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

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[‡]
[†]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: [†]{sugawara,oshiro,saito,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24年1月の時点で1,719[1]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，政治学では合併前後の違いに関する研究が数多く行われている [1, 2]．さらに，地方政治に関する研究は政治学以外にも経済学や社会言語学，情報工学の分野においても行われている [3, 4, 5]．これらの研究において，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及び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．本プロジェクトの全体像を図1に示す．構築する地方議会会議録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，地方議会会議録における議員の発言を中心とした政治情報システムに関する研究を行っている [6]．この研究では利用者の考えに近い議員を探し出すことができるシステムの構築を目指している．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．

地方議会会議録コーパスの構築に当たっては，我々が木村ら [7] 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，全都道府県の県庁所在地と政令指定都市の計51市の会議録について調査し，収集と整形を行った [8]．その成果を踏まえ，さらに地方議会会議録コーパスを拡充するために，51市が使用している会議録検索システムと共通した会議録検索システムを使用している自治体についても会議録の収集と整形を行うことにした．

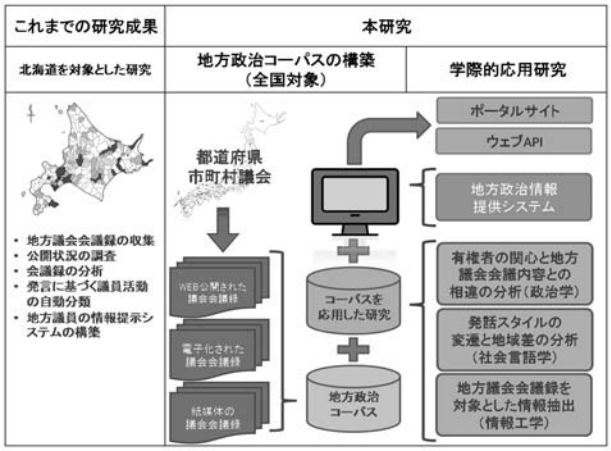

図 1: プロジェクトの全体像

本稿では地方議会会議録コーパスを構築する際の収集手法や整形手法を概観し，それら手法を用いる過程で発生した問題点を整理・分析をすることにより，どのように対処したかについて述べる．本稿の構成は次の通りである．2章では，関連研究について述べる．3章では，地方議会会議録検索システムの調査と収集・整形対象について述べる．4章では，収集手法における問題点とその対応について述べる．5章では，収集した会議録の整形における問題点とその対応について述べる．6章では，その他の市町村に収集を拡大した場合における取るべき収集・整形手法を考察する．7章で本稿のまとめとする．

2　関連研究

国会の会議録については国会会議録検索システム[2]が公開されている．一方で，地方議会会議録の公開形式は市町村毎に異なっているため，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に統一的な方法で閲覧することが困難であった．これに対し，複数の市町村を対象に横断的に検索することができる会議録ナ

[1] http://www.soumu.go.jp/gapei/gapei2.html

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

表 1: 主な自治体の会議録検索システム

会社名	自治体数 (政令指定都市 県庁所在市数)
会議録研究所[5]	379(23)
大和速記情報センター[6]	110(14)
フューチャーイン[7]	63(10)
神戸綜合速記[8]	66(3)
合計	**618(50)**

ビ[3]が公開されている．しかし，検索可能な市町村は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必要があり，検索条件もキーワードのみの文書検索である．類似する検索システムとして DiscussNet Cross Search[4]があるが，これは商用で自治体職員のみ利用可能である．そこで本研究では研究者が利用可能な，統一した書式に整形した地方議会会議録コーパスの構築を目指す．

これに関連して，乙武ら [9] 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形式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，本研究では全国規模の会議録の収集を目指す．

3 自治体が利用する会議録検索システム

自治体の会議録は，ウェブ上で専用の会議録検索システムを通して公開されている場合が多い．我々の調査の結果，多くの自治体は会議録検索システムに既存の市販システムを用いており，表 1に示すように大きく分けて 4 つの会社の会議録検索システムが使用されていることが分かった．我々は，政令指定都市 51 市の会議録を収集・整形した際に開発したプログラム [8] を活用して，この 4 つの会議録検索システムを使用している 51 市以外の自治体より会議録を半自動的に収集・整形することを試みた．なお，会議録検索システムで提供される会議録の多くは HTML 文書であり，HTML 文書の会議録を収集・整形対象としている．

