地方議会会議録コーパスと 地方議会会議録を用いた学術研究の現状†

高丸 圭一*

1. はじめに

地方議会では住民の中から直接選挙で選出された議員や首長が、予算や条例の制定等についての審議を行っている。議会における議員、首長および職員の発言はすべて文字に書き起こされ、地方自治法 123 条に基づき会議録として保存される。従来は製本され役所に保管されるだけのものであったが、情報公開の進展およびインターネットの普及により、地方議会会議録をウェブに公開する自治体が増加している。

地方議会会議録は各自治体によって個別に公開されている。複数の自治体の議会会議録を横断的に利活用するために、ウェブ上に公開された地方議会会議録を収集し、コーパスとして学術研究に利用することを目指した研究プロジェクトが進められている。

住民が積極的に地方自治へ参画するためには、議会を傍聴したり、会議録を閲覧し、議会における議論の内容を知る必要がある。議会での議論を知ることが、投票の判断基準としても有用であろう。しかし、都道府県議会会議録のうち、もっとも文字数の少ない山形県議会でも、本会議 4 年分で 200 万文字以上あり、住民がすべてを読むことは現実的ではない。そこで、自然言語処理、人工知能技術の支援による要約や可視化といった研究の進展が期待される。さらに、首長や議員による発言の記録は、地方自治におけるファクトチェックの一次情報としての利用が期待される。

議会での発言内容は地方政治研究や選挙研究にも有用なデータであることはいうまでもない。また、議会で言及されるさまざまな問題に関連する種々の研究分野において分析の対象となり得ると考えられる。発言の記録を言語資料（談話資料）として捉え、言語学的な研究に活用することも可能である。

表 1 に自治体数、地方議会会議録のウェブ公開率、議員定数、議員実数（男女別）をそれぞれ「都道府県」「市区」「町村」別に集計した値を示す。このうち、会議録のウェブ公開率は小樽商科大学木村研究室による調査の結果であり、2017 年 3 月現在のものである。議員数は総務省「地方公共団体の議会の議員及び長の所属党派別人員調等（平成 29 年 12 月 31 日現在）」

表 1 全国の自治体数、会議録のウェブ公開率、議員定数、および、議員数

	都道府県	市区	町村	合計
自治体数	47	814	927	1,788
会議録のウェブ公開率	100%	100%	73%	86%
議員定数	2,687	19,399	11,166	33,252
男性議員数（実数）	2,350	16,248	9,906	28,504
女性議員数（実数）	264	2,855	1,092	4,211

に基づくものである。

会議録の公開状況については、都道府県レベルでは 2007 年 2 月の時点ですべてウェブ公開されていることが指摘されている [1]。その 10 年後にあたる 2017 年 3 月時点の調査では、都道府県議会に加えてすべての市・区議会の会議録がウェブ公開されている。一方、町・村議会では、全 927 町村のうち 246 町村でウェブサイトにおける会議録の公開が確認されず、公開率は 73% である。会議録を公開している自治体においても、公開範囲が本会議（定例会・臨時会）にとどまっている自治体も多く存在し、委員会の会議録については、非公開であったり、概要のみの公開である場合がある。

全国 1,788 地方自治体における地方議会議員の総数は 33,165 人（欠員 349）である。会議録を公開していない町村の割合で町村議員数を按分して差し引いても、ウェブに公開された地方議会会議録には約 3 万人の議員の発言が掲載されている計算になる。議会では、議員に加え、首長や理事者（役所の職員）、参考人等も発言を行うため、公開された会議録に収録される発言の発言者総数はさらに多い。

表 1 に示した通り、議員に占める女性の割合は低く、全体で約 13% である。自治体区分ごとに見ると、市区は 14% であり、都道府県、町村（いずれも 10%）と比べやや高い。地方議会議員の年齢構成および出身地については、竹安 [2] の 2002 年時点での地方議会議員に対するアンケート調査がある。[2] の表 3-1 に基づいて計算すると、25 歳以上 39 歳未満が 2%（有効回答 16,842 のうち 339 人）、40 歳以上 59 歳未満が 47%（同 7,787 人）、60 歳以上が 52%（同 8,716 人）である。また、[2] の表 3-2 に基づいて計算すると、出生地と現在居住する都道府県が等しい議員は 88%（有効回答 16,844 のうち 14,740 人）であった。このうちの 86%（同 12,709 人、全体の 75%）の議員は出生地と現在居住する市区町村が等しかった。地方議会議員は高年層の男性が多く、自分の生まれ育ったまちにおいて議員

† Corpus of Local Assembly Minutes and Recent Academic Research using Local Assembly Minutes
Keiichi TAKAMARU

* 宇都宮共和国大学シティライフ学部
Faculty of City Life Studies, Utsunomiya Kyowa University

を務めているということは直感的に明白であるが、このことがこれらのデータにより裏付けられる。

先に述べたように地方議会会議録は、議会における発言をすべて記録することを目的に作成される。しかし、可読性を考慮して言い直しや単純な言い間違い、助詞の誤用、冗長な言い回しなどそのまま文字化すると読みづらい表現は修正されることが知られている。この作業は整文と呼ばれる [3]。整文については、高丸ら [4, 5] において具体的な分析を行った。

図 1 に [6] を引用し、整文の具体例を示す。議会の映像記録と会議録を照合して、会議録で削除されている発言には打消線を付し、発言にはなく会議録作成時に挿入された文字は [ ] 内に示した。また、修正箇所にはこれらを併記した。

