地方新聞を用い、中央と地方を跨いで展開される熾烈なエリート競争の実態を明らかにしている。著者らは、まず 143 紙の省級新聞(うち党報は 49 紙)から腐敗と産業災害を報道する記事を収集し、さらにその中から別の省の関連事項に言及している記事を特定した。このように構築された省間相互言及のテキストデータをもとに、著者らは、当該省の指導者(党書記と省長)と同一の中央リーダーをパトロンとする(つまり同じ派閥に属している)指導者がいる別の省に対して、ネガティブな報道が増加する傾向があることを示した。地方新聞は昇進をめぐるエリートたちの権力闘争のアリーナとなっているのである。日本においても工藤・于[2016]や于[2023]が地方新聞を用いて中央と地方の権力動態を分析している。

関連した研究としてメディアバイアスに関する研究もある[Piotroski, Wong, and Zhang 2017; Qin, Strömberg, and Wu 2018]。ピオトゥロスキらは中国の新聞を中央レベル地方レベルで分け、それぞれが取り扱うトピックやトーンを比較することで、中央の影響をより強く受ける新聞は、センシティブな話題を掲載せず、より明るいトーンであることを指摘した。トピックの面では、例えば中央の新聞は要人の動向を報道する傾向がある一方で、地方新聞は、より企業や市場に関する報道を重視している。

(4) 政策文書を用いた中央・地方関係の検討

第四の研究群は、北大法宝[9]に代表される政策文書データベースから得た文書を用いて、中央・地方関係のダイナミズムや政策実施の効果を分析したものである。中国のガバナンスを理解する上で、政府間コミュニケーションの態様を理解することの重要性は早くから指摘されていたものの[Oksenberg 1974]、コミュニケーションの主要な手段や政策実施上の効果に関する知見の蓄積は乏しかった。

Ang [2024]は政府間コミュニケーションの態様を、政策文書の「言葉」の分析を通じて解明している。具体的には、北大法宝のデータベースから収集した 1978 年から 2017 年までの「中央文献」4923 件を対象に、政策内容や実施義務の明確さや曖昧さを示唆する言葉を特定し、リストを作成した。このリストをもとに、教師なし機械学習モデル(K-means アルゴリズム)を使い各文書のタイプを分類した。それによると政策内容が曖昧な「グレー」の文書が全体の 23%、政策実施の対象を明確に規定している「ブラック」の文書が 20%、逆に実施すべきでない内容を明確にしている「レッド」の文書は 14%であった。中

[9] https://www.pkulaw.com/ (最終アクセス 2024 年 11 月 2 日)

央からの指示が通常曖昧で抽象的な「スローガン」に過ぎず、それゆえ政策内容の解釈と実施は地方政府に委ねられているとの通説に対して再検討を促す分析結果である。

Wang and Yang [2024]は、地方での政策実験と全国的拡大に特徴づけられる中国特有の政策モデル[Heilmann, 2018]のメカニズムの解明に挑んだ。著者らは政策実験を表すキーワード(「試点」と「試点区」)を含む中央と地方の政策文書約 2 万件を対象に、総計 633 個の政策実験の例を抽出した。それを指標として、政策実験地域の選定メカニズムや政策実験の成否を決める要因などを分析している。同論文によれば、政策実験の対象地域はそもそも恵まれた条件を有しており、いわゆる自然実験のようにランダム選ばれているわけではない。加えて、地方幹部には試験区政策の対象になっている政策に追加的な財政資源を投入するインセンティブがあり、従って当該地域での政策実験の成功確率が高くなる。その結果、政策実施の範囲を広げる(全国化する)と実験地域のような高い政策効果は現れにくい、という興味深い知見を提供している[10]。

(5) 新聞を用いた市場への影響の検討

第五の研究群は、テキストに含まれるトーンがどのように市場参加者に影響を与えているかを検討するもので、主に経済学の分野で研究が蓄積されている。

これらの研究がとくに興味深いのは、メディアバイアスを含む中国の新聞テキストからでも、経済指標に影響を与えるという意味において、影響力のある指標が作成可能であることを示しているという点である。前提にあるのは Baker, Bloom, and Davis [2016]の経済政策不確実性の研究である。同論文を嚆矢として Davis, Liu, and Sheng [2019]は『人民日報』と『光明日報』から中国版の経済政策不確実性指標を作成している。また Huang and Luk [2020]は中国国内の主要地方新聞 10 紙をベースラインとする一方、合計で地方新聞 114 紙を用いて結果の検証を行っている。

