地方議会会議録の収集とコーパスの構築

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[‡]
[†]横浜国立大学 大学院 環境情報学府
[‡]横浜国立大学 大学院 環境情報研究院
[§]小樽商科大学 商学部 社会情報学科
E-mail: [†]{saito,oshiro,sugawara,nagadon}@forest.eis.ynu.ac.jp,
[‡]{shib,mori}@forest.eis.ynu.ac.jp,
[§]kimura@res.otaru-uc.ac.jp

1　はじめに

総務省の発表によれば，日本政府が平成11年から進めてきた「平成の大合併」と平成17年に施行された「合併特例新法」の影響により，平成11年3月末の時点で3,232存在した市町村の数は，平成21年10月の時点で1,772にまで減少している．

この平成の大合併は地方政治に関する研究に多大な影響を与えており，政治学においては合併の前後での違いに関する研究が数多く行われている．例えば，平野[5]は平成の大合併前後に行われた市長選挙についての分析をしており，合併を行った市と行わなかった市の違いを当選者の属性から比較している．また，森脇[3]は合併が地方議会や議員の活動に対して与えた影響を856議員にアンケート調査することで分析を行っている．

さらに，地方政治に関する研究は政治学以外の分野でも行われており，経済学の分野においては川浦[4]による「小規模自治体の多選首長は合併に消極的」という仮説を検証するために，北海道の地方議員、首長の情報を人手で調査している．また，我々も以前から，一人ひとりの住民の興味や関心にマッチした政治情報を提供する住民本位型政治情報システムに関する研究開発を木村ら[6]やTakamaru et al.[1]において行っており，北海道の64市町村を対象とした地方議会会議録のデータを収集，利用している．

こういった研究において，対象となるデータを独自に収集することは大きな負担であり，結果として小規模なデータに限定されてしまうといった研究遂行上の障害となることが多い．また，人文科学や社会科学の分野においてもコンピュータ上での処理が一般的になっているが，各研究者間で重複するデータの電子化作業など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，我々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**の構築を目指している．コーパスの構築に当たっては，我々が木村ら[6]等において行った，北海道の地方議会会議録データの自動収集や加工の技術を活用し，全国の市町村の議会会議録を対象としたプロジェクトを行うこととした．

本稿では，このプロジェクトにおける地方議会会議録の収集とそのコーパスの構築について述べる．2章では，プロジェクトの概要について述べる．3章では，関連研究について述べる．4章では，会議録検索システムについて説明し，それぞれの会議録検索システムに対しての自動的な収集方法について述べる．5章では，我々の構築するコーパスについて述べる．6章はまとめである．

2　プロジェクトの概要

図1は本プロジェクトの全体像を示したものである．今回構築するコーパスは将来的に，政治学，社会言語学，情報工学などにおいて利用される予定であり，一例として，地方議会会議録とWebを用いた議員情報の収集に関する研究を行っている．この研究では地方議員の情報を自動的に収集・整理して提示するシステムの構築を目指しており，会議録から発言単位の抽出が行えるとこのような研究に役立つと考えている．

また，上記の研究で得られるであろう知見は，我々がこれまでに行ってきた住民本位型政治情報システムの研究開発においても役立つことが期待され，これらの知見を学際的に応用した研究成果として全国の市町村を対象とした政治情報システムの研究開発を行う予定である．

3　関連研究

国会会議録検索システム[1]というシステムが公開されており，国会の会議録を自由に検索・閲覧することができる．この会議録の書式は全て統一されている．これに対して，地方議会会議録のもつ会議録検索システムは，市町村ごとに書式が異なっているため，複数の市町村の会議録を対象にした研究を行おうとした場合，そのまま利用することは難しい．そこで，地方議会会議録を収集して統一された書式に整形を行う必要がある．

それに関連して，乙武ら[2]は，北海道内の各市町村を対象に地方議会会議録の自動収集に向けた公開パタンの分析を行っている．51種類の収集パタンによる自動収集プログラムを用いて約94%の自治体から会議録の収集に成功している．これを受けて，本稿では全国規模の会議録の収集を目指す．

4　会議録の収集

本プロジェクトの目標は全市町村の会議録を収集し，議会名や発言者名などの付随情報を付与したコーパスを作成し，その後，利用しやすいようにデータベースに登録・整備することである．そこで，まず全都道府県の県庁所在地と政令指定都市の計51市町村の会議録について合併特例新法が施行された平成17年から平成22年を対象に収集を行うことにした．市町村の会議録の多くは，Web上で専用の会議録検索システムを通して公

