

## 政党や政治家の政策的な立場を推定する コンピュータによる自動コーディングの試み

上 神 貴 佳  
佐 藤 哲 也

### <要約>

本研究の目的は、政党や政治家の政策的な立場を推定するために考案されてきた様々な方法論を概観し、その現状と発展の可能性について考察することである。

政党や政治家の政策的な立場を推定する方法を大別すると、公約の内容分析とアンケート調査の二種類に分けることができる。アンケート調査とは異なり、内容分析には、政党や政治家の立場を明らかにすべき争点の選択が客観的であるという長所がある。しかし、この方法には分析コストの高さや結果の信頼性について改善の余地がある。

そのための手段として、本研究はコンピュータによるコーディングの自動化を提案する。具体的には、政策の内容分析に必要であるコーディング作業について、「教師付き学習に基づく分類の自動化」を行う。実際に、2005 年総選挙と 2007 年参院選における各党のマニフェストにこの方法を適用し、その有用性と可能性を示す。

### イントロダクション

本研究の目的は、政党や政治家の政策的な立場を推定するために考案されてきた様々な方法論を概観し、その現状と発展の可能性について考察することである。その前に、このような方法論の研究と開発が求められる理由を明らかにするため、選挙研究、より広くは政治学研究における利用の実績を紹介することから始めたいと思う。

まず、これらの方法論を用いることにより、選挙における争点とそれに対する政党や政治家の立場をより明確にすることができる。一例を挙げよう。総選挙の性格については、従来から、特に解散の経緯に基づいて様々な特徴付けが行われてきた。2005 年 9 月の総選挙であれば、「郵政解散」が人口に膾炙しているようである。このネーミングはもっともらしいものであり、郵政民営化の是非が圧倒的に重要な争点となったことを示唆している。この見解に異議を唱えるものではないが、後述の手法により各党のマニフェストを分析すると、郵政分野への言及割合は各党の平均で 2.7% に過ぎないことが分かる。直感に依らず、

計量的な手法に基づく客観的な方法論を適用することにより、違った一面が見えてくる。また、ヨーロッパにおける公約分析は左右の対立軸に基づく政党の政策的な立場の変化を時系列的に明らかにしている (Budge et al., 2001)。このような分析手法は、あらゆる選挙研究が依拠しなければならない科学的かつ経験的な基礎データを提供する。

実際に、このようなデータは政党に関する理論的な研究の検証に用いられてきた。日本における代表的な応用例をいくつか挙げる。ポートフォリオ・アロケーション・モデルの研究は、連立政権交渉における政党間の役職配分を説明するものであるが、主要な次元における政党の政策的な立場を専門家に対する調査によって特定し、理論の妥当性を検証している (加藤・レイヴァー, 1998)。また、候補者の公約を分析することにより、中位投票者理論の検証を中心とする空間競争モデルの研究に役立てることもできる (堤・上神, 2007)。さらに、政党組織論にも応用可能である。党活動家 (ないし中核的な支持者) の政策に関する意見は議員や一般の支持者よりも急進的であるという予想が存在するが (メイの法則)、政党の候補者とその支持者に同一の政策質問を行うことにより、その適否について確かめることができる (谷口, 2006a)。

このような方法論による研究成果を社会貢献に用いることも可能である。マニフェストの分析とアンケート調査を用いて、各党の政策的な立場を明らかにすることにより、政策争点に対する賛否に基づく投票を支援するインターネット上のプログラムが開発されている (上神・堤, 2008)。2007年の参院選においては、20万人以上がこのプログラムを利用しており、今後とも利用の拡大が見込まれる。

以上、選挙研究のための基礎的なデータの提供、政治学における理論の検証、応用研究の観点から、政党や政治家の政策的な立場を推定する方法論を研究し開発する必要性を指摘した。現代日本政治の変化もこのような方法論の発展を促進する重要な背景要因である。国政選挙と地方選挙におけるマニフェストの急速な普及に見られるように、選挙における政策の比重が増してきていることは否定し難い事実であり、分析手法の洗練が求められていることは論を俟たない。

しかしながら、日本の選挙研究ないし政治研究において、政党や政治家の政策的な立場を推定する各種の方法論を自覚的に検討する論考は、管見の限り、存在しない。ヨーロッパの政党研究においては、それぞれの方法論を提唱する研究者の相互の批判により、可能性と限界を明らかにする試みが積み重ねられてきた (邦語の文献としては、レイヴァー・ブノア, 2006)。このような学術共同体における論争は、方法論の更なる発展のために不可欠であるが、言語の壁もあり、日本における有力な学派である公約研究は海外からの批判的な検討にさらされることもなく、独自の進化を遂げてきた。そこで、本研究は欧米における議論の積み重ねを参照しつつ、これらの方法論の特徴を明らかにする。その上で、長所を活かし、短所を矯正するためにコンピュータを利用する方策を提案する。

本研究の行論をまとめると、以下のようなになる。まず、政党や政治家の政策的な立場を推定する各種の方法論を検討し、公約の内容分析とアンケート調査の二種類に大別する。前者の代表例は人間による分類作業である。この方法には、きめ細かい分析という長所が

あるが、膨大な時間と人件費を要し、特に複数の作業者をを用いる場合、分類の一貫性が失われがちという短所もある。一方のアンケート調査は、専門家に対するものであれ、政党や政治家に対するものであれ、即時かつ低コストで実施できるが、質問文の妥当性や低回収率など調査固有の問題を免れない。

そこで、本研究は公約の内容分析に注目して、日本において開発されてきた方法論を紹介する（第1節）。その上で、この方法が抱えるコストや信頼性の問題を解決するための手段として、コンピュータによるコーディングの自動化を提案する（第2節）。

具体的には、「教師付き学習に基づく分類の自動化」を用いて、内容分析に必要であるコーディング作業を行う。自動化の対象となる作業手順は、品田裕が開発した方法を参考とする（品田, 1998a, 2006）。また、本研究が扱う学習データは、主に2005年総選挙と2007年参院選における各政党のマニフェストである。そのマニフェストをコーダーが分類したものをベンチマークとして利用する。

## 1. 政策立場の推定に関する先行研究

本節では、政党や政治家の政策立場を推定する方法について、先行研究を紹介し、その特徴を検討する。この研究分野においては、様々な方法論が提案されているが、大別すると、公約の内容分析、専門家に対するアンケート調査、政治家に対するアンケート調査の三つに分けることができる。それぞれに特徴があり、本研究における議論の出発点として検討する必要がある。

### 1.1. 公約分析, 専門家調査, 政治家調査

政党や政治家の政策立場を推定する方法については、公約分析、専門家調査、政治家調査の他にも、有権者調査や点呼投票を利用する方法も存在する（Benoit and Laver, 2006, Table 3.1）。本研究では、後二者を取り扱わない。点呼投票の事例自体が政党規律の強い日本の議院内閣制では稀であり、少なくとも議員の政策的な立場を推定するには適さない。また、有権者調査における回答者の政治的知識の程度はバラツキが大きいと考えられるからである。