その結果、新聞記事としては掲載内容に一定の偏りがある国内新聞からでも、マクロ経済変数や企業レベルの行動といった実体経済に影響を与える指標を作成可能であることが明らかになっている。また Huang and Luk [2020]は、地方新聞 10 紙をベースライン、合計 114 紙を頑健性チェックのために利用しているが、メディアバイアスがあるとされる新聞を用いても、経済政策不確実性指数への影響が軽微であることを報告している。また関連して、新聞記事一般としてではなく、習近平国家主席に関する報道を用いることで、政治的リーダーの認識を計測し、それが企業レベルの投資行動に影響を与えていることを報告する論文もある[Ito, Lim, and Zhang 2023]。

(6) 応用例の貢献の濃淡

このように量的テキスト分析は、従来は扱えなかった規模の多様なテキストを対象とし

[10] 北大法宝データの利用例としては Tan [2020]も参照。

て、従来の手法では必ずしも分析が容易でなかった問題や論点に新しい視点と洞察を提供している。ただし、これらの研究のなかでも、学術的貢献の度合いと範囲には差があることも認識しておく必要がある。論文の引用回数から判断すれば、キングらの研究は突出しており[11]、SNS 時代の情報統制という社会科学全般に広がる新しい実証研究の系譜を作り出した。実際、近年の政治経済学では、権威主義体制の生存戦略、中でも情報の収集と統制戦略の重要性に注目が集まっており[Guriev and Treisman 2019; Egorov and Sonin 2023; Carter and Carter 2023]、キングらの研究は重要な実証研究と位置づけられている[Zhuravskaya, Petrova, and Enikolopov 2020]。さらには、中国の事例をより広くデジタルな抑圧(digital repression)として概念化する動きもあり[Earl, Maher, and Pan 2022]、この論点に限っていえば、中国の事例は中国研究の領域を超えたインパクトを与えている。

なお、かつて主流であった対外政策の研究も継続している。近年、中国のグローバルな影響力の増大に伴い、対外向けのイデオロギー・宣伝工作が活発化しており、外交部報道官のデータを用いた分析や[Mochtak and Turcsanie 2021; Dai and Luqiu 2022]、ツイッターデータを用いた分析もある[Brazys, Dukalskis, and Müller 2023]。

IV　可能性と課題

１．可能性

上記の応用例を踏まえたうえで、本節では同手法が持つ更なる可能性を指摘する。

第一に、社会科学一般に該当することだが、機械学習の手法的進化と更なる応用によって一層多様な研究を遂行する潜在力がある。まずテキストに含まれる潜在的な話題を推定するトピックモデルでは多くの改良が提案されている。例えばトピックモデルの推定の際に共変量を含める構造トピックモデルや[Roberts, et al. 2014]、キーワードを与えることによって特定トピックを抽出できる手法が提案されている[Eshima, Imai, and Sasaki 2023]。また自然言語処理の分野で単語の意味を高密度の行列として表現する単語埋め込み(word embeddings)が開発され、政治学領域では単語埋め込みを利用して、意味の違いを示す埋め込み回帰も提案されている[Rodriguez, Spirling, and Stewart 2023]。また近年では深層学習の進展が著しく、同一単語の文脈による意味の変化も捉えるようになっている。2018 年にグーグルが提案した自然言語処理技術 Bidirectional Encoder Representations from Transformers (BERT)を用いて、例えば経済学では求人情報から当該求人がリモートワークを許容している否かを判定し、既存の方式よりも高い精度でデータを作成するといった事例が報告されている[Hansen et al. 2023]。深層学習を用いた画像認識精度向上による歴史資料の高精度デジタル化や[Bryan et al. 2023]、ChatGPT を利用してテキストの感情やトピックの判定に使う動きもある[Ornstein, Blasingame, and Truscott 2023] 。

---

[11] King, Pan, and Roberts [2013]の 2024 年 10 月 30 日時点での Google Scholar 論文引用件数は 3,035 である。

第二に、分析データの拡張と新鮮な問いによる研究の余地の大きさである。利用されるテキストの種類は拡大しているが、例えば近現代の歴史資料を用いた社会科学的分析は依然として少数である。歴史資料では、例えば、地方誌(県誌など)をデジタル化して自然言語処理することで、1950 年代の土地改革の成果を県レベルの指標を作成することで検証しようとする事例がある[Alesina et al. 2020]。またかつての研究が華僑・華人研究にも視野を広げていたことを考えれば[岡部 1971, 299-335]、華字新聞等を用いることで分析の視野はさらに広がる可能性がある。加えて中国語以外のテキストも含めた多言語での比較分析の可能性もある。

