形態素N-gramを用いた地方議会会議録コーパスの地域変異検出の試み
―文末表現を例に―

高丸圭一[1]　乙武北斗[2]　渋木英潔[3]　木村泰知[4]　森辰則[5]
[*1]宇都宮共和大学　[*2]福岡大学　[*3*5]横浜国立大学　[*4]小樽商科大学
[*1]takamaru@kyowa-u.ac.jp　[*2]ototake@fukuoka-u.ac.jp　[*3]shib@forest.eis.ynu.ac.jp
[*4]kimura@res.otaru-uc.ac.jp　[*5]mori@forest.eis.ynu.ac.jp

1.　はじめに

ウェブに公開された言語資源の活用や，検索サイトが提供する検索エンジン等のサービスを利用した言語研究が盛んになっている[1][2]。筆者らは，地方自治体の議会事務局がウェブに公開している会議録を収集・整形し，関係データベースに登録することにより，コーパスとして利用することを目指した研究を進めている[3]。地方議会会議録全文検索の言語研究への応用例として，[4]では「去った○日」という表現（「去る○日」の意）が那覇市の会議録に見られることを指摘しており，[5]では「めっちゃんこ」が名古屋市の会議録に見られることを指摘している。また，[6]では，「終わす」（「終わらせる」の意）が栃木県内の複数の議会会議録で観察されることを指摘している。このように全文検索を用いることで，既知の方言語彙が会議録に含まれていることを個別に確認することが可能である。

地方議会会議録コーパスは，全国各地の議員らの発言を収集したものである。発言者の居住地が明確であるため，表現や用語の地域差を研究するのに適している。大規模言語資源である会議録コーパスに内在する特徴を明らかにするために本研究では，コーパスから作成した形態素N-gramを用いて会議録の地域差を捉える方法について検討する。会議録には議論の内容（主題）を含めて，様々な地域差があると思われるが，地域差を捉える方法自体を検討することが目的であるので，差が見えやすいと考えられる文末表現に対象を限定する。例えば，関西方言の終助詞は関西で多く使われて，他の地域ではほとんど使われていないといった，文末表現と使用地域との関係を知ることができる手法を使って，全国の自治体を分析すれば会議録に内在した様々な地域差を捉えることができると考えられる。

2.　研究対象

2.1.　地方議会会議録

本研究では，プログラムによる自動処理によって収集・データベース化した地方議会会議録コーパス[7]を利用する。地域差のみを見るために2010年の会議録に対象を限定した。分析対象はすべての都道府県を網羅した405自治体である。

地方議会会議録は，議会での発言をすべて記録することを目的としている。しかし，議会を円滑に運営する目的で，議員の発言（質問）内容は事前に通告されており，読み上げ原稿が存在する発言が含まれる。また，整文の作業によって話しことばの特徴の一部が書きことば的に修正されている[8]（整文の指針[9]には「訛りは標準語に直す」という項目も存在する）。この2つの点において，会議録は厳密には自由会話の書き起こし資料であるとはいえない。地域差の検討には，議会会議録がもつこれらの性質を考慮に入れる必要がある。

2.2.　ひらがなの文末の形態素4-gram

本研究では，収集した会議録を形態素解析ツールMeCab[10]によって形態素に分割した。形態素解析辞書にはUnidic[11]を用いた。形態素解析の結果から，自治体別に会議録の形態素N-gramを作成した。

このN-gramから文末表現の地域差を検討するが，どの範囲が文末表現に相当するかは文によって異なる。大規模データから特徴抽出を行うためには統一した基準が必要である。終助詞等の文末表現の多く

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表3　出現頻度の高いフレーズ

順位	フレーズ	全国計	最大	最小
1	て／おり／ます／。	866,450	佐賀県 (0.1861)	和歌山県 (0.0687)
2	で／ござい／ます／。	652,018	高知県 (0.1883)	和歌山県 (0.0259)
3	で／あり／ます／。	323,651	富山県 (0.1504)	神奈川県 (0.0230)
4	て／い／ます／。	172,671	秋田県 (0.0537)	長崎県 (0.0097)
5	て／いただき／ます／。	110,858	奈良県 (0.0371)	鹿児島県 (0.0033)
6	ませ／ん／か／。	109,905	鳥取県 (0.0620)	佐賀県 (0.0067)
7	を／いたし／ます。	87,083	鳥取県 (0.0419)	青森県 (0.0030)
8	いたし／まし／た／。	70,804	鹿児島県 (0.0306)	石川県 (0.0061)
9	て／ござい／ます／。	62,891	東京都 (0.0258)	富山県 (0.0000)
10	あり／まし／た／。	61,340	和歌山県 (0.0233)	鹿児島県 (0.0041)

