内容抽出のためのテキストマイニング手法の比較研究
――日本の歴代首相の所信表明演説の内容分析を例に――

毛　　文　偉

1．はじめに

テキストを計量的に分析することでテキストの内容を抽出する研究手法として、TF-IDF、対応分析、トピックモデルなどが挙げられるが、それぞれに一長一短があり、すべての研究ニーズに対応できるわけではない。本稿はこれらの方法をもって2000年から2021年までの日本の歴代首相の所信表明演説を分析し、内容と通時的変化を比較し、各研究手法の長短を探求する。

2．従来の研究

まず、政治テキストを計量的に分析する研究を概観する。東照二（2006）は、東條英機から小泉純一郎まで、戦中・戦後の歴代首相の国会での演説や答弁を調査し、「あります」や「いたしました」などそれぞれの話し方の特徴や時代による変化を分析して、政治とことばの関係を詳しく考察している。鈴木崇史・影浦峡（2011）は演説の語彙の多様性（TTR）と語彙の偏り（シンプソンのD）という視点から、中曽根康弘と小泉純一郎の施政方針演説と所信表明演説を比較し、両者に差異が観察されないと述べている。また、他の総理大臣と比較すると、圧倒的に小泉において特徴的な値を示しており、小泉独自の特徴をあらわすものであると指摘している。河瀬彰宏・吉原秀樹（2020）は戦後の東久邇宮内閣から2019年の第四次安倍内閣までの歴代首相の演説を昭和前期、昭和後期、平成前期、平成後期の四つの時代に分けた上で、TF-IDF 値を算出することによって特徴語を抽出している。そして全演説に対してLDA によるトピックモデルを作成し、首相ごとのトピックの推定を実施し、各時代区分で政治思想の変化がみられることを指摘している。すなわち、戦後から現在にかけて「国家制度の整備」、「経済の発展」、「国家制度の再建」へと演説のトピックが遷移していることを明らかにしている。

一方、政治テキストの分析ではないが、対応分析の応用例として、金明哲（2018）が挙げられる。金明哲（2018）は大学生の作文から抽出した名詞のデータセットを対象に対応分析を行い、作文をグループ分けして、キーワードとの対応関係を探っている。また、原田朋子

（2019）は、日本語の母語話者と学習者の作文を対応分析で考察し、「すなわち」、「つまり」、「だが」、「したがって」、「ところが」などの接続表現は母語話者の作文で相対的に多く使用される傾向があり、「次に」、「それに」、「ですから」などの接続表現は相対的に学習者の作文でよく現れることを指摘している。

以上のように、先行研究ではTF-IDF値、対応分析、トピックモデルをもって政治テキストや作文などを分析し、その内容を抽出する研究手法が行われている。しかし、管見の限り、これらのデータ分析法の比較研究はいまだ不十分であると思われる。以下、筆者は日本の歴代首相の所信表明演説を考察し、その内容と通時的変化を明らかにするとともに、この三つのデータ分析法を比較し、その特性を究明したいと思う。

3．データの概要

日本の総理大臣の国会演説は施政方針演説と所信表明演説に分けられる。施政方針演説はその内閣の施政全体を示す包括的な内容を含むのに対して、所信表明演説は臨時国会に行われ、含まれる政策トピックは選択的なものとなり、総理大臣個人の意向がより強く反映され（鈴木崇史・影浦峡 2011）、個人の言語特徴が色濃く出ている（ソジエ内田恵美 2018）。そこで本研究では後者に注目し、データベース「世界と日本」[1)]に載せられた2000年以降の歴代首相の所信表明演説を計17部ダウンロードして、研究対象とした。

まず、各演説のTTR（Type-Token Ratio）とSTTR（Standardized Type-Token Ratio）を集計した（表1）。TTRとSTTRはいずれもテキストの語彙の多様性を測定するための指標で、値が大きいほどテキストの語彙の多様性が高いことを示す。しかし、TTRはテキストの長さに影響を受けやすいという欠点があり、同じ質のテキストでも長くなると、TTRは低下してしまう。一方、STTRはテキストを同じサイズのセグメントに分割し、各セグメントのTTRを計算して平均を取ることで計算される。これにより、テキストの長さによる影響を減らし、異なる長さのテキスト間で語彙の多様性を比較することが可能になる。所信表明演説には2000語ぐらいのもの（小泉純一郎2005、安倍晋三2017など）と8000語を超えるもの（鳩山由紀夫2009）があり、その長さにかなりの差があるため、STTRで語彙の多様性を評価したほうがよいと考えらえる。

STTRが最も高いのは小泉純一郎（2004）、菅義偉（2020）、安倍晋三（2016）であり、ほかの演説に比べ語彙がより多様である。最も低いのは麻生太郎（2008）であり、当該演説で麻生首相は「民主党に要請します」「民主党に問うものです」と繰り返して民主党に協力するように呼び掛けている。比喩や修辞技巧を多用するよりも、わかりやすく直接的な言葉遣いである点で特徴的である。

