様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 ２６ 日現在

機関番号：３２２０７
研究種目：基盤研究(C)（一般）
研究期間：2017～2020
課題番号：１７Ｋ０２７３９

研究課題名（和文）地方議会会議録を核とした発言地域情報付きテキストコーパスの定量分析

研究課題名（英文）Quantitative Analysis of Textual Corpus with Geometric Information Focused on the Local Assembly Minutes

研究代表者
高丸　圭一（TAKAMARU, Keiichi）

宇都宮共和大学・シティライフ学部・教授

研究者番号：６０３８３１２１

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：延べ600人分の会話が記録された方言談話資料の電子化（テキストデータ構築）を行った。また、全国の地方議会会議録と発言位置情報付きツイートを収集し、整形を行った。文節単位または文単位のメタ情報付きデータを作成し、データベース化した。方言談話資料と地方議会会議録については発言者情報についてもデータベース化した。これを用いて、語彙の出現分布、オノマトペの出現頻度、文末表現などの変異を明らかにした。会議録の分析では、政治用語の出現頻度も比較し、地方政治学への寄与も検討した。

研究成果の学術的意義や社会的意義
本研究課題の特色は分析対象データの選定にある。一つは、既に研究を進めている地方議会会議録コーパスを分析データの核として引き続き利用している点である。これは本研究プロジェクトのもっとも大きな強みである。二つ目は既存の方言資料のデータベース化である。従来の方言研究の成果を活かしつつ、これまで定量的な評価を行っていない談話資料を総合的に分析する価値は高い。三つ目はＳＮＳデータの発信地情報推定手法の言語研究への応用である。これは新しい試みであり、現代もっとも多く使われているコミュニケーション手段であるＳＮＳにおける日本語の地域変異を捉えることを目指すことで新規性の高い研究となった。

研究成果の概要（英文）：We digitized (constructed text data) dialect discourse materials in which a total of 600 people's conversations were recorded. In addition, we collected and formatted local assembly minutes and geotagged tweets from all over Japan.
These data were devided into clauses or sentences and were added meta-information. Then They are stored to the database. For dialect discourse materials and local assembly minutes, we also created a database of speaker information. Using these data, we clarified variations in the distribution of lexical occurrences, the frequency of onomatopoeia, sentence-final expressions and so on. In the analysis of the minutes, the frequency of political terms was also compared, and their contribution to the study of local politics was also examined.

研究分野：社会言語学

キーワード：地方議会会議録　位置情報付きツイート　方言談話資料　電子化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本応募課題では､発言地域情報付きの大規模テキストデータベースを構築する｡これを用いて、言語情報の地域差に着目した定量的分析を行う。応募者はこれまでに地方議会会議録を方言資料として活用する研究に取り組んだ経験を有し、地方議会会議録コーパス、および、分析システムを構築済みである。本研究課題では、構築済みの地方議会会議録コーパスと電子化された方言資料（目的Ⅰ）および発言地域情報付きＷＥＢデータ（目的Ⅱ）を同じ枠組みでデータベース化し、分析する。地方議会会議録コーパスにおける取り組みでは、応募者らが構築した分析システムを介して、方言語彙の出現分布（図①)、オノマトペの出現頻度（図②)、2 人称の使用（図③)、文末表現（図④）などの変異を明らかにしてきた。また，議会会議録の分析では，政治用語の出現頻度も比較し，地方政治学への寄与も検討した。

２．研究の目的

(1) 方言学分野では、これまでにもコンピュータを用いた分析を目指して、方言資料の電子化を行うプロジェクトが複数行われてきた｡それぞれのプロジェクトを経て、電子化そのものはかなり進んできていると考えられるが、この電子化資料を活用した分析が十分に行われたとはいえない｡本応募課題では､応募者がこれまでの研究課題で積み上げてきた横断検索、KWIC 表示、クロス集計、出現頻度の自動可視化（地図化）の方法を利用し、方言資料をコーパス言語学的手法によって実証的・定量的に分析する。これにより、既存の方言資料に埋もれている知見を掘り起こすことを目指すものである。なお、現在も国語研究所を中心とした方言コーパス（方言音声コーパス）の構築が進められている。本応募課題は、それとはアプローチを異にし、コーパスに品詞や共通語訳のアノテーションをせずに、文字列を直接探索する方法で分析を進める。