[5] の分析では、整文により追加・削除・置換された文字数は会議録 100 文字あたり平均約 11 文字であった。このうち、「あの一」等のフィラーの削除が 53.4 % を占める。この他の整文の具体例として、省略した名称や単位の補完、言い間違い、言い直し、繰り返し等の削除、くだけた口語表現の修正などが見られる。

議論の内容を把握する目的で会議録を利用する分には整文は大きな影響を与えないと考えられるが、フィラーの頻度や「ら」抜きことばの使用分布といった整文が影響を及ぼす言語現象の研究を行う際には注意が必要である。

本稿では、筆者が参加する地方議会会議録コーパスプロジェクト¹が構築を進めている地方議会会議録コーパスの概要について述べる。次に、地方議会会議録を用いた学術研究の現状について具体的な事例を挙げて概観する。

2. 地方議会会議録コーパス

2.1 第 I 期コーパス：大規模な収集を目指して

2010 年から斎藤ら [7]、菅原ら [8] により、全国の自治体の地方議会会議録の収集・整理が開始された。この研究における調査の結果、地方自治体向け議会会議録検索ウェブシステムは主に 4 社から提供されていることが分かっている。都道府県議会における会議録検索ウェブシステムの採用状況を表 2 に示す。42 都道府県は大手 4 社のシステムを採用しているが、独自のシステムを採用している県が 5 県存在している。また、このほか市町村レベルでは、検索システムを採用せずに、PDF ファイルや HTML ファイル、テキストファイルによって会議録を公開している自治体も存在する。

第 I 期コーパスでは、大手 4 社のシステムを採用している市町村議会を対象に、会議録検索ウェブシステムごとにリンク解析手法やパラメータ生成手法を用いて、会議録データを収集した。

地方議会会議録コーパスでは、会議録に記録された発言を原則として句点によって区切られた文に相当する単位に分割し、1 文を 1 レコードとして、自治体コード、年月日、会議名、発言者などのメタ情報を付与して記録している。これにより、コーパスからいつ、どこで、誰が、何を発言したかを明らかにすることができる。

プログラムでは財政 [調整] 基金を八十億円を見込んで、各種基金の活用を六十四 [億円] として、~~まー~~その六十四 [億円] を基盤として~~まーあの一~~地域福祉基金と土地開発基金、これは十九億 [円]、三十九億 [円] ~~ということ~~でなりまして、それで百二十億円になって、(以下略)

(2010 年 2 月 24 日栃木県議会)

えーおやじに怒られるにしても、「はい」って返事すれば、おやじも怒られ~~ん~~[てる] のに何で返事す [る] んだ~~ん~~べな~~って~~いうよう[というふう] な感覚的な、~~まー~~あるぶん[意味]ではお怒られるとこが[ことも] 少し和らぐかなというようなの[こと] もあった~~ん~~[の] かもしれませんし、わかりませんけれども、(以下略)

(2010 年 6 月 7 日那須塩原市議会)

そうしたら、おめほうに 2 人も市 [議] 会議員がいるんだ~~は~~んで[のだから]、そういう話はあーうちさ[に] 出さねー[ない] で市 [議] 会議員さ[に] しゃべろと。

(2013 年 3 月 6 日 弘前市議会)

図 1 会議録整文の例

表 2 都道府県議会における会議録検索システムの採用状況

システム名	採用都道府県数	販売元
discuss	18	NTT-AT
DB-Search	16	大和速記情報センター
VOICES	6	フューチャーイン
Sophia	2	神戸総合速記
その他	5	(不明)

第 I 期では、423 自治体の議会会議録を収録した地方議会会議録コーパス (約 1 億 3400 万文、約 80GB) を構築した (図 2)。このコーパスに収録された各会議録の期間の分布は図 3 に示すとおりである。

乙武ら [9] によって、このコーパスの単語 n-gram データや検索用ウェブ UI が構築され、学術研究利用に限定した公開を行った。これらを利用し、自然言語処理分野における情報抽出の研究や、言語学分野における言語使用や言語変異の実証的研究が進められてきた。地方議会会議録を利用した研究の成果については 3 章にまとめる。

第 I 期コーパスの取り組みに続くように、2016 年頃からビッグデータの利活用が盛んになり、「議事ボックス」²や「chiholog 地方議会議事録横断検索」³など、地方議会会議録を横断検索できる一般向けウェブサービスが登場した。これらのサービスではスクレイピング技術により大規模なデータ収集を実現しているものの、発言をレコードに分けて発言者情報を付与するといったデータ整形は行われていない。

2 <http://giji.rocks/> (※ 2018 年に閉鎖)

3 <https://chiholog.net/>

1 <http://local-politics.jp/>

図2 第Ⅰ期コーパスのデータ量

2.2 第Ⅱ期コーパス：統制が取れたコーパスを目指して

2.2.1 収集

種々の分野で会議録を研究対象とするとき、各種の統計量（たとえば、都道府県 P での 1 会議あたりの平均発言者数）を算出したり、ある一定の条件下での発言（たとえば、単語 w を含む議員 A の発言）を抽出したりするという利用方法が考えられる。これを実現するためには、会議録を単に大規模に収集したコーパスではなく、収集範囲および収集期間の統制がとれたコーパスサブセットが必要である。そこで、第Ⅱ期コーパスとして、統制が取れたコーパスの構築に取り組んでいる。最初のサブセットとして「47pref_2011-2015」を構築した。このサブセットの収集範囲は全国 47 都道府県議会の本会議（定例会および臨時会）であり、収集期間は 2011 年 4 月から 2015 年 3 月までの 4 年間である。2011 年および 2015 年の 4 月には統一地方選挙が実施されている。統一地方選挙の時期以外に議員選挙を行う議会も存在するが、収集範囲をこの 4 年間に設定することにより、サブセット内での議員の入れ替わりがもっとも少なくなるという利点がある。

