3B3-NFC-4-3

地方議会会議録コーパスの構築とその利用

Creation and Use of Regional Assembly Minutes Corpus

木村泰知[*1] 渋木英潔[*2] 高丸圭一[*3] 乙武北斗[*4] 森辰則[*2]
Yasutomo Kimura Hideyuki Shibuki Keiichi Takamaru Hokuto Ototake Tatsunori Mori

[*1]小樽商科大学 Otaru University of Commerce
[*2]横浜国立大学 Yokohama National University
[*3]宇都宮共和大学 Utsunomiya Kyowa University
[*4] 福岡大学 Fukuoka University

Our aim is to construct regional assembly minutes corpus that many researchers would use. This paper reports results of the construction of the regional assembly minutes corpus. We explain Web services such as information retrieval, cross table retrieval and KWIC. Furthermore, we describe some related researches, which used the regional assembly minutes corpus.

1. はじめに

近年，国会や地方議会などの会議録がウェブ上に公開されている．会議録は，首長や議員の議論が書き起こされた話し言葉のデータであり，長い年月の議論が記録された通時的データでもあることから，政治学，経済学，言語学，情報工学などの分野において研究対象のデータとして利用されている．国会会議録を利用した研究は，会議録の整備が進んでいることもあり，多くの分野で行われている [松田 2008]．一方，地方議会会議録を利用した研究では，各分野で研究が行われているものの，自治体によりウェブ公開されているフォーマットが異なるため，収集作業や整形作業に労力がかかっている．また，各研究者が重複するデータの電子化作業を個別に行っているといった非効率な状況も招いている．

このような背景から，本研究では，多くの研究者に地方議会会議録データを利用してもらうことを目的として，地方議会会議録コーパスの構築を行っている．地方議会会議録コーパスは，ウェブに公開されている全国の地方議会会議録を対象として,「いつ」「どの会議で」「どの議員が」「何を発言したのか」を検索可能な形式で収録する．本稿では，地方議会会議録コーパス構築の進捗状況を報告するとともに，その課題や各分野でのコーパスを利用した研究について述べる．

2. では，地方議会会議録コーパスの構築について述べる．3. では，コーパス提供のためのウェブサービスについて述べる．4. では，コーパス公開の課題について述べる．5. では，コーパスを利用した各分野での研究について述べる．6. では，まとめと今後の課題について述べる．

2. 地方議会会議録コーパスの構築

2.1 会議録の公開状況

我々は，平成 22 年度に全国の自治体 1,727 の市町村と東京都の区議会 23 区を対象に，地方議会会議録に関するアンケート調査を実施した [高丸 2011a]．アンケートの結果，993 市区町村中 729 市区町村（73.4%）の自治体が会議録をウェブ公開していることがわかった．公開されている期間は，インターネットが普及し情報公開が進められた平成 11 年以降が多いが，大阪市のように 1966 年からデータをすべて電子化して公開している例もある．したがって，ウェブ公開されている会議録を収集することで，比較的十分な量の会議録を確保できると考え、まずはウェブ公開されている会議録を対象として地方議会会議録コーパスを構築することとした．

連絡先: 木村泰知，小樽商科大学，小樽市緑 3 丁目 5-21，0134-27-5388，kimura@res.otaru-uc.ac.jp

2.2 フォーマット

各自治体がウェブ上に公開している会議録は，自治体によって形式が異なり，PDF, 静的な HTML, 検索機能付きのサーバーサイドで生成されている HTML などの形式がある．研究で会議録を利用する場合には，どこの市町村の，何年度の，何の会議の，何回目の，何日目（日付）に，誰が，どのような発言をしたのかといった情報が，利用される分野を問わず最低限必要になると考えられる．そこで，我々の地方議会会議録コーパスでは，様々な形式で公開されているこれらの情報に一元的にアクセスできるよう，表 1 の項目を付与することとした．