4 会議録の自動収集における問題点

我々は大きく分けて次の二つの方法を用いて会議録の自動収集を行っている．

リンク解析手法 ページ中に含まれるリンクを解析し，リンク先のページを取得する．

パラメタ生成手法 ページ中に含まれる語句より CGI プログラムに与えるパラメタを生成し，そのページを取得をする．

51 市の会議録を収集する際に，会議録研究所や大和速記情報センターの会議録検索システムから会議録を収集するために主にリンク解析手法を用い，フューチャーインや神戸綜合速記の会議録検索システムから会議録を収集するために主にパラメタ生成手法を用いた．何故なら，それぞれの会議録検索システムに次のような特徴があったからである．まず，前者は市毎に使用している CGI やパラメタが異なる場合や，ページ遷移が生じる構成となっており，そのためにセッションを保つ必要がある場合がありパラメタ生成手法を適用することが難しい．その一方で，会議録が階層的な構造を有しており，それがリンクによって構成されているため，リンク解析手法を用いるのに適している．また，後者は JavaScript などを用いているためにリンク解析手法を適用することが難しいが，市毎に使用している CGI やパラメタが大きく違わず特定のページよりそのパラメタを推定できるため，パラメタ生成手法を用いるのに適している．

51 市の会議録を収集したこれらのプログラムをその他 568 の自治体の会議録の収集に適用した所，例えば会議録研究所の会議録検索システムを利用している自治体においては，リンク解析手法を用いたプログラムにおいて 63.3 %の自治体の会議録を収集することができた．これに対して，フューチャーインの会議録検索システムを利用している自治体においては，パラメタ解析手法を用いたプログラムにおいて 100 %の自治体の会議録を収集することができた．ここで，収集に失敗した自治体の会議録検索システムについて調査を行ったところリンク解析手法には次の様な問題点があることが分かった．まず，特定の自治体では，リンクやフォームなどリンク解析を行う対象である箇所に JavaScript が埋め込まれているために解析が正しく行われないことがあった．また，自治体毎に例えば日単位や発言者単位など会議録の公開単位が違う場合があり，その判定が正しく行われずに全ての会議録を取得できていないことがあった．さらに，リンク解析手法やパラメタ生成手法において文書より年一覧や議会名を正しく取得できる統一的なパタンを定義することは，自治体毎の差異が存在するため困難である．そこで，51 市の会議録を収集した際には，その HTML の構造や CGI のリンクに含まれる文字列をパタンとして含め利用していた．そのため，そのパタンに照合できない例外が存在すると正しくリンクを取得することができない．例えば，パラメタ生成手法を用いている神戸綜合速記の会議録検索システムでは議会名を取得するためにリンクに含まれる JavaScript をパタンの一部として用いていた．しかし，JavaScript の僅かな差異により議会名を取得できないことがあった．

我々はこれらの問題に解決するために次の 3 つの対応を行うことにした．1 つ目はリンク解析において JavaScript の解析も行うことの出来る既存ライブラリ，ここでは HtmlUnit[9]を利用することにした．その結果，リンクやフォームなどに想定していない JavaScript が埋め込まれていても，リンク解析を正しく行うことができた．2 つ目は公開単位の判定を行うのではなく，そのページに含まれる語句を用いてパラメタ生成手法を部分的に適用することにした．例えば，会議録研究所の会議録検索システムでは会議録のページにリンクを持たないアンカータグが埋め込まれており，その属性値よりパラメタを推定しパラメタ生成手法を適用することで，日単位で会議録を取得することができた．最後にいくつか

[3] http://www.db-search.com/oudan/
[4] http://www.kaigiroku.co.jp/contents/public03/dncs/
[5] http://www.yamatosokki.co.jp/
[6] http://www.kaigiroku.co.jp/
[7] http://www.futureinn.co.jp/
[8] http://www.sogosokki.co.jp/
[9] http://htmlunit.sourceforge.net/