第三に、テキストデータと地域研究の相性の良さである。テキストデータは情報量が多いという意味で高次元なデータであり、特定のデータから抽出しうる論点は幅広い。例えば先行研究が実施していたように、一定の限界のもとで、『人民日報』記事から外交や日本への関心を測ることができる。内政的論点、例えば反腐敗運動に関心があれば、その指標も作成できる。つまり量的テキスト分析は、高次元なデータの性質上、実装の過程で多様な定性的知見の活用が可能であり、かつそうした知見が必要な研究手法である。経済学や政治学のディシプリンとの対比において、地域研究は、理論的に想定される限られた変数に着目するのみならず、特定社会における歴史的経緯や経路依存性に由来する変数や関係性に興味を持つ。手法によっては事前の仮定を置かずにテキストから意味や論点を抽出できる量的テキスト分析と地域研究は相性がよく、応用の可能性が大きい。

２．課題

上述した可能性がある一方で、今後の更なる研究の遂行においては課題もある。

第一に、テキスト分析一般が直面している課題がある。より多くの量的テキスト分析が蓄積されることで、分析手続きの際に注意すべき点も明らかになっている。例えばテキストの統計処理の際には、一般的なテキストデータのクリーニング(不要な記号や URL 等の削除)に加えて、計算上の負荷を軽減する目的から、不要と考えられる単語(いわゆるストップワード)を削除することによって次元削減を行うことが多い。これらは前処理と呼ばれており、この過程での選択の差異は、例えばトピックモデルの推定結果にも影響を与える[Denny and Spirling 2018]。また文書間の類似性を計測する際にも、特徴を文書単語行列で作成するのか、トピックモデルで作成するのか、またその手続き次第で結果が大きく変わる[Ash and Hansen 2023, 676-679]。加えて、テキストデータを用いた指標の測定がより容易となるなかで、計測された指標の妥当性の検証は特に重要な作業である[Grimmer, Roberts, and Stewart 2022, 211-218; Grimmer and Stewart 2013, 271]。

第二に、中国研究が直面しうる課題として、そもそもテキストデータへのアクセス、とりわけ中国国外からのアクセスが技術的あるいは法的に制限されるリスクがある。一般に情報を提供するウェブサイトが利用規約上、情報の抽出技術(自動スクレイピング等)の使用を禁止している場合がある。この背景には接続先サーバーへの過重な負荷をかけるケー

スや著作権侵害を理由とする場合もある。ログイン要件やアンチスクレイピング技術を導入して情報の抽出を技術的に規制する動きは一般化している。加えて中国では法的な規制リスクも近年明確となってきている。例えば論文データを扱う中国最大手のプラットフォーム「中国知網」(CNKI)に対しては中国当局が締め付けを強化している[12]。国家インターネット情報弁公室は 2023 年初めに CNKI の運営会社に対して学位論文(修士・博士)、学術会議録(国内会議、国際会議)、特許、統計の 4 種類のコンテンツの提供に対して新たな規制が適用されることを通知した。この結果、米国、日本、台湾、香港の複数の大学や研究機関は、2023 年 4 月 1 日からこれらのフルテキストへのアクセスが「一時的に停止」されている[13]。

第三に、量的テキスト分析が万能薬ではないことを自戒し、そのうえで中国研究の蓄積を有効に生かすことである。量的テキスト分析は、あくまでも研究者の理解を補助するものであり、「言語は複雑であるため、自動化された内容分析手法がテキストの注意深い精読に取って代わることは決してない」[Grimmer and Stewart 2013, 268]。研究者は質的な読解や事後的な妥当性検証を蔑ろにできない。量的テキスト分析は質的な内容分析や文献調査と相互補完関係にあることを分析者は意識する必要がある。また利用するテキストや測定しようとする尺度の設定等では、中国研究の蓄積を生かしていくべきだろう。例えば岡部[1964]は特定日に掲載される社説に特化することで分析サンプル数を減らし、関連情報への深い理解に基づく解釈を行った。またオクセンバーグは『人民日報』を用いた量的分析を行う際に、仮に毛沢東路線を計測したいのであれば、文化大革命期の一時期には『解放軍報』を使う必要があると指摘している[Oksenberg 1969, 597]。このような質的な知見は有効な研究デザインを策定するうえで有用である。また近年では、ロボットによる自動的な情報発信が一般化し、収集したテキストにそもそもバイアスが多く含まれるリスクがある。業者がデータベースを作成する際に検閲または自己検閲を行うことで、幾重にもデータにバイアスがかかった状態で分析を行う危険性もある[14]。このように、テキストデータを収集する際にはバイアスやノイズが含まれうるため、研究者はデータ生成過程や内在的なバイアスに自覚的でなければならない。