表 1: 地方議会会議録コーパスのフォーマット

項目	備考
発言 ID	自動採番
市町村コード	全国地方公共団体コード (6 桁)[*1]
議会種別コード	独自のコード
年	西暦
回	開催数
月	開催月
議会名	例. 定例会，予算委員会など
号	会議が何日目なのか
日付	開催日 4 桁
表題	議会名の情報を含む文字列
段落番号	発言の段落番号
役職名	議員の役職
議員フラグ	議員ならば 1, それ以外は 0
発言者名	議事録より抽出されたもの
発言者表層	発言者と役職が分けられない場合全文字列
議員 ID	議員リストに対応が無い場合 -1
パス	原ファイルの保存場所を表す相対パス
発言	文単位
その他	発言以外の文字列

2.3 収集作業と整形作業

会議録をウェブ公開している自治体は，システム開発業者に委託していることが多く，主に 4 社（会議録研究所，大和速記情報センター，フューチャーイン，神戸綜合速記）の会議録検索システムが採用されている．我々の調査では，618 の自治体（会議録を公開している 85 %）が上記の 4 社の会議録検索システムを利用していることを確認した [斉藤 2011][菅原 2012]．

会議録の収集は，ページに含まれるリンクを解析し，ページ中に含まれる語句から CGI プログラムに与えるパラメータを生成することで，自動収集を行っている．会議録の整形は，収集した発言に表 1 の項目を付与することである．

ここで，整形作業の課題について述べる．本研究では，「どの議員が」「何を発言したのか」を検索できるようにするために，議員の同定が必要となる．具体的には，会議録に含まれる「議長（山田太郎君）」のように「役職名（姓名（＋敬称））」の統一的な表記で記述されている表現を見つけ，機械的に処理している．しかし，常任委員会などの会議録では，役職名や姓名の区切りが明確でない表記が存在するため [菅原 2012]，同一人物の判定を行う必要がある．

また，議員の発言は，質問単位あるいはトピック単位が「改行」で区切られている傾向にあったため，「改行」を手がかりとして，段落番号を付与した．しかし，段落の判定は，下記の例に示すように，改行だけで判断することが困難な例もあることから，改善の余地がある．

1 発話を改行で判断することが困難な例

以上のような考え方に基づいて編成いたしました結果、予算案の総額は、
一般会計 2 兆 9089 億 6400 万円
特別会計 8154 億 3700 万円
合 計 3 兆 7244 億 100 万円
となりました。

さらに，コーパスに付与する基本情報には，政党・会派の情報が必要と考えられる．例えば，松本は，地方議会会議録に含まれる発言を政党ごとに分類している [松本 2008]．ただし，政党や会派については，所属政党名の変更や新しい政党の誕生などにより，変更時期や所属している期間を正確に把握しなければならず，会議録以外の情報を利用する必要があり，簡単に付与することができない．

3. コーパス提供のためのウェブサービス

3.1 情報検索

地方議会会議録データは，議会における議論を要約せずに書き残しているため，データサイズが大きくなる．例えば，政令指定都市と県庁所在地の市を合わせた 51 市の会議録を収集・整形をしたデータベースの dump ファイルは約 17GB である．そこで，本研究では，コーパスを利用するユーザが，データサイズを意識せずに利用できるように，ウェブサービスによる提供を検討している．

我々は，ウェブサービスの一つとして，地方議会会議録コーパスから簡単に情報をみつけられるように，情報検索を構築した．図 1 に情報検索の画面を示す．この情報検索では，検索条件として，発言本文に含まれる単語，発言者，対象市町村，対象会議録，対象年度を設定することができる．対象市町村の条件では，法律関係者の要望から，政令指定都市や市に限定した検索を選択できる．また，対象会議録の条件では，本会議（定例会），予算特別委員会，決算特別委員会 などを入力することにより，対象範囲を絞ることができる．さらに，対象年度の条件では，検索対象の開始年度と終了年度を選択することができる．