毛文偉（2022）は国会会議録、日常会話、論説文、小説、新聞記事のSTTRを計算し、そ

表１　歴代首相の所信表明演説の一覧

年	氏　名	代	延べ語数	異なり語数	TTR	STTR
2004	小泉純一郎	88	4,378	1,074	0.2453	0.5210
2020	菅 義偉	99	4,491	1,110	0.2472	0.5168
2016	安倍晋三	97	4,800	1,142	0.2379	0.5134
2011	野田佳彦	95	6,330	1,282	0.2025	0.5094
2013	安倍晋三	96	2,944	802	0.2724	0.5090
2000	森 喜朗	86	5,581	1,121	0.2009	0.5078
2000	森 喜朗	85	3,194	746	0.2336	0.5046
2017	安倍晋三	98	2,341	669	0.2858	0.5046
2006	安倍晋三	90	5,279	1,162	0.2201	0.5042
2021	岸田文雄	101	5,733	1,204	0.2100	0.5002
2005	小泉純一郎	89	2,123	618	0.2911	0.4998
2001	小泉純一郎	87	4,151	913	0.2199	0.4909
2010	菅 直人	94	7,085	1,334	0.1883	0.4896
2021	岸田文雄	100	4,547	950	0.2089	0.4888
2009	鳩山由紀夫	93	8,373	1,422	0.1698	0.4882
2007	福田康夫	91	4,113	874	0.2125	0.4839
2008	麻生太郎	92	3,876	929	0.2397	0.4710
—	全　体	—	—	—	—	0.4996

表２　ジャンル別テキストの STTR の平均値

ジャンル	国会会議録	日常会話	論説文	小　説	新聞記事
STTR	0.4379	0.44884	0.4551	0.4656	0.4995

の平均値を算出した（表２）。それによると、国会会議録の STTR の平均値は0.4379で、表１に示された諸演説を大幅に下回っている。これは、首相の所信表明演説は国会でのほかの応酬と違い、言葉の繰り返しが少なく、より明快に自分の意図を述べていることを示唆している。

４．TF-IDF

TF-IDF はコーパスや収集された文書群において、ある単語がいかに重要なのかを反映させることを意図した統計量である。その値は文書内におけるある単語の出現回数に比例して増加し、またその単語を含むコーパス内の文書数によってその増加が相殺される。歴代首相の所信表明演説における単語の TF-IDF を計算して上位10語を観察すると、各演説におけるキーワードの特徴が窺える。ここでは紙面の都合により、第85代森喜朗から第88代小泉純一郎までの

表3 TF-IDF 値の上位10語一覧（一部）[2)]

	2000森85		2000森86		2001小泉87		2004小泉88	
1	新生	0.207	新生	0.680	構造	0.176	選手	0.255
2	小渕	0.155	サミット	0.188	世紀	0.175	民営	0.123
3	推進	0.140	世紀	0.129	以内	0.125	イラク	0.117
4	サミット	0.138	少年	0.123	小泉	0.125	会社	0.109
5	密接	0.136	二十	0.106	処理	0.111	当たり	0.095
6	広範	0.136	推進	0.095	作成	0.109	業務	0.095
7	施政	0.136	倫理	0.094	報償	0.109	習慣	0.095
8	農村	0.136	九州	0.094	聖域	0.109	解禁	0.095
9	倫理	0.125	発出	0.092	二十	0.108	スポーツ	0.088
10	考える	0.125	プラン	0.080	相応しい	0.108	増える	0.088

所信表明演説のデータ（表3）について論じる。その他のデータは本稿末の附表を参照されたい。森首相の演説では、二回とも「新生」がトップになっている。これは二回にわたる演説で森首相がいずれも「日本新生」を多くのスペースを割いて強く唱えていたことを示している。しかし、その後の小泉首相の演説ではこの言葉が姿を消している（例1、2）。そのほか、「(九州・沖縄）サミット」「倫理（の向上)」なども森首相の二回の演説の共通内容として観察される。また、「推進」という言葉はほかの首相演説の中にも現れたが、森首相の演説ほど頻出していないため、森首相の演説における TF-IDF 値が高くなり、上位3位となった。このことから森首相が意欲的に内閣の施策を推し進めようとしている姿が窺える。一方、病に倒れた小渕首相に次いで首相になった森氏は一回目の所信表明演説で小渕氏のことを称え、その志を引き継いで国政に取り組む旨を伝えて、二回目の演説で少年犯罪の防止を重点に置いたため、「小渕」「少年」もキーワードとして現れている。

(1) 私は本内閣を「日本新生内閣」として、「安心して夢を持って暮らせる国家」、「心の豊かな美しい国家」、「世界から信頼される国家」、そのような国家の実現を目指してまいります。(2000森85)

(2) 私は、決意新たに「日本新生」に取り組み、活力ある進路を開き、国際社会の中で名誉ある地位を占める「ジャパニーズ・フロンティア」の実現を目指してまいります。(2000森86)

表3のデータを見ると、小泉首相の一回目の演説においては「聖域（なき）構造（改革)」「二十（一）世紀」[3)]「(不良債権）処理」「小泉（内閣)」などが、二回目の演説では「(日本人）選手」「民営（化)」「イラク（復興)」などが独特な内容として捉えられていることが窺われる。