の例外に対応するためにパタンの拡充を行うようにした．

5 会議録の自動整形における問題点

本プロジェクトでは地方議会会議録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，特に政治学や社会言語学の観点から考えて会議録中の発言を様々な粒度で参照できると良いと考える．そこで収集した会議録を発言の意味を保った最小の粒度と考える文単位に整形しコーパスとして保存している．しかし，会議録の収集・整形の作業が進行していくと，自治体毎や議会種別によって会議録の表記の違いが存在し，従来の整形形式を適用することが困難になる場合があることが分かった．

そこで，その問題に対応するために構築済みの 51 市の会議録コーパスと新たに収集した会議録を表 2に示す形式に整形・統合し，新たにデータベース化することにした．ここでは新項目についてのみ後述し，その他の項目についての詳細は文献 [8] を参照して頂きたい．本章では，まず従来の整形形式とその手法における問題点について述べ，その後でその対応と新項目の関係について述べる．

定例会や臨時会の会議録中の発言者については「議長（北市朗君）」といった「役職名（姓名 (+ 敬称)）」の統一的な表記で書かれていることが多い．敬称は「君」や「議員」など網羅可能なパタンであるため，そのパタンと記号を含めたパタンにより機械的に役職名と姓名を整形することが可能である．しかし，常任委員会などの会議録中では「佐藤市民生活部長」といった役職名や姓名の区切りが明確でない表記が存在する．役職名は多種多様であるため，統一的なパタンを定義し，機械的に役職名と姓名を整形することは困難である．また，発言者が議員であると推定できる場合には，その発言者の議員 ID の特定を行っている．ただし，議員 ID とは党派など議員の詳細情報を利用する為に我々が各議員に割り当てた ID である．しかし，発言者の中には議員 ID が割り振られていない議員が存在し，発言者は必ずしも議員であるとは限らないことより，その発言者が議員であることを識別することは困難である．

さらに，従来の整形手法にはいくつか不適切な場合があった．まず，従来は段落を単純に BR タグなどにより決定していた．そのため，例えば図 2の様な文は，従来の整形手法では行毎に異なる段落だと見なされていた．しかし，これらは同一の段落として参照可能であると考えるのが適当である．また，従来は HTML タグを全て取り除き整形を行っていた．しかし，会議録検索システム上で資料などを公開している自治体では表のデータを HTML の TABLE タグを用いて表現していたり，罫線などの文字列記号を用いて表現していたり，画像ファイルを用いて表現していたりする．そのため，TABLE タグを取り除かれた場合に表であったという情報が失われてしまっていた．

我々はこれらの問題に対応するために発言に付与する情報の追加と整形手法の改善を行うことにした．まず，役職名と姓名がパタンによる整形が困難である問題に対し，整形する際に利用した手がかりとなる文字列を発言者表層という新項目に登録するようにした．これは本コーパスを利用するシステム等でさらなる整形ができるようにすることも意図している．また，議員であるかどうかを識別する為に議員フラグという新項目を追加するようにした．さらに自動整形する際に，正しく整形されていない場合にその問題箇所を特定するために，整形プログラムとそのバージョンを整形プログラム名という新項目に登録するようにした．また，句点で終了しない文については同一の段落であるとし，HTML の TABLE に関連するタグは取り除かず同一の段落であるとして整形を行うようにした．そのため，図 2の様な文や HTML の TABLE を用いた表は段落番号を用いて一つのまとまりとして参照可能となった．

6 ウェブ上の会議録の収集・整形のための考察

我々は 4 つの市販された会議録検索システムを使用している自治体について会議録の収集と整形を行っている．2012 年 1 月時点で 618 件中 422 件の自治体の会議録について収集・整形が完了し，残りの自治体の会議録の収集・整形を継続している．しかし，『市町村議会会議録のウェブ公開とデータ提供に関するアンケート報告書[11]』によればウェブ上に会議録を公開している市町村は少なくとも 729 存在することが分かっており，会議録コーパスを拡充するためにさらに 100 以上の市町村の会議録が収集・整形の対象となる可能性がある．そのため，本研究の成果を活用し，どのようにウェブ上に公開されている会議録の収集・整形を試みるべきであるかを考察する．