[1] http://kokkai.ndl.go.jp/

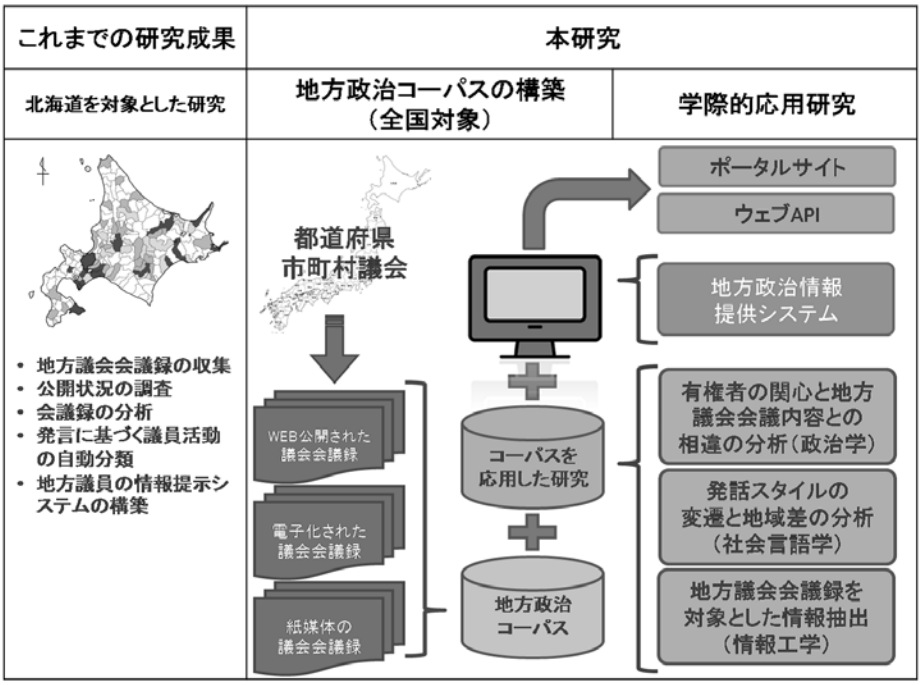

図 1: プロジェクトの全体像

表 1: 51 市町村の会議録検索システム

会社名	市町村数
大和速記情報センター[2]	14
会議録研究所[3]	23
フューチャーイン[4]	10
神戸綜合速記[5]	3
その他	1

開されている．調査の結果，表 1に示すように大きく分けて 4 つの会社が会議録検索システムを作っていることが分かった．その他に該当するのは秋田市のみで，独自の会議録検索システムを作っていたため，人手により会議録を収集した．

以下の節では，それぞれの会議録検索システムについて会議録の自動的な収集方法を説明する．収集した会議録は全て HTML 形式のファイルであり，HTML タグを残したまま保存している．コーパスに収録する際の形式は現在検討中であるが，収集した HTML 文書からタグ等を取り除いた本文と 5章で説明する表 2に示す項目の情報を収集する予定である．

4.1 大和速記情報センター

大和速記情報センターの会議録検索システムを導入している市町村の Web ページでは，トップページもしくはトップページから直接リンクが張られているページに，各年度の会議録へのリンク一覧が存在するものがほとんどである．リンク一覧が存在しない場合，検索用の入力フォームに未記入で検索を実行することにより，全会議録の検索結果が表示される．これらのリンクを我々が構築したクロールプログラムに辿らせることにより会議録のページを自動的に取得した．人手による調整は，リンクの自動抽出部に埋め込まれたパタン等だけで済んだ．これは，市町村によりリンク一覧の掲載方法が若干異なるためである．

4.2 会議録研究所

会議録研究所の会議録検索システムを導入している市町村の Web ページでは，4.1節と同様に各年度の会議録へのリンク一覧が存在するものがほとんどである．これらのリンクを我々が構築したクロールプログラムに辿らせることにより会議録のページを自動的に取得した．人手による調整は，リンクの自動抽出部に埋め込まれたパタン等だけで済んだ．これは，市町村によりリンク一覧の掲載方法が若干異なるためである．

4.3 フューチャーイン

フューチャーインの会議録検索システムを導入している市町村の Web ページでは，会議録検索システムの出力が CGI プログラムにより自動生成されていて，その CGI プログラムに渡すパラメタにより出力内容を制御できる．例えば，会議録の検索は，CGI プログラムに以下のようなパラメタを渡すことにより行われている．

```
ACT=100&KENSAKU=0&SORT=0&KTYP=0,1,2,3,4&KG
TP=0,1,2,3,4&PAGE=1
```

CGI プログラムに渡すパラメタ PAGE の値を順次変えることで全ての検索結果を見ることができる．パラメタ ACT，KTYP の値はそれぞれ，ページの表示方法，会議種別に対応する．また，会議録は発言毎に分割されており，同じ CGI プログラムにおいて，次のようなパラメタを渡すことにより各発言を取得できる．