表 3 と所信表明演説の内容を照らし合わせると、TF-IDF 値による内容分析には次のようなメリットとデメリットのあることがわかる。

メリット：

① 特定の文脈でテキストごとに特徴語を抽出できる。

TF-IDF は、特定の文書内で頻繁に出現し、他の文書ではあまり出現しない単語ほど高い重要度を割り当てる。従って、その単語が当該文書の特徴語であり、主題を特徴づける可能性が高いと判断できる。

② 一般的な高頻度語の影響を減らす。

TF-IDF はすべての文書で頻繁に出現する単語（名詞、動詞、助詞、助動詞など）に低い重要度を割り当てる。これにより、一般的な高頻度語が文書の主題を特徴づける能力を過大評価することを防げる。

デメリット：

① 複数テキストに共通した主題を見いだせない。

TF-IDF は多数の文書で頻繁に出現する単語の影響を抑えることによって特定の文書の特徴語を抽出できるメリットがある一方で、一部のキーワードの重要性を過小評価し、複数のテキストに共通した主題を見逃す恐れがある。たとえば、後で行うトピックモデル分析でわかるように、「国民」「経済」「社会」などはいずれの演説においても重要な話題であるが、TF-IDF の上位語リストには現れていない。

② 単語の意味や文脈が無視される。

TF-IDF は単語の出現頻度に基づいて計算されているため、単語の意味や文脈を考慮しない。従って、同じ単語でも異なる文脈で異なる意味を持つ場合、その違いを捉えることができない。例えば、小泉首相の2001年の演説では、「処理」は不良債権の処理とごみの処理という二つの主題で使用されている（例 3 、4 ）が、TF-IDF ではその違いが無視され、一括して取り扱われてしまう。

(3) 第一に、二年から三年以内に不良債権の最終処理を目指します。（2001小泉87）

(4) 例えば、大量のゴミの廃棄で処理の限界に至っている大都市圏を、新しいゴミゼロ型の都市に再構築する構想について、具体的検討を行います。（2001小泉87）

③ レアな単語を過大評価する傾向がある。

TF-IDF は文書の集合全体で稀にしか出現しない単語に高い重要度を割り当てる。しかし、これらの単語が必ずしも有用な情報を提供するわけではないため、誤った解釈を引き起こす可能性がある。例えば、「作成」は計17部のテキストの中で小泉首相の2001

年の演説においてしか使われていないため、TF-IDF 値の上位 6 位になっている。しかし、使用回数は 2 回だけで、重要な内容との関連性も薄いため、その重要度が過大評価されている（例 5 ）。

(5) その実現を確かなものとするため、「e-Japan 重点計画」を着実に実行するとともに、中間目標を設定する「IT 二〇〇二プログラム」を作成したいと考えます。（2001小泉87）

5．対応分析

対応分析はコレスポンデンス分析とも呼ばれ、1970年以降に広まった分析手法である。データ表の行と列に含まれる情報を少数の成分に圧縮し、それらの関係を散布図上に布置することで、視覚的なデータの俯瞰を可能にする（石川慎一郎・前田忠彦・山崎誠 2010）。

歴代首相の所信表明演説から合計出現数30を超える一般名詞と固有名詞を抽出し、「こと」「ため」「の」などの形式名詞を削除してから、SPSS V. 27で対応分析を行った。その結果、16の次元が抽出され、第 2 次元までの累積寄与率は33.2％であった（表 4 ）。分析に用いたカテゴリー数が135種類あったため、各次元の寄与率が低くなっているが、このような場合においても、最も寄与率の高い第 1 次元と第 2 次元を見ることで、演説の間のつながり及び演説とキーワードとの間に潜む対応関係が観察できる。

表4 対応分析の要約表

次元	特異値	イナーシャ	イナーシャの寄与率	
			説 明	累 積
1	0.371	0.137	0.214	0.214
2	0.275	0.076	0.118	0.332
3	0.260	0.068	0.105	0.437
⋮	⋮	⋮	⋮	⋮
16	0.105	0.011	0.017	1.000

図 1 において、各演説を代表するポイントは右下―右上―左上―左下というルートをたどって連続的につながっている傾向が見られる。これによって、17本の演説を2000～2007年、2008～2009年、2011～2017年、2020～2021年の四つのグループに分けることができる（表 5 ）。この分析は演説と名詞の対応関係を探るものであり、両者の間に強い関連性が見られる。今回分析の対象とした首相のうち、鳩山由紀夫、菅直人、野田佳彦は民主党出身で、その他は自民党出身であるが、図 2 から、所属政党の違いとは関係なく、どの党出身の人であれ、それ以前の内閣が抱えた課題を継承しながら、国内外の情勢に合わせ、少しずつ新しい議題を打ち出していくことがわかる。この傾向は第 6 節で行うトピックモデルによる考察でも裏付けられる。唯一の例外は菅直人（2010）であり、その内容は第 2 グループか第 3 グループのかわりに、第 1 グループで注目される政治的課題に戻っている。