表 2 に示すように、都道府県議会の会議録全文検索ウェブサイトは、主要 4 社が 89% のシェアを持つ。そこでまず、この 4 システムに対応したクローラを作成する。ただし、自治体ごとに検索システムや表示部分をカスタマイズされているため、個別対応するようにプログラムを調整したり、部分的に人手による作業を挟みながら収集作業を進める。さらに、全都道府県の会議録を網羅するために、主要 4 社以外のシステムによって会議録を公開している 5 県については、1 自治体ごとにクローラを作成して、収集を行っている。この作業によって収集された全国 47 の都道府県の本会議会議録は約 400 万文、1.5GB である。

2.2.2 データ整理の流れ

統制されたコーパスを構築するため、会議録を統一されたデータ形式に整理しなおす必要がある。第Ⅱ期コーパスでは、

図3 第Ⅰ期コーパスの収集期間

以下の方針で進めている。

- 発言は 1 文単位に分割し、非発言と明確に区別すること
- 会議録に含まれる情報を過不足なく抽出すること
- 議員に対して、選挙時に公表されている属性情報を付与すること

以上の方針を踏まえて、都道府県議会会議録を「発言テーブル」と「発言者テーブル」で構成される関係データベースの形で整理する。発言テーブルは発言文 1 文を 1 レコードとし、図 4 に示す 17 のフィールドをもつ。

発言者テーブルは発言者 1 名を 1 レコードとし、図 5 に示す 9 項目のフィールドをもつ。「発言者 ID」に基づいて、発言テーブルから発言者テーブルを参照する。

議員の年齢や性別、対象行政区（選出選挙区）が分かると、発言内容の分析や言語的特徴の分析にとって有益であると考えられる。しかし、これらの情報は会議録中には記載されていないため、「4. 氏名」～「8. 性別」は選挙の立候補者情報がまとめられたサイト（選挙情報サイト、例えば「選挙ドットコム」や「政治山」）から情報を収集し、格納する。なお、発言者が議員である場合、「9. 職名」は空欄となる。一方、発言者が議員ではない場合、「4. 氏名」～「8. 性別」は空欄とし、職名に対して発言者 ID を付与する。

会議録は従来印刷物として作成していたものであるため、Web 上で公開されている会議録も、紙に出力した時の体裁に

表3 会議録における発言者名表記の例

	発言者パターン	例
北海道	○〔〔役職〕〔姓名〕君〕〔全角空白〕〔本文〕 ◆〔〔議席番号〕番〔姓名〕君〕〔全角空白〕〔本文〕 ◎〔〔役職〕〔姓名〕君〕〔全角空白〕〔本文〕	○（議長石井孝一君） 御異議なしと… ◆（14 番堀井学君） （登壇・拍手）… ◎（知事高橋はるみ君） （登壇）ただいま…
岩手県	○〔〔役職〕〔姓名〕君〕〔全角空白〕〔本文〕 ○〔〔議席番号〕番〔姓名〕君〕〔全角空白〕〔本文〕 ○〔姓〕〔名〕〔〔役職〕〕〔全角空白〕〔本文〕 ○〔姓名〕〔〔役職〕（続）〕〔全角空白〕〔本文〕	○議長（佐々木博君） 次に、石川… ○41 番（伊藤勢至君） それぞれ御答弁を… ○佐々木大和年長委員 御異議なしと… ○千葉伝委員（続） 今、お話を伺い…
長野県	○〔〔役職〕〔姓名〕〕〔全角空白〕君〕〔本文〕 ◆〔〔議席番号〕番〔姓名〕〕〔全角空白〕君〕〔本文〕 ◎〔〔役職〕〔姓名〕〕〔全角空白〕君〕〔本文〕	○副議長（竹内久幸 君）休憩前に… ◆6 番（清水純子 君）県民クラブ… ◎環境部長（荒井英彦 君）地下水の…

準じたフォーマットになっていることが多い。また、公開されたデータをさらに別の用途で用いることは考慮していないため、データとしての構造化は行われていない。さらに、会議録は自治体ごとに独自に作成されるものであるため、各自治体でフォーマットが異なる。したがって、会議録を収集し、発言テーブルの各フィールドに値を格納する作業を行うには、それぞれの自治体における会議録のフォーマットを逐一確認する必要がある。完全に自動化することはできない。特に、発言テーブルの「11. 発言者 ID」～「15. 発言以外の文」を完成させるためには、「発言の抽出」「発言者の抽出」「発言者の名寄せ」を適切に行う必要がある。

以下に、発言者に関するフィールドを埋める際に必要となり、発言者の抽出および発言者の名寄せの処理について述べる。

2.2.3 発言者の抽出

会議録から発言者の氏名と職名を抽出し、全発言文の発言者を同定し、発言テーブルの「12. 発言者名」および「13. 発言者の役職」に値を格納する。発言者の氏名と職名の抽出処理は、発言者を示す記号（多くの場合「○」「◎」「◆」）を手がかりに、正規表現を利用して行う。表3に正規表現の例を挙げる。