地政学的な環境変化により中国研究が冷戦期のようなアプローチに「原点回帰」しつつあるとの Shambaugh [2024]の指摘は重いものだが、我々は以上の諸点を踏まえて、同じようにテキストを用いたとしても、そしてデータアクセスの面を筆頭として新たな困難があ

---

[12] 安全保障貿易情報センター[2023]および日本貿易振興機構[2023]参照。

[13] UC Berkeley Library, March 21st, 2023, “Important service announcement for CNKI resources”. (https://update.lib.berkeley.edu/2023/03/21/important-service-announcement-for-cnki-resources/)を参照(最終アクセス 2024 年 11 月 2 日)。

[14] 例えば Roberts et al. [2023]は中国の司法裁判所が公開する判例データは、その一部が削除されていることを明らかにしている。

りつつも、社会科学研究としての進歩がありえると考えている。

V　おわりに

本稿は中国研究を事例として量的テキスト分析の初期の研究と近年の増加傾向、そして応用例と課題を検討してきた。最後に、本稿の考察が量的テキスト分析と地域研究の関係についてどのような示唆を与えているかを述べておきたい。

前述したように、量的テキスト分析は、データの性質や分析モデルの柔軟性という点において、多様な変数間の複雑な関係を想定する地域研究との相性が良い。しかし同時に、両者間の距離や緊張関係についても認識しておく必要がある。

両者の緊張関係は、第一に、近年の量的テキスト分析を牽引してきた機械学習の性質と効用に起因する。量的テキスト分析は定性的な知見を活かせる定量的な手法であるものの、そのすべての手法が地域研究の重視する内在的変数の析出や特定の文脈に根ざした関係性の解明に向いているわけではない。むしろ機械学習の効用は、複雑な構造を有する大量のデータから一定の大まかなパターンを見出すことを基盤としている。つまり地域研究が重んじる個別事象の詳細な説明は、機械学習一般に求められる役割ではない。

両者の緊張関係のもう一つの根源は、量的テキスト分析の実装に必要なスキルの習得コストである。量的テキスト分析は対象データの特質上(他の機械学習に不必要な)追加的な処理を行う必要があり、データの構築と操作のためのスキルの習得がどうしても必要になる。加えて分析結果の解釈には統計的素養も求められる。こうした学習コストも地域研究にとって量的テキスト分析を実装する上でのハードルになっている。

さらに言えば、機械学習の進歩、中でも領域固有型の大規模言語モデルの開発が、地域研究者の優位性を掘り崩す可能性もある。例えば北京大学の研究者は中国の法ドメインに特化した ChatLAW を開発している[Cui et al. 2023]。深層学習分野における技術革新のなかで、いかに地域研究者がその特性を生かしていくかが問われている。

これらを踏まえたうえで、我々は機械学習の特性と今後の発展が、量的テキスト分析と地域研究の間の距離を遠ざける方向に作用するとは限らないと考えている。機械学習にはデータの生成過程に事前に強い仮定をおく必要のない多様な手法が備わっており、その分析結果を地域研究の定性的知見を生かして深く解釈する余地が大きい。また、近年の生成系 AI を含む大規模言語モデルの発展は、両者間のハードルを下げる追い風となる潜在性がある。2022 年 11 月 30 日にリリースされた ChatGPT に代表される生成系 AI が Python や R のコーティングをサポートするようになっている(コパイロット、すなわち「副操縦士」機能)。これにより機械学習の実装のためのスキル習得コストが低下することで、今後、地域研究が強みを持つ特定の文脈的知見または領域固有知識の相対的な価値が増す可能性もある。いずれにしても地域研究と技術的フロンティアの距離が縮まってくることは不可避だと思われる。今後、地域研究者はディシプリンの研究者に加えて計算機科学者も含めて、異なる分野の研究者と一層交流し、協力していくことが求められるだろう。

参考文献

※日本語

天児慧 1982.「政治転換期における大衆動員――一九五六―五八年の『人民日報』投書欄分析を中心として」衛藤瀋吉編『現代中国政治の構造』日本国際問題研究所: 201-237 頁.