公約の内容分析による政策立場の推定方法については、ヨーロッパと日本において、それぞれ蓄積がある。後者については、第1.3節の説明に譲るとして、まずは前者を中心に紹介する。代表的な研究としては、主にヨーロッパの研究者が構成するマニフェスト・リサーチ・グループ（MRG）<sup>1</sup>による政党公約の比較研究プロジェクトが挙げられる（Budge et al., 2001）。このMRGによるプロジェクトには、日本から猪口孝が参加している（猪口, 1983, 第2章）。

そのデータ・セットの分析対象は日本を含むOECD諸国における国政レベルの諸政党の公約であったが（1945年から1998年まで）、最も新しいものは、中東欧諸国における国政

---

<sup>1</sup> 後年、Comparative Manifestos Project (CMP)に改編された。便宜上、本研究ではMRGの表記を用いる。

レベルの諸政党の公約をもカバーする(1990年から2003年まで, Klingemann et al., 2006)。これら諸政党の公約を共通のコーディング方法に基づいて分類する訳であるが, それは, 対外関係(10), 自由と民主主義(4), 政治システム(5), 経済(16), 福祉と生活の質(7), 社会のあり方(8), 社会グループ(6)の計7領域56範疇から構成されている(括弧内は各領域に含まれる範疇の数)。これらのコーディング・カテゴリーは左右のいずれかに分類されており, 集計を経て, 諸政党はイデオロギー空間上に位置付けられる。

コーディングの担い手としては, 人間と機械の二通りがある。人間のコーダーによる作業は時間と手間が掛かるため, 近年, コンピュータによるコーディングが試みられている。人間の作業手順をコンピュータによって再現したい場合には, あらかじめ定められた辞書を用いて単語を分類する方法が自然である(Bara, 2001a, 2001b; Kleinnijenhuis and Pennings, 2001; Ray, 2001など)。しかし, 言語の解釈には固有の難しさがある。また, プログラムが参照する辞書を作成する際には, 人間によるコーディングの結果が必要となるなど, この方法論は未だ発展途上にあるといえる。

これらの限界を克服するために, Laver and Garry (2000), De Vries et al. (2001), Garry (2001), Laver et al. (2003)らは, (専門家調査などにより)政策的な位置があらかじめ判明している文書 reference text との単語の頻出度合いの類似性に注目して, 新しい文書 virgin text の政策的な位置の推定を試みている。文章や単語の解釈という困難な課題を避けられる一方, どのような文書を reference text として用いるかによって結果が大きく左右されるという特徴がある。いずれにせよ, コンピュータを用いるテキスト分析はますます発展していくものと考えられる。最新の動向については, Political Analysis 誌の16巻4号などを参照されたい。

続いて, 専門家調査について簡単に紹介する。専門家調査とは, 政党や選挙の研究を専門とする政治学者を対象に, アンケートによって政党の政策位置と争点の重要度を尋ねる調査である。2002年から2003年に実施された調査においては, 日本を含む47カ国の387政党が対象となり, 1,491人の専門家から有効回答が得られた(Benoit and Laver, 2006, Chapter 4)。

日本において初めて実施された専門家調査は, 1989年にマイケル・レイヴァーとW・ハントが実施したものである(Laver and Hunt, 1992)。日本における利用例としては, 前述のポートフォリオ・アロケーション・モデルの研究がある(加藤・レイヴァー, 1998)。また, 2000年総選挙後に実施された専門家調査については, その結果が報告されている(加藤・レイヴァー, 2003)。

政治家調査とは, その名の通り, 政治家に対するアンケートを実施することにより, その政策位置を推定しようとするものである。日本においては, 新聞社による調査が知られており, 2003年に実施された東京大学と朝日新聞社による共同調査を代表的なものとして挙げる事ができる(蒲島・山本, 2005)<sup>2</sup>。この東大・朝日共同調査を利用した分析結果

<sup>2</sup> これに加えて, 2004年参議院議員・参院選候補者調査, 2005年衆院選候補者調査, 2007年参議院議員・参院選候補者調査が公開されている。詳細については, 『日本政治研究』5巻1・2合併号に所収のコード

としては、谷口（2006a, 2006b）を挙げられる。衆議院議員と衆院選候補者の政策的な立場をそれぞれの次元で表現し、各次元における位置を決める要因を探っている。

このように、政党や政治家の政策立場を推定する方法は多様である。それぞれの特徴を踏まえた上で、利用の仕方を考える必要がある。以下では、問題の所在と解決策を提案する。

## 1.2. 各方法論の検討

本小節では、内容分析に基づく政策分析の可能性を主張する。内容分析にはデメリットもあるが、本研究が提唱するコンピュータによる自然言語処理と機械学習を用いて、その影響を軽減することができるからである。

まず、（専門家調査、政治家調査を問わず）アンケート調査と内容分析のアプローチにおける大きな違いの一つは、前者が先験的、後者が帰納的という点にある（レイヴァー・ブノア, 2006）。つまり、政党や政治家の政策立場を明らかにすべき争点の選び方が全く異なる。アンケート調査の場合、質問文作成の段階において、その材料となる争点の客観的な根拠を方法論に内在する形で示すことはしない。内容分析は、政策に関する資料を分析することにより、重要な争点や（もしあるとすれば）政策次元を示す。あらかじめ設定されたコーディング方法の枠内ではあるが、研究者による分析視角を先験的に設定せず、資料をして語らしめる帰納的なアプローチに特徴がある<sup>3</sup>。

また、内容分析の長所は、分析に用いられる材料の性質にもある。政党のマニフェストや政治家の選挙公報は公開されており、当初より入手可能性は確保されている。アンケート調査のように、調査の品質が回収率に大きく依存することはない。例えば、加藤・レイ

---

ブックを参照されたい。その他の政治家調査については、1998年11月から12月にかけて衆参両院議員に対して実施された読売新聞社政治部と東京大学法学部の蒲島郁夫研究室が共同で実施した調査がある（蒲島, 2000）。また、2001年参院選においては、関東の1都6県から出馬した候補者に対してアンケートを行った佐藤（2003）がある。2007年参院選については、政党に対してアンケートを実施し、政策立場の回答を求めた上神・堤（2008）もある。

