民主的討論の構造と動態：トピックモデルによる日米議会スピーチデータの比較分析

阪本拓人（東京大学）・瀧川裕貴（東北大学）

要約

トピックモデルに代表される統計的潜在意味解析は、政治的・公共的討議を規定するフレームやアジェンダの構造・動態を自動的に抽出できる手法として、政治学や比較政治学において注目されつつある．本研究では、トピックモデルを日米両国の議会における過去 20 年間の議事録のデータに適用した．分析の結果、米国の議会では、具体的な政策課題よりも、連邦政府や連邦議会の運営・機能に関する一般的・理念的な議論に、より多くの言葉が割かれていることが明らかになった．これに対して、日本の国会では、外交や安全保障、景気と雇用、税と社会保障など多様な政策課題が高い頻度で議論される傾向がある．また、日本の国会の方がその時々の政治・社会情勢に応じてトピックの頻度分布が時間的に変動する傾向が強いこと、日本の方が政党間のトピック分布の違いが際立っていること、などが明らかにされた．両国間のこうした違いは、立法府における政策的な関心の差異を反映しているだけでなく、議会政治そのものの制度的な位置付けの違いにも起因していると考えられる．

はじめに

政治的・公共的討議をめぐる政治学や比較政治学の研究領域では、米国を中心に、近年、自然言語処理など計算社会科学的手法を用いた研究が盛んに行われている(Frimer, Aquino, Gebauer, Zhu, & Oakes, 2015; Grimmer & King, 2011; Grimmer & Stewart, 2013; Jacob, Ethan, Suresh, & Laurence, 2012; Lucas et al., 2015; Rule, Cointet, & Bearman, 2015)．その内容は多岐にわたるが、先端的な手法の一つとして活用されているのが、トピックモデルに代表される統計的潜在意味解析である(Blei, Ng, & Jordan, 2003; 佐藤 & 奥村, 2015)．トピックモデルは、公共空間における討議を規定するフレームやアジェンダの構造・動態を自動的に抽出できる、有益な手法として注目されている(Grimmer, 2010; Lucas et al., 2015)．本研究では、この手法を、日米両国の議会における過去 20 年以上に及ぶ討議を記録した議事録データに適用した．米国議会の議事録をトピックモデルで分析した研究としては、Quinn らの研究があるが、対象が上院に限定されており、期間も 10 年程度にとどまっている(Monroe, Colaresi, & Quinn, 2008; Quinn, Monroe, Colaresi, Crespin, & Radev, 2010)．これに対し、本研究では、米国については上下両院 20 年以上の討議をカバーしている点、さらに、同時期の日本の衆参両院のデータの分析をこれに加えて比較を試みている点などが、既存の研究に見られない新たな点である．

データ

分析対象となる議事録データは、米国については、政府印刷局（GPO）のサイトから入手できる *Congressional Record* のテキストデータを活用した[1]．対象時期は 1994 年 1 月から 2016 年 12 月まで、上院・下院両方のデータを用いた．GPO のサイトからのダウンロードの際には、

[1] https://www.gpo.gov/fdsys/browse/collection.action?collectionCode=CREC

第三者の作成したパーサーを活用し[2]、発言者の名称を抽出するなど xml 形式でのデータの構造化を行なった．発言者の所属政党といった *Congressional Record* 本体に含まれない情報は、ProPublica の API サービスが提供する米国議会の議員データベースから取得した[3]．

得られたデータは、python ベースの自然言語処理ツール nltk を利用して、前処理した．具体的には、スペースや句読点等で全テキストを単語に分割し、1-gram の bag of words を作成した．その際、冠詞や前置詞等のストップワードを取り除いたほか、nltk の Snowball ステマーによって単語を対応する語根へと変換した．こうして得られたデータ、すなわちコーパスは、6,423,453 の発言（パラグラフ単位）から構成され、合計 183,622,386 語（重複を除いたトークンの数は 187,452 語）を含む．トピックモデルの適用にあたっては、これらからさらに極端な高頻度語（50%以上の発言にわたって出現するトークン）・低頻度語（コーパス全体で 50 回未満しか出現しないトークン）を取り除いた 24,981 語に対して分析を行うことで、処理の高速化を図った．