[2] http://www.yamatosokki.co.jp/
[3] http://www.kaigiroku.co.jp/
[4] http://www.futureinn.co.jp/
[5] http://www.sogosokki.co.jp/

表 2: 発言に付与する項目

項目	型	備考
発言 ID	int	自動採番
市町村コード	varchar	総務省により割り当てられた地方公共団体コード[6]
議会種別コード	varchar	定例会 0010, 臨時会 0020, その他 1000
年度	int	西暦
回	int	開催数
月	int	開催月
議会名	varchar	例：定例会，定例市会，予算委員会，経済環境協議会
号	int	会議が何日目か
日付	varchar	開催日
表題	mediumtext	ページのタイトル
段落番号	int	発言の段落番号
役職名	varchar	議員の役職
発言者名	varchar	発言者の氏名
議員 ID	int	あらかじめ議員に割り当てられた番号
ファイルのパス	varchar	元ファイルの保存場所
発言	mediumtext	1 文
その他	mediumtext	会議録内の発言以外の内容

```
ACT=203&KENSAKU=0&SORT=0&KTYP=2,3&KGTP=1,2
&TITL_SUBT=%95%BD%90%AC%82Q%82Q%94N%81@%82Q
%8C%8E%92%E8%97%E1%89%EF%81%7C03%8C%8E03%93
%FA-04%8D%86&HUID=46845&KGNO=&FINO=655&HATS
UGENMODE=0&HYOUJIMODE=0&STYLE=0
```

パラメタ TITL_SUBT, HUID の値はそれぞれ，URI エンコードされた表題，発言 ID に対応する．パラメタ ACT は，この例の「203」では「発言」，1 つ前の例の「100」では「検索結果」を指している．パラメタ TITL_SUBT は，この例では「平成 22 年 2 月定例会 -03 月 03 日 -04 号」を指している．我々のクロールプログラムにおいては，これらパラメタの値を順次変えることにより CGI プログラムを経由して会議録を自動的に取得した．

4.4 神戸綜合速記

神戸綜合速記の会議録検索システムを導入している市町村の Web ページでは，会議録検索システムの検索結果の出力が CGI プログラムにより自動生成されていて，その CGI プログラムに渡すパラメタにより出力内容を制御できる．そして，検索結果のページには会議録のページへのリンクが存在するため，そのリンクを辿ることで会議録の収集が可能となる．例えば，会議録の検索は，CGI プログラムに以下のようなパラメタを渡すことにより行われている．

```
treedepth=%95%BD%90%AC22%94N%20%95%BD%90%AC
22%94N%203%8C%8E%92%E8%97%E1%89%EF%20
```

パラメタ treedepth は URI エンコードされた和暦と表題に対応する．例では「平成 22 年 平成 22 年 3 月定例会」を指している．我々のクロールプログラムにおいては，このパラメタの値を変え，CGI プログラムが生成したページに張られたリンクを辿ることにより会議録のページを自動的に取得した．

5 会議録コーパスの付随情報

前章の方法により収集した会議録に対して表 2に示す付随情報を付与する．また，利便性を考えてデータベース化する．データベース化する際には，必要な発言のみを簡単に参照できるように会議録を発言単位に分割する．発言単位の分割については，句点や括弧等を区切りにしており，その際に HTML タグは全て取り除いている．以下，発言に付与する項目について説明する．発言 ID は各発言の識別を行うため，市町村コードは市町村ごとの検索のため，議会名は議会ごとの検索のため，議会種別コードは市町村によって名称の違う議会名を分類するためにそれぞれ必要となる．年度，回，月，号，日付については時間情報として重要なため必要である．表題はページのタイトルとして，段落番号は段落ごとの抽出を容易にするため，役職名は会議によって議員の役職が変わることがあるため，発言者名は会議録中の文字列をそのまま保持するため，また，発言者が議員であるとは限らないため，議員 ID は議員の識別のためにそれぞれ必要である．ファイルのパスは元ファイルを参照することを容易にするため，その他は発言とそれ以外の内容を区別するため，それぞれ必要となる．