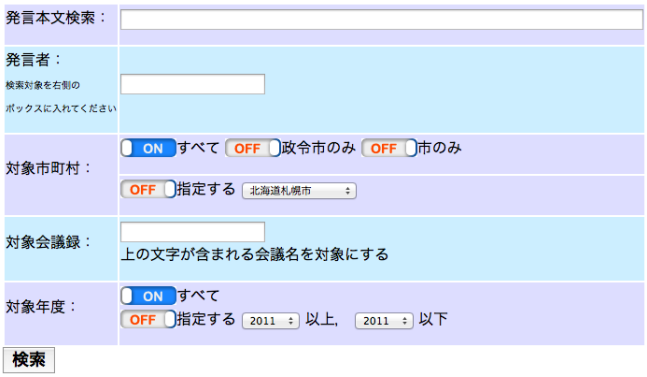

図 1: 情報検索の画面

3.2 クロス表検索

会議録を分析する場合には，キーワードが含まれる発言を表示する情報検索だけではなく，他の自治体や他の議員との違いをキーワードの出現回数で比較することが考えられる．そこで，我々は，キーワードが含まれる回数を「議員と議会」の観点と「年度と自治体」の観点からクロス表を作成するクロス検索の構築を行った．議員と議会のクロス検索では，議会名を縦軸に，議員名を横軸にして，入力されたキーワードが含まれる回数を表示する．議員単位でキーワードが含まれる回数をクロス表として表示することは，検索キーワードに関連した事柄に興味を持っている議員などを見つけることに役立つ．また，年度と市町村のクロス検索では，年度を縦軸に，市町村名を横軸にして，入力されたキーワードが含まれる回数を表示する．自治体単位でキーワードが含まれる回数をクロス表として表示することは，検索キーワードに関連した事柄に力を入れている自治体を見つけることに役立つ．図 2 と図 3 には，キーワードとして「除雪」を入力した場合の議員と議会のクロス検索と年度と市町村のクロス検索の結果を示す．

今回の検索では，完全一致の単語のみを対象としているため，同義語への対応として，シソーラスなどの利用が考えられる．また，会議録に記述されている議員名は，姓名の両方が記述されているとは限らず，議員の名字だけの場合や首長の代わりに職員の名前が記述されている場合があり，議員の特定が必要となる．現段階では，「ザ・選挙」が保有している議員リストとの照合を行っているが，前述したように，名字だけの記述も多いことから，議員とそれ以外の識別を自動で行えない例も数多く存在した．

3.3 KWIC

KWIC とは，KeyWord In Context の略で，キーワードの前後にある文脈を表示する方法である [吉平 2004]．ここで文脈とは，文を超えた表現ではなく，キーワードの前後にある表現のことである．地方議会会議録において，前方の表現にはキーワードを修飾する単語が存在していることが多いため，キーワードを含む複合名詞の表現や例外の表現を見つけることができ，内容の詳細を知ることができる．また，後方の表現には，助詞などが多いことから，キーワードが他の単語とどのような関係になりやすいのか見つけることができる．図 4 に「対策」をキーワードとした KWIC による検索結果の例を示

検索語： 除雪
市町村： 北海道札幌市
年度： 2011 以上, 2011 以下 検索

検索クエリ = "除雪"

	新谷	宝本英明	松浦忠	吉沢	坂本恭子	上田文雄
北海道札幌市 2008年度 税財政・地方分権調査特別委員会	12	4	1	1	1	0
北海道札幌市 2008年度 定例会	0	0	0	0	0	8
北海道札幌市 2008年度 第一部予算特別委員会	0	0	1	0	0	0

図 2: **議員と議会のクロス検索結果の例**

検索語： 除雪
検索

検索クエリ = "除雪"

	2011 73	2010 1,129	2009 754	2008 846	2007 1,153	2006 1,661	2005 967	2004 1,185
北海道札幌市 5,857	-	136	422	421	478	493	264	216
青森県青森市 1,891	-	235	14	22	88	212	242	294
岩手県盛岡市 1,827	-	59	84	47	76	207	82	120
宮城県 212	-	13	6	12	6	40	9	11
宮城県仙台市 916	-	13	35	52	96	28	53	95