はひらがなであることから，ひらがなで構成される形態素に限定する。また，終助詞の連続を考慮してN=4とする。すなわち，本研究の分析対象は，第4形態素が句点であり，かつ，第1から第3形態素がすべてひらがなで構成されている4-gramである（以下，この4-gramを「フレーズ」と呼ぶこととする)。また，1都道府県あたり平均1回程度の出現が見込まれる表現，すなわち総出現頻度（全自治体の和）が50回以上のフレーズに対象を限定する。

3.　特徴の分析

3.1.　出現確率

405自治体の会議録からひらがなで構成された文末の形態素4-gram(フレーズ)の出現頻度を求めた。全自治体を合計すると4,433,834（異なり数25,341パタン）のフレーズがあった。このうち，合計出現頻度50以上のフレーズは，4,341,447（異なり数1,331パタン）であった。収集した会議録は自治体ごとに全体の量が異なるため，出現頻度そのものを自治体間で直接比較することは適当ではない。このため，次式によってフレーズの出現確率を求めた。

$$\text{出現確率} = \frac{\text{そのフレーズの出現頻度}}{\text{第4形態素が句点である4-gramの総出現頻度}}$$

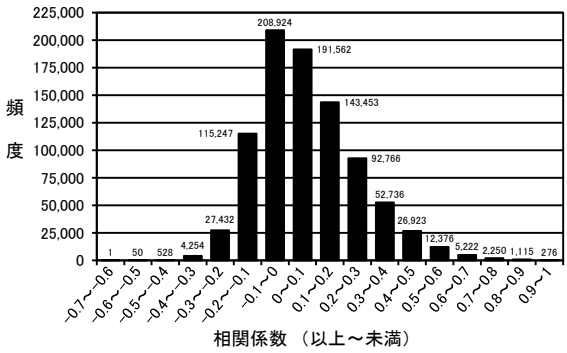

図1　フレーズ間の相関係数の分布

次に，出現確率を都道府県単位で集計した。表3に出現頻度が高い10フレーズを示す。また，このフレーズの出現確率がもっとも高い都道府県ともっとも低い都道府県を併せて示す。出現頻度が最も高い「ております。」と10番目の「ありました。」との間に約14倍の頻度の開きがある。また，この10パタンで,全1,331パタン出現頻度の和の約58%(2,517,671)を占めている。このことから，少数の文末表現が高頻度で繰り返し使用されていることが分かる。ただし，出現確率をみると，都道府県ごとにばらつきがある。例えば「ております。」では，佐賀県 (0.1861) と和歌山県 (0.0687) の間に約3倍の開きがある。

3.2.　フレーズ間の相関

フレーズの出現傾向（どの都道府県で多く出現するか）の類似性を明らかにするために，フレーズ間の相関係数を用いる。1,331フレーズの組み合わせ885,115組について，47都道府県の出現確率をパラメータとして求めた相関係数の分布を図1に示す。相関係数が最も高い組み合わせは「やけどね。」と「ますやん。」の0.992であった。一方，相関係数の絶対値が最も小さい組み合わせは「だけですね。」と「ないんでしょう。」の-6.77×10^{-8}であった。それぞれの散布図を図2(a)(b)に示す。

図2(a)では，関西方言である「やけどね。」と「ますやん。」が京都府，兵庫県，大阪府，およびその近県で出現し,その他の都道県では出現頻度が0である。図2(b)では，縦軸や横軸上にプロットされた都道府県―すなわち,「ないんでしょう。」と「だけですね。」

