# 国会会議録データを用いた自然災害に関する集合的認知ダイナミクスの分析

瀧川裕貴（東北大学）・阪本拓人（東京大学）

**要約**

ある社会における集合的意識や認識、認知は集合体のあり方を規定する主要な構成要素であり、それらがいかなるものでどのように変化していくかは社会科学にとって重要な問いをなしている．本研究では、自然災害についての集合的認知を対象にして、大規模な歴史的テクストの自然言語処理等を用いた量的テクスト分析により、とくに変化に着目して、集合的認知のダイナミクスを帰納的に捉える方法論を提案する．データとしては、1947-2016 年の間の日本の衆議院・参議院における本会議と常任委員会におけるほぼすべてのスピーチデータを用いる．トピックモデルを用いて災害関連スピーチを同定したあと、Glove による単語分散表現とクラスタ分析により、単語クラスタを災害関連スピーチにおける潜在的テーマとして特定する．その上で、変動係数を用いて変化の大きい単語クラスタを抽出し検討する．この方法によって、被災や安全をめぐる認知をめぐる認識の転換など、自然災害についての集合的認知の変化についていくつかの実質的発見をすることができる．

**キーワード**

量的テクスト分析、自然言語処理、単語分散表現、クラスタ分析

## １．はじめに

E.デュルケム(Durkheim 1985=1978)が集合表象を社会学の中心的対象としたように、国家や共同体を構成する人々の集合的な認識や認知は集合体のあり方やふるまいを規定する主要な構成要素である．例えば、どのような行為が犯罪で、どのような行為が犯罪でないか、に関する人々の集合的な認識はその社会の刑法の基本を形作っている．また、自然災害に関する集合的な認識は、社会の防災のあり方や災害発生時の責任帰属のあり方などを様々な仕方で規定している.そこで、ある社会において集合的認識がどのようであるか、そして時代を通じていかに変化していくかという問いは社会科学にとって重要な問いをなしている．

このような集合的認識の研究は伝統的には主として、法規や外交文書、文学テクストなどの同時代的・歴史的文書の質的読解に基づいていた．しかしながら近年、様々な歴史的文書をデジタル化して保存する動きが生じるのに伴い、これらの、しばしば人的に読解することが不可能な大規模の量の文書を自然言語処理の手法などを通じて量的に解析することで、集合的意識とそのダイナミクスを解明しようとする試みが現れつつある(Evans and Aceves 2016; Grimmer and Stewart 2013).

本論文もこうした流れに棹さし、大規模な歴史的デジタル文書の自然言語処理に基づき、集合的認知のダイナミクスを分析するための方法を提案する．特に本論文では、大規模なテクストデータを効果的に縮約し、かつトレンドの変化に着目する手続きを生みだすことに焦点をあてることにする．

具体的なトピックとしては自然災害についての国政政治家の集合的認知に焦点を当てる．

いうまでもなく、自然災害を政治的権力のある人々がどのように認識しているかは、防災、緊急対応、復興のプロセスに大きな影響を与える．したがって、そのあり方とダイナミックな変化を分析することは、研究上だけでなく、政策上も重要な課題となりうる．

## ２．先行研究

デジタル化された歴史的文書を利用して集合的意識の変化を分析した研究はいくつかある．この中で近年のものを 2 つ取り上げよう．

1 つは、アメリカ大統領教書演説を用いてアメリカ人の政治意識の決定的変化の時期を同定しようとした Rule らによる研究(Rule et al. 2015)である．彼らの目標は、文書をいくつかの時代によって分割し、それぞれの時代ごとの概念の変化を検討することにある．概念の同定の方法は、各文書を共起ネットワークとして表現した上でコミュニティ検出法を応用して、共起ネットワーク上のコミュニティとして概念を同定するという形をとる．

2 つ目の研究としては、イギリスの刑法裁判所の裁判記録を分析することにより、N.エリアスの文明化の過程仮説、とりわけヨーロッパ史のある段階において人々の暴力に対する集合的意識が決定的に変化したという仮説（Elias 1969=1977,1978）を検討しようとした Klingenstein らの研究(Klingenstein et al. 2014)である．彼らは、裁判記録を暴力に関連したものと暴力に関連しないものとに区別した上で、その語彙の類似度を JS ダイバージェンスによって測定し、両者のかい離が大きくなることをもって、暴力に関する人々の集合意識の変化の瞬間として同定している．