表3に示すように、発言者名の表記方法は統一されておらず、自治体ごとにパターンが異なる。このため、それぞれの自治体のフォーマットに合わせた正規表現を用意する。会議録作成者のミスや、書式の変更により、発言者パターンと一致しない記述がある（たとえば、半角スペースと全角スペースが混在しているなど）ため、目視によって確認を行い、個別に手作業で対応する必要がある。

2.2.4 発言者の名寄せ

各発言者に固有の「発言者 ID」を与えるため、会議録から抽出した発言者を対象として、同一人物を特定する。ここでの「名寄せ」作業には2つの側面がある。

1 つ目は、会議録内での同一人物の特定である。会議録から抽出した発言者は、同一人物でも複数の表記をもつことがある。例えば、「梅澤佳一自然再生・循環社会対策特別委員長」「梅澤佳一議会運営委員長」「梅澤佳一委員」「五十番（梅澤佳一議員）」「五十六番（梅澤佳一議員）」は、いずれも同一人物の「梅澤佳一」氏を示している。このような場合、発言者の氏名に付加された敬称や肩書、議席番号等を正規化する処理を行う。

1. 発言 ID 主キー、発言文ごとに振られた固有 ID
2. 自治体名 例：「北海道」「青森県」「宮城県」
3. 回 定例会の回数（第〇回定例会）
4. 号 定例会の号数（〇日目）
5. 年 開催年（西暦）
6. 月 開催月
7. 日 開催日
8. 開催期間 日数
9. 会議名表層 例：「平成24年 2月 定例会（第335回）-02月27日-04号」
10. 発言名表層 会議録に記載された発言者名
11. 発言者 ID 発言者テーブルを参照する外部キー
12. 発言者の氏名 「発言者名表層」から抽出した発言者の氏名
13. 発言者の役職 「発言者名表層」から抽出した発言者の役職
14. 発言文 会議録に記録された発言文
15. 発言以外の文 例：「〔…君登壇〕」「-----」
16. 取得元の URL
17. HTML ファイル名

図4 発言テーブルのフィールド名

1. 発言者 ID 主キー、発言者ごとに振られた固有 ID
2. 都道府県名 例：「北海道」「青森県」「宮城県」
3. 自治体名 例：「宮城県」
4. 氏名 例：「畠山 和純」
5. フリガナ 例：「ハタケヤマ カズヨシ」
6. 対象行政区 例：「気仙沼市」
7. 生年 投票日の西暦から投票日時点の年齢を減じて算出
8. 性別 「男」または「女」
9. 職名 例：「議会事務局議事課長」「知事」「会計管理者」「教育長」

図5 発言者テーブルのフィールド名

2 つ目は、会議録と選挙情報サイトとの間での人物の同定である。発言者テーブルの「4. 氏名」は、選挙情報サイトにおける表記を採用しているが、この表記が会議録中の表記と異なることがある。

表4に、議員名の表記パターンと選挙情報サイトとの一致の結果を示す。2,793人の議員に対して、会議録中に5,229種類の表記が見られた。このうち、87%にあたる4,563の表記は選挙

柳ヶ瀬 裕文 やながせ ひろふみ	
	選挙 東京都議会議員選挙 大田区選挙区
	投票日 2017年7月2日
	氏名 柳ヶ瀬 裕文 やながせ ひろふみ
	得票数 21460 当
 	年齢 42
	性別 男性
	党派 維新
	新旧 現
主な肩書き 都議会議員2期	
ウェブサイト http://www.yanagase.org	

図6 選挙の際に公開される情報の例

<https://seiji.yama.jp/area/card/14040/9gG415/M?S=qenel0shoit>

○副議長（小磯善彦君） 十番やながせ裕文君。 〔十番やながせ裕文君登壇〕 〔副議長退席、議長着席〕	121
○十番（やながせ裕文君） まず、豊洲新市場の問題について伺います。 一つ目は、盛り土についてであります。 豊洲での土壌汚染対策として、専門家会議は四・五メートルの盛り土を提言し、これが実行されなかったわけでありませうけれども、さまざまな資料を読み込んで、四・五メートル必要だという数字の出どころがよくわかりません。	122

図7 会議録に記録される発言者名の例

東京都議会 平成28年第3回定例会 第13号

情報サイトと完全一致していたが、13%にあたる666の表記は選挙情報サイトと一致しなかった。この不一致の内訳を、表5に示す。不一致の3分の2は漢字表記とかな表記の差異である。氏名の一部あるいは全部をひらがなやカタカナで表記する議員が多く見られる。これは「通称」として公職選挙法で認められており、選挙情報サイトにおいても、会議録においても通称が使用されることがある。

選挙情報サイトで戸籍名が使用され、会議録で通称が使用される例を図6、7に示す。これとは逆に、会議録中では戸籍名が使用され、選挙情報サイトでは通称で表記されている例もある。例えば、「瑞慶覧功」議員は会議録には左記のと通りの戸籍名で表記されているが、選挙情報サイトには「ズケラン 功」と表記されている。字面が一致しない文字列は機械では同一人物と判定されないため、漢字の読みを推定するなどしてマッチングを行う必要がある。

このほか、「崎」や「高」「渡辺」の「辺」などの異体字・旧字体の不一致や会議録における入力誤りが多く見られる。芸名の使用や選挙情報サイトの誤りも少数ながら存在する。なお、複合的な要因により不一致が生じている例があるため、内訳の和は不一致表記数よりも多い。

2.2.5 都道府県議会会議録検索システム「ぎ〜みる」の公開

構築したコーパスを学術研究だけで使用するのではなく、一般の市民が関心を持ったことばを検索できるようにすることで、地方政治への関心をさらに高め、研究成果の社会還元を行うことができる。このため、収集した会議録の全文検索