田中明彦 1983.「「教科書問題」をめぐる中国の政策決定」岡部達味編: 193-219 頁.

安全保障貿易情報センター(CISTEC) (2023)「中国の最近の輸出規制とその関連動向(第 2 版)　2022 年秋以降の動向を中心として」安全保障貿易情報センター, 2023 年 2 月 27 日掲載.

猪口孝著・衛藤瀋吉監修 1970.『国際関係の数量分析 : 北京・平壌・モスクワ 1961-1966 年』巌南堂書店.

于海春 2023.『中国のメディア統制: 地域間の「不均等な自由」を生む政治と市場』勁草書房.

衛藤瀋吉・岡部達味 1965.「中華人民共和国対日発言の内容分析―1958 年の 2 つの時期における人民日報を材料として」『外務省調査月報』第 6 巻第 1 号: 1-17 頁.

岡部達味 1964.「内容分析による中共対外政策の研究」『アジア研究』10(4)、28-58 頁。

岡部達味 1971.『現代中国の対外政策』東京大学出版会.

岡部達味編 1983.『中国外交　政策決定の構造』日本国際問題研究所。

岡部達味 2002.『中国の対外戦略』東京大学出版会.

岡部達味 2011.「同時代研究としての中国研究」平野健一郎・土田哲夫・村田雄二郎・石之瑜編『インタビュー戦後日本の中国研究』平凡社: 129-177 頁.

カタリナック・エイミー、渡辺耕平(2019)「日本語の量的テキスト分析」『早稲田大学高等研究所紀要』第 11 号: 133-143 頁.

工藤文・于海春 2016.「党報と都市報の「両会」に関するアテンション分析――中国の新聞を用いたテキストマイニングから」『早稲田政治公法研究』第 112 号: 1-17 頁.

工藤文・中山敬介 2022.「『人民日報』における報道内容の変容－1950 年から 2020 年を対象とした計量テキスト分析－」『メディア研究』第 101 巻: 233-253 頁.

高木誠一郎 1982.「「国民経済発展十か年計画要綱」の後退と因果関係の認識――認知構造図法による接近――」衛藤瀋吉編『現代中国政治の構造』日本国際問題研究所: 238-275 頁.

高木誠一郎 1983.「中国の対外認識の展開(一九七二―一九八二)―国連総会一般演説の内容分析－」、岡部達味編: 29-63 頁.

田中明彦 1983.「「教科書問題」をめぐる中国の政策決定」岡部達味編、193-219 頁。

中兼和津次 1998.『中国経済発展論』有斐閣.

唐亮 1997.『現代中国の党政関係』慶應義塾大学出版会.

中兼和津次 2003.「わが国における中国経済研究の回顧と展望」『中国経済研究』第 1 巻第

1 号: 51-64 頁.
日本貿易振興機構 2023.「中国当局、個人情報保護法違反の知網(CNKI)に約 10 億円の罰金」『ビジネス短信』2023 年 9 月 8 日.
丸川知雄 2008.「21 世紀型の産業政策――中国の事例を中心に」武田康裕・丸川知雄・厳善平共編『現代アジア研究 3 政策』慶應義塾大学出版会: 209-230 頁.
ヤーコブソン、リンダ&ディーン・ノックス『中国の新しい対外政策―誰がどのように決定しているのか』岩波現代文庫.