次に、図 2 から各グループに対応する特徴語を観察する。まず、第 1 グループ（2000～

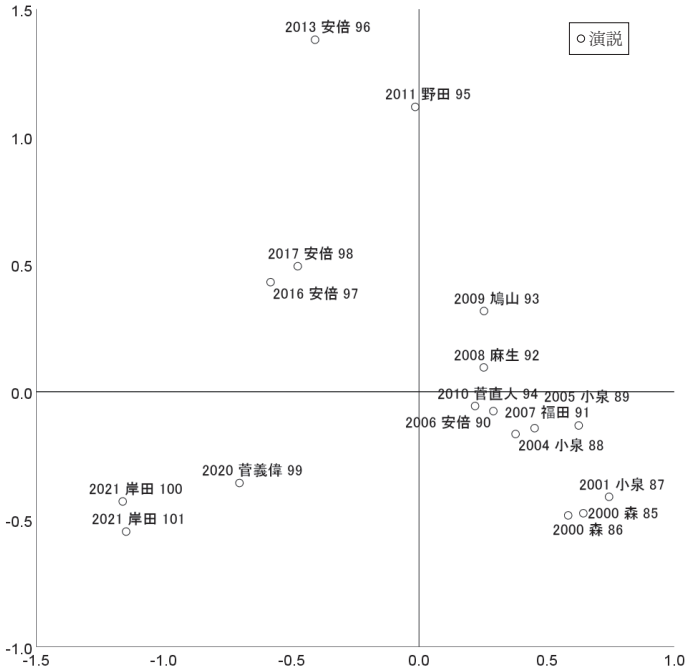

図1　第1アイテムスコア散布図

表5　対応分析によるグループ分け

	期　間	演　説
1	2000～2007年	2000森85、2000森86、2001小泉87、2004小泉88、2005小泉89、2006安倍90、2007福田91、2010菅直人94
2	2008～2009年	2008麻生92、2009鳩山93
3	2011～2017年	2011野田95、2013安倍96、2016安倍97、2017安倍98
4	2020～2021年	2020菅義偉99、2021岸田100、2021岸田101

2007年）では、「（二十一）世紀」「構造改革」「民間」「教育」「見直し」「（科学）技術」「国際」などがトピックになった。第2グループ（2008～2009年）のトピックとしては、「雇用」「政治」「（地域の）再生」「予算」「協力」などが挙げられる。また、2011年3月11日に発生した東日本大震災の影響を受け、第3グループ（2011～2017年）のトピックには「大震災」「被災」「危機」「復興」「挑戦」「被害」など震災と深くかかわるものが多い。「拉致」「（経済の）成長」もその時期の特徴的なトピックである。第4グループ（2020～2021年）において一番時代を反映できるトピックはいうまでもなく「新型コロナ」「（感染）拡大」「医療」であり、「デジタル」もかつてないほど重要視されている。また「（新しい資本）主義」「高齢（者）」が注目される課題になって、首相は「支援」「活用」「実現」などを頻繁に使用することによって内閣が

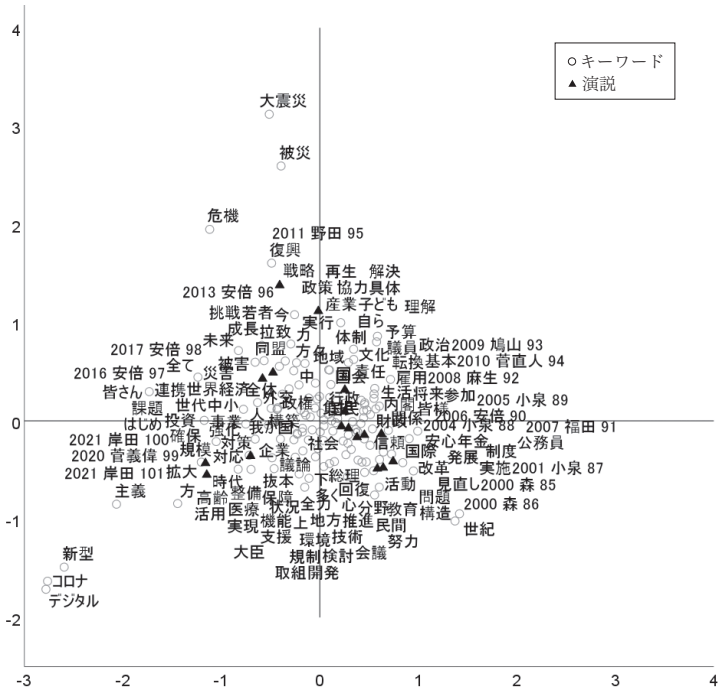

図 2　対応分析のパイプロット

難局にむけ、進んで対策を撮る姿を示している。

本節の考察を通して、対応分析には以下のメリットとデメリットがあることがわかる。

メリット：

① 非数値データを分析可能である。

対応分析はカテゴリーデータの分析に適している。これは、多くの他の多変量分析手法が連続的な数値データを必要とするのとは対照的に、変数とケースの違いが問題にならない。

② データの次元を削減できる。

対応分析は、データの次元を減らすことでデータの冗長性を削減する。これにより、データの主要なパターンや構造をより明確に捉えることができる。また、第 1 カテゴリーと第 2 カテゴリーを同時に分類できるのが大きな利点である。