表4 議員表記のパターン数

議員数	表記パターン	完全一致	不一致
2,793	5,229	4,563	666

表5 表記パターン不一致の内訳

かな表記	芸名	異体字	会議録の誤り	選挙情報の誤り
447	3	190	190	1

図8 「ぎ〜みる」のシステムアウトライン

および検索結果の可視化が可能なWebシステムの実装を進め、第II期コーパスを「都道府県議会会議録検索システム」(<http://local-politics.jp/47pref/>)として一般公開した。本節では乙武ら[10, 11]に基づき、2018年2月に一般公開したシステムの概要について述べる。

このシステムは47都道府県議会の本会議会議録を検索・視覚化するウェブアプリケーションである。ユーザは一般的なウェブブラウザを用いて、会議録を検索し、結果を視覚化することができる。システムのアウトラインを図8に示す。システムが検索クエリを受け取ると、クエリは全文検索エンジンに送られる。検索エンジンから検索結果を受け取った後、システムはユーザのリクエストに応じて検索結果を可視化して出力する。全文検索エンジンにはElasticsearchを用いた。

一般公開システムのトップページの様子を図9に示す。はじめに訪れたユーザが試用しやすいように、検索文字列に「オリンピック」をデフォルト値として入力している。右側の「検索」ボタン、または、下部のアイコンをクリックすることにより、検索を開始することができる。また、上部には、「発言検索」「マップ検索」「クロス表検索」「時系列グラフ」「統計」の5つのメニューがあり、ここから本システムが有するすべての機能へアクセスすることができる。「発言検索」メニューには「発言検索」「KWIC」「TF・IDFキーワード検索」のサブ項目がある。「マップ検索」メニューには「発言マップ検索」「TF・IDFマップ検索」のサブ項目がある。「統計」メニューをクリックすると、本システムが提供するデータに関する統計情報を表示することができる。

3. 地方議会会議録を用いた研究

地方議会会議録を用いた研究事例の現状について、自然言語処理分野、言語学分野、その他の分野に分けて述べる。

図9 「ぎ〜みる」のメニュー画面

図10 議員発言特徴語の例

3.1 自然言語処理分野

自然言語処理分野では、地方議会会議録の利活用を目指した要素技術または応用技術に関する研究が進められている。木村ら [12] では、北海道内の市議会を対象として、議員と市民をマッチングするために、議員の議会での発言に基づいた質問を生成するシステムが提案された。このほか、議会において誰がどのような政治課題に取り組んでいるのかを明らかにすることを目的として、木村ら [13]、葦原ら [14]、田中ら [15]、坂地ら [16]、乙武ら [17] などによって、政治課題抽出や意見抽出の研究が進められている。

桧森ら [18] はコーパス構築支援のため、会議録中の発言部分と非発言部分（資料やト書きなど）の自動分類を試みている。内田ら [19] では、議会会議録コーパスから議員の活動を可視化するための基礎分析が行われている。木村・佐々木 [20] では、東京都議会を対象に議会会議録から、政党を自動判別する *Stance Classification* が提案されている。木村ら [21] では、政治課題について議論の背景（新聞記事）、議論の過程（議会会議録）、議論の結果（条例の制定）の3つを結びつける地方政治コーパスの構築について述べている。

高丸ら [22] では、議員の発言に対して TF・IDF 技術を適用することで、議員発言特徴語の抽出を進めた。通常の形態素解析では、例えば「離島航空路線」は「離島」「航空」「路線」の3語に分割され、「地域医療」は「地域」「医療」の2語に分割される。このまま「路線」や「地域」といった語の頻度を数えても、発言の特徴を捉えることは難しいため、複合名詞⁴を基準として特徴語を求めている。議員ごとの特徴語を選挙情報サイトに提供し、順次公開を進めている（図10）。

このほか、NTCIR の評価型タスク QALAB PoliInfo において、地方議会会議録を用いたファクトチェックに向けて、発言の引用に対する原文の自動抽出（Segmentation Task）、発言文に含まれる政治的態度の自動分類（Classification Task）、会議

図11 「離合」の使用分布

録からの要約文の自動生成（Summarization Task）といった取り組みが進められている [23, 24]。

3.2 言語学分野

言語学分野では、地方議会会議録を大規模言語資源として取り扱うことで、言語使用や言語変異の実証的研究が進められている。一つの地方議会の会議録を遡れば通時的な言語変化を辿ることができ、全国の地方議会会議録を横断的に調査すれば地域差を分析することができる。地方議会における発言者はすべて現在その地域に居住する人であり、その多くはその地域の出身の人であるため、議会会議録は方言談話資料の一つととらえることができる。

第Ⅰ期・第Ⅱ期検索システムの機能として実装されているマップ検索機能によって、語の使用頻度の地域差を容易に可視化できるようになった。

「離合」のマップ検索結果を図11に示す。「離合集散」とい

4 Comainu (<http://comainu.org/>) を用いて作成した長単位を用いた

図 12 「あんた」の使用分布

図 13 「おめえ」の使用分布

う語の一部として全国的に使用される語である。九州地方を中心に「離合」は「車や列車のすれ違い」という意味で使用されることがあるため、九州地方で使用頻度が高い傾向にある。