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

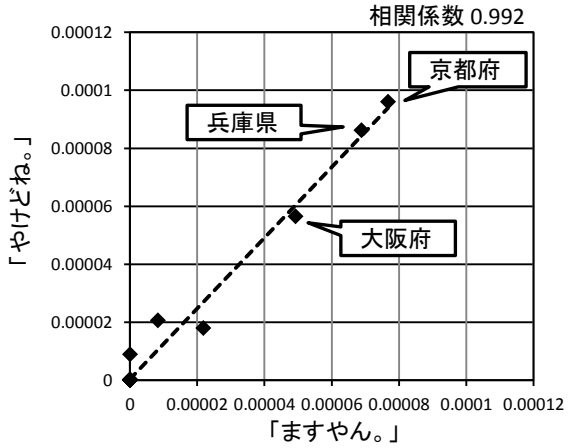

図2(a)　相関の高いフレーズ間の出現確率

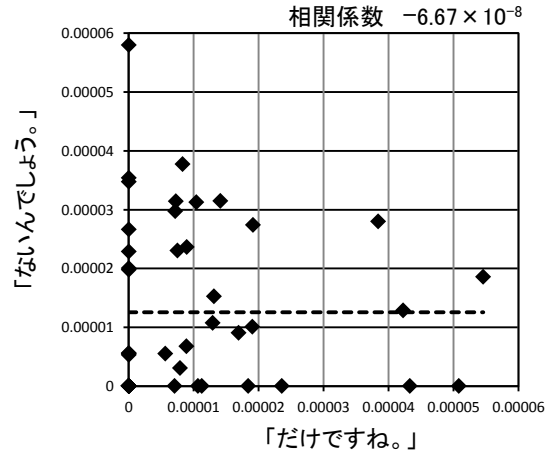

図2(b)　相関の低いフレーズ間の出現確率

のどちらか一方だけが出現し，他方は出現していない都道府県―も多く存在し，両方が出現している都道府県においても，出現確率に相関が見られないことがわかる。したがって，フレーズ間の相関係数を求めた上で，そのフレーズ対の出現確率が高い地域を調べることによって出現傾向に地域差のあるフレーズとそのフレーズが出現する地域を知ることが可能である。表4に相関係数の高い上位10組を示す。

上位10組には関西方言が多く観察される（①③④⑤⑧）。このほか，長崎県において「～である。」という文末表現が多く（②⑥⑩），宮城県では「～のです。」という文末表現が多いことが分かる（⑦⑨）。自治体別の4-gramを調査したところ，長崎県佐世保市議会の会議録に「だ／である」の文末表現が多いため，長崎県での出現確率が突出した。また，「～のです。」は宮城県議会や宮城県石巻市議会等の会議録に比較的多く見られた。これらの表現は他の都道府県でも出現し，出現地域が連続的に分布しているわけではないことから，いわゆる方言的特徴というよりも，議会発言における表現技法の偏り，または，整文規則の偏りによるものであると考える方が適切である。しかしこれらも会議録の興味深い地域差であることには違いない。

表4　相関係数の高いフレーズ（上位10組）

	フレーズ対	最頻都道府県	相関係数
①	「やけどね。」⇔「ますやん。」	京都府	0.992
②	「わけである。」⇔「ことである。」	長崎県	0.989
③	「わけやな。」⇔「ことやな。」	三重県	0.986
④	「んやな。」⇔「ことやな。」	三重県	0.985
⑤	「わけやね。」⇔「どないですか。」	兵庫県	0.979
⑥	「わけである。」⇔「のである。」	長崎県	0.978
⑦	「なのです。」⇔「あるのです。」	宮城県	0.976
⑧	「ねんけれども。」⇔「とるわけや。」	兵庫県	0.975
⑨	「なのです。」⇔「ないのです。」	宮城県	0.974
⑩	「ことである。」⇔「ていた。」	長崎県	0.971

3.3.　フレーズ間のネットワーク表現

3.2節で求めたフレーズ間の相関係数を利用して，出現傾向（出現の地域差）が類似しているフレーズをネットワーク図によって表現することを試みる。相関係数が高いフレーズ同士をネットワーク状に結ぶことで，出現傾向が類似したフレーズがクラスタを形成する。本研究で対象とする「文末のひらがな」という限定的な条件下であっても，フレーズのパタン数は膨大であるため，このような可視化手法を用いた分析は有用であると考えられる。