日本のデータは、国立国会図書館が管理する「国会会議録検索システム」を利用して取得した議事録データのうち[4]、1994 年 1 月から 2016 年 10 月までの衆議院・参議院の本会議および予算委員会のものを用いた．各発言に関わる発言者やその所属会派といった情報も同じデータベースから得た．テキストは Mecab で形態素解析を行なった上で、米国の議事録と同様、ストップワード（「てにをは」や「委員、大臣」など）を取り除き、1-gram の bag of words とした．コーパスに含まれる発言数（発言者単位）は合計 482,051、総語数は 27,413,942（重複を除いたトークン数は 98,660 語）であるが、分析の際には、極端な頻度を示す語を除いた 19,889 語のみを用いた．日米間でコーパスの「発言数」に乖離があるのは、*Congressional Record* と「国会会議録」とで発言の集約単位（前者はパラグラフ単位・後者は個別の発言者単位）が異なっているからであり、この点は注意する必要がある．

方法

これらテキストデータの生成モデルとして、最も単純なトピックモデルである潜在ディリクレ配分法（Latent Dirichlet allocation, LDA）を仮定した(Blei, Ng, & Jordan, 2003)．LDA では、テキストの背後に与えられた数 K の潜在的なトピック k (k=1,⋯,K)が存在すると考える．コーパスを構成する個々の文書 d (M を文書の総数として d=1,⋯,M)はこれらトピックの一定の混合 θ_d=($\theta_{d,1}$,⋯, $\theta_{d,K}$) によって特徴付けられ、各トピック k は V 種類の単語の生成確率を指定する分布 ϕ_k=($\phi_{k,1}$,⋯, $\phi_{k,V}$) によって意味付けられる．これらの確率ベクトルは、それ自体、以下のようにディリクレ（Dirichlet）分布から生成されるものとする．

$$\left.\begin{aligned}\theta_d &\sim Dirichlet(\alpha) \quad (d=1,\cdots,M),\\ \phi_k &\sim Dirichlet(\beta) \quad (k=1,\ldots,K).\end{aligned}\right\}$$

ここで、パラメータ α と β は、それぞれ K 次元、V 次元のベクトルである．

この時、d 中の i 番目に現れる単語 $w_{d,i}$ ($w_{d,i}$ ∈{1,⋯,V}) は、次のような生成過程によって確

2 https://github.com/unitedstates/congressional-record

3 https://propublica.github.io/congress-api-docs/

4 http://kokkai.ndl.go.jp/

率的に決定されるものとする．まず、θ_dをパラメータとする多項分布によって、単語に対応するトピックがK種類の中から一つ選ばれる．これを潜在変数 $z_{d,i}$ ($z_{d,i} \in \{1,\cdots,K\}$) で表す．次に、$\theta_{z_{d,i}}$をパラメータとする多項分布によって、$V$種類の単語の中から一つを選ぶことになる．以上をまとめると、

$$\left.\begin{aligned} z_{d,i} &\sim Multinominal\left(\theta_d\right), \\ w_{d,i} &\sim Multinominal\left(\phi_{z_{d,i}}\right). \end{aligned}\right\}$$

この階層的なモデルにおけるパラメータ（$\alpha, \beta, \theta, \phi$など）を日米の議事録データを用いて推定することで、モデルを学習する．推定にあたっては、python の自然言語処理パッケージ gensim（gensim.models.ldamulticore.LdaMulticore）を活用した．用いた学習アルゴリズムは、オンライン変分ベイズ法という高速な確率的最適化手法である(Hoffman, Blei, & Bach, 2010)．学習は、トピック数 K を 10 から 100 の範囲で変化させながら行った．特に断らない限り、以下で報告する結果は K=40 のケースに基づく．日米間のトピック分布の差異に関する定性的な特徴は、Kを多少変化させても大きく変わることはない．

結果

(1)トピック分布の日米比較

図 1 は、米国（左）と日本（右）の議会における発言のトピック頻度 θ_dの推定値を、発言内の語数を考慮して、1994 年から 2016 年の全期間における全ての発言にわたって集約したものである．図中のトピック番号と対応する色は恣意的に割り当てたものであり、日米間で特に対応があるわけではないので注意されたい．各トピックの意味的な内容を記述する単語頻度ベクトル ϕ_kの推定値については、コーパス中での出現頻度の高い上位 10 のトピックについて、表 1（米国）および 2（日本）に、各トピック上位 10 個の単語（米国については語根）とそれらの頻度を記してある．