(2) ＳＮＳ（ソーシャルネットワーキングサービス）における個人による情報発信は、いわゆる書きことばではなく、話しことば的な表現、および、ＳＮＳ特有の表現が用いられる。近年、このようなＳＮＳ上の言語現象の解明に取り組む研究は多く見られる。本研究では、それらとは目的を異にし、マーケティングなどにおいて盛んに利用されるＳＮＳの発信地情報推定手法を応用し、ＳＮＳにおいて個人が発信する情報に発言地域の情報を付与し、言語的地域差の発見を目指す。

３．研究の方法

地方議会会議録に加えて、既存の方言資料（方言談話資料や俚言集、方言辞典）を収集する。既に電子化されているものは、そのまま使用することができる。未電子化の資料は、本研究期間内に電子化作業を進める。方言資料は、方言研究のために作成されたデータ、または、方言研究の成果としてとりまとめられたデータであるため、言語的な地域差を分析するためのデータとして有用性が高い。

目的Ⅱのため、本研究期間に発言地域情報付きＷＥＢデータ（ツイートやブログ）を新たに収集整備する。ＷＥＢデータは大規模なテキストデータではあるものの、発言者属性情報が精緻に付与されたデータではないため、信頼性については留意が必要である。しかしながら、現代の活きた日本語であるＷＥＢデータから得られる、言語的な地域差に関する知見は、これからの日本語研究において、重要な価値を持つものであるといえる。研究計画（３・４）

方言資料、地方議会会議録、発言地域情報付きＷＥＢデータを、データベースに登録し、ウェブベースでの横断検索、KWIC 表示、クロス集計、出現頻度の可視化（地図化）を行える仕組みを整える。研究計画（５）

上記の仕組みを利用して､収集したデータを対象として語彙分布の定量的な検証を行う。語の出現頻度（この段階では形態素解析を行わない予定であるため、実際には部分文字列の出現頻度）を地図化することで、目視による効率的な洗い出しを進める。さらに、多変量解析によって地域差を検討する。また、方言資料から得られる結果と、地方議会会議録コーパスから得られる結果を比較分析することで、会議録に見られる地域差を方言的要素とその他の特徴（自然環境，地域特性に基づく使用語彙の差）に分離する。

共通語に存在する語が異なる用法・語義で用いられている場合（=気づかない方言)、表層的な出現頻度には明らかな違いがみられない。そこで、本応募課題ではコロケーションに着目する。これまでの研究から、語と語義の関係を明らかにする共起語として、(1)単文の述語、(2)（修飾語の場合）係り先名詞、(3)単文中の表層格（ガ格、ヲ格等）にあたる名詞が有効であることが分かっている。形態素解析器、および、係り受け解析器を用いてこれらを全探索し、比較することにより、気づかない方言の洗い出し、および、実態の解明を進める。研究計画（６・７）

WEB データを主たる対象として、出現頻度に地域差がある部分文字列を抽出する。言語学的知見になり得る地域差を手作業で洗い出し、検索システムで再度定量的に分析する。ここで得られる結果には、言語学的な特徴だけでなく、地域特有の固有名詞などが含まれる。これは言語学的な知見ではないものの、WEB データを発言地域別に分析した事例は多くないため、地域研究等で学際的に有効利用できる可能性がある。研究計画（８・９）

４．研究成果

(1)電子化・データベース化

本研究課題において「全国方言資料ＣＤ－ＲＯM版」（ＮＨＫ出版）に収録されている画像資料の文字化，データベース化を進めた。まず，データ形式について検討を行った。この結果，収録内容については，「ファイル名」「大分類」「番号」「小分類」「ページ」「出演者」「括弧内」「発音」「標準語」の９つのフィールドを持つテーブルに格納することとした。また「ファイル名」「大分類」「番号」「小分類」「出演者」「氏名」「生年」「職業」の８つのフィールドを持つテーブルに格納することとした。また、収録地は「地名」に加えて「GEO タグ」を付与した。