図 3: **年度と市町村のクロス検索結果の例**

す.「対策」というキーワードの前の表現には,「経済対策」「いじめ・不登校の対策」など政治課題に対する対策が記述されており，どのような政治課題があるのかを知ることができる．一方,「対策」の後ろの表現には,「対策の強化」「対策を促進」「対策のあり方」などの表現のように,「対策」に対する「対応」が記述されている．このような前後の表現を容易に検索できるようにすることで，会議録の分析を効率的に行えると考えられる．

対策 search

" したがって、国として次の 対策 を講ずる必要があると考えます。
会意見書案第１８号季節労働者 対策 の強化を求める意見書（案）秋
、しかも自治体には過去の「経済 対策 」による公共事業の地方債償還が
くりのために、いじめ・不登校の 対策 をより充実させるよう強く要望し
せ、あわせて省エネルギーなどの 対策 を促進させること。"
の究明をはじめ、いじめや不登校 対策 として教育条件整備を進めるとと
ものは認められないため、有効な 対策 となりません。加えて、国・厚生
予算で措置された「燃油高騰緊急 対策 基金」と同様の基金を再度創設し
就労と所得保障など実効ある追加 対策 を講ずること。"
意見書案第１４号いじめ・不登校 対策 のための施策の充実を求める意見
成８年４月に提出された「ウタリ 対策 のあり方に関する有識者懇談会報
油製品の価格を引き下げるための 対策 として、政府において、次の事項
灯油等石油製品価格引下げの緊急 対策 を求める緊急動議」が提起され、
確立するなど安定供給 に万全の 対策 を講ずること。"

図 4: **KWIC の例**

4. コーパス公開の課題

本研究では，地方政治に関する学際的な応用研究の活性化を目指し，地方議会会議録を収集しているが，多くの研究者に利用してもらう場合，著作権について考える必要がある．国会会議録の FAQ サイト[*2]には，著作権についての説明があり，個々の発言内容の著作権は発言者に帰属し，データベースの編集著作権は国立国会図書に帰属していると記述されている．本研究では，今後も，著作権法などの問題を解決するために，法律の専門家と協議しながら，提供方法について検討する．

5. 会議録コーパスを利用した各分野での研究

5.1 情報工学の研究

情報工学の分野では，会議録に含まれるテキストから，政治問題の表現や要求表現の自動抽出，抽出データの関係推定などを利用して，住民，自治体職員，政治家などに有益な情報を提供する研究が行われている．

会議録は定例会だけでも膨大な量であり，北海道小樽市の市議会会議録の場合，定例会 1 回分の会議録だけで A4 版にすると 200 ページを超える．木村らは，大量のテキストデータに対して能動的にアクセスして，これらのデータを読む住民が少ないと考え，政治的問題の関心を明確にするための質問をシステムから利用者に行うことで，利用者の考えに近い議員を提示する方法を提案している [木村 2010][木村 2011].

また，大城らは，利用者の考えに近い議員を提示するために，住民が関心を持つ施策・事業に関する意見を会議録から容

*2 http://kokkai.ndl.go.jp/KENSAKU/www_faq_top.html

易に得られるようにすること，住民自身に近い考えを持つ議員を探し出すことを目的として，地方議会会議録への注釈タグ付けセットを提案している [大城 2012].

他には，会議録に含まれる重要部分を抽出する研究が行われている．木村らは，会議録に含まれる政治問題を自動で抽出することを目指し，会議録に含まれる政治問題の定義を行い，主辞に着目することを提案している [木村 2012]．また，葦原らは，会議録に含まれる重要な内容が議員からの質問に含まれることが多いことに着目し，議員の質問から要求表現を抽出する研究を行っている [葦原 2012].

5.2 言語学の研究

地方議会会議録は，社会言語学，日本語学，方言学などの研究に寄与する言語資源であると考えられる．しかし，会議録は議会における発言を一字一句厳密に記録しているわけではなく，文章としての読みやすさを考慮して，意味内容が大きく変わらない範囲で修正（整文）が加えられている．高丸らは，地方議会会議録の言語資源としての性質を明らかにするための基礎研究として，複数の地方議会会議録における整文の状況を分析し，実態を比較した [高丸 2010][高丸 2011b].