ネットワーク図の作成には，Cytoscape[12]を用いる。相関係数の下限を0.85に設定し，相関係数が0.85以上の696組（異なりフレーズ数207）とその相関係数から表5のようにクラスタが生成された。生成されたクラスタの中でもっとも大きいクラスタ（ノード数51）は，3.2節で述べた長崎県の会議録に見られる常体の表現である。また，次に大きいクラスタ（ノード数36）は関西方言を中心としたまとまりである。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表5　ネットワーク図の概要

※本表はネットワーク図内に，あるノード数のクラスタがいくつ生成されたかを示している。（例えば，ノード数が2のクラスタは15個生成された。）

クラスタ内のノード数	作成されたクラスタ数
2	15
3	6
4	2
5	2
6	3
12	1
24	1
36	1
51	1

また，宮城県に多い「～のです。」はノード数12のクラスタを形成した。このほか尊敬表現の「～してみえる」が愛知県，岐阜県，三重県，滋賀県に出現しており，クラスタを形成した。一例を図3に示す。紙幅の都合から本論文の範囲では，他のクラスタについて具体的に述べることは控えるが，クラスタ一つ一つが詳細に分析するに価値を持つ地域差であると考えられる。

4.　まとめ

形態素N-gramを用いて地方議会会議録の地域差を捉える方法について検討した。フレーズ間の相関係数は出現地域の類似性にしたがって値が変化するため，出現傾向の地域差を見つけることに適していた。また，ネットワーク表現は出現傾向が共通するフレーズを視覚的に概観するのに有用であった。

関西方言が整文で修正されずに京都府や兵庫県，大阪府などの会議録に出現することは予想通りであったといえるが，関西方言のうちどのような文末表現が会議録に多く出現するかは新たな知見となる。尊敬表現「～してみえる」が東海地方において整文されずに残されていることが確認された。また，長崎県で常体の表現が，宮城県で「～のです」がそれぞれ他地域と比べて多くみられることを発見できた。

この方法で明らかになった地域差が，発言の地域差であるか整文規則の地域差であるかまでは分からない。しかし，元の会議録に戻って調査することが可能であるし，議会の録画をウェブ上でオンデマンド配信している自治体もある。今後，本調査手法を出発点として地域差が見られるそれぞれの表現について詳細な分析を進める。さらに，同じ手法を応用して，一つの自治体の年代差の分析をすることで通時的な変化についても検討を進める。

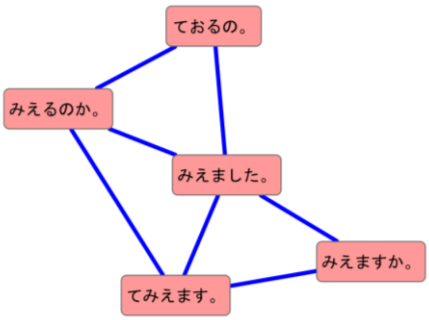

図3　ネットワーク図の例

謝辞　本研究の一部は，科研費 No:22300086 による。

参考文献

[1] 松田謙次郎・編(2010)『国会会議録を使った日本語研究』ひつじ書房
[2] 荻野綱男，田野村忠温・編 (2011)『コーパスとしてのウェブ』明治書院
[3] 木村泰知他(2012)「地方議会会議録コーパスの構築とその利用」第26回人工知能学会全国大会，3B3-NFC-4-3
[4] 井上史雄(2012)「[ことばの散歩道] 171 去った○日」『日本語学』Vol.31-10
[5] 山下暁美(2012)「なりすましの方言「めっちゃんこ」」地域語の経済と社会，第209回 (http://dictionary.sanseido-publ.co.jp/wp/2012/07/07/)
[6] 高丸圭一，木村泰知 (2010)「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』第10号, pp.74-86
[7] 齋藤誠他(2011)「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
[8] 高丸圭一 (2011)「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
[9] 野村稔・鵜沼信二(1996)『地方議会実務講座 第3巻』ぎょうせい
[10] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[11] 伝康晴他(2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22号 pp.101-122
[12] http://www.cytoscape.org/

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.