# 地方政治コーパス構築における従来の成果と現在の課題

## - 政治・経済分野の応用研究に向けたパネルデータの構築 -

◯ [1] 木村 泰知　[2] 小林 暁雄　[3] 坂地 泰紀　[4] 内田 ゆず
[5] 高丸 圭一　[6] 乙武 北斗　[2] 吉田 光男　[7] 川浦 昭彦
[1] 小樽商科大学　[2] 豊橋技術科学大学　[3] 成蹊大学　[4] 北海学園大学
[5] 宇都宮共和大学　[6] 福岡大学　[7] 同志社大学

kimura@res.otaru-uc.ac.jp

## 1　はじめに

近年，地方分権や地方創生などの議論が盛んになっており，地方政治が注目を集めている．このため，地方議会における発言が書き記された地方議会会議録の分析も重要性を増している．幾つかの研究分野において，自治体がウェブ公開している地方議会会議録を対象とした調査・分析が行われている [1]．このような研究では，自治体によって個別に公開された議会会議録を逐一収集しており，調査・分析を始めるまでの準備に時間と労力がかかるという問題がある．

そこで，我々は，種々の分野における地方議会会議録を用いた全国規模の研究を推進することを目指して，平成 22 - 25 年度に，全国の自治体の地方議会会議録を収集・整理する手法を確立し，「地方議会会議録コーパス」の構築を進めた[1]．その結果，自然言語処理の分野において，本コーパスを利用した応用研究が成果をあげている [2][3][4][5]．また，社会言語学の分野においても，我々の構築したコーパスを利用したさまざま研究が進められている [6][7][8][9].

しかしながら，政治学・経済学等の分野における研究では，これまでに収集したコーパスをそのまま利用することが困難であることが明らかになった．

そこで，本稿では，従来の地方議会会議録コーパスの成果および課題を説明するとともに，政治学・経済学分野での利用を目指して，新たに進めている「地方政治コーパス」構築の目的について述べる．

## 2　従来の会議録収集の成果と課題

本章では，従来（平成 22〜25 年度）の会議録収集の目的と成果について述べる．従来の会議録収集の目的は，地方政治に関する研究の活性化・学際的応用を目指して，研究者が利用可能な地方議会会議録コーパスを全国規模で構築し，ウェブ上で提供することであった．全国を対象として収集した結果，20 の都道府県および 405 の市区町村（341 市，13 区，43 町，8 村）の会議録を集めた．収集した地方議会会議録の容量は，約 80GB である．

自然言語処理の分野では，地方議会会議録コーパスを利用した成果として情報抽出，アノテーション，政治情報システムに関する研究がある．[2][3] では，名詞句および節を対象として，議員の発言から政治問題を抽出する効果的な方法を明らかにした．政治問題の末尾には，「負担」「促進」のような特定の単語が頻繁に利用される特徴がある．そこで，それらの単語を利用することで「施設利用料の負担」「復旧費用の負担」「建築工事経費の負担」のような政治問題を獲得できることを確認した [2][2]．また，名詞句より長い範囲を抽出する場合には，文単位よりも，節単位で抽出するのが効率的であり，モダリティ表現の利用が効果があることを確認した [3]．さらに，筒井らは，地方議会会議録コーパスの構築および政治情報システム構築を目的としてアノテーションを行っている [4]．木村らは，地方議員マッチングシステムにおける能動的質問のための質問生成手法を構築した [5][3]．議員マッチングシステムでは，利用者に対して政治的問題の関心を明確化するための質問を行うことで，その利用者の考えに近いと思われる複数の議員の組み合わせを提示した．

社会言語学の分野では，地方議会会議録コーパスを利用した成果として整文，方言，オノマトペ等に関する研究がある．高丸らは，言語学的研究において会議録を利用するために，どの程度忠実に議会の発言を書き起こしているかを明らかにすることを目的として，

[1] http://local-politics.jp

[2] http://www.radiobots.link/NY-2011/show.cgi

[3] http://hokkaido-politics.net/

整文の調査を行っている [9]．オノマトペに関する研究では，地方議会会議録コーパスを用いて，現代の話しことばにおけるオノマトペの出現傾向を明らかにしている [8]．

以上のように，自然言語処理や社会言語学の分野では，従来の地方議会会議録コーパスが有効に利用されていることを確認した．

しかしながら，従来のコーパスは，政治・経済の分野で利用されなかった．その主な原因の一つに，収集する条件や期間が揃えられていないことがあげられる．つまり，収集条件を満たしていない部分的なデータでは，母集団から恣意的なサンプル選択が行われているのではないかという疑いを持たれる可能性がある．従来の地方議会会議録の収集では，全国の市町村[4]のうち，技術的に収集が容易な地方議会会議録のみを収集しており，収集する会議録も期間はまちまちであった．すなわち，条件を決めて「網羅的に収集する」という観点が欠けていた。