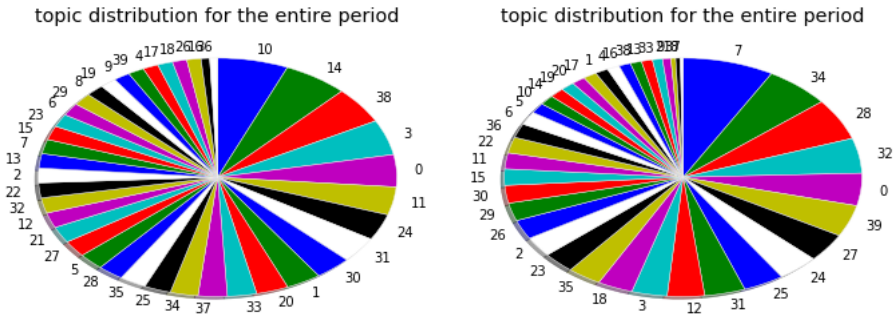

図 1　米国の議会（左）と日本の国会（右）におけるトピック分布

まず、米国について見ると、外交上や内政上の具体的な政策課題が上位 10 の高頻度トピックの中にほとんど見られないのが特徴的である．これは、こうした課題がモデルの中でトピックとして捕捉されていないわけではなく、たとえば、税制と国民への負担に関するトピック 24 (第 8 位)、景気や雇用に関わるトピック 1（第 10 位）をはじめ、戦争、核兵器や安全保障（トピック 33、第 12 位)、石油やガスなどの資源（トピック 34、第 14 位)、医療と社会保障（トピッ

ク 33、第 16 位）など、トピックのリストを広く精査すれば、一通りの課題は抽出されていることが分かる．だが、米国の上下両院の過去 20 年以上におよぶ討議の中で、より高い頻度で言及されるトピックとして抽出されたのは、これらとは別種のものであった．一つは、多くの発言において頻繁に使われる語彙（want, change, good など）や数字をまとめたトピック（トピック 10、14、38、3）であり、これらが上位 4 位を占めている．これに続くのが、トピック 0、11、24 であるが、これらは、連邦準備制度（FED）を含む連邦政府諸機関の規制や機能（トピック 10、11）、あるいは連邦議会内の議事進行や立法（トピック 24）に関わる、行政的ないし手続き的な内容である．最後に、第 9 位に入っているトピック 30 は、典型的には、米国における自由や民主主義の価値を論じるような、理念的な内容を持つトピックである．

表 1　上位 10 のトピックにおける単語頻度分布（米国）

topic 10 (freq: 0.063708)
go: 0.063160
get: 0.039826
want: 0.036432
say: 0.032317
peopl: 0.030697
make: 0.023039
know: 0.020871
side: 0.016968
way: 0.014549
sure: 0.013191

topic 14 (freq: 0.058032)
think: 0.029345
one: 0.022143
talk: 0.020018
peopl: 0.019002
would: 0.017698
thing: 0.015739
know: 0.015184
issu: 0.015018
come: 0.012744
time: 0.011758

topic 38 (freq: 0.049960)
would: 0.020397
chang: 0.017087
believ: 0.014164
bill: 0.012807
make: 0.012476
-: 0.011107
polici: 0.010080
rule: 0.009178
issu: 0.009134
process: 0.008826

topic 3 (freq: 0.047785)
$: 0.073436
year: 0.056961
percent: 0.045077
million: 0.039433
billion: 0.034174
1: 0.032662
000: 0.031484
2: 0.019142
5: 0.019083
10: 0.015042

topic 0 (freq: 0.043503)
feder: 0.025188
govern: 0.021433
requir: 0.020442
busi: 0.019005
would: 0.017395
-: 0.014418
agenc: 0.014315
inform: 0.013931
regul: 0.013762
small: 0.012805

topic 11 (freq: 0.038052)
program: 0.066105
-: 0.025023
need: 0.021958
provid: 0.019717
develop: 0.019257
fund: 0.017592
research: 0.017520
help: 0.016203
state: 0.013223
grant: 0.011621