③ 分析結果は視覚化され、わかりやすい。

対応分析は、高次元のカテゴリーデータを低次元（通常は 2 次元または 3 次元）の空間にマッピングすることで、データの構造を視覚化することができる。これにより、データのパターンや関係性を直感的に理解することが可能になる。

④　操作が簡単である。

分析者が調整すべきオプションが少なく、操作しやすいうえに分析者の判断によって結果が左右されにくい。

デメリット：

①　パターンの一部が欠落する可能性がある。

対応分析は、データの主要なパターンや構造を捉えるために、データの次元を削減する。しかし、この次元削減の過程で、データの一部の詳細や微妙なパターンが失われる可能性がある。特に、データが非常に高次元である場合や複数の重要なパターンが存在する場合には、これらのパターンの一部が視覚化する段階で欠落する可能性がある。例えば、本研究で使った次元 1 と 2 は合わせて33.2％の変数間の関係しか説明しておらず、これは図 2 から三分の二の情報が観察できないということを意味する。

したがって、対応分析の結果を解釈する際には、この限界を理解し、結果を適切に解釈することが重要である。また、必要に応じて他の分析手法を補完的に使用することも考慮すべきである。

②　明示できる項目の数に限界がある。

視覚化した結果、各項目は画面にプロットされるが、データの数が多くなると、点が重なってしまい、その位置を明確に観察できなくなってしまう。例えば、本節で使用したデータセットに名詞は135語あるが、図 2 ではその一部しか明示されていないだけでなく、どの点がどの単語を代表しているかということもはっきりしない。

③　外れ値に影響されやすい。

対応分析は外れ値の影響を受けやすい。外れ値が存在すると、分析の結果が歪められ、誤った解釈を引き起こす可能性がある。

6．トピックモデル

トピックモデルとは、文書の内容が複数の話題（トピック）から構成されるものと仮定して、確率モデルを用いて単語の出現頻度から文書ごとのトピックとその割合を推定する方法である。代表的な手法として、学習データに依存せずトピックの推定が行える LDA（Latent Dirichlet Allocation）が挙げられる（河瀬彰宏・吉原秀樹 2020）[4]。筆者は本稿の考察対象である所信表明演説から一般名詞と固有名詞を抽出して、内容と関係が薄い「こと」、「ため」、「の」、「もの」などを排除し、MTMineR V. 5.4をもってトピックモデル分析を行った。

分析にあたっては、まずトピックの数を決めなければならない。トピック数を四つに設定する場合、分類の結果は対応分析によるグループ分けの結果とほぼ同じである（図 3 ）。しかし、対応分析では研究対象を最大 4 種類に分けられるのに対し、トピックモデル分析ではトピッ

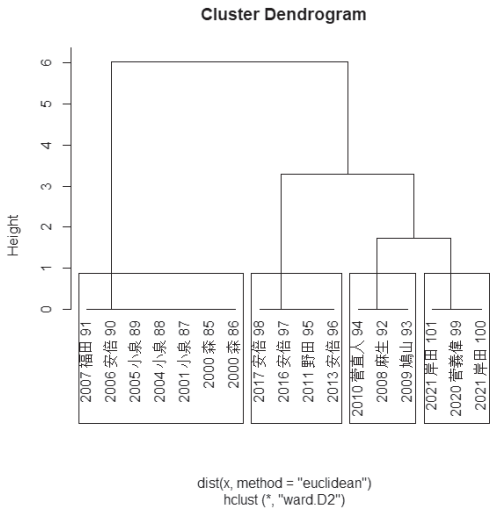

図3 テキストのトピックのクラスター樹形図

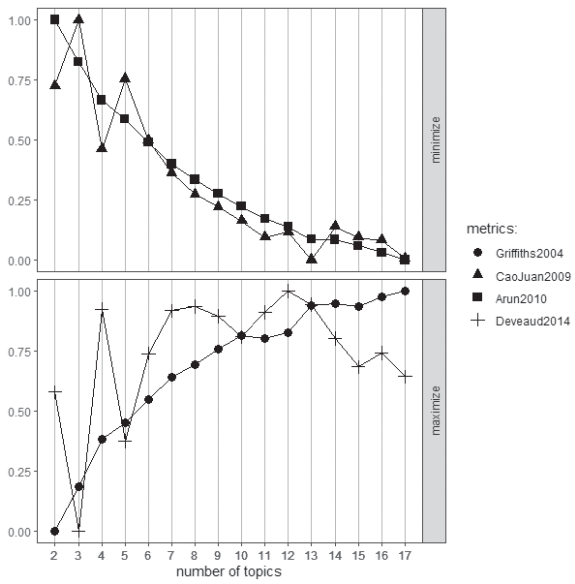

図4 トピック数を決める諸指標

ク数をより柔軟に設定できる。トピック数を決める参考値として、MTMineRは CaoJuan 2009、Arun 2010、Griffiths 2004、Daveaud 2014の四つの指標を出力しうる。そのうち、Griffiths 2004は対数尤度を用いて、CaoJuan 2009はトピック構造とトピック間の距離関係を利用した密度に基づいて LDA モデルの適応性を判断する。Arun 2010はトピックに基づいた対称的な KLD を用いて評価し、Deveaud 2014はすべてのトピック間の情報のダイバージェンス D を最大にすることによって潜在的なトピック数の数を推定する（金明哲 2018）。CaoJuan 2009と Arun 2010は小さいほど、Griffiths 2004と Deveaud 2014は大きいほどよいとされ、四つの指標を総合して判断すると、トピック数を七つに設定するのが最適なように思われる（図4）。