<sup>3</sup> このような内容分析とアンケート調査の違いは、政策立場の推定方法にも表れている。イアン・バッジによると、MRGによる公約分析は顕出性 saliency の理論に依拠しており、対立性 position を重視する立場とは相容れないという（Budge et al., 2001, Chapter 3）。曰く、論理的には、争点に対する立場は賛成か反対に分けられるが、人気のある立場はいずれかに偏っている。例えば、税金については、増税と減税という2つの立場が想定できるが、誰も税金は安いに越したことはないと思うのであれば、それは合意争点といえる。政党は公約においていずれかの立場を打ち出せるが、敢えて増税を訴えることを想定しない。どれだけの割合を用いて減税を訴えているかについて測定することにより、その政党の政策立場を推定できる。小さな政府を支持する立場であれば、減税を主張するであろうし、大きな政府を支持する立場であれば、増税を取り上げることはないであろう。この違いから、政党間の政策的な違いを検出できる。MRGは争点の顕出性を重視するアプローチであり、対立性を重視するアンケート調査のアプローチとは異なるという。しかしながら、MRGのコーディング方法は、賛成か反対に分類しなければならないカテゴリーも複数含んでおり、バッジの主張は一貫していない。残りのカテゴリーについて賛成と反対の分類を追加しない理由の1つは、再コードに伴うコストの問題であり、理論的に内容分析が顕出性のみに依拠する必要はないことが分かる。実際、第1.3節で紹介する日本の公約研究グループによるコーディング方法は、全ての政策について現状維持か改革かという立場の違いを分類するように設計されている。合意争点と対立争点、それぞれにおける政党間の立場の違いを検出するよう設計できる内容分析とは異なり、アンケート調査における立場の違いとは対立争点のそれに限定されている点に注意が必要である。

ヴァー（2003）によると、2000年総選挙後に実施された専門家調査の回収率は17%に留まる。政治家調査に目を転じると、先述の東大・朝日調査は82%と高く、ハードルをクリアしているといえよう。

一方、内容分析に伴う困難の多くは、コーディング作業にある。これらは専門家調査や政治家調査の難しさとは異なる性質のものである。まず、コーダーの熟練度や個性によって、作業結果が左右されてしまうことが課題として挙げられる。なぜなら、言葉の解釈は人によって様々であり得るからである。2007年参院選のマニフェストにおいて、二番目に多く言及されていた文部科学省の管轄に該当する政策分野を例に挙げよう。「教育」について言及されていても、教育の「環境」を整備するということなのか（53言及数）、教育の「内容」を改革するということなのか（27言及数）、コーダーの判断が問われることになる。

政治一般や政策に関する知識の有無は、作業結果に影響を及ぼす重要な要因と考えられるため、コーダーのトレーニングが欠かせない。複数のコーダーを用いる場合、コーダーによって判断が分かれるような難しい政策領域については、あるコーダーはAと振る傾向があり、別のコーダーはBと振る傾向があるというように、個性が反映されることもある。単一のコーダーを用いても、作業が一貫しているという保証はない。コーダーが迷うことのないように、分類の詳細について、あらかじめ決めておくことが重要である<sup>4</sup>。抜本的な解決策はコンピュータによる均質なコーディングである。この場合、文脈依存的な言葉の微妙なニュアンスをどの程度までプログラムに読み込ませることができるかが鍵となる。

上記のようなコーディングに内在する問題に加えて、内容分析を実施するには、多大な時間と費用を要することも挙げておかねばならない。実際に作業を進める上で、この問題は決定的な障害となり得る。例えば、2003年総選挙の候補者公約におけるコーディングの対象者は1,004人、対象箇所は26,000余りに上る（泡沫候補を除く）。分析の対象となる公約が長いほど、当然のことながら、コーディングにも時間が掛かる。コーダーのトレーニングにも時間を割かねばならない。時間に比例して、コーダーの人件費も掛かることになる。その費用は莫大なものであり、潤沢な研究資金の確保ができない場合、作業は滞ってしまう。コーダーによる分類に依拠する内容分析は、専門家調査や政治家調査のように即時かつ安価に実施できないといえる。コストを低減するためにも、コンピュータを利用する自然言語処理の応用を図る必要がある。

さらに、政策は時代と共に変化するため、事前に決められた分類に従ってコーディングを行う場合、カテゴリーの改廃も避けられない。従って、分析結果の時系列的な比較には注意が必要となる。もし新しいカテゴリーに従って全てのテキストの再コーディングを行うならば、比較可能性を確保できるであろうが、コスト面から難しいといわざるを得ない。この問題についても、コンピュータを用いるコーディングによる解決が望まれる。

以上、政党や政治家の政策立場を推定する方法論の違い、各々の特徴について検討して

<sup>4</sup> また、複数のコーダーを用いることによって、コーダーの個性に起因する分類結果のバラツキをある程度までは中和できるはずである（品田，2006）。

きた。公約分析には様々な長所があるが、克服すべき短所も多いといえる。

### 1.3. 日本における公約研究グループのコーディング方法

日本において、政党や政治家の公約に内容分析を適用した研究は数多く存在する。この内、品田（1998a, 2006）の方法論に依拠する研究としては、1990年、1993年、1996年の各総選挙における候補者公約を分析対象とする品田（1998b, 2001, 2002）、1990年、1993年、1996年、2000年の各総選挙における候補者公約を分析した堤（2002）、2003年総選挙における候補者公約を分析対象とする堤・上神（2007）、2003年総選挙における自民党と民主党のマニフェストを分析対象とする上神（2006）、2005年総選挙における各党のマニフェストを分析対象とする上神・堤（2008）を挙げられる<sup>5</sup>。

以下では、品田が開発した内容分析の方法について、その概略を説明する。この分類方法は、選挙公約が一般に「 の皆さんのために を××します」という形式で述べられることに注目し、政策対象（ に対応）、政策分類（ に対応）、政策賛否（××に対応）についてコード化する<sup>6</sup>。

まず、政策対象とは、字義の通り、公約の対象であり、国民や市民等々である。特に記述がない場合は対象なしとコードを振る。実際には、対象がない場合が多いようである。

政策分類は、原則として旧省庁の職掌に対応してコード化され、旧省庁を単位とする政策分野（省庁の職掌外であるもの、それとの対応が不明確であるものについては内閣、政治、その他にまとめられている）と、その政策分野を細分化した政策内容の2段階から成り立っている。具体的には、内閣（9）、自治（7）、安全保障・外交（10）、大蔵（10）、文部・科学技術（8）、厚生（10）、労働（5）、農水（8）、通産（10）、運輸（5）、郵政（4）、建設（8）、環境（6）、政治（8）、その他（10）、構造改革（3）、計16政策分野121政策内容から構成されている（括弧内は各政策分野の政策内容数）。