例えば，図 2のような会議録が与えられたとき，下線部に対して表 3のように情報が付与される．発言 ID は会議録で 2 番目の発言であることを表している．市町村コードは総務省により割り当てられた地方公共団体コードを指す．議会種別コードは定例会と臨時会には個別のコードが割り当てられているが，その他の委員会は市町村によって名称が異なるためその他と一括りにしている．年度は表題に含まれる和暦を西暦に直している．回は表題に「第○回定例会」のように書かれているものもあるが，例のように「○月定例会」と書かれているものは，同一ディレクトリ内の定例会の開催月を比較して何回目であるかを推定している．議会名は表題から日付や回などを省くことで生成される．号はファイル名より会議が 1 日目であることを表している．日付はこの例の中には現れていないが，会議録の HTML タグの中に現れるものを抽出している．表題は会議録の HTML の title であるが，title がない場合は，ファイル名から「平成〜年○○会」までを抽出している．段落番号は
 タグをパタンで見つけて区切り，2 番目の段落であることを表している．役職名と発言者名は，発言者の発言の最初に「○役職名（発言者名　君）」のような形で表れるものを抽出している．議員 ID は全国の地方議員の一覧を用意しているので，全ての議員に

[6] http://www.stat.go.jp/index/seido/9-5.htm

○議長（桜井正富君）　次に、日程第３　議案第１号から日程第５５　議案第５３号までの５３案を一括議題といたします。
市長から、上程議案全部に対する提案理由の説明を求めます。
なお、提案理由の説明にあわせて、新年度に臨む所信の表明を行いたい旨の申し出がありますので、これを許します。
市長　宮島雅展君。
（市長　宮島雅展君　登壇）
○市長（宮島雅展君）　本日ここに、３月市議会定例会を開会するに当たり、私の市政運営に対する所信の一端と、議案第１号から議案第１５号までの平成２２年度予算案の概要につきまして述べさせていただきたいと存じます。

図 2: 甲府市議会会議録の例

表 3: 情報を付与した例

項目	値
発言 ID	2
市町村コード	19201
議会種別コード	0010
年度	2010
回	1
月	3
議会名	定例会
号	1
日付	0301
表題	平成 22 年 3 月定例会
段落番号	2
役職名	議長
発言者名	桜井正富
議員 ID	1
ファイルのパス	/ 甲府市議会 /2010/ 平成 22 年 3 月定例会 / 平成 22 年 3 月定例会（第 1 号）.txt
発言	市長から、上程議案全部に対する提案理由の説明を求めます。
その他	

割り振られている．ファイルのパスはコーパスのファイルの保存場所を示している．発言は 1 文の発言を示している．その他には発言以外の会議録の内容，例えば「（市長　宮島雅展君　登壇）」のような記述が入れられる．

2011 年 1 月時点で，会議録の収集は 51 市町村中 36 市町村まで完了しており，残りの会議録の収集と並行してデータベースの整備に取り掛かり始めている段階である．

6　おわりに

本稿では，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の構築の先駆けとして，全都道府県の県庁所在地と政令指定都市の計 51 市町村について会議録の収集とコーパスの構築を行っている．51 市町村の会議録は Web 上で主に 4 社の会議録検索システムにより提供されており，ページに張られたリンクを辿っていく方法と CGI のパラメータを変えていく方法等により，会議録を自動的に収集することを行い，2011 年 1 月時点で 36 市町村までの会議録の収集が完了し，残りの市町村の会議録の収集を継続している．会議録コーパスには 17 項目の情報を付与しており，また，発言単位に分割してデータベース化を行っている．今後は，全市町村の会議録の収集を目指すと共に，議員の政党や個人の Web ページ，ブログのような新たな情報についても収集することを検討している．

謝辞

本研究の一部は，科学研究費補助金 (No.22300086) の助成を受けたものである．

参考文献

[1] Keiichi Takamaru, Hideyuki Shibuki, Yasutomo Kimura, Dai Hasegawa, Hokuto Ototake, and Kenji Araki. Extraction of Political Activity of Assemblyman from Minutes of Municipal Assemblies Using the Political Category. *Paciling2009*, 2009.

[2] 乙武北斗, 高丸圭一, 渋木英潔, 木村泰知, 荒木健治. 地方議会会議録の自動収集に向けた公開パタンの分析. 言語処理学会第 15 回年次大会発表論文集 pp.192–195, 言語処理学会, 2009.

[3] 森脇俊雅. 合併と地方議会活動：議員アンケート調査の分析を中心にして. 日本選挙学会年報 選挙研究 第 23 巻, pp. 82–90, 2008.

[4] 川浦昭彦. Self-Serving Mayors and Local Consolidations in Hokkaido. 小樽商科大学・地域研究会 報告論文, 小樽商科大学, 2009.

[5] 平野淳一. 「平成の大合併」と市長選挙. 日本選挙学会年報 選挙研究 第 24 巻第 1 号, pp. 32–39, 2008.

[6] 木村泰知, 渋木英潔, 高丸圭一. 地方議員と住民間の協働支援に向けたウェブの利用. 選挙研究第 25 巻第 1 号, pp. 100–118, 2009.