トピック数を七つに設定した場合、所信表明演説のグループ分けは図5に示したとおりである[5)]。表5と比べると、森喜朗首相の二回にわたる演説、菅直人（2010）、また安倍晋三（2016）と安倍晋三（2017）がそれぞれ一グループに分かれ、対応分析の結果と多少異なるが、全体を見ると、両者の結果は似通っていることがわかる。つまり、対応分析とトピックモデルのグループ分けの結果には共通するところが多い。また、自民党出身と民主党出身の首相（麻生太郎（自）と鳩山由紀夫（民）、安倍晋三（自）と野田佳彦（民））の演説は相変わらず同じグループに分けられる。つまり、どの政党の首相であっても、政策の主眼に継続性があるということがトピック分析においても実証された。

各グループを時系列で並べると、表6のようになる。表7では各トピックを構成する主な

語句が示されている。これを概観してわかるように、各トピックに「国民」「経済」「社会」など共通する語句がある。これは2000年以降の歴代首相がそれぞれの優先事項を持つ一方で、国民、経済、社会を重視するという点で共通していることを示唆する。そのほか、各グループに属する首相は所信表明演説でおのおの「教育」「構造改革」「地域（の再生）」「（強い）財政」「大震災（からの）復興」「世界と未来」「（新型）コロナ（対応）とデジタル」などに焦点を当てて見解を述べていることが窺える。

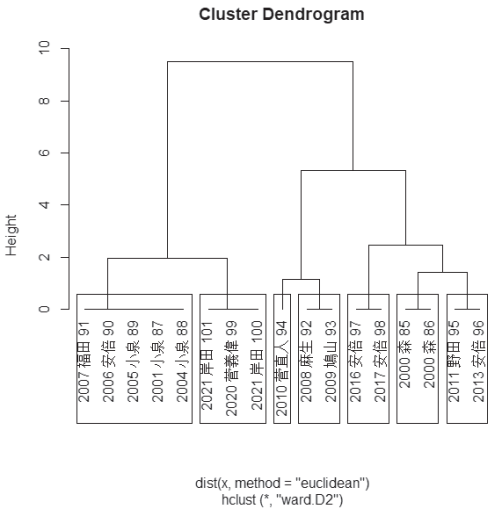

図5　テキストのトピックのクラスター樹形図

表6　トピックモデル分析によるグループ分け

Topic	所信表明演説
5	2000森85、2000森86
6	2001小泉87、2004小泉88、2005小泉89、2006安倍90、2007福田91
1	2008麻生92、2009鳩山93
2	2010菅直人94
4	2011野田95、2013安倍96
3	2016安倍97、2017安倍98
7	2020菅義偉99、2021岸田100、2021岸田101

表7　各トピックに属する上位語一覧

	T5	T6	T1	T2	T4	T3	T7
1	推進	改革	**国民**	*経済*	*経済*	世界	*経済*
2	**国民**	社会	政治	社会	被災	未来	実現
3	教育	**国民**	地域	**国民**	危機	*経済*	コロナ
4	皆様	*経済*	*経済*	財政	大震災	我が国	社会
5	政策	地方	国	実現	復興	皆さん	**国民**
6	*経済*	構造	社会	保障	社会	改革	成長
7	総理	国際	世界	内閣	**国民**	社会	デジタル
8	世界	実現	人	政策	世界	**国民**	新型
9	国家	問題	問題	成長	方々	地方	保障
10	心	関係	行政	我が国	再生	今	主義

本節の考察を通して、トピックモデルによる分析には次のようなメリットとデメリットがあることがわかる。

メリット：

① 非監督学習で学習データはいらない。

トピックモデル分析は非監督学習の一種で、ラベル付けされていないテキストデータからトピックを抽出することができる。これにより、大量のテキストデータを効率的に分析することが可能になる。

② データの次元を削減できる。

トピックモデル分析は、テキストデータの次元を削減することができる。各文書はトピックの分布を手掛かりにして表現され、これによりデータの構造をより簡潔に表現することができる。

③ 分析結果は視覚化され、文書の構造と内容の共通点を理解しやすい。

トピックモデル分析は、文書のトピック構造を理解するのに役立つ。これにより、文書の主要なテーマやテキスト間の意味の関連性を理解することが可能であり、トピックによってテキストを分類できる。