井上 [25] では、議録を用いて二人称の使用分布を分析した。地方議会における「あんた」と「おめえ」の使用分布を図 12, 13 に示す。

高丸 [26] は地方議会議録コーパスを用いて、オノマトペ（擬音語・擬態語）の使用分布を明らかにした。たとえば、関西地方を中心に物事を大掴みに捉えるさまを「ばくっ」と表現することや、九州地方を中心に「びしゃっ」が共通語とは異なる語義で頻出することを明らかにした（図 14）。

井上 [27]、二階堂ら [6]、佐藤 [28]、高丸 [29] により、第 I 期コーパスを用いた言語の地域差（方言）に関する分析が行われている。

- まず、何ていうんですかね、議会としてのばくっとした考え方なりを聞きたいというそういった思いで委員会、特別委員会だと思っております。（愛知県尾張旭市議会）
- どういう課題が出てくるのかということについてはばくっとは聞きましたけれども、たればという話がいっぱいあってなかなか定かになっていない（滋賀県大津市議会）
- 平均値がどれくらいちゅのがもしわかれば、ばくっとでも結構です。（大阪府羽曳野市議会）
- こうなってくると、ばくっと私が考えるのは、逆にその責任体制が分散してしまって、だれがじゃあこのプロジェクトを回していくのか（兵庫県豊岡市議会）

- だから、審査会の権限とか、そういうものについてびしゃっと明らかにして示す必要があるんじゃないかというふうに思うんですが、そこら辺についてどう考えますか。（福岡県嘉麻市議会）
- そりゃ、職員は黒字が出ようが赤字が出ようがびしゃっとボーナスも出ておる。（長崎県雲仙市議会）
- それから、そのときも出たんですけども、資料の提出方法を、もうちょっとびしゃっと定めた方がいいのかなというような気はいたしました。（熊本県熊本市議会）
- こら辺をびしゃっと整備しなければいけないというふうに思っておりますし、住民から信頼される自治体となるべく努力をしていかなければいけない（宮崎県小林市議会）

図 14 使用に地域差のあるオノマトペの例

このほか議録を用いた方言研究の事例には、桑戸ら [30] などがある。

3.3 その他の分野

まず、地方議会議録コーパスプロジェクトが構築したコーパスを用い、複数の地方議会を対象とした横断的な統計分析の事例について述べる。大山ら [31] は第 I 期コーパスを用いて、地方議会における女性議員の比率と福祉や子育てに関する語の出現頻度の相関を分析した。川浦ら [32] は第 II 期コーパスを用いて、都道府県議会における発言量と選挙における得票数の関係についての計量的に分析を行っている。

テキストマイニング（計量テキスト分析）によって、自治体の取り組みなどを明らかにしようとする試みが、高丸 [33]、小田切 [34]、増田 [35] によって進められている。

このほか、いくつかの学術分野において、地方議会議録を利活用した研究が見られるようになってきた。例えば、図書館の問題（松本 [36]）、環境問題（上田他 [37]）、景観問題（森下 [38]）、学校教育への応用（國原 [39]）が挙げられる。

4. おわりに

本稿ではまず、地方議会議録コーパスプロジェクトの取り組みについて述べた。第 I 期コーパスの取り組みにより、大規模な学習データを必要とする自然言語処理分野の研究や大規模言語資源に基づく言語学的な実証研究における研究成果が得られた。第 II 期の統制コーパスサブセットへの取り組みにより、計量的、統計的に精緻な分析が可能となり、さらに学術研究の

裾野が広がりを見せている。

第 II 期コーパスの特徴の一つである、発言者の名寄せ(発言者テーブルとの結びつけ)を活用した研究が今後進展することが期待される。現在のところ、議員の所属する党派(会派)に関する情報は付与していないが、政治学的な分析を進めるために必要な情報であると考えられるため、今後、会派情報の付与についても検討を進めたい。

地方議会議録コーパスプロジェクトでは 2019 年 2 月現在、新たな統制コーパスサブセットとして、「47pref_2015-2019」(2015 年 4 月から 2019 年 3 月までの都道府県議会本会議会議録)および「tokyo23_2015-2019」(2015 年 4 月から 2019 年 3 月までの東京 23 区議会本会議会議録)の構築を進めており、検索システム等を近日公開できる見込みである。今後、県庁所在地市議会サブセット、政令指定都市市議会サブセットなどを順次構築する予定である。

地方議会議録は自治体ごとに公開されているが、国会会議録については国立国会図書館によって一元的にデータベース化が行われている。公開時期も長く、帝国議会にまで遡って発言を検索することが可能となっている。国の政策を決定づける重要な議論の過程が記録されているため、様々な研究分野で利用されている。たとえば、外国人看護師・介護士(布尾[40])、労働者のストレス(内山ら[41])、就労支援(金川[42])などの問題について国会会議録にもとづいてその経緯が分析されている。また、技術面でも映像検索(鈴木[43]、増山[44])、人工知能(佐藤[45])などの分野で研究が盛んに進められている。簡便に検索できる大規模言語資源であることから、松田編[46]をはじめとして、種々の言語現象の検証に利用されている。また、計量的な言語研究に向けた検討も進んでいる(岡田[47])。

地方議会と国会の関係に目を向けると、国会においても地方自治に関係することが議論され、法律・予算などが決められる。また、地方議会議員を経て国会議員になる人、国会議員から地方の首長になる人などがある。国会会議録と地方議会議録の間で、政治課題のつながりや、議員の発言の追跡など様々な観点で、連携しつつ研究を進める価値があると考えられる。地方議会と国会における政治課題の結びつけ、国会における地域政策と地方財政の可視化、地方議員と国会議員を網羅した発言者の名寄せなど、地方と国を関連づける研究課題は山積しているものの、ほとんど手を付けられていないのが現状である。