topic 24 (freq: 0.035700)
year: 0.044441
last: 0.029132
week: 0.027551
work: 0.024090
congress: 0.022932
pass: 0.022328
day: 0.018650
month: 0.017989
bill: 0.013717
time: 0.013685

topic 31 (freq: 0.035158)
tax: 0.071294
pay: 0.036513
famili: 0.019897
-: 0.018502
american: 0.018220
money: 0.017371
credit: 0.016611
peopl: 0.015539
taxpay: 0.013590
incom: 0.013537

topic 30 (freq: 0.032753)
american: 0.040173
nation: 0.023472
serv: 0.021039
peopl: 0.020569
right: 0.017514
countri: 0.014327
state: 0.012047
constitut: 0.011248
histori: 0.010957
citizen: 0.010761

topic 1 (freq: 0.031339)
job: 0.048844
economi: 0.027754
product: 0.023038
compani: 0.018071
market: 0.017803
econom: 0.017680
price: 0.017237
industri: 0.017228
worker: 0.017086
busi: 0.016781

表２　上位 10 のトピックにおける単語頻度分布（日本）

topic 7 (freq: 0.082337) 総理: 0.0218813089056 聞く: 0.0139463628957 質問: 0.00924200389238 出る: 0.00896140565745 問題: 0.00795885680566 国民: 0.00784557110227 政治: 0.00779755887472 書く: 0.00737224917264 わかる: 0.00685754390804 見る: 0.00642762045438	topic 39 (freq: 0.042926) 復興: 0.0169644680141 被災: 0.0127732581936 支援: 0.0104728914328 取り組む: 0.00782930538825 日本: 0.00780000992131 向ける: 0.00772082851642 進める: 0.00642505577297 地域: 0.00599623939203 重要: 0.00596559257296 国: 0.00595467285995
topic 34 (freq: 0.058567) 県: 0.0147891476559 地域: 0.00876387125611 問題: 0.0077226671395 市: 0.00698989513106 聞く: 0.00674722785032 地元: 0.00645617353845 日本: 0.00642739678129 出る: 0.00618692185649 状況: 0.00589326317857 見る: 0.0056212613285	topic 27 (freq: 0.038675) 議論: 0.0440207475313 問題: 0.0296397476764 意味: 0.0147174245547 国民: 0.0122826186242 やっぱり: 0.00898964831586 理解: 0.00883716763016 国会: 0.00774748620998 必要: 0.00762790438034 いろんな: 0.00729075066675 基本: 0.00663788846154
topic 28 (freq: 0.056299) 日本: 0.0214073444521 問題: 0.0184747326734 総理: 0.0167338623033 国: 0.0100261484416 中国: 0.0084450333543 アメリカ: 0.00807574071888 聞く: 0.00778712494163 関係: 0.00695707526845 北朝鮮: 0.00607749060995 交渉: 0.00544712324404	opic 24 (freq: 0.037245) 円: 0.0217940058204 問題: 0.0147376827542 出す: 0.0085339752429 お金: 0.00762035469855 資料: 0.00747670397289 聞く: 0.00736319618287 責任: 0.00686583714771 出る: 0.00662242609578 国民: 0.00659167170225 見る: 0.00570573155081
topic 32 (freq: 0.047318) 円: 0.0226866511494 消費: 0.00958680577016 税: 0.00947587717844 総理: 0.00868442324217 財政: 0.00859285415208 負担: 0.00804228747625 見る: 0.00785501366071 予算: 0.00770399744903 上がる: 0.00704191266317 国民: 0.00655183720132	topic 25 (freq: 0.036744) 米: 0.0180281072896 自衛隊: 0.0115095981196 日本: 0.0114148929917 軍: 0.00969856735201 アメリカ: 0.00835739323339 行う: 0.00822864410579 沖縄: 0.00734820191624 政府: 0.00714080776053 行使: 0.00652830513021 憲法: 0.00636869418373
topic 0 (freq: 0.043844) 経済: 0.025488192227 状況: 0.0111209858033 日本: 0.0110529793471 デフレ: 0.0104321960465 景気: 0.00990851583782 企業: 0.0086474096129 物価: 0.00808648118273 政策: 0.00802570878784 対策: 0.00773272018109 見る: 0.00729451373492	topic 31 (freq: 0.036683) 働く: 0.010538903528 労働: 0.0104916320801 女性: 0.00882288368771 介護: 0.00879056642131 雇用: 0.00874613263366 保育: 0.0086591086943 子供: 0.00835175615475 制度: 0.00779776405705 歳: 0.0073439300586 支援: 0.00706138137918