量的研究ではないが、われわれの対象と深く関連した研究として、Samuels の研究(Samuels 2013=2016)を挙げておきたい．彼の研究は、2011 年の東日本大震災に関わる人々の言説について、国家安全保障、エネルギー政策、地方自治という 3 つのテーマに注目して分析を進めている．自然災害時において、とくにこれら 3 つのテーマをめぐり競合する複数のナラティブが活発化するという指摘は本研究においても参照点として利用可能である．

本研究では、Rule らの研究とは異なり、ある政治的文書全体における複数の概念の構成の変化を問うのではなく、自然災害に対する集合的認識という一つの具体的なテーマについての内的構成の変化を検討する．検討する概念やそれに基づく集合的認識を 1 つにしぼることで、その内的構成の詳細をよりきめ細かく検討することができる．また、Kligenstein らの研究とは異なり、あらかじめラベルづけられた複数の文書の間の類似度の変化を問うのではなく、ある集合的認識の内的構成の変化をむしろ帰納的に発見することを重視する．最後に、Samuels の研究とは対象を共有するが、われわれは日本の政治言説の戦後から現在にいたるより長期の変化を検討している．その際、大規模なテクストデータを検討する必要があるため、質的方法よりも量的方法をとる．

## ３．データと方法

本研究で利用するデータは、1947 年以後の戦後日本の国会で議論された言説内容である．具体的には、1947-2016 年の間の衆議院・参議院における本会議と常任委員会におけるほぼすべてのスピーチデータである．データ取得は、インターネット上の国会会議録検索システムを利用して行われた．

まず、スピーチデータの全体から災害に関連するスピーチを同定するために、トピックモデルを用いた．トピック数を 70 と設定し標準的な潜在ディリクレ配分法（LDA）による推定を行い(Blei et al. 2003)、高頻度語のリストを検討する．こうして、災害に関連するトピックを同定できる．次に、このトピックの比率が 10%以上のスピーチを災害関連スピーチとして特定した．以後の分析は、すべてこの災害関連スピーチに限定して行った．

災害関連スピーチはさらに災害に関連する多くの潜在的な言説的テーマによって構成されていると考えられる．災害についての集合意識や集合的認知はこうした潜在的テーマの構成によって特徴付けることができる．そこで本論文ではこうした潜在的テーマを抽出するために次のような方法を採用した．まず、スピーチにおいて用いられている単語の分散表現を求める．これには GloVe（Global Vectors for Word Representation）という手法を用いた(Pennington et al. 2014)．次に、分散表現から計算される単語どうしの距離（類似度）に基づいて、単語クラスタを計算した．これは、クラスタ数を 1000 とし、k-means 法によって求めた．このようにして、災害関連スピーチを構成する 1000 の潜在的テーマ（単語クラスタ）を同定できる．

本研究の特徴は、集合的認知のダイナミクスを描き出すために、潜在的テーマの中で戦後を通じて大きな変化のあったテーマを特定し、これに注目することにある．この目的のためには、いかなるテーマが対象期間を通じて大きく変化したかを特定する方法が必要となる．そこで本論文では、変動係数をこの変化の指標として用いることを提案する．具体的には、まず、年ごとの災害関連スピーチデータ全体において各単語クラスタの占める比率を計算する．その上で、この比率についての変動係数、つまり全期間におけるクラスタ比率の標準偏差を平均で除した指標を算出する．この変動係数の高い潜在的テーマを、期間を通じて変化の大きかったテーマとして同定するのである．

ただし、変動係数にはいくつかの限界がある．例えば、変動係数は頻度（比率）の低いクラスタほど高くなる傾向がある．そこでわれわれは検討の対象を頻度上位 100 のクラスタにしぼった．これに加えて、変動係数では捉えきれない注目すべき変化も存在する．そこで、これら頻度上位 100 のクラスタについて比率の変化のプロットを定性的に検討することで分析を補完した．