政策賛否とは、その公約が現状肯定と方針転換のいずれに該当するか判断して振られるコードである。ヨーロッパのMRGとは異なり、日本の公約研究グループは合意争点のみな

<sup>5</sup> その他の方法論に依拠する研究としては、まず、日本の公約分析におけるパイオニア的な存在である小林良彰によるものがある（小林, 1997, 2008）。分析対象は、1947年から1990年までの主要政党の公約（1997, 第3章）、1986年、1990年、1993年の各総選挙における候補者公約（1997, 第4章）、2001年参院選、2003年衆院選、2004年参院選、2005年衆院選における候補者公約である（2008, 第2章）。戦後の政党公約と2000年代の国政選挙における候補者公約は、人的サービス（4）、防衛（1）、外交・貿易（1）、物理的資産（6）、特別なカテゴリー（4）、計5政策分野16政策領域に分類される（括弧内は各政策分野の政策領域数）。1986年、1990年、1993年の各総選挙における候補者公約については、別のコーディング方法が用いられている。1996年総選挙における候補者公約を分析した堤（1998）は、行政改革、経済構造改革、地方分権、中小企業対策、福祉、農林水産業対策、社会資本整備、医療、消費税反対、護憲、環境、女性参画、教育、政治改革の計14項目についてコーディングを行った（下位のカテゴリーを含めると約200項目）。また、2003年総選挙における自民党と民主党のマニフェストを政治改革、経済政策、社会政策、安全保障の政策分野に分類したMiura et al. (2005)、2003年衆院選における候補者のホームページを分析し、掲げられた公約を景気・経済、教育、安心と安全、環境、改革、雇用、制度、高齢者、政治、医療に分類した山本（2005）も挙げておきたい。このように、既に紹介したMRGによる分析以外にも、日本においては公約研究の蓄積がある。その分析手法も様々である。

<sup>6</sup> コード表の詳細については、品田（2006）を参照されたい。

らず、対立争点もコーディングの対象とする。この政策賛否によって、それが可能となっている。

分析の対象となる政党や政治家の公約は、マニフェストや選挙公報という形で入手できる。これらをテキスト入力したものを文節毎に区切り、前述の公約形式と合致するものにコードを逐次振っていく。これら一連の作業を一人でこなすことは困難であるため、複数の作業員で分担することになる。特にコード振りについては、作業員の個性が反映されやすく、作業結果を安定させるために対策を講じる必要がある。具体的には、頻出する公約や判断が難しい公約について振るべきコードを定めた辞書をコーダー間で共有している。

以上が公約研究グループの方法論の概要である。より詳しくは、品田（2006）を参照されたい。

## 2. プログラムの概要と分類結果

日本の政治研究におけるコンピュータを利用した内容分析の例としては、田中らによる、戦後における内閣総理大臣の国会演説の研究がある（田中，1999）。この研究において開発されたプログラム（Content Analyzer 2.0）は、単語の出現頻度を計測し、その情報を元に主成分分析やクラスター分析を行う。また、鈴木・影浦（2007）も、総理大臣による国会演説の形態素解析を行い、彼らのスタイルの違いを析出する<sup>7</sup>。しかし、日本において、コンピュータによる公約の分析を行った研究を寡聞にして知らない。以下では、前小節のコーディング方法を元にして、コンピュータによる自動化を試みる。

### 2.1. 機械学習に基づくコーディングの概念

機械学習とは、人工知能の一つで、コンピュータに学習させるためのアルゴリズムに関係がある。すなわち、アルゴリズムによって、与えられたデータの特質を推測できるような情報を生み出し、その情報を利用して将来現れるであろうデータの予測をするということの意味している。

このようなことが可能な理由は、データには何らかのパターンが含まれており、それを機械によって一般化することができるからである。一般化するためには、データのどの面が重要なのかを決定づけるモデルを学習しなければならない。しかし、その学習されたアルゴリズムが、それまでに見たことのないパターンに遭遇する場合、誤って解釈してしまうことが少なくない。また、過度の一般化という問題を抱えており（過学習）、少ない例を元に一般化を行っても正確であるとは限らない点には注意しなければならない。

本研究では、機械学習アルゴリズムの一つである教師付き学習に基づくアルゴリズムを使って、コーディング作業システムを実現する。教師付き学習には、事前に分類のお手本となるデータ（学習データ）が必要であり、これをコンピュータでアルゴリズムに基づいて計算させ、分類器と呼ばれる対象を分類するための最適なアルゴリズムを作る。これに基

<sup>7</sup> その他には、感性語に注目して、政治家のホームページを分析した村井他（2008）がある。

づいて、未知のデータ（テストデータ）が与えられたときに、コンピュータが分類する。

本研究は、2005年総選挙<sup>8</sup>と2007年参院選<sup>9</sup>のマニフェストを対象にした手作業によるマニフェストの分類データを学習データとして、分類器を生成する。また、分類器を生成する際に、適切に分類するために学習データを加工する。図1のように、マニフェストのテキストデータを構造的に表した後に、テキストを単語に区切っていく作業（形態素解析<sup>10</sup>）を行い、マニフェスト全体から見た単語の特徴度を評価値（TFIDF<sup>11</sup>）として計算し、適切な分類アルゴリズムを選択した上で、分類器を生成し、機械学習を行う。