今後、コーパスの拡充、データ形式の整備などを網羅的に進め、地方議会議録の横断的研究、さらには、国会会議録など他の言語資源との連携による研究がますます進展することが期待される。

謝辞

本稿は JSPS 科研費 JP16H02912, JP17K02739 による。

参考文献

- 大山英久: “地方議会の公開と会議録をめぐる,” レファレンス, Vol.57, No.6, pp. 31-46, 2007.
- 竹安栄子: “地方議員のジェンダー差異 — 「2002 年全国地方議員調査」結果の分析より —,” 京都女子大学現代社会研究, Vol.7, pp. 99-118, 2004.
- 野村稔, 鶴沼信二: 地方議会実務講座 第 3 巻, ぎょうせい, 1996.
- 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, Vol.10, pp. 74-86, 2010.
- 高丸圭一: “規模の異なる自治体における地方議会会議録の整文の比較,” 社会言語科学会第 27 回研究大会発表論文集, pp. 256-259, 2011.
- 二階堂整, 川瀬卓, 高丸圭一, 田附敏尚, 松田謙次郎: “地方議会議録による方言研究 — セミフォーマルと気づかない方言 —,” 方言の研究, Vol.1, pp. 299-324, 2015
- 齋藤誠, 大城卓, 菅原晃平, 永井隆広, 渋谷英潔, 木村泰知, 森辰則: “地方議会議録の収集とコーパスの構築,” 言語処理学会第 17 回年次大会論文集, pp. 2-21, 2011.
- 菅原晃平, 大城卓, 齋藤誠, 永井隆広, 渋谷英潔, 木村泰知, 森辰則: “地方議会議録コーパスの拡充における問題点の分析と対処,” 言語処理学会第 18 回年次大会論文集, pp. 1-15, 2012.
- 乙武北斗, 高丸圭一, 渋谷英潔, 木村泰知, 森辰則: “地方議会議録コーパスの学際的応用を目的とした n-gram データの構築およびウェブ UI の試作,” 言語処理学会第 19 回年次大会発表論文集, pp. 733-736, 2013.
- 乙武北斗, 高丸圭一, 内田ゆず, 木村泰知: “一般公開版「都道府県議会議録検索システム」の概要,” 第 32 回人工知能学会全国大会, pp. 1-28, 2018.
- H. Ototake, H. Sakaji, K. Takamaru, A. Kobayashi, Y. Uchida, and Y. Kimura: “Web-based system for Japanese local political documents,” *Int. J. of Web Information Systems*, Vol.14, No.3, pp. 357-371, 2018.
- 木村泰知, 渋谷英潔, 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.
- 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.
- 葦原史敏, 木村泰知, 荒木健治: “地方議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.
- 田中琢真, 小林暁雄, 坂地泰紀, 内田ゆず, 乙武北斗, 高丸圭一, 木村泰知, 増山繁: “都道府県議会議録を対象とした議題・議案表現の自動抽出に向けた検討,” 第 31 回人工知能学会全国大会, 3G2-2, 2017.
- 坂地泰紀, 酒井浩之, 小林暁雄, 内田ゆず, 乙武北斗, 高丸圭一, 木村泰知: “都道府県議会議録からの意見や意志を表す発言の抽出,” 言語処理学会第 23 回年次大会, pp. 426-429, 2017.
- 乙武北斗, 高丸圭一, 内田ゆず, 木村泰知: “述語項構造に基づく地方議会議録からの意見抽出手法の検討,” ファジィシステムシンポジウム講演論文集, Vol.34, pp. 222-225, 2018.
- 松森拓真, 木村泰知, 坂地泰紀: “地方議会議録における発言文の推定,” 言語処理学会第 24 回年次大会, P1-3, pp. 125-128, 2018.
- 内田ゆず, 高丸圭一, 乙武北斗, 木村泰知: “都道府県議会議録コーパスを用いた議員の議会活動の可視化に向けて,” 第 32 回人工知能学会全国大会, 1E3-03, 2018.
- Y. Kimura and M. Sasaki: “Stance Classification Using Political Parties in Tokyo Metropolitan Assembly Minutes,” *The Seventh Int. Conf. on Data Analytics*, pp. 46-49, 2018.
- 木村泰知, 小林暁雄, 坂地泰紀, 内田ゆず, 高丸圭一, 乙武北斗, 吉田光男, 荒木健治: “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の試み,” 第 31 回人工知能学会全国大会, 3G2-4, 2017.
- 高丸圭一, 内田ゆず, 木村泰知: “地方政治コーパスにおける都道府県議会議録パネルデータの基礎分析,” 宇都宮共和大学シテライフ学論叢, Vol.18, pp. 136-155, 2017.
- 木村泰知, 渋谷英潔, 乙武北斗, 内田ゆず, 高丸圭一, 阪本浩太郎, 石下円香, 三田村照子, 神門典子: “NTCIR-14 QA Lab-PoliInfo の Formal Run Dataset の構築,” 言語処理学会第 25 回年次大会, C2-2, 2019.
- 渋谷英潔, 木村泰知, 乙武北斗, 内田ゆず, 高丸圭一, 阪本浩太郎, 石下円香, 三田村照子, 森辰則, 神門典子: “NTCIR-14 QA

Lab-PoliInfo Formal Run における Summarization Task の結果と評価,” 言語処理学会第 25 回年次大会, pp. 8-11, 2019.