## ４．結果

トピック数を 70 としてスピーチデータ全体について LDA によるトピック推定を行った．その結果、表 1 のような高頻度語からなる災害関連トピックが同定された．図 1 はスピーチ全体における災害関連トピックの比率の時系列変化を示している．みてとれるように、阪神・淡路大震災のあった 1995 年、東日本大震災のあった 2011 年にはこのトピックの比率が爆発的に増大していることが分かる．

表 1 災害関連トピックの高頻度語

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| なかった | 被害 | 対策 | 対応 | 安全 | 事故 | 被災 | 発生 | 復興 | 災害 |
| 起こる | 防止 | 起きる | 原因 | 避難 | 事態 | 状況 | 取り組む | 地震 | 震災 |

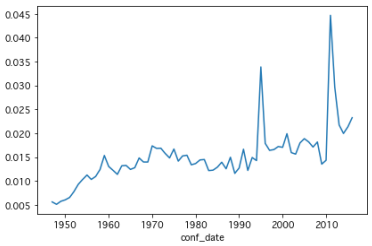


図 1　災害関連トピックの時系列変化

次に、災害関連トピックの比率が 10%以上となるスピーチを災害関連スピーチとして同定し、このサブデータに対して GloVe により単語分散表現を計算し、単語を k-mean により 1000 のクラスタに分類した．表 2 に対象期間を通して高い頻度でスピーチデータに現れたクラスタを示す．ここでは、表 3 に示す変動係数の高いクラスタをいくつか検討する（表 3）．まず、単なる言葉遣いの変化を反映しているクラスタ、「或いは」クラスタ（単語は当該クラスタの最頻語を表す、以下同様)、「災害」クラスタ、などについては考察の対象外とする．すると、変動係数で上位にくるクラスタの潜在的テーマのいくつかは、阪神・淡路大震災と東日本大震災という 2 つの戦後の巨大災害を反映するものとなっていることがわかる．「震災」クラスタや「原子力」クラスタがそれである．また、1960 年代前半までは「台風」が自然災害においてメインのテーマであったが、その後は議論の比重が小さくなっている．これに対して「地震」はやや意外なことに 1960 年代前半まではそれほど大きなテーマとなっていなかったようである（以上図 2)．次に「省」や「機能」、「予算」といった行政に関わるテーマを反映したクラスタが上位に来ている（図 3)．「予算」クラスタについては、「公共事業」や「国庫補助」などの語が含まれているが、これらは 1960 年代前半までに特に頻出している．この時代には災害に関して、地方行政よりも国や国家予算に関わる問題として理解されていた可能性がある．「省」クラスタは「国土交通省」や「財務省」などの 2001 年の中央省庁再編により新たに登場した省の名前を反映している．興味深いのは、阪神淡路大震災後に水準を増した「機能」クラスタで、これには「センター」、「広域」、「担う」、「人材」、「バックアップ」、「ネットワーク」といった語が含まれており、災害に際して果たすべき行政や社会の機能についての期待の変化が捉えられているといえる．変動係数の上位でもう 1 つ注目すべきは、「被災」クラスタである（図 4)．これは、「被災」、「復興」、「支援」、「住民」、「生活」、「地元」、「自治体」といった語からなるが、このクラスタは 1995 年の阪神淡路大震災以前にはほとんど出現していない．つまり、阪神淡路大震災が被災に関わる集合意識についての決定的な変化のポイントであったということである．具体的には、被災が復興や支援の問題として、また住民の生活や地元の自治体の関わる問題として焦点化されるという認識の転換が生じたと考えられる．そしてその転換は東日本大震災時に再度、さらに巨大なテーマとして立ち現れることになる．