※英語
Alesina, Alberto F., Marlon Seror, David Yang, Yang You, and Weihong Zeng 2020. “Persistence despite revolutions.” NBER Working Paper No. w27053. National Bureau of Economic Research.
Ang, Yuenyuan 2024. “Ambiguity and Clarity in China's Adaptive Policy Communication.” *The China Quarterly* 257: 20-37.
Ash, Elliott and Stephen Hansen 2023. “Text algorithms in economics.” *Annual Review of Economics*, 15: 659-688.
Baker, Scott R., Nicholas Bloom, and Steven J. Davis 2016. “Measuring economic policy uncertainty.” *The Quarterly Journal of Economics* 131(4): 1593-1636.
Berelson, Bernard 1952. *Content Analysis in Communication Research*. Free Press.
Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo 2018. “quanteda: An R package for the quantitative analysis of textual data.” *Journal of Open Source Software* 3(30): 774-774.
Benoit, Kenneth 2020. “Text as data: An overview.” *The SAGE handbook of research methods in political science and international relations*, 461-497.
Blei, David M., Andew Ng, and Michael Jordan 2003. “Latent dirichlet allocation.” *Journal of Machine Learning Research*, 3(Jan), 993-1022.
Brazys, Samuel, Alexander Dukalskis, and Stefan Müller 2023. “Leader of the Pack? Changes in “Wolf Warrior Diplomacy” after a Politburo Collective Study Session.” *The China Quarterly* 254: 484-493.
Bryan, Tom, Jacob Carlson, Abhishek Arora, and Melissa Dell 2023. “EfficientOCR: An Extensible, Open-Source Package for Efficiently Digitizing World Knowledge.” arXiv preprint arXiv:2310.10050.
Carlson, Allen, Mary E. Gallagher, Kenneth Lieberthal, and Melanie Manion eds. 2010. *Contemporary Chinese Politics: New sources, methods, and field strategies*. Cambridge University Press.
Carter, Erin Baggott and Brett L. Carter 2022. “When autocrats threaten citizens with violence:

Evidence from China." *British Journal of Political Science* 52: 671-696.

Carter, Erin Baggott and Brett L. Carter 2023. *Propaganda in Autocracies: Institutions, information, and the politics of belief*. Princeton University Press.

Chen, Ting and Ji Yeon Hong 2021. "Rivals within; Political factions, loyalty, and elite competition under authoritarianism." *Political Science Research and Methods* 9: 599-614.

Cui, Jiaxi, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan 2023. "Chatlaw: Open-source legal large language model with integrated external knowledge bases." arXiv preprint arXiv:2306.16092.

Davis, J. Steven, Dingqian Liu, and Xuguang S. Sheng 2019. "Economic policy uncertainty in China since 1949: The view from mainland newspapers." In Fourth Annual IMF-Atlanta Fed Research Workshop on China's Economy Atlanta, Vol. 19: 1-37.

Dai, Yaoyao and Luwei Rose Luqiu 2022. "Wolf Warriors and Diplomacy in the New Era." *The China Review* 22(2): 253-283.

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 26*(2): 168-189.

Earl, Jennifer, Thomas V. Maher, and Jennifer Pan 2022. "The digital repression of social movements, protest, and activism: A synthetic review." *Science Advances* 8(10), eabl8198.

Egorov, Georgy and Sonin, Konstantin 2023. "The political economics of non-democracy." *Journal of Economic Perspective* 62(2): 594-636.

Eshima, Shusei, Kosuke Imai, and Tomoya Sasaki 2022. "Keyword-Assisted Topic Models." *American Journal of Political Science* 68(2): 730-750.

Frankenreiter, Jens and Michael A. Livermore 2020. "Computational methods in legal analysis." *Annual Review of Law and Social Science* 16: 39-57.

Gallagher, Mary and Blake Miller 2021. "Who not what: The logic of China's information control strategy." *The China Quarterly* 248: 1011-1036.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy 2019. "Text as data." *Journal of Economic Literature* 57(3): 535-574.

George, Alexander L. 1959. *Propaganda Analysis*. Evanston, Illinois: Row, Peterson & Co.

Guriev, Sergei and Daniel Treisman 2019. "Informational autocrats." *Journal of Economic Perspectives* 33(4): 100-127.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart 2022. *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, Justin and Brandon M. Stewart 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis 21*(3): 267-297.

Hansen, Stephen, Peter John Lambert, Nick Bloom, Steven J. Davis, Raffaella Sadun, and Bledi Taska 2023. "Remote work across jobs, companies, and space." NBER Working Paper No.

w31007. National Bureau of Economic Research.

Heilmann, Sebastian 2018. *Red Swan: How unorthodox policy-making facilitated China's rise*. The Chinese University of Hong Kong Press.

Hobbs, William R. and Margaret E. Roberts 2018. "How sudden censorship can increase access to information." *American Political Science Review 112*(3): 621-636.

Holsti, Ole R. 1966. "External conflict and internal consensus: The Sino-Soviet case." In *The General Inquirer: A computer approach to content analysis*, 343-358.

Holsti, Ole R. 1969. *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley Publishing Company.

Hu, Yue 2020. "Refocusing Democracy: The Chinese government's framing strategy in political language." *Democratization* 27(2): 302-320.

Huang, Yun and Paul Luk 2020. "Measuring economic policy uncertainty in China." *China Economic Review* 59, 101367.