## 3　現在の会議録収集の状況と課題

本章では，前章で述べた課題を踏まえて，現在の会議録収集[5]の目的について説明する．現在の会議録収集の目的は，政治・経済分野の研究対象として有用な地方政治コーパスを構築することである．

経済学の分野では，欠損があるデータは研究対象に利用できず，ある条件を満たす全てのデータを集める必要がある．計量経済学の分野では，時系列データ，横断面データ，あるいは，それらを合わせたパネルデータが利用される [10]．

時系列データ (タイムシリーズデータ) は，ある現象の時間的な変化を，連続的に観測して得られた値であり，横断面データ (クロスセクションデータ) は，時間を一時点に固定して各地点で起こっているデータを記録した値である．パネルデータは，時系列データと横断面データを合わせたものであり，同一の対象を継続的に記録したデータであることから，より多くの変数を同時に織り込んだ分析が可能となる．

地方政治コーパスは，統一地方選挙の任期期間として平成 23 年度 (平成 23 年 4 月) から平成 26 年度 (平成 27 年 3 月) までの 4 年間の「時系列データ」と 47 都道府県議会会議録の「横断面データ」を合わせたパネルデータを構築していることになる．図 1 に本研究におけるパネルデータのイメージを示す．

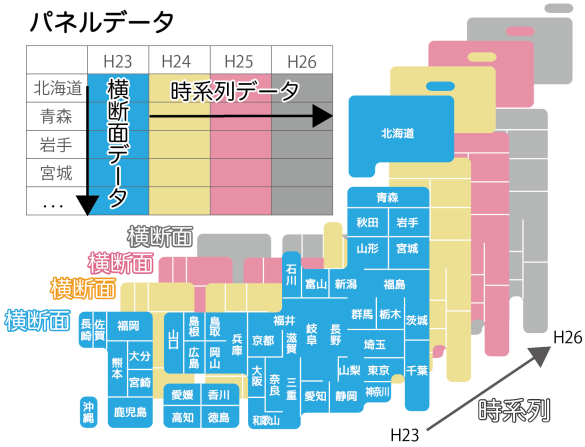


図 1: 本研究におけるパネルデータのイメージ

計量経済学で利用するパネルデータは「数値」のデータセットであることから，収集した会議録テキストを「文字数」「単語の出現頻度」などへ変換する必要がある．また,「自治体ごとの文字数」「発言者ごとの文字数」などに分類して表示することが求められる．このような数値を計算するためには，異なる形式で記述された会議録を統一したフォーマットで管理する必要があり，自治体や発言者などの項目へ分けられなければならない．そこで，我々は，下記のデータベースのスキーマを作成し，このスキーマに正しくデータを格納することを目指している．

1. 自治体名
2. 会議の回
3. 会議の号
4. 会議の開催月
5. 会議の開催日
6. 会議の開催日数
7. 会議名 ... 開催回，号などを含む会議の種別
8. 発言者名 ... 役職を含む発言者名
9. 発言者の役職
10. 発言内容 ... 発言は句点・改行で区切る
11. 発言以外の文

現在までに，47 都道府県の地方議会会議録の 4 年間分を漏れなく収集し，おおまかなルールに従い，データベースに格納した．表 1 は，データベースに基づいて，発言者数，総文字数を計算した 47 都道府県議会会議録の統計データである．

しかしながら，現状のデータベースでは，“発言者名”，“発言者の役職”，“発言内容”，“発言以外の文”

[4]平成 22 年の時点では 1,727 市町村であったが，平成 26 年 4 月の時点では 1,718 市町村となっている．

[5]科学研究費補助金・基盤研究 (B) 「議論の背景・過程・結果を関連づける地方政治コーパスの構築とその学際的応用」

の 4 つのスキーマにノイズが含まれている．つまり，「発言者の同定」と「発言と非発言の切り分け」という課題がある．上記の 2 つの課題に共通している点は「自治体による違い」と「サービス提供会社による違い」である．会議録は，規則性の高い記述であるが，サービス提供会社や自治体によって記述方法が異なる．表 1 のサービスの列からわかるように，4 つのサービス提供会社が，42 都道府県に会議録検索システムを提供しているが，同じサービスを利用しても異なる記述パターンが存在する．4 つのサービス会社を下記に示す．

1. Discuss Net Premium（以降，discuss と略記）株式会社議事録発行センター
2. VOICES 株式会社フューチャーイン
3. DB-Search 株式会社大和速記情報センター
4. Sophia 神戸綜合速記株式会社

また，サービスが「その他」になっている自治体については，全く別の収集方法とデータベースへの格納方法を考える必要がある．下記では，2 つの課題についての具体例を示すとともに，解決方法についても検討する．