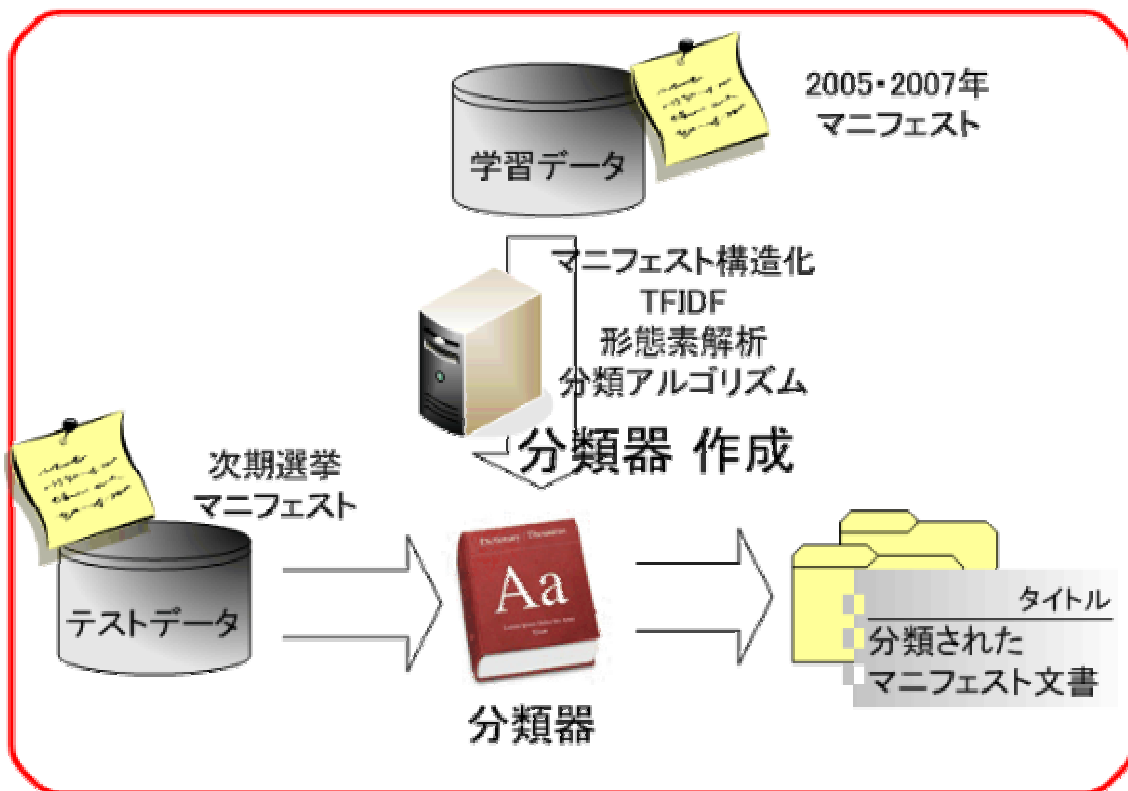
## 図1 マニフェスト自動分類に用いる機械学習の概要

---

<sup>8</sup> 2005年衆院選のマニフェストとして用いたデータは、自民党の「2005 自民党の約束」、民主党の「政権公約権 500 日プラン」、公明党の「マニフ 2005」、共産党の「衆議院選挙にのぞむ日本共産党の各分野の政策」、社民党の「総選挙公約 2」、国民新党の「公約」である。院選のマニフェストとして用いたデータは、自民党の「参議院選挙公約・2007」、民主党の「政権公約・マニフェスト」、公明党の「参院選重点公約・マニフェスト 2007（マニフェスト 2005 改定）」、共産党の「2007 参議院選挙政策 / 12 の重点政策」、社民党の「参議院選挙公約 2007」、国民新党の「第 21 回参議院議員選挙・わが党の選挙公約」、新党日本の「新しい日本宣言」である。

<sup>10</sup> 形態素とは言語で意味を持つ最小単位のこと。形態素解析とは、与えられた文を形態素に分割する作業である。このとき、辞書（「品詞」などの情報付きの単語リスト）中の情報を参照することで、「品詞」、「活用形」、「読み」などの情報を得ることが可能である。

<sup>11</sup> TF：Term Frequency と IDF：Inverse Document Frequency の略。文章中における特徴的な単語を抽出するためのアルゴリズムである。図 4 を参照。



本研究では、機械学習を行うために、Weka<sup>12</sup>と呼ばれるデータマイニングのライブラリ<sup>13</sup>を使用している。また、自然言語処理の際には、GoSen<sup>14</sup>と呼ばれる形態素解析のツールを使用している。

## 2.2. 学習データの作成

以下、順を追って本研究の機械学習に用いるデータの作成プロセスを述べる。

### 2.2.1. データのクリーニング

まず、手作業でコーディングを行ったデータの全項目について、スペルミスやコードの単純ミスがないかを点検する。全く同じコードが含まれていないか確認し、人手の作業によるミスをできる限り取り除く。

このようにして作られたデータを機械学習の学習データとして使う。以後、機械学習するためにデータを加工していくが、元のデータと区別するために、ここで作られたデータを

<sup>12</sup> Weka のホームページは、<http://www.cs.waikato.ac.nz/ml/weka/>である。ワイカト大学（ニュージーランド）を中心に開発が進められている。本システムでは、Weka の Development Version の 3.5.7 を使用している。

<sup>13</sup> 汎用性の高い複数のプログラムを、再利用可能な形でまとめたもの。

<sup>14</sup> GoSen は、Java で開発された形態素解析エンジンの Sen (<http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html>) を改良したものである (<http://www.itadaki.org/wiki/index.php/GoSen>)。

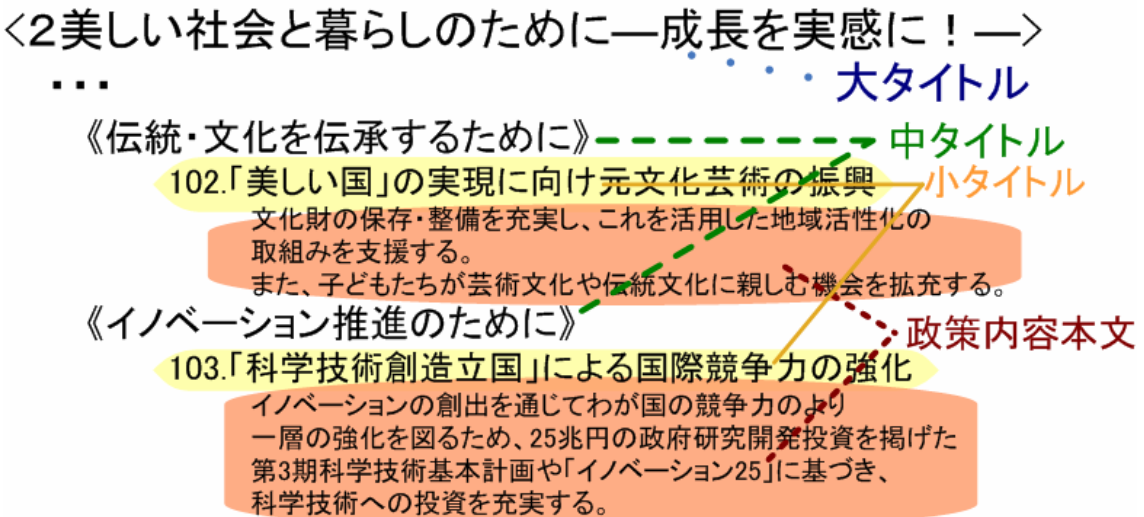
コーパス（辞書）と呼ぶことにする。

### 2.2.2. コーパスの構造化

次に、マニフェストのテキストの特徴を明らかにするために構造化を行う。マニフェストを分析する単位は句読点で分けられたテキストであるため、テキスト情報を解析してコーディングを試みる際、一つの単位だけではテキストが短すぎる場合もあり、文脈を捕らえるのが難しくなる。

そこで、対象となるテキストが、元のマニフェストでは、どのような構造に位置しているかを把握することが重要である。例えば、マニフェストの文章構造は図 2 のような章と段落に分かれていて、ある程度、政策としてまとまって書かれている。その章や段落に注目し、句読点区切りの一単位の文章に加えて、学習データを構成するテキストとして解析の対象とする。

図 2 マニフェストの構造的な章立ての例



(注)2007年自民党マニフェストより

表 1 マニフェストの構造化を表現したデータ例

大タイトル	中タイトル	小タイトル	テキスト
<2美しい社会と暮らし	《伝統・文化を継承す	102.「美しい国」の実	文化財の保存・整備を充実し、
<2美しい社会と暮らし	《伝統・文化を継承す	102.「美しい国」の実	これを活用した地域活性化の取組
<2美しい社会と暮らし	《伝統・文化を継承す	102.「美しい国」の実	また、
<2美しい社会と暮らし	《伝統・文化を継承す	102.「美しい国」の実	子どもたちが芸術文化や伝統文化
<2美しい社会と暮らし	《イノベーション推進	103.「科学技術創造	イノベーションの創出を通じてわか
<2美しい社会と暮らし	《イノベーション推進	103.「科学技術創造	25兆円の政府研究開発投資を掲
<2美しい社会と暮らし	《イノベーション推進	103.「科学技術創造	科学技術への投資を充実する。