デメリット：

① 解釈が困難な場合がある。

トピックモデル分析の結果は、特に高次元のデータに対しては、解釈が難しい場合がある。また、モデルが抽出したトピックが人間の直感的な理解と一致しない場合もある。

② パラメータの選択が困難である。

トピックモデル分析では、トピックの数などのパラメータを事前に設定する必要がある。しかし、最適な値を選択するのが難しいこともある。

③ モデルが複雑で、訓練と評価には時間がかかる。

トピックモデル分析は、計算が複雑で、大量のデータを扱う場合、多くの計算リソースを必要とする。したがって、モデルの訓練と評価をするには時間がかかる。

7．おわりに

本稿は TF-IDF 値、対応分析とトピックモデルを用いて、2000年以降の日本の歴代首相の所信表明演説を分析し、グループ分けと内容の要約を試みると同時に、各分析手法の特性について考察した。その結果、日本政府の政策には一貫性があり、どの党派出身の首相も、前任の内閣が抱えていた課題を引き継ぎつつ、国内外の状況に応じて新たな政策を打ち出しているこ

とが明らかになった。その政策は国民、経済、社会の重視という安定した部分を終始持ちつつ、時勢に従い、「教育」「構造改革」「地域（の再生）」「（強い）財政」「大震災（からの）復興」「世界と未来」「（新型）コロナ（対応）とデジタル」などに焦点が移されてきた。

各分析手法の特徴をまとめると、どれも一長一短があることが明らかになった。TF-IDF 値による内容分析は、特定の文脈でテキストごとに特徴語を抽出でき、一般的な高頻度語の影響を減らすメリットがある。しかし、複数のテキストに共通する主題を見いだせず、単語の意味や文脈が無視され、レアな単語を過大評価する可能性があるというデメリットがある。対応分析は非数値データの分析が可能で、データの次元を削減できる。操作が簡単で、分析結果は視覚化され、理解しやすい。しかし、パターンの一部が欠落する可能性があり、明示できる項目の数に限界があり、外れ値に影響をうけやすいという弱点を持つ。一方、トピックモデルによる分析は対応分析と同様にデータの次元を削減でき、分析結果は視覚化され、わかりやすいという利点がある。非監督学習としての機械学習で、学習データは不要である。しかし、解釈とパラメータの選択が困難な場合があり、訓練と評価に時間がかかるという問題がある。

したがって、テキストの内容を抽出する際には、これらの分析手法を総合的に活用することが望ましい。データ量が少ない場合は対応分析を用い、データ量が大きい場合はトピックモデルを使用して、テキストをグループ分けし、共通点をまとめる。さらに TF-IDF の計算を行うことで、各テキストにおける特徴語を特定し、文書の主題や内容を要約することを推奨する。

注

1）データベース「世界と日本」：https://worldjpn.net/（2023年 1 月26日参照）。
2）紙面の関係で、「年・姓・代」を歴代首相の所信表明演説の略称とする。例えば、「2000森85」は森喜朗が2000年に85代目の首相として発表した所信表明演説を指し示す。以下同様。
3）「二十（一）世紀」は森首相の二回目の演説においても重要なキーワードとなっている。
4）トピックモデル LDA の詳細は金明哲（2018）を参照されたい。
5）トピックの番号は MTMineR によってふられたもので意味がない。

引用文献

石川慎一郎・前田忠彦・山崎誠（2010）『言語研究のための統計入門』，くろしお出版
河瀬彰宏・吉原秀樹（2020）「戦後の歴代首相の施政方針演説と所信表明演説の計量分析」『情報知識学会誌』30(2)，pp. 200–205
岸江信介・田畑智司（2014）『テキストマイニングによる言語研究』，ひつじ書房
金明哲（2018）『テキストアナリティクス』，共立出版
ソジエ内田恵美（2018）「戦後日本首相による所信表明演説の研究」『年報政治学』69(2)，pp. 177–199
杉村泰（2023）「中国人日本語学習者における日本語の「乗り物＋で」と「乗り物＋に乗って」の選択」『名古屋大学人文学研究論集』(6)，pp. 23–39
鈴木崇史・影浦峡（2011）「名詞の分布特徴量を用いた政治テキスト分析」『行動計量学』38(1)，pp. 83–92

山田太造（2017）「新聞記事に対するトピックモデルの適用とトピックの時系列変化に関する考察」『研究報告人文科学とコンピュータ（CH）』(1), pp. 1–5
原田朋子（2019）「日本語母語話者と上級日本語学習者の小論文の比較―テキストマイニング手法と目視による分析を通して―」『同志社大学日本語・日本文化研究』(16), pp. 1–15
東照二（2006）『歴代首相の言語力を診断する』, 研究社
東照二（2022）「菅政権と政治言語力―言葉はどのように政治を動かしたのか―」『立命館食科学研究』7, pp. 87–102
藤本一美（2022）「戦後日本政治と「首相演説」①」『社会科学年報』(56), pp. 259–281
毛文偉（2022）「数据挖掘技术在学习者作文特征分析中的应用研究」『日语学习与研究』219(02), pp. 72–81
毛文偉・梁鵬飛・蒋夏夢（2022）「进展 问题 展望 数据挖掘技术在日语语言研究中的应用」『日语学习与研究』(6), pp. 76–94