こうしたトピックの配列と対照的な様相を示すのが、日本の国会での討議である．上位 10 の高頻度トピックには、第 1 位のトピック 7 や第 7 位のトピック 27 といった、本会議や委員会における質疑や議事進行に関わる手続き的な内容のトピックが見られるものの、大半は、外交上や内政上の政策課題によって占められている．たとえば、外交と安全保障の分野については、アジアの安全保障環境に関わるトピック 28（第 3 位）や、沖縄問題を含む日米同盟に関するトピック 25（第 9 位）が挙げられる．税制や財政に関するトピック 32（第 4 位)、景気や物価に関するトピック 0（第 5 位）も高い頻度で言及される．最後に、トピック 39（第 6 位）のように、災害や復興に関するトピックが上位にある点も、米国には見られない特徴的な点である．

(2)トピック分布の時間変化や政党間比較に見られる違い

その他の違いについても簡単にまとめておく．図 2 は、上述した米国（左）と日本（右）それぞれの高頻度トピックの言及頻度の経年変化を図示したものである．両国とも各トピックの頻度の変動やそれに伴う順位の入れ替えが見られるが、相対的に日本の方がより激しい変動を経験してきたことが分かる．たとえば、日米同盟や集団的自衛権に関するトピック 25 について見ると、いわゆる安保法制をめぐって国内世論が沸騰した 2015 年においては、言及頻度が 6.5%（全期間を通じた平均は 3.7%）に達し第 4 位のトピックになったのに対し、翌年には（沖縄における基地問題の継続にもかかわらず）その数値は 3.5%にまで落ち込み、順位も第 10 位まで下降している（図表省略）．これは一例に過ぎないが、日本の国会におけるトピック分布は、その時々の政治・社会情勢に敏感に反応して変化しやすいということは言えそうである．

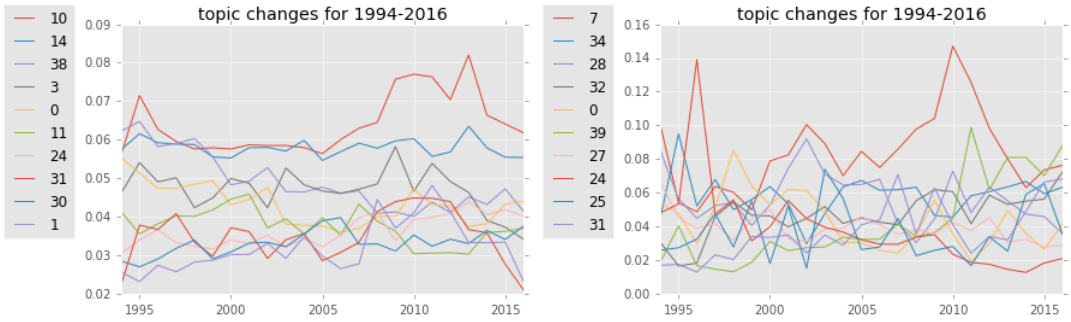

図 2　米国の議会（左）と日本の国会（右）におけるトピック分布の経年変化

最後に、トピック分布を発言者の所属政党ごとに集計し直したところ、ここでも日米間に大きな違いが観察された．一般的に言って、日本では、国会での討議において取り上げるトピックの中身も、これらへの関心の割き方も、政党間で相当の違いがある．図 3 は極端な例として自由民主党と日本共産党のトピック分布の違いを図示したものであるが、多様な政策課題をいわば広く浅く論じる傾向のある自民党と、消費税など税負担の問題（トピック 32）と日米安保や沖縄問題（トピック 25）を集中的に論じる傾向がある共産党との違いが、浮き彫りになっている．これに対して、米国の二大政党である共和党と民主党の間には、トピック分布上、それほど大きな差は見られない（図省略）．言い換えると、個別の課題をめぐる政策上・イデオロギー上の相違は別にして、米国の議員は、関心の所在やフレーミングの仕方といった、討議を行うための「土台」を、政党間の違いを超えてある程度共有しているということになる．