このデータをデータベースに入れて、処理を行うと、表 1 のように大タイトル・中タイト

ル・小タイトル・本文という構造となる。

### 2.2.3. 自然言語処理とTFIDFによる単語の評価

本研究では、形態素解析を適用した後、図3のように特定の品詞（名詞，形容詞，形容動詞，動詞）の形態素のみを取り出して、特徴語とした。このとき、活用語は終止形に揃えられている。また、マニフェストの特徴を表す品詞に適さない単語を除外している（一文字のみの形態素，番号など）。

図3 テキスト解析の説明図

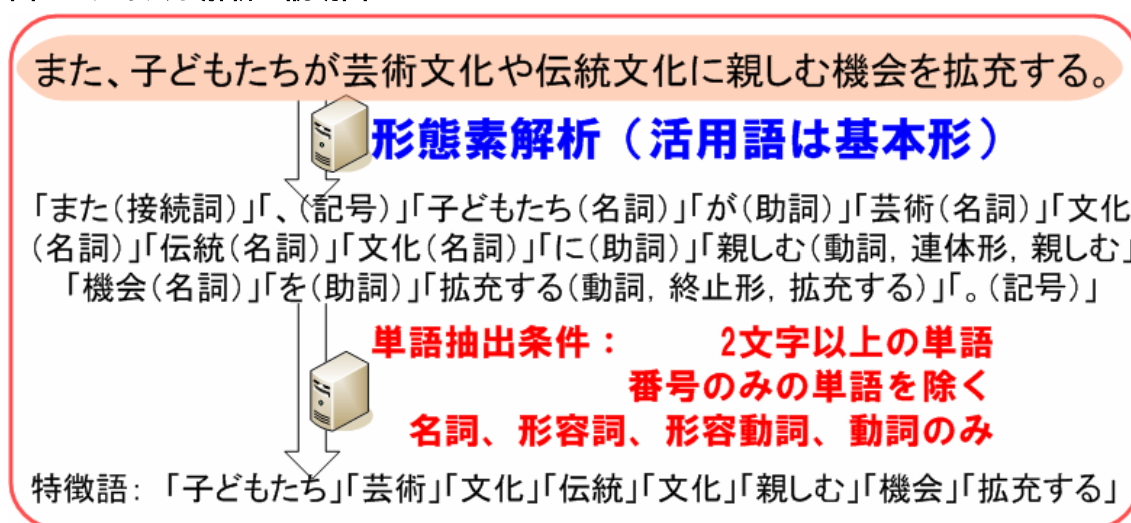
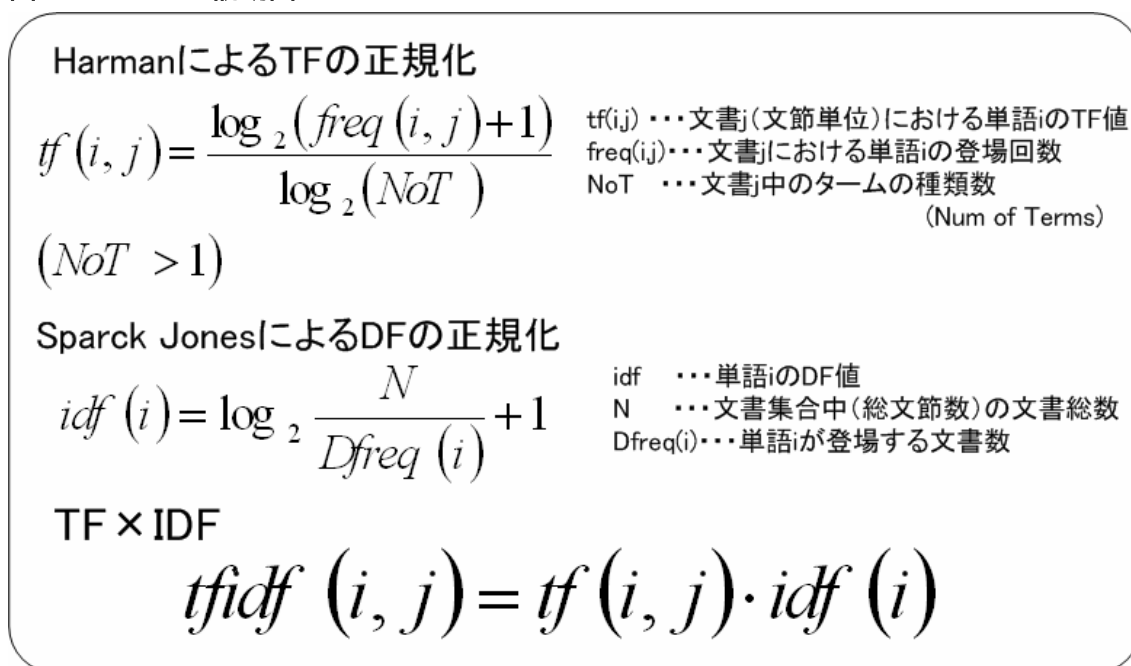


図4 TFIDFの説明図



さらに、図4のTFIDFを使用して特徴語を数値で評価し、特徴語ベクトルを作った。その結果、得られた特徴語ベクトル群をマニフェストの特徴語データベースという。

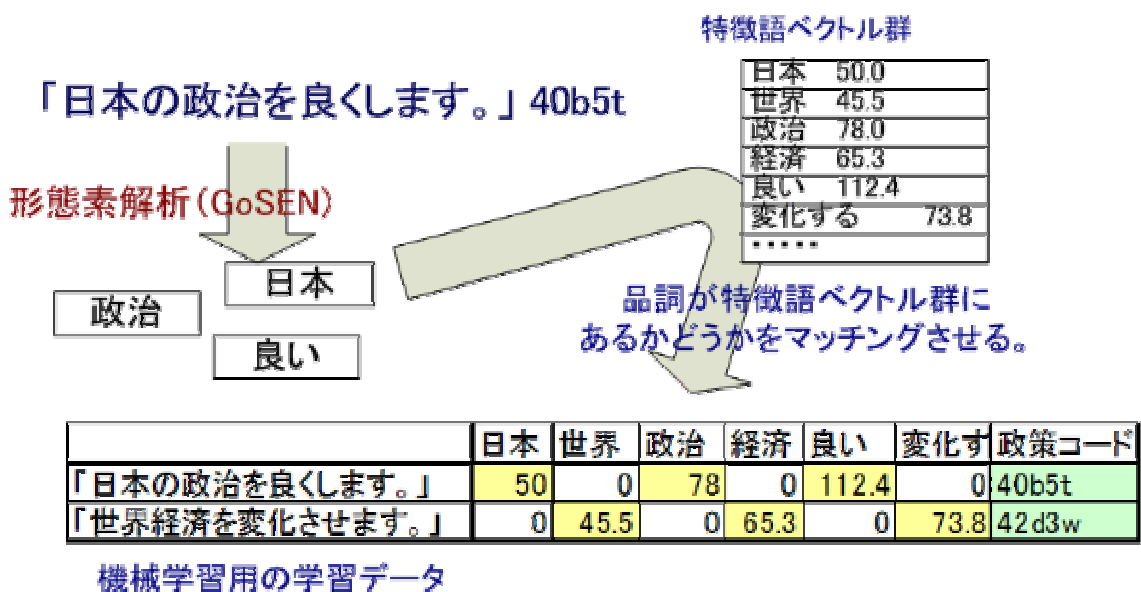
特徴語ベクトルは与えられた文章群における特定の文章の特徴をより良く表現する程度を表す。具体的には、特徴的で珍しい単語ほど値が大きくなる。それを利用して単語の重み付けを行うことにより、機械学習の際に効率的な分類が可能となる。一般的にTFIDFの値を採用することが多く、本研究でもそれを踏襲している。