キーワード：テキストマイニング、TF-IDF、対応分析、トピックモデル

附表　TF-IDF 値の上位10語一覧

	2000森85		2000森86		2001小泉87		2004小泉88		2005小泉89		2006安倍90	
1	新生	0.207	新生	0.680	構造	0.175	選手	0.254	民営	0.407	官邸	0.214
2	小渕	0.155	サミット	0.188	世紀	0.175	民営	0.123	郵政	0.285	ゼロ	0.187
3	推進	0.140	世紀	0.129	小泉	0.125	イラク	0.117	賛成	0.166	チャレンジ	0.153
4	サミット	0.138	少年	0.123	以内	0.125	会社	0.109	アスベスト	0.145	美しい	0.124
5	密接	0.136	二十	0.106	処理	0.111	当たり	0.095	反対	0.145	固定	0.098
6	広範	0.136	推進	0.095	作成	0.109	業務	0.095	止め	0.145	規律	0.098
7	施政	0.136	倫理	0.094	報償	0.109	習慣	0.095	為さる	0.145	フリーター	0.086
8	農村	0.136	九州	0.094	聖域	0.109	解禁	0.095	給与	0.134	ブランド	0.086
9	倫理	0.125	発出	0.092	二十	0.107	スポーツ	0.088	民間	0.112	保つ	0.086
10	考える	0.125	プラン	0.080	相応しい	0.107	増える	0.088	判断	0.111	勝ち組	0.086

	2007福田91		2008麻生92		2009鳩山93		2010菅直人94		2011野田95		2013安倍96	
1	共生	0.133	臨む	0.194	幸せ	0.189	鳩山	0.131	原発	0.332	少女	0.239
2	一歩	0.124	存ずる	0.146	列島	0.126	続行	0.131	震災	0.226	手紙	0.228
3	不便	0.124	不安	0.113	友愛	0.126	閉塞	0.131	福島	0.204	危機	0.207
4	支払う	0.124	寧ろ	0.111	社長	0.126	先生	0.123	収束	0.168	突破	0.184
5	職務	0.114	当然	0.111	本当	0.116	立て直し	0.105	放射	0.168	円高	0.110
6	地方	0.106	手段	0.111	彼の	0.096	無駄遣い	0.105	事故	0.129	銀行	0.103
7	行政	0.106	議会	0.111	掛け橋	0.096	市川	0.103	空洞	0.128	震災	0.096
8	自立	0.104	段階	0.105	鳩山	0.096	パーソナル	0.103	原子	0.123	さん	0.091
9	年金	0.096	不足	0.099	市民	0.096	リストラ	0.103	危機	0.107	会派	0.091
10	立場	0.096	目的	0.099	掃除	0.096	打ち破る	0.103	被災	0.103	入学	0.091

	2016安倍97		2017安倍98		2020菅義偉99		2021岸田100		2021岸田101	
1	さん	0.169	ケン	0.277	ウイルス	0.353	コロナ	0.448	デジタル	0.379
2	リオ	0.163	革命	0.247	コロナ	0.308	新型	0.289	コロナ	0.284
3	創生	0.155	無償	0.211	デジタル	0.237	デジタル	0.249	ワクチン	0.189
4	蒲鉾	0.122	人作り	0.158	新型	0.196	分配	0.248	新型	0.189
5	難民	0.122	トランプ	0.138	グリーン	0.118	ワクチン	0.124	賃上げ	0.188
6	クマモト	0.093	引き	0.138	ポスト	0.116	主義	0.124	感染	0.171
7	セブン	0.093	生産	0.136	年末	0.116	分断	0.122	接種	0.133
8	俯瞰	0.093	連続	0.127	洪水	0.116	遠く	0.122	資本	0.129
9	施設	0.088	力強い	0.097	感染	0.114	資本	0.117	協働	0.124
10	活躍	0.088	年間	0.097	交代	0.098	接種	0.099	承認	0.124

Summary

A Comparative Study of Text Mining Techniques for Content Extraction:
Using Analysis of Policy Speeches by Japanese Prime Ministers as an Example

Wenwei Mao

This paper applies TF-IDF values, correspondence analysis, and topic modeling to analyze the speeches of Japanese prime ministers since 2000, grouping them and summarizing their contents while also examining the characteristics of each analysis method. As a result, it was found that there is continuity in Japanese government policy, and regardless of the prime minister's party affiliation, they continue to address issues inherited from previous cabinets, adapting to domestic and international situations and formulating new policies. These policies consistently prioritize the people, economy, and society. Over time, the focus has shifted through issues such as "education," "structural reform," "regional (revitalization)," "(robust) fiscal policy," "reconstruction (from the great earthquake)," "the world and the future," and "(responding to) new coronavirus and digitalization."

A review of the distinct characteristics of each analytical method revealed that all of them have their own unique strengths and limitations. Hence, when attempting to extract content from a text, it is advisable to employ these analytical methods in a comprehensive manner. For smaller data sets, the use of correspondence analysis is recommended, whereas for larger sets, topic modeling should be utilized to categorize texts and compile common points of view. Moreover, calculating TF-IDF values enables the identification of distinctive words for each text, thereby facilitating the summarization of document themes and content.

Keywords: Text Mining, TF-IDF, Correspondence Analysis, Topic Modeling