第一の課題は「発言者の同定」であり,「発言者の曖昧性解消」「名寄せ処理」とも呼ぶことができる．現時点では，おおまかなルールとして，正規表現による規則を用いて発言者を抽出しているが，正しく抽出できていない．抽出が困難な理由としては，上記の自治体とサービス会社の違いに加えて，同一人物の異なる表記，政治家以外の発言者名，発言者名の記述ミスなどがある．同一人物の異なる表記名は「加藤鉱一議員」「加藤 (鉱) 委員」のような例がある．政治家以外の発言者名は，首長の代理として説明する自治体の職員名である.「質問者」は議員であるのに対して「答弁者」は首長の代理として「副知事」「総務部長」などが登壇することがあり，役職と一緒に名前が記述されている．発言者の記述ミスは「遠藤達雄」「遠藤達雄」「遠藤達夫」のような例がある．表 1 から明らかなように，発言者数が不自然に多い自治体がある[6]．この課題については，予め議員リストを手作業で作成し，議員か議員ではないかを区別した後で，議員の同定を行うことを検討している．

第二の課題は「発言と非発言の切り分け」である．会議録には発言以外にも「目次」「状況・様子　例. (拍手) 〔…君登壇〕」「時刻表現　例. 午後 1 時 2 分 開会・開議」「発言者名・出席者リスト」「参考資料」など

[6]ちなみに，東京都議会の議員数は，最も多く 127 人である．

記

議案第12号　人事委員会の委員の選任に関し同意を求めることについて
議案第13号　公安委員会の委員の任命に関し同意を求めることについて
（議案及び報告の登載省略）
○議長（佐々木一榮君）　次に、発議案１件が提出になっております。お手元に配付いたしてありますから、御了承願います。
発議案第１号
平成23年６月30日

目次
状況・様子
発言
参考資料

図 2: **発言と非発言の切り分けが難しい例**

どが含まれている．現在の「発言内容」は,「発言」と「非発言」を切り分けられていない [11]．図 2 に発言と非発言の切り分けが難しい例を示す．この課題については，切り分ける「手がかり表現」をみつけることを考えている．手がかり表現として “ー”の連続の例を示す．

発言と非発言を “ー”の連続で区切っている例

諸般の報告は、お手元に配付してあります議長報告のとおりでありますので、朗読を省略いたします。
ーーーーーーーーーーーーーーーーーーーー
議　長　報　告　（朗読省略）

## 4　おわりに

本稿では，従来の地方議会会議録コーパスの成果および課題を説明にするとともに，現在行っているコーパス構築の目的および現状について述べた．従来のコーパスは，20 の都道府県および 405 の市区町村（341 市，13 区，43 町，8 村）の会議録を収集したものであり，自然言語分野の情報抽出，アノテーションや社会言学の分野の整文，オノマトペ，方言で研究の成果をあげている．しかしながら，従来のコーパスは，収集が容易な自治体の会議録のみを収集しており，収集が困難な自治体の会議録を対象外としたことから，経済学の分野で利用されなかった．そこで，現在のコーパス構築は，計量経済学の分野で利用可能な地方議会会議録のパネルデータ作成することを目的としており，47 都道府県議会会議録を対象に期間を定めて収集することが重要であることを確認した．また，網羅的に収集する以外にも，データを正確に整理することが重要であり,「発言者の同定」と「発言と非発言の切り分け」の 2 つ課題があることを述べた．

本研究の学術的な特色は，学際的な応用研究を目指し，複数分野で利用できるコーパスを構築する点であり，本コーパスを政治学・経済学・言語学・情報工学の分野で利用してもらうために，異なる専門分野の研究者と連携して作成を進めることは，大きな意義があると考えている．