地方議会会議録については、分析期間，分析範囲の統制が取れるように，サブセットの作成を行った。議員については選挙において公開された情報に基づいて，発言者属性（生年，性別，選挙区）を付与し、多角的な分析を行えるようにした。

位置情報付きツイートについては，研究分担者の吉田が大規模な収集を行った。収集時に得られる位置情報に基づき，地域別の発言の分類を行った。

(2)分析と成果の公表

地方議会会議録データベースのサブセットを対象に，議員の議会活動の可視化の研究を実施した。この研究では，議員の年齢･性別構成を明らかにしたうえで各議員の発言文字数を集計し，議員の属性による発言量の違いを考察した。また，対数尤度比と tf-idf を用いて議員の属性ごとの特徴語を抽出し，議員の属性による発言内容の違いを考察した。その際，政治に関連する用語を抽出する指標として「政治語彙度」を新たに提案した。このほか，都道府県議会を経て国会議員になった議員を対象に，地方議会と国会での発言の比較について基礎的な検討を行った。地方に関係する国の施策について，地方議会で言及し，国会では自身の取り組みを紹介したり，更なる推進を求めたりする例や，ある特定の政治課題に関心を持ち続け発言していると見られる例が観察された。

全国方言資料および地方議会会議録に加えて，位置情報付きツイートデータを利用して，地方議会会議録，方言談話資料，位置情報付きツイートの横断的定量分析を進めた。この成果の一部を，人工知能学会全国大会，国際会議（LREC2018）等で発表した。地方議会会議録については横断分析システムをウェブサイトで一般公開した。また、分析結果の一部についてウェブサイトで公開している。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　0件／うちオープンアクセス　3件）

〔学会発表〕　計18件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 高丸圭一
２．発表標題 地方自治体の行政文書を対象としたLinked Open Documentsの考え方
３．学会等名 2019年 社会情報学会（SSI）学会大会
４．発表年 2019年

１．発表者名 木村泰知，渋木英潔，高丸圭一，秋葉友良，石下円香，内田ゆず，小川泰弘，乙武北斗，佐々木稔，三田村照子，横手健一，吉岡真治，神門典子
２．発表標題 NTCIR-15 QA Lab-PoliInfo2 のタスク設計
３．学会等名 言語処理学会第26回年次大会
４．発表年 2020年

１．発表者名 高丸圭一，内田ゆず，木村泰知，松田謙次郎
２．発表標題 地方議会と国会における同一議員による発言の比較に向けた検討
３．学会等名 第35回ファジィシステムシンポジウム
４．発表年 2019年

１．発表者名 Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Keiichi Takamaru, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura, Noriko Kando
２．発表標題 Influence of Classifiers and Encoders on Argument Classification in Japanese Assembly Minutes
３．学会等名 The fifth Linguistic and Cognitive Approaches to Dialog Agents Workshop
４．発表年 2019年

１．発表者名
竹内京子，高丸圭一

２．発表標題
IPA学習のためのカルタアプリ製作の検討

３．学会等名
第32回日本音声学会全国大会

４．発表年
2018年

１．発表者名
乙武北斗，高丸圭一，内田ゆず，木村泰知

２．発表標題
一般公開版「都道府県議会会議録検索システム」の概要

３．学会等名
第32回人工知能学会全国大会

４．発表年
2018年

１．発表者名
内田 ゆず，高丸 圭一，乙武 北斗，木村 泰知

２．発表標題
都道府県議会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

３．学会等名
第32回人工知能学会全国大会

４．発表年
2018年

１．発表者名
Yasutomo Kimura, Yuzu Uchida, Keiichi Takamaru

２．発表標題
Speaker Identification for Japanese Prefectural Assembly Minutes

３．学会等名
Proceedings of the Eleventh International Conference on Language Resources and Evaluation（国際学会）

４．発表年
2018年

１．発表者名
Keiichi Takamaru

２．発表標題
Demonstration of the Online Local Assembly Minutes

３．学会等名
Workshop 5: Hansards as a dialectal resource, the Sixteenth International Conference on Methods in Dialectology (METHODS XVI)（国際学会）