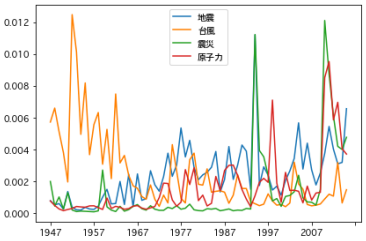


図 2　「震災」、「原子力」、「台風」クラスタの比率の時系列変化

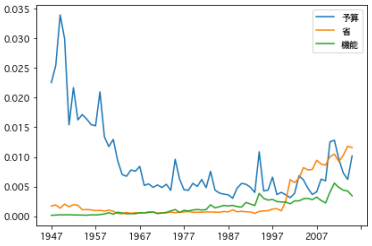


図 3　「予算」、「省」、「機能」クラスタの比率の時系列変化

表 2 高頻度のクラスタ

| クラスタ（上位五つの高頻度語） | 頻度 |
|---|---|
| 問題，聞く，なかった，事態，得る | 0.05697 |
| 安全，必要，対応，行う，検討 | 0.05352 |
| 出る，起こる，見る，起きる，わかる | 0.04721 |
| 関係，調査，具体，承知，段階， | 0.03652 |
| 被害，状況，発生，受ける，大きな | 0.03416 |
| 対策，措置，政府，防止，緊急 | 0.03145 |
| 努力，立場，行政，強い，終わる | 0.02008 |
| 法，法律，制度，改正，救済 | 0.01614 |
| 事故，事件，起こす，重大，事例 | 0.01492 |
| 報告，説明，内容，確認，資料 | 0.01392 |

表 3 変動係数上位のクラスタ

| クラスタ（上位五つの高頻度語） | 変動係数 |
|---|---|
| 或いは，又，申，行，かく | 2.28935 |
| 震災，津波，東日本大震災，大震災，神戸 | 1.89475 |
| 省，庁，是非，国土交通省，様々 | 1.29407 |
| 原子力，原発，福島，原子力発電所，稼働 | 1.22675 |
| 被災，復興，地，支援，避難 | 1.00702 |
| 災害，参る，きわめて，起る，方面 | 0.97331 |
| 台風，水害，冷害，激甚，雪 | 0.92089 |
| 機能，役割，センター，広域，担う | 0.85266 |
| 地震，予知，震度，圏，首都 | 0.77159 |
| 予算，復旧，負担，費，補助 | 0.73818 |

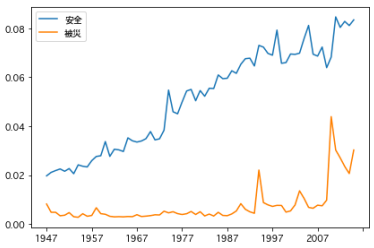


図 4 「被災」、「安全」クラスタの比率の時系列変化

以上は変動係数が上位のクラスタに着目して解釈を与えてきたが、変動係数による変化の測定には限界もある．まず、「震災」クラスタや「原子力」クラスタが 2 つの大震災という突発的出来事を強く反映していたように、変動係数は、平均からの爆発的かい離を強く反映する傾向がある．いいかえると、漸進的だが重要な変化は過小評価される可能性がある．また、変動係数は言及される頻度（比率）の平均が小さいクラスタほど上位にくる傾向があるため、この点でも解釈に注意が必要である．

クラスタ比率の変化を定性的に検討した結果、漸進的変化として目につくのは「安全」クラスタの変化である．図 4 にあるように、このクラスタの比率は、戦後一貫して漸進的に上昇している．このクラスタは多くの語からなるが、例えば「対応」、「検討」、「確保」、「防災」といった語が含まれる．このことは集合的認知において、自然災害が人々の「対応」や事前の「検討」を要すべき事象として、社会の防災対応によって扱われるべきリスクとして捉えられるようになったこと、つまりはリスク社会化の傾向として解釈できる．

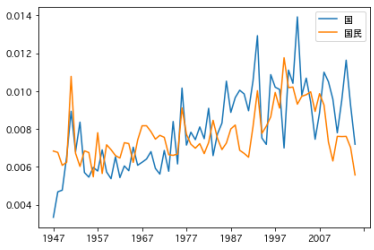


図 5 「国」、「国民」クラスタの比率の時系列変化

最後に Samuels の研究(Samuels 2013=2016)で着目されていた国の安全保障をめぐるナラティブと関連する単語クラスタとして、「国」と「国民」クラスタの変化を検討したい．まず、「国」クラスタを構成するのは、「世界」、「国際」、「国連」、「貢献」といった語である．このクラスタは 80 年代以降にやや水準を上げている．このことは、自然災害が世界の文脈に位置づけられて語られることが多くなったこと、そしておそらく、災害時の援助や支援などが国際貢献の一環として意識される傾向が強くなったことと関連している．「国民」クラスタは、「守る」、「保護」、「人命」、「国