表 1: 47 都道府県議会議録の概要

| | 都道府県 | サービス | 発言者数 | 総文字数 |
|---|---|---|---|---|
| 1 | 北海道 | VOICES | 202 | 6,690,771 |
| 2 | 青森県 | DB-Search | 230 | 6,075,989 |
| 3 | 岩手 | その他 | 115 | 5,265,209 |
| 4 | 宮城県 | discuss | 107 | 7,180,505 |
| 5 | 秋田県 | その他 | 61 | 3,563,875 |
| 6 | 山形県 | discuss | 103 | 2,432,496 |
| 7 | 福島県 | discuss | 595 | 4,338,535 |
| 8 | 茨城県 | DB-Search | 388 | 4,378,426 |
| 9 | 栃木県 | VOICES | 150 | 2,563,498 |
| 10 | 群馬県 | VOICES | 173 | 5,713,893 |
| 11 | 埼玉県 | discuss | 1,081 | 6,281,006 |
| 12 | 千葉県 | DB-Search | 173 | 3,392,658 |
| 13 | 東京都 | DB-Search | 409 | 5,746,806 |
| 14 | 神奈川県 | discuss | 156 | 5,896,670 |
| 15 | 新潟県 | discuss | 899 | 15,889,237 |
| 16 | 富山県 | DB-Search | 125 | 4,694,955 |
| 17 | 石川県 | VOICES | 158 | 4,413,772 |
| 18 | 福井県 | DB-Search | 149 | 4,568,260 |
| 19 | 山梨県 | DB-Search | 164 | 4,274,363 |
| 20 | 長野県 | VOICES | 524 | 19,389,877 |
| 21 | 岐阜県 | discuss | 494 | 6,490,646 |
| 22 | 静岡県 | その他 | 241 | 5,376,236 |
| 23 | 愛知県 | DB-Search | 305 | 5,881,919 |
| 24 | 三重県 | discuss | 116 | 5,554,262 |
| 25 | 滋賀県 | VOICES | 250 | 8,626,244 |
| 26 | 京都府 | DB-Search | 766 | 14,714,871 |
| 27 | 大阪府 | discuss | 536 | 17,318,822 |
| 28 | 兵庫県 | Sophia | 155 | 3,892,397 |
| 29 | 奈良県 | discuss | 108 | 4,349,724 |
| 30 | 和歌山県 | その他 | 102 | 3,473,289 |
| 31 | 鳥取県 | DB-Search | 169 | 10,844,070 |
| 32 | 島根県 | DB-Search | 123 | 6,010,462 |
| 33 | 岡山県 | discuss | 104 | 6,336,143 |
| 34 | 広島県 | DB-Search | 171 | 3,357,629 |
| 35 | 山口県 | discuss | 92 | 5,260,382 |
| 36 | 徳島県 | discuss | 86 | 3,812,198 |
| 37 | 香川県 | DB-Search | 389 | 8,749,248 |
| 38 | 愛媛県 | Sophia | 204 | 4,198,966 |
| 39 | 高知県 | discuss | 93 | 6,526,952 |
| 40 | 福岡県 | DB-Search | 178 | 4,948,309 |
| 41 | 佐賀県 | DB-Search | 126 | 5,740,329 |
| 42 | 長崎県 | discuss | 675 | 12,804,182 |
| 43 | 熊本県 | discuss | 117 | 4,837,486 |
| 44 | 大分県 | discuss | 175 | 4,593,789 |
| 45 | 宮崎県 | discuss | 98 | 7,416,181 |
| 46 | 鹿児島県 | DB-Search | 152 | 7,266,842 |
| 47 | 沖縄県 | その他 | 153 | 7,553,407 |

## 謝辞

本研究は JSPS 科研費 JP16H02912 の助成を受けたものです.

## 参考文献

[1] 増田正, 地方議会の会議録に関するテキストマイニング分析 - 高崎市議会を事例として -『地域政策研究』(高崎経済大学地域政策学会),Vol.15 No.1, pp.17-31, 2012.

[2] 木村泰知, 関根聡, 主辞に基づく政治問題抽出手法, 人工知能学会論文誌, Vol.28, No.4, pp.370-378, 2013.

[3] 葦原史敏, 木村泰知, 荒木健治, 地方議会会議録における節単位による議員の要望抽出, 電子情報通信学会論文誌, Vol.J98-D, No.11, pp.1390-1401, 2015.

[4] 筒井貴士, 我満拓弥, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則, 地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案. 自然言語処理, Vol.21, No.2, pp.125-156,2014.

[5] 木村泰知, 渋木英潔, 高丸圭一, 乙武北斗, 小林哲郎, 森辰則, 地方議員マッチングシステムにおける能動的質問のための質問生成手法, 人工知能学会論文誌, Vol.26, No.5, pp. 580-593, 2011.

[6] 二階堂整, 川瀬卓, 高丸圭一, 田附敏尚, 松田謙次郎, 地方議会会議録による方言研究の可能性, 日本方言研究会第 99 回研究発表会発表原稿集, 2014.

[7] 井上史雄 ,「去った〇日」『ことばの散歩道』明治書院，pp.154-155, 2013.

[8] 高丸圭一, 内田ゆず, 乙武北斗, 木村泰知, 地方議会会議録コーパスにおけるオノマトペ -出現傾向と語義の分析-, 人工知能学会論文誌, Vol.30, No.1, pp.306-318, 2015.

[9] 高丸圭一, 木村泰知, 栃木県の地方議会会議録における整文についての基礎分析― 本会議のウェブ配信と会議録との比較―, 都市経済研究年報, (10) pp.74-86, 2010.

[10] 北村行伸, パネルデータの意義とその活用, 日本労働研究雑誌 551, pp.6-16, 2006.

[11] Yasutomo Kimura et al., Creating Japanese Political Corpus from Local Assembly Minutes of 47 Prefectures, Coling 2016, The 12th Workshop on Asian Language Resources, pp.78-85, 2016.