Ito, Asei, Jaehwan Lim, and Hongyong Zhang 2023. "Catching the Political Leader's Signal: Economic policy uncertainty and firm investment in China." *China Economic Review* 102035.

Jaros, Kyle and Jennifer Pan 2017. "China's Newsmakers: Official Media Coverage and Political Shifts in the Xi Jinping Era." *The China Quarterly* 233.

Johnston, Alastair Iain 2013. "How new and assertive is China's new assertiveness?" *International Security* 37(4): 7-48.

Johnston, Alastair Iain and Daniela Stockmann 2007. "Chinese Attitudes toward the United States and Americans." In Peter J. Katzenstein and Robert O. Keohane (eds.), *Anti-Americanism in World Politics*. Ithaca, NY: Cornell University Press: 157–195.

King, Gary, Jennifer Pan, Margaret E. Roberts 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(2): 326-343.

King, Gary, Jennifer Pan, Margaret E. Roberts 2017. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American Political Science Review* 111(3): 484-501.

Kozlowski, Austin C., Matt Taddy, and James A. Evans 2019. "The geometry of culture: Analyzing the meanings of class through word embeddings." *American Sociological Review* 84(5): 905-949.

Kringen, John A. 1975. "An exploration of the "red-expert" issue in China through content analysis." *Asian Survey* 15(8): 693-707.

Leites, Nathan, Elsa Bernaut, and Raymond L. Garthoff 1951. "Politburo images of Stalin." *World Politics* 3.3: 317-339.

Lieberthal, Kenneth and Michel Oksenberg 1988. *Policy Making in China: Leaders, Structures, and*

Processes. Princeton University Press.

Mochtak, Michal and Richard E. Turcsanie 2021. “Studying foreign policy narratives: Introducing the Ministry of Foreign Affairs press conference corpus.” *Journal of Chinese Political Science* 26: 743-761.

Oksenberg, Michel 1969. “Sources and methodological problem in the study of contemporary China.” In A. Doak Barnett ed. *Chinese Communist Politics in Action*, Seattle and London: University of Washington Press. pp.577-605.

Oksenberg, Michel 1974. “Method of communication within the Chinese bureaucracy.” *The China Quarterly* 57: 1-39.

Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott 2023. “How to Train Your Stochastic Parrot: Large Language Models for Political Texts.” mimeo.

Piotroski, Joseph D., T. J. Wong, and Tianyu Zhang 2017. “Political bias in corporate news: the role of conglomeration reform in China.” *The Journal of Law and Economics* 60(1): 173–207.

Qin, Bei, David Strömberg, and Yanhui Wu 2017. “Why does China allow freer social media? Protests versus surveillance and propaganda.” *Journal of Economic Perspectives* 31(1): 117-140.

Qin, Bei, David Strömberg, and Yanhui Wu 2018. “Media bias in China.” *American Economic Review* 108(9): 2442-2476.

Roberts, Margaret E. 2018. *Censored: Distraction and diversion inside China's Great Firewall*. Princeton University Press.

Roberts, Margaret E., Benhamin Liebman, Rachel Stern, and Xiaohan Wu 2023. “Rolling Back Transparency in China's Courts.” American Political Science Association, Presentation.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand 2014. “Structural topic models for open-ended survey responses.” *American Journal of Political Science* 58(4): 1064-1082.

Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart 2023. “Embedding regression: Models for context-specific description and inference.” *American Political Science Review* 117(4): 1255-1274.

Shambaugh, David 2024. “The Evolution of American Contemporary China Studies: Coming Full Circle?” *Journal of Contemporary China* 33(146): 314-331.

Stockmann, Daniela 2010. “Who believes propaganda? Media effects during the anti-Japanese protests in Beijing.” *The China Quarterly* 202: 269-289.

Tan, Yeling 2020. “Disaggregating “China, Inc.”: The hierarchical politics of WTO entry.” *Comparative Political Studies* 53(13): 2118-2152.

Wang, Shaoda and David Yang 2024. “Policy experimentation in China: The political economy of

policy learning." *Journal of Political Economy*, forthcoming.

Wong, Paul 1967. "Coding and Analysis of Documentary Materials from Communist China." *Asian Survey* 7(3): 198-211.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov 2020. "Political effects of the internet and social media." *Annual Review of Economics* 12: 415-438.