## 2.2.4. 構造化文章と特徴語の一致判定

学習データを作成するためには、構造化したコーパスのテキストが上記の特徴語に存在するかどうかを示す特徴語ベクトルが必要である。

図5のように、特徴語が存在するものは、その特徴語ベクトルの値を用いる。存在しない場合は、値は0となる。

図5 構造化文章と特徴語の一致判定の例



## 2.2.5. Weka 用の学習データの生成

ここまでの手順で作られたデータに基づいて、Wekaで機械学習を行うためにデータを加工する。

機械学習においては、特徴語ベクトル群に加えて、政策対象、政策分野、政策内容、政策賛否、それぞれの政策コード別にデータが必要となる。

本研究では、句読点区切りの文章と、それに割り振られた政策コード、特徴語ベクトルとの一致判定による値、これら一行分のデータをインスタンスと呼ぶ。インスタンスはWekaで扱う学習データを構成する一単位となる。また、対象、分野、内容、賛否、四つの政策コード別に分けられた、総体としての学習データをインスタンス群と呼ぶ。

### 2.3. 分類器の作成と評価

最後に Weka を使って，マニフェストのデータを機械学習させる．分類器はそれを使って予測するものに応じて作る必要がある．そのため，政策対象，政策分野，政策内容，政策賛否の各政策コードに対応する四つの分類器を作る．インスタンスに必要なデータは，識別 ID (分析単位の区別)，特徴語ベクトルの単語の値，政策コードの大きく三つの要素となる．

加えて，本研究では，政策分野と政策内容を求めるに際して政策対象も学習データに入れる．なぜなら，政策分野と政策内容は政策対象のコードに依存すると考えられるからである．また，政策賛否を求める際には政策対象と政策分野を学習データに加える．

既存研究<sup>15</sup>によると，機械学習のアルゴリズムとして Weka のライブラリから J4.8<sup>16</sup>を選択し，インスタンス数と特徴語ベクトル数が最大となるときに条件が最も良くなる．そこで，本研究の実験では，マニフェストのコーパス全てを解析して得られたデータを使って分類器の一致率を検証する．今回の実験では，6,073 のインスタンスと 5,563 次元の特徴語ベクトルを用いることになる．

なお，一致率とは人間のコーダーによる分類結果とコンピュータによるコーディングとの合致の程度を指す．人間による分類には信頼性の点で問題が残ることを既に指摘したが，他に代わりとなる適切なベンチマークが存在しないので，便宜的に用いる．従って，一致率が低いからといって，コンピュータによるコーディングの精度が低いとは必ずしもいえないことに注意が必要である．

上記の仕様で実験機<sup>17</sup>を用いて計算すると，一つのインスタンス群につき，9 時間ほどで計算が終わり，全ての学習データを計算するのにおよそ 36 時間かかる．分類器の一致率は表 2 のとおりである．

表 2 実験結果による分類の再現率 (%)

10-folds Cross Validation		Training-mode	
1 政策対象	80.2	1 政策対象	91.5
2 政策分野(大分類)	71.0	2 政策分野(大分類)	88.8
3 政策内容(小分類)	53.5	3 政策内容(小分類)	81.6
4 政策賛否	84.0	4 政策賛否	93.6

分類器を作る際には，学習データ全てを使う Training-mode と，一致率をより詳しく調

<sup>15</sup> 上神・佐藤 (2008, 15-20) を参照．

<sup>16</sup> 機械学習のためのアルゴリズムの一つである．枝刈りを任意にできるように拡張された C4.5 アルゴリズムであり，プログラム化しやすいように改良されている．C4.5 アルゴリズムについては以下を参照のこと．

Ross Quinlan .1993. *C4.5: Programs for Machine Learning*, Morgan aufmann Publishers, San Mateo, CA.

<sup>17</sup> 実験機は，Dell Precision Workstation 380 (CPU : Pentium D 950-3.40GHz，メモリ : 4.0GB，Java 1.5.11)．

べる際に用いる 10-folds Cross Validation という二つの機能を用いて、それぞれの一致率を調べた。Training-mode では、学習データ全部を使って分類器を作り、テストデータも学習データで使うデータと同じものを使用するため、過学習を伴う指標値になる可能性がある。一方、10-folds Cross Validation は、学習データを 10 分割して、その一つから分類器を作り、残りの学習データをテストデータとして計算するという試行を 10 回行うため、分類器の性能をある程度正確に測ることができる。

さて、実験結果によると、10-folds Cross Validation では、政策対象と政策賛否で約 8 割、政策分野で約 7 割、政策内容が約 5 割の一致率になった。また、Training-mode では、政策対象と政策賛否で約 9 割、政策分野が 8 割強、政策内容が 8 割弱となった。過学習の可能性のあるものの、機械学習としては、かなり精度の良い分類器ができたと考えられる。

## 結語

本研究では、政党や政治家の政策的な立場を推定するために考案されてきた様々な方法を概観し、改善の方策について検討してきた。これらの方法論は選挙研究や理論研究の基礎的なデータを得るために必要不可欠であるが、日本においては方法論の検討が十分に行われてきたとは言い難い。

そこで、主に日本を事例とする先行研究を参照しつつ、各方法論の特徴を考察した。アンケート調査とは異なり、公約の内容分析には政党や政治家の立場を明らかにすべき争点を客観的な手続きに基づいて選び出すことができるという特徴がある。しかしながら、公約分析には多大なコストが掛かり、分析の信頼性にも問題が残る。

これらの短所を解決する方策として、日本においてもコンピュータを用いる内容分析の研究を進める必要性を指摘し、その概念と作業手続き、人間によるコーディングとの一致率を示すことにした。2005 年総選挙と 2007 年参院選における各党のマニフェストにコンピュータによる自動コーディングを適用した結果を見る限り、その有用性と可能性を示すことができたのではないと思われる。

近年のマニフェスト選挙に見られるように、日本における政策研究の重要性は増している。また、技術革新に伴って、コンピュータによる自動コーディングの精度はさらに向上すると予想される。本研究が日本における政策研究と情報技術を架橋し、方法論の更なる発展に寄与することを願う。