整文の過程において，冗長な表現の削除，言い間違いや方言語彙の修正などが行われているため，地方会議録コーパスを用いて話し言葉に含まれる非流暢性を分析することは困難であると考えられる．しかし，本コーパスは通時性・共時性を併せ持つ言語資源であるため，新しい文法表現の受容の実態や議会用語の変遷，整文が加えられていない方言（気づかない方言）などを分析することが可能であり，現在これらの研究への本コーパスの活用を進めている．

5.3 政治学の研究

地方議会では，自治体ごとに議会の進め方や公開方法が異なる．議会改革を進めている自治体では，議会機能の強化や議会の組織・構成及び運営の改革を行っている．例えば，北海道の栗山町は，議会改革を進めている自治体の一つであり，議会ライブ中継，議会録画配信に加えて，質問方式も異なる[*3]．議会の質問方式には，一括質疑・一括答弁方式と一問一答方式があり，発言回数の制限，質問回数の制限などの違いもある．本コーパスでは，質問方式による議会への影響について，議論議会改革を進めている自治体とそれ以外の自治体との違いを分析することができる．

6. おわりに

我々は，地方議会会議録を利用する研究者に対して，利用しやすいコーパスを提供することを目的として，地方議会会議録コーパスの構築を行ってきた．

本稿では，全国の自治体を対象として，ウェブに公開している会議録の収集方法と整形処理について報告した．また，収集した地方議会会議録コーパスを利用するためのウェブサービスとして，情報検索，クロス表検索，KWIC 検索について説明した．さらに，現在行われている地方議会会議録コーパスを利用した各分野での研究について紹介した．

今後は，収集対象の会議録を増やすとともに，地方議会会議録コーパスを多くの研究者に利用してもらう予定である．

謝辞

本研究の一部は科研費 22300086 の助成を受けたものである．

∗3 http://www.town.kuriyama.hokkaido.jp/gikai/term/index.html

参考文献

[葦原 2012] 葦原史敏, 木村泰知, 荒木健治, “地方議会会議録における要求・要望表現抽出の提案”, 言語処理学会第 18 回年次大会論文集, P1-27, 2012.

[大城 2012] 大城卓, 渡邊裕斗, 渋木英潔, 木村泰知, 森辰則, “地方政治情報システムのための地方議会会議録への注釈付けタグセットの提案”, 言語処理学会第 18 回年次大会論文集, P3-9, 2012.

[木村 2010] 木村泰知, 渋木英潔, 高丸圭一, 小林哲朗, 森辰則, “北海道を対象とした地方議員と住民間の協働支援システムのユーザインターフェース評価”, 第 24 回人工知能学会全国大会論文集, No. 2J2–NFC2–3. 人工知能学会, 2010.

[木村 2011] 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.

[木村 2012] 木村泰知, 関根聡, “主辞に着目した政治問題の定義と注釈付け”, 言語処理学会第 18 回年次大会論文集, F1-6, 2012.

[斉藤 2011] 齋藤誠, 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, P2-21, 2011.

[菅原 2012] 菅原晃平, 大城卓, 齋藤誠, 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, P1-15, 2012.

[高丸 2010] 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.

[高丸 2011a] 高丸圭一, 木村泰知, 渋木英潔, “全国の市町村議会会議録のウェブ公開とデータ提供の状況”, 宇都宮共和大学 都市経済研究年報第 11 号, 2011.

[高丸 2011b] 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, pp.256-259, 2011.

[松田 2008] 松田謙次郎 編, 国会会議録を使った日本語研究 ひつじ書房, 2008.

[松本 2008] 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会議録分析をもとに”, 日本図書館情報学会誌, Vol. 54, No. 1, pp.39-56, 2008.

[吉平 2004] 吉平健治, 武田善行, 関根聡, “WEB 文書を対象とした KWIC システム”, 言語処理学会第 10 回年次大会, pp.137-139, March 2004.