補足資料　量的テキスト分析による現代中国研究論文データセットの構築

第三節第一項の集計で用いている論文は、以下の 2 段階で抽出した。

第一段階　「現代中国に関する社会科学論文」の抽出

クラリベイト・アナリティクス社が提供する Web of Science データベースを用いて、論文の書誌情報(要旨を含む)をダウンロードした。検索画面から①論文要旨に China OR Chinese を含み、なおかつ Web of Science Categories (WC)に下記の補表 1 の「社会科学 4」に該当するカテゴリーの、③論文または書籍所収の章(Article or book chapter)で、④1990 年 1 月 1 日から 2022 年 12 月 31 日に刊行されたものを抽出し、該当する 137,037 件をダウンロードした。この段階のデータには、近刊(forthcoming)論文や、英語以外の他言語論文、一部重複データもあったため、これらをクリーニングし、分析対象の「現代中国に関する社会科学論文」は 126,555 本となった。

補表 1 Web of Science 抽出カテゴリー

名前	社会科学1：Academy of Social Sciences, UK https://acss.org.uk/what-is-social-science/	社会科学2：The National Institute of Social Sciences, USA https://www.socialsciencesinstitute.org/what-is-social-science	社会科学3：独自分類	社会科学4:１～３のいずれかでも該当したらYES
Agricultural Economics Policy	YES	YES	YES	YES
Anthropology	YES	YES		YES
Area Studies	YES		YES	YES
Asian Studies	YES		YES	YES
Business	YES		YES	YES
Business Finance	YES		YES	YES
Communication			YES	YES
Development Studies	YES			YES
Economics	YES	YES	YES	YES
Education Educational Research	YES	YES		YES
Education Scientific Disciplines	YES	YES		YES
Geography	YES			YES
History		YES		YES
Hospitality Leisure Sport Tourism	YES			YES
International Relations	YES		YES	YES
Language Linguistics	YES			YES
Law		YES	YES	YES
Linguistics	YES			YES
Management	YES		YES	YES
Operations Research Management Science	YES			YES
Political Science	YES	YES	YES	YES
Psychology	YES	YES		YES
Psychology Applied	YES	YES		YES
Psychology Clinical	YES	YES		YES
Psychology Educational	YES	YES		YES
Psychology Experimental	YES	YES		YES
Psychology Multidisciplinary	YES	YES		YES
Psychology Social	YES	YES		YES
Regional Urban Planning	YES			YES
Social Sciences Interdisciplinary	YES	YES	YES	YES
Social Sciences Mathematical Methods	YES	YES		YES
Sociology	YES	YES	YES	YES
Statistics Probability	YES			YES
Telecommunications			YES	YES
Urban Studies	YES			YES

（出所）筆者ら作成。

第二段階　定量テキスト分析論文と分析対象分野論文の抽出

第一段階で抽出された論文は、複数分野にまたがった論文が残っており、とくにコンピューターサイエンス系や心理学系の論文など、本稿の主要な問題意識から外れている分野の論文も数多く残っている。このため WoS 書誌情報の Research Areas(SC)欄を用いて、“ECONOMICS”, “DEVELOPMENT STUDIES”, “INTERNATIONAL, RELATIONS”, “GOVERNMENT & LAW”, “SOCIOLOGY”, “SOCIAL SCIENCES”, “COMMUNICATION”, “AREA STUDIES”, “ASIAN STUDIES”を含む論文を抽出し、そのうえで、“COMPUTER SCIENCE”と“ENGINEERING”を含む論文を除外した。この結果、71,889 本が残った。

更に、量的テキスト分析を用いた論文を抽出するために、下記のキーワードのいずれかを①論文タイトル、②論文要旨、③著者指定キーワード、④キーワードプラスに含む論文を検索し 400 本を特定し、これらを第三節第一項のコーティング分析の対象とした[15]。

テキスト分析キーワード:

"text analysis|textual analysis|quantitative content analysis|topic model|sentiment analysis|text-as-data|text data|textual data|natural language processing|text mining|computer-assisted text|computer-assisted read|corpus linguistics|computational linguistics|word embeddings|latent dirichlet allocation|emotion detection|text classification|text clustering|part-of-speech tagging|lemmatization|tokenization|syntax analysis|text coherence|text readability|text summarization|speech processing|n-gram|ngram|mining text|text regression|million social media|billion social media|million blog posts|billion blog posts|automated text analysis|based on Chinese newspaper|based on newspaper|full text"

[15] キーワードプラスは WoS がアルゴリズムによって作成したキーワードで、当該論文の参考文献のタイトルに頻繁に出現するが、記事自体のタイトルには出現しない単語またはフレーズを抽出している。