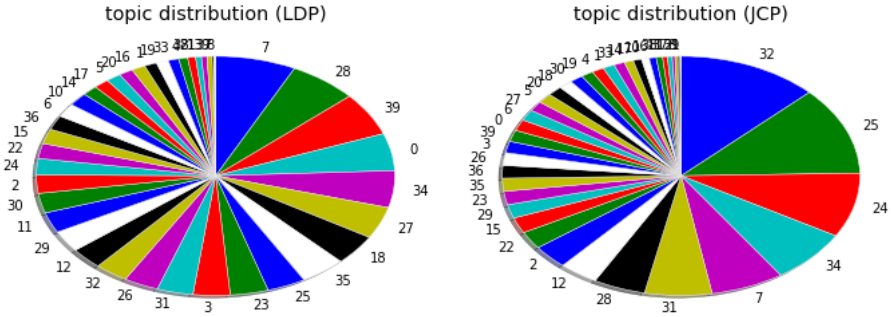

図 3　自民党（左）と共産党（右）のトピック分布

むすび

以上の分析の結果、米国の議会と日本の国会とでは、政治的な討議の構造において大きな違いがあることが明らかになった．すなわち、前者では、連邦政府・議会の機能や手続きに関わる議論や、自由と民主主義に関する理念的な議論に、多くの言葉が割かれる一方、後者では、内外の幅広い政策課題が取り上げられ議論される傾向がある．また、両国間には、討議の構造の時間的な変動や政党間の差異に関しても、顕著な違いが見られた．こうした違いが、いかなる要因によってもたらされているのかについては、慎重な検討を要するが、一つ留意すべきは、日米間で議会政治の制度的な位置付けが大きく異なる点であろう．日本の国会においてより具体的な政策課題がより高頻度で議論される傾向があるのは、行政府の閣僚と立法府の議員との相互作用が密な日本の議院内閣制のあり方を考慮すれば、それほど驚くべき結果ではない．こうした制度的な要因と、政党や個人のレベルの政策選好がどう絡まり合って議会における討議を形作るのかについては、今後より深い分析が必要となるであろう．

本研究のその他の課題として以下を指摘しておく．ストップワードの選択や発言の集約単位の調整などを通じて日米間の比較をより厳格に遂行する点が一つである．また、本研究では、トピック分布の時間的な変動を追う際（図 2）、各トピックの内容を表す単語頻度分布 ϕ_k を所与のものとしたが、これはかなり無理のある想定である．ϕ_k の時間変化を明示的に組み込んだモデル、たとえば動的トピックモデルによる分析が求められるであろう(Blei & Lafferty, 2006)．最後にトピックの抽出だけではなく、トピックの「論じられ方」の政党間や個人間の違いにも目を向ける必要がある．意見分析や感情分析との組み合わせが有益であろう．

引用文献

Blei, D. M., & Lafferty, J. D. (2006). "Dynamic topic models." *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania, USA.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res., 3*, 993-1022.

Frimer, J. A., Aquino, K., Gebauer, J. E., Zhu, L., & Oakes, H. (2015). A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences, 112*(21), 6591-6594. doi:10.1073/pnas.1500355112

Grimmer, J. (2010). A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. *Political Analysis, 18*(1), 1-35. doi:10.1093/pan/mpp034

Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences, 108*(7), 2643-2650. doi:10.1073/pnas.1018067108

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*. doi:10.1093/pan/mps028

Hoffman, M. D., Blei D. M., & Bach, F. (2010). "Online learning for Latent Dirichlet Allocation." *Proceedings of the 23rd International Conference on Neural Information*

Processing Systems, Vancouver, British Columbia, Canada.

Jacob, J., Ethan, K., Suresh, N., & Laurence, W.-S. (2012). Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. *Brookings Papers on Economic Activity, 45*(2 (Fall)), 1-81.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis, 23*(2), 254-277. doi:10.1093/pan/mpu019

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis, 16*(4), 372-403. doi:10.1093/pan/mpn018

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science, 54*(1), 209-228. doi:10.1111/j.1540-5907.2009.00427.x

Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. *Proceedings of the National Academy of Sciences, 112*(35), 10837-10844. doi:10.1073/pnas.1512221112

佐藤, 一., & 奥村, 学. (2015). *トピックモデルによる統計的潜在意味解析* (Vol. 8): コロナ社.