まず，収集に関して述べる．乙武ら [9] の調査により，北海道内のウェブ上で会議録を公開している 63 市町村の内，4 つの市販された会議録検索システムを導入している市町村を除いた約 87 %が静的な HTML，又は PDF で会議録を公開していることが分かっている．それらの市町村のウェブページには会議録の全てのリンクが一つのページにまとまっているものも存在し，その収集は比較的に容易であると考える．また，ファイル名に命名規則が存在する市町村もあり，場合によっては会議録の URL を推定できる．それ以外の市町村においては，多くの会議録が階層的な構造を有していることから，本研究の様なより複雑な収集手法を検討しなければならない．その場合には，4 つの会議録検索システムからの会議録の収集において汎用性の高かったパラメタ解析手法の適用をまず検討し，その次にリンク解析手法の適用を検討するべきであると考える．ただし，リンク解析手法を適用しなければならない場合でも，HTML の構造や CGI のリンクに含まれる文字列などをパタンとして含め利用することは再利用可能性が低いことから避けるべきである．そのため，既に構築された会議録コーパスの「表題」や「議会名」の利用し，年一覧や議会名を取得するための統一的なパタンの発見を検討すると良いと考える．

最後に，整形に関して述べる．会議録の形式は「記号 役職名（姓名 (+ 敬称)）」によって発言者の情報が記述され，その後に発言が続くという形式が多い．そのため，PDF や HTML であっても不要な空白やタグを

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

[11] http://politics.kimura-s.otaru-uc.ac.jp/

表 2: 発言に付与する項目

新項目	項目名	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[10]
	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，予算委員会
	号	int	会議が何日目か
	日付	varchar	開催日
	表題	varchar	議会名の情報を含む文字列
	段落番号	int	発言の段落番号
	役職名	varchar	議員の役職
○	議員フラグ	int	議員ならば 1，それ以外は 0
	発言者名	varchar	発言者の姓名
○	発言者表層	varchar	発言者名を含む文字列
	議員 ID	int	あらかじめ議員に割り当てられた番号，対応がない場合 -1
	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
○	整形プログラム名	varchar	使用したプログラムとそのバージョン

さて、今回提出いたしました補正予算案は、国・県補助の確定に伴う経費、職員の給与改定に要する経費、その他緊急所要の経費の補正が主なる内容であります。これらの会計ごとの補正予算額は、
一般会計　　16 億 3,464 万 7,000 円
公営企業とそれ以外の特別会計
４億 2,850 万 6,000 円
合計　　20 億 6,315 万 3,000 円
でありまして、全会計の補正後の予算総額は、前年度の同期に比して 5.5 %の増加と相なっております。

図 2: 金沢市議会会議録中の一例

除くことができれば統一的なプログラムで整形を行えるため，本研究の成果を活用できると考える．さらに会議録コーパスの「役職名」を利用し，統一的なパタンの発見を検討すると良いと考える．

7　おわりに

本稿では地方議会会議録コーパスの拡充を行い，その収集手法や整形手法を概観し，それら手法を用いる過程で発生した問題点とその対応について述べた．2012 年 1 月時点で 618 件中 422 件の自治体の会議録について収集・整形が完了し，残りの自治体の会議録の収集を継続している．収集においてパラメタ生成手法よりリンク解析手法の方が問題が発生することが多かったため，リンク解析手法の改善や部分的なパラメタ生成手法の適用を検討した．また，整形においては会議録中の発言を新たに 20 項目の情報を付与しデータベース化を行い，既存の整形手法を改善した．最後に，更なるコーパスの拡充について，その方針を検討した．今後は，コーパスの拡充を目指すと共に，そのコーパスを利用した利用者の考えに近い議員を探し出すことができる政治情報システムの研究開発を行っていきたいと考えている．

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] 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.

[2] 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.

[3] 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.

[4] 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．第 27 回社会言語科学会研究大会, 2011．P-31.

[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, 第 26 巻, pp. 580–593, 2011.

[6] 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, 2012．P3-9.

[7] 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.

[8] 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 回年次大会発表論文集, 2011．P2-21.

[9] 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.