- [25] 井上史雄: “日本語 2 人称代名詞の地方議会議録における出現,” 明海日本語, Vol.21, pp. 1-16, 2016.
- [26] 高丸圭一著, 小林隆編: “地方議会におけるオノマトベの使用分布,” 感性の方言学, pp. 119-146, 2018.
- [27] 井上史雄: “去った〇日,” ことばの散歩道, 明治書院, pp. 154-155, 2013.
- [28] 佐藤亜実: “議録における接尾辞ラヘンの用法拡張,” 国語学研究, Vol.56, pp. 192-208, 2017.
- [29] 高丸圭一: “形態素 N-gram を用いた地方議会議録における地域差の分析手法の検討 — ひらがなで構成された文末の 4-gram に着目して —,” 明海日本語, Vol.18, pp. 1-10, 2013.
- [30] 桑戸孝子, 上野誠司: “長崎方言話者による「イチゴが売ってある」という表現について — 質問紙および地方議会議録調査から,” 長崎総合科学大学紀要, Vol.58, No.2, pp. 115-125, 2018.
- [31] 大山礼子, 五ノ井健: “女性議員は地方議会を変えるのか? — 地方議会議録による分析の試み,” 女性展望, Vol.691, pp. 6-10, 2018.
- [32] A. Kawaura, Y. Kimura, K. Takamaru, and Y. Uchida: “Elected Officials in the Local Assembly: Analysis of Prefectural Plenary Session Transcripts,” Doshisha University Center for the Study of the Creative Economy Discussion Paper Series, 2018.
- [33] 高丸圭一: “地方議会では何が話題になっているのか — 宇都宮市議会議録のテキストマイニング,” 宇都宮共和大学都市経済研究年報, Vol.13, pp. 162-173, 2013.
- [34] 小田切康彦, “地方議会における協働言説 — 関西地方を例として,” 同志社政策科学研究 (20 周年記念特集号), pp. 45-57, 2016.
- [35] 増田正: “我が国地方議会における政治・行政関係の計量テキスト分析,” 地域政策研究, Vol.20, No.3, pp. 1-19, 2018.
- [36] 松本直樹: “地方議員の図書館への関心に関する予備的考察 — 埼玉県市議会の議会議録分析を元に,” 日本図書館情報学会誌, Vol.54, No.1, pp. 39-56, 2008.
- [37] 上田翔, 八木田浩史: “地方議会議事録における環境用語の出現頻度に基づく自治体の環境問題対応の解析,” 環境情報科学学術研究論文集, Vol.26, pp. 283-288, 2012.
- [38] 森下壽典: “川崎駅西口に存在した赤れんが倉庫のオブジェ化 — 川崎市議会議録からその過程をみる,” 産業考古学, Vol.150, pp. 2-9, 2013.
- [39] 國原幸一朗: “地方議会における争点をふまえた公民の授業 — 東海豪雨と東日本大震災を事例として,” 名古屋学院大学論集 人文・自然科学篇, Vol.53, No.2, pp. 93-106, 2017.
- [40] 布尾勝一郎: 迷走する外国人看護・介護人材の受け入れ, 2016.
- [41] 内山夕風, 吉本知江子: “改正労働安全衛生法に基づくストレスチェック制度に関する一研究 — 国会議録の分析を通して,” 純心現代福祉研究, Vol.22, pp. 11-40, 2018.
- [42] 金川めぐみ: “母子家庭の母及び父子家庭の父の就業の支援に関する特別措置法成立までの歴史的経緯 — 国会議録からみる「就労支援」へのシフトに着目して,” 法学雑誌, Vol.64, pp. 396-362, 2018.
- [43] 鈴木泰山, 内山雄司, 青木保一, 相良毅, 秋田祐哉, 河原達也, 竹田香織, 増山幹高: “音声認識技術の活用による国会審議映像検索システムの実現,” 情報処理学会研究報告音声言語情報処理, Vol.2014-SLP-103, No.6, 1-4, 2014.
- [44] 増山幹高, 竹田香織: “いかに見たい国会審議映像に到達するか? — 国会審議映像検索システムの概要,” レヴアイアサン, Vol.56, pp. 54-79, 2015.
- [45] 佐藤美沙, 柳井孝介, 柳瀬利彦, 是枝祐太, 丹羽芳樹: “国会議録を用いたディベート人工知能による意見生成,” 第 31 回人工知能学会全国大会論文集, 4Q19in2, 2017.
- [46] 松田謙次郎: 国会議録を使った日本語研究, 2008.
- [47] 岡田祥平: “日本語コーパスとしての「国会議録検索システム 検索用 API」 — 計量的研究の精緻化・深化の可能性,” 新潟大学教育学部研究紀要 人文・社会科学編, Vol.11, No.1, pp. 31-51, 2018.

(2019 年 2 月 16 日 受付)

[問い合わせ先]

〒320-0811 栃木県宇都宮市大通り 1-3-18
宇都宮共和大学
高丸 圭一
E-mail: takamaru@kyowa-u.ac.jp

—— 著 者 紹 介 ——

高丸 圭一 [非会員]

2004 年北海道大学大学院工学研究科博士後期課程単位修得退学。同年那須大学都市経済学部専任講師, 2013 年宇都宮共和大学シティライフ学部准教授, 2018 年同教授 (現在)。修士 (工学), 博士 (応用言語学)。自然言語処理, 音声学, 社会言語学, コーパス言語学などの研究に従事。電子情報通信学会, 人工知能学会, 社会言語科学会, 日本音声学, 日本語学会各会員。