４．発表年
2017年

１．発表者名
井原 大将，内田 ゆず，高丸 圭一，木村 泰知，江崎 浩

２．発表標題
全地方議会会議録の横断検索に向けたデータ収集とデータ構造の検討

３．学会等名
情報処理学会第233回自然言語処理研究会

４．発表年
2017年

１．発表者名
木村泰知，内田ゆず，高丸圭一

２．発表標題
都道府県議会会議録のパネルデータ作成に向けた発言者情報の付与

３．学会等名
第33回ファジィシステムシンポジウム講演論文集

４．発表年
2017年

１．発表者名
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

２．発表標題
都道府県議会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

３．学会等名
第31回人工知能学会全国大会

４．発表年
2017年

１．発表者名
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

２．発表標題
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

３．学会等名
第31回人工知能学会全国大会

４．発表年
2017年

１．発表者名
Keiichi Takamaru, Yasutomo Kimura, Hideyuki Shibuki, Hokuto Ototake, Yuzu Uchida, Kotaro Sakamoto, Madoka Ishioroshi, Teruko Mitamura and Noriko Kando

２．発表標題
Extraction of the Argument Structure of Tokyo Metropolitan Assembly Minutes: Segmentation of Question-and-Answer Sets

３．学会等名
Proceedings of The 12th Language Resources and Evaluation Conference（国際学会）

４．発表年
2020年

１．発表者名
高丸圭一

２．発表標題
テキストデータの公開による実現可能な応用技術に向けたタスク設計

３．学会等名
2020年 社会情報学会（SSI）学会大会

４．発表年
2020年

１．発表者名
高丸 圭一，木村 泰知，内田 ゆず，佐々木 稔，吉岡 真治，秋葉 友良，渋木 英潔

２．発表標題
東京都議会会議録における議案への賛否を表明する発言の分析

３．学会等名
第34回人工知能学会全国大会，4Q3-GS-9-01（2020-06）

４．発表年
2020年

１．発表者名
内田ゆず；高丸圭一；乙武北斗；木村泰知

２．発表標題
都道府県議会会議録コーパスの拡張―2011期と2015期の比較―

３．学会等名
言語処理学会第27回年次大会（NLP2021）

４．発表年
2021年

１．発表者名
木村泰知；渋木英潔；高丸圭一；内田ゆず；乙武北斗；石下円香；三田村照子；吉岡真治；秋葉友良；小川泰弘；佐々木稔；横手健一；神門典子；森辰則；荒木健治；関根聡

２．発表標題
NTCIR15 QA Lab-PoliInfo-2の報告およびデータセット公開

３．学会等名
言語処理学会第27回年次大会（NLP2021）

４．発表年
2021年

〔図書〕　計1件

１．著者名	４．発行年
浜野祥子　小林隆　定延利之　半沢幹一　竹田晃子　高丸圭一　平田佐智子　友定賢治　小野正弘　川﨑めぐみ　田附敏尚　小西いずみ　有元光彦	2018年
２．出版社 ひつじ書房	５．総ページ数 256
３．書名 感性の方言学	

〔産業財産権〕

〔その他〕

地方議会会議録コーパスプロジェクト
http://local-politics.jp/
高丸研究室
http://www.takamaruzemi.com/
オノマトペを分析する研究グループの情報発信サイト
http://ono-collo.com/

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	木村　泰知 (Kimura Yasutomo) (50400073)	小樽商科大学・商学部・教授 (10104)	
研究分担者	内田　ゆず (Uchida Yuzu) (80583575)	北海学園大学・工学部・教授 (30107)	
研究分担者	乙武　北斗 (Ototake Hokuto) (20580179)	福岡大学・工学部・助教 (37111)	
研究分担者	吉田　光男 (Yoshida Mitsuo) (60734978)	豊橋技術科学大学・工学（系）研究科（研究院）・助教 (13904)	
研究分担者	井上　史雄 (Inoue Fumio) (40011332)	東京外国語大学・その他部局等・名誉教授 (12603)	

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関