家」、「公共」、「治安」、「秩序」などの語からなり、国民国家のナラティブと強く関連している．トレンドは微妙であるが、あえていえば、戦争直後を除いて、90 年代と 2000 年代前半にやや比率が大きくなっている．やや意外なことに東日本大震災時にはむしろ、それほど高い水準にはない．このデータからみると、東日本大震災では直接的な国民ナラティブはむしろ喚起されなかったということになる．

## ５．結論

本研究では、集合的認知のダイナミクスにおいてとくにその変化を帰納的に特定するための方法について提案した．すなわち、単語分散表現に基づくクラスタ頻度比率の変動係数を変化の指標とする方法である．これによって、質的な検討が不可能な大規模データにおいて、集合的認識の布置の変化、とくに大きく変化した潜在的テーマの特定が可能となった．

自然災害についての集合的認知のダイナミクスという主題についてもいくつかの実質的発見をすることができた．以下は、政治家の言説の変化から推定された集合的認知の変化であるため、その妥当性の検討をめぐっては今後、多くの努力が払われる必要があるが、現時点での仮説として列挙しておく．

1) 災害をめぐる集合的認知は、2 つの大震災のような巨大な出来事に対応して大きく変化した．
2) 戦後の初期には、災害は国家予算や公共事業との関連で認識されることが多かったのに対して、阪神大震災以後は、ネットワークやバックアップなどより分権的な仕方で対処すべきものと考えられるようになりつつある．
3) 被災した地元の復興や住民の生活への支援といった今日われわれが当然のように考える災害時の課題は、阪神淡路大震災以前はそれほど言及されていない．復興や支援への着目という認識の転換は 1995 年に生じた．
4) 自然災害が人々の対応すべき課題であり、安全の確保が社会の側の防災対応によって積極的に進められるべきだという意識が戦後を通じて徐々に大きくなっている．これはわれわれの社会のリスク社会化という認識を反映している．
5) 国民国家をめぐるナラティブは、1980 年代以降、災害との結びつきを強めた．ただし、東日本大震災においては、こうした国民ナラティブはそれほど動員されなかった．

上で述べたようにここで得られた結論は、かなり暫定的な仮説として捉えられるべきである．この仮説の妥当性を検討するためには、例えば、新聞記事等をはじめとする別のデータによる多角的検証が求められるだろう．

## 引用文献

Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003, "Latent Dirichlet Allocation." *J. Mach. Learn. Res.* 3: 993-1022.

Durkheim, E., 1985, *Les Règles de la méthode sociologique*（=1978, 宮島喬訳『社会学的方法の規準』岩波書店）

Elias, N., 1969, *Über den Prozeß der Zivilisation*, Suhrkamp(=1977,1978, 赤井慧爾他訳『文明化の過程』法 政 大 学 出 版 局)

Evans, J. A., and P. Aceves. 2016, "Machine Translation: Mining Text for Social Theory", *Annual Review of Sociology*, 42: 21-50.

Grimmer, J., and B. M. Stewart. 2013, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts", *Political Analysis*, 21: 267-97.

Klingenstein, S., Hitchcock, T. and DeDeo, S. 2014, "The Civilizing Process in London's Old Bailey." *Proceedings of the National Academy of Sciences* 111, 26: 9419-24.

Pennington, J., Socher, R. and Manning, C.D., 2014, "GloVe: Global Vectors for Word Representation." Paper presented at the Empirical Methods in Natural Language Processing (EMNLP).

Rule, A., Cointet, J.P. and Bearman, P. 2015, "Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014." *Proceedings of the National Academy of Sciences* 112, 35: 10837-44.

Samuels, R.J., 2013. *3.11: Disaster and change in Japan*. Cornell University Press (=2016, プレシ南日子ほか訳『3.11：震災は日本を変えたのか』英治出版)