## 謝辞

本研究は、2008 年 5 月に開催された日本選挙学会研究会のワークショップにおける報告を元に行っている。共同研究者の津江宣寿氏、匿名の査読者の先生方、品田裕先生を始めとするワークショップの参加者各位、コメントをお寄せ下さった福元健太郎先生に記して謝意を表したい。

## 参考文献

- 猪口孝．1983．『現代日本政治経済の構図』東洋経済新報社．
- 上神貴佳．2006．「投票支援ツールと『政策中心の選挙』の実現 オランダの実践と日本における展望」『選挙学会紀要』6号，43-64．
- 上神貴佳・堤英敬．2008．「投票支援のためのインターネット・ツール 日本版ポートマツチの作成プロセスについて」『選挙学会紀要』10号，27-48．
- 上神貴佳・佐藤哲也．2008．「政党や政治家の政策的な立場を推定する 方法論の過去・現在・未来」日本選挙学会研究会報告論文．
- 加藤淳子／マイケル・レイヴァー（福島啓之訳）．1998．「96年日本における政党の政策と閣僚ポスト」『レヴァイアサン』22号，106-114．
- 加藤淳子／マイケル・レイヴァー（杉之原真子訳）．2003．「2000年総選挙後の日本における政策と政党間競争」『レヴァイアサン』33号，130-42．
- 蒲島郁夫．2000．「全国会議員のイデオロギー調査 連立時代の議員と政党」東大法・蒲島ゼミ（編）『現代日本の政治家像 第I巻』木鐸社，25-46．
- 蒲島郁夫・谷口将紀・菅原琢．2005．「2003～04年東京大学・朝日新聞社共同世論調査コード」『日本政治研究』2巻1号，190-208．
- 蒲島郁夫・山本耕資．2005．「2003年東京大学・朝日新聞社共同政治家調査コードブック」『日本政治研究』2巻2号，392-418．
- 小林良彰．1997．『現代日本の政治過程 日本型民主主義の計量分析』東京大学出版会．
- 小林良彰．2008．『制度改革以降の日本型民主主義 選挙行動における連続と変化』木鐸社．
- 佐藤哲也．2002．「電子投票エージェント作成を目的とした選挙争点抽出手法」『日本社会情報学会学会誌』14巻1号，31-44．
- 佐藤哲也．2003．「争点投票支援システムの提案とその評価 2001年参院選を対象として」『選挙研究』18号，43-64．
- 品田裕．1998a．「選挙公約政策データについて」『神戸法学雑誌』48巻2号，541-72．
- 品田裕．1998b．「90年代日本の選挙公約」水口憲人・北原鉄也・久米郁男（編）『変化をどう説明するか：政治篇』木鐸社，147-171．
- 品田裕．2001．「地元利益指向の選挙公約」『選挙研究』16号，39-54．
- 品田裕．2002．「政党配置：候補者公約による析出」樋渡展洋・三浦まり（編）『流動期の日本政治：「失われた十年」の政治学的検証』東京大学出版会，51-72．
- 品田裕．2006．「選挙公約政策データについて」『日本政治研究』3巻2号，63-91．
- 鈴木崇史・影浦峡．2007．「総理大臣演説における語彙多様性の変化」行動計量学会大会報告論文．
- 田中明彦（研究代表者）．1999．『政治テキストの内容分析システムの構築』（課題番号07552001・平成7年度～平成9年度科学研究費補助金・基盤研究（A）（1）・研究成果報告書）．
- 谷口将紀．2006a．「衆議院議員の政策位置」『日本政治研究』3巻1号，90-108．
- 谷口将紀．2006b．「衆議院総選挙候補者の政策位置」『年報政治学』2005-II，11-24．

- 堤英敬．1998．「1996年衆議院選挙における候補者の公約と投票行動」『選挙研究』13巻，89-99．
- 堤英敬．2002．「選挙制度改革と候補者の政策公約 小選挙区比例代表並立制導入と候補者の選挙戦略」『香川法学』22巻2号，90-120．
- 堤英敬・上神貴佳．2007．「2003年総選挙における候補者レベル公約と政党の利益集約機能」『社会科学研究』58巻5・6合併号，33-48．
- 村井源・松本斉子・山本竜大・往住彰文．2008．「Webの計量言語学的分析からみた政治的感性の特徴」『日本感性工学会研究論文集』7巻3号，561-569．
- 山本竜大．2005．「2003年衆議院選挙における候補者ホームページとその政策・公約に関する分析」『選挙学会紀要』5号，79-95．
- レイヴァー，マイケル／ケネス・ブノア（上ノ原秀晃訳）．2006．「政党の政策位置を推定する：比較の中の日本」『日本政治研究』3巻1号，109-133．

- Bara, Judith. 2001a. "Using Manifesto Estimates to Validate Computerized Analyses." In Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and, Eric Tanenbaum. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1988*. Oxford University Press, 143-156.
- Bara, Judith. 2001b. "Tracking Estimates of Public Opinion and Party Policy Intentions in Britain and the USA." In Michael Laver ed. *Estimating the Policy Position of Political Actors*. Routledge, 217-236.
- Benoit, Kenneth and Michael Laver. 2006. *Party Policy in Modern Democracies*. Routledge.
- Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-1988*. Oxford University Press.
- De Vries, Miranda, Daniela Giannetti, and Lucy Mansergh. 2001. "Estimating Policy Positions from the Computer Coding of Political Texts: Results from Italy, the Netherlands and Ireland." In Michael Laver ed. *Estimating the Policy Position of Political Actors*. Routledge, 193-216.
- Garry, John. 2001. "The Computer Coding of Political Texts: Results from Britain, Germany, Ireland and Norway." In Michael Laver ed. *Estimating the Policy Position of Political Actors*. Routledge, 183-192.
- Kleinnijenhuis, Jan and Paul Pennings. 2001. "Measurement of Party Positions on the Basis of Party Programmes, Media Coverage and Voter Perceptions." In Michael Laver ed. *Estimating the Policy Position of Political Actors*. Routledge, 162-182.
- Klingemann, Hans-Dieter, Andrea Volkens, Judith L. Bara, Ian Budge, and Michael D. McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and*

- Governments in Eastern Europe, European Union, and OECD 1990-2003*. Oxford University Press.
- Laver, Michael and W. Benn Hunt. 1992. *Policy and Party Competition*. Routledge.
- Laver, Michael and John Garry. 2000. "Estimating Policy Positions from Manifestos and Other Political Texts." *American Journal of Political Science* 44 (3): 619-634.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311-331.
- Miura, Mari, Kap Yun Lee, and Robert Weiner. 2005. "Who Are the DPJ?: Policy Positioning and Recruitment Strategy." *Asian Perspective* 29 (1): 49-77.
- Ray, Leonard. 2001. "A Natural Sentence Approach to the Computer Coding of Party Manifestos." In Michael Laver ed. *Estimating the Policy Position of Political Actors*. Routledge, 149-161.