

〈論 文〉

形態素 N-gram を用いた地方議会会議録における
地域差の分析手法の検討

— ひらがなで構成された文末の 4-gram に着目して —

高 丸 圭 一

キーワード：地方議会，コーパス，形態素，相関係数，地域差

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（松田 2010，萩野・田野村 2011，井上 2011 など）。科研費採択課題「地方議会会議録コーパスの構築とその学際的应用研究」では，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（議会の議事録）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が行われている（木村・洪水・高丸・乙武・森 2012）。また，学際的な研究利用に向けて自治体横断全文検索や形態素 N-gram の整備が進められている。全文検索の言語研究への応用例として，井上（2012）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，山下（2012）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，高丸（2011）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の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の研究に適していると考えられる。これを日本語研究に有効活用するために，大規模言語資源に内在する特徴を明らかにする方法を確立する必要がある。本論文では，地方議会会議録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ことが予想される。関西方言が関西で使われていることを捉え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であろう。

2章で使用するデータについて述べた上で，3章で N-gram に基づく出現確率（3.1節），都道府県間またはフレーズ間の相関係数（3.2～3.3節），ネットワーク図による結果の可視化（3.4節）につい

て述べる。

2. 研究対象

2.1. 地方議会会議録

本研究では、齋藤・大城・菅原・永井・渋谷・木村・森（2011）が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。2010年の会議録が収集されている自治体の内訳を表1に示す。この405自治体はすべての都道府県を網羅している。

表1 分析対象の会議録

自治体分類	対象数
市	323
区	13
町	42
村	8
道県	19
合計	405

一般に会議録には、発言の要点だけをまとめたものと、発言の逐語録的なものがある。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おり、後者に該当する。しかし、「読み上げ原稿が存在する発言が含まれる」「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」という2つの点において、完全な自由会話の書き起こし資料であるとはいえない。

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、答弁者も質問通告に対する答弁をあらかじめ用意している。事前に発言内容が用意された発話では原稿を読み上げている部分が確かに存在する（高丸 2011）。原稿を読み上げた発話を書き起こしても話しことばの性質が含まれないが、発言者は必ずしもそのまま読み上げるわけではない。また、答弁に対する再質問など朗読によらない自由な議論も多く含まれる。

会議録は読みやすさの観点から整文と呼ばれる文章の修正が行われている。（高丸・木村 2010）。整文の作業では、言い直しや単純な言い間違い、助詞の誤用、冗長な言い回しなどそのまま文字化にすると読みづらい表現等は修正される。整文の指針（野村・鶴沼 1996）には「訛りは標準語に直す」という項目も存在する。ただし、すべてを書きことばのように修正するのではなく、発言の意図や議論の様子を記録するために、話しことばのまま記載される場合もある。高丸（2011）における議会発話音声と会議録の比較によると、整文前後のレーベンシュタイン距離（整文によって削除、追加、置換される文字数）は、会議録1,000文字あたり約60文字であった。会議録の作成は各自治体の中で行われるため、地域差の中でも特に気づかない方言はそのまま残る可能性が高い。

議会会議録がもつこれらの性質を考慮に入れて地域差を検討する必要がある。

2.2 形態素N-gram

形態素N-gramとは、ある文書においてN個の連続する形態素列が出現する頻度を求めたものであり、文書中の形態素の共起頻度や共起確率を知ることができる。本研究では、収集した会議録を形態素解析ツールMeCab⁽¹⁾によって形態素に分割した。形態素解析辞書にはUnidic（伝・小木曾・小椋・山田・峯松・内元・小磯 2007）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。形態素N-gramの例を表2に示す。

表2 形態素N-gramの例（北海道札幌市議会の7-gramの一部）

第1形態素	第2形態素	第3形態素	第4形態素	第5形態素	第6形態素	第7形態素	頻度
本市	の	出資	団体	数	に	つい	1
の	出資	団体	数	に	つい	て	1
団体	数	に	つい	て	ご	説明	1
数	に	つい	て	ご	説明	さ	1
に	つい	て	ご	説明	さ	せ	14
つい	て	ご	説明	さ	せ	て	14
て	ご	説明	さ	せ	て	いただき	25
ご	説明	さ	せ	て	いただき	ます	41
説明	さ	せ	て	いただき	ます	。	47
さ	せ	て	いただき	ます	。	</S>	117

※「</S>」は文末を示すマークでありN-gramでは形態素と同等に扱われる。

2.3 ひらがなで構成された文末の形態素4-gram

地方議会会議録の形態素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⁽²⁾。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、N-gramには出現頻度が小さいパターンが大量に含まれることが知られている⁽³⁾。このため、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パターン）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パターン）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。

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

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を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10倍以上の頻度の開きがある。また、この10パターンで、全1,331パター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。

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)

3.2 都道府県間の相関

フレーズ出現頻度の地域差をみるために、1,331フレーズの出現確率をパラメータとして都道府県間の相関係数を求める。相関係数は都道府県間における文末表現の使用傾向の類似度を表すと考えられる。全都道府県の組み合わせ1,081組の相関係数の分布を図1に示す。都道府県間の相関係数は全体的に高い正の相関を示しており、0.9を超える組み合わせが785組あった。最大値は0.997 (大阪府と兵庫県)、最小値は0.566 (新潟県と高知県) であった。相関が最も高い大阪府と兵庫県の出現確率の散布図を図2(a)に、相関が最も低い新潟県と高知県の出現確率の散布図を図2(b)にそれぞれ示す。

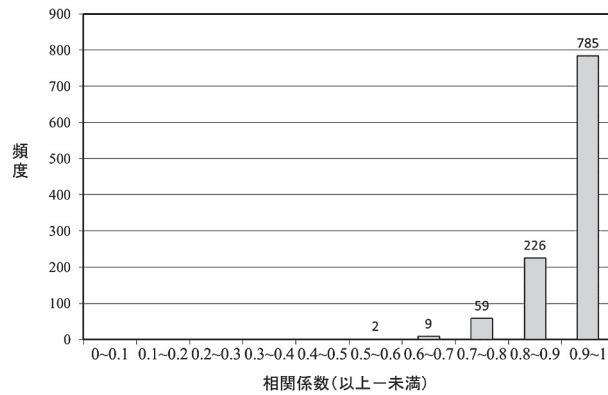


図1 都道府県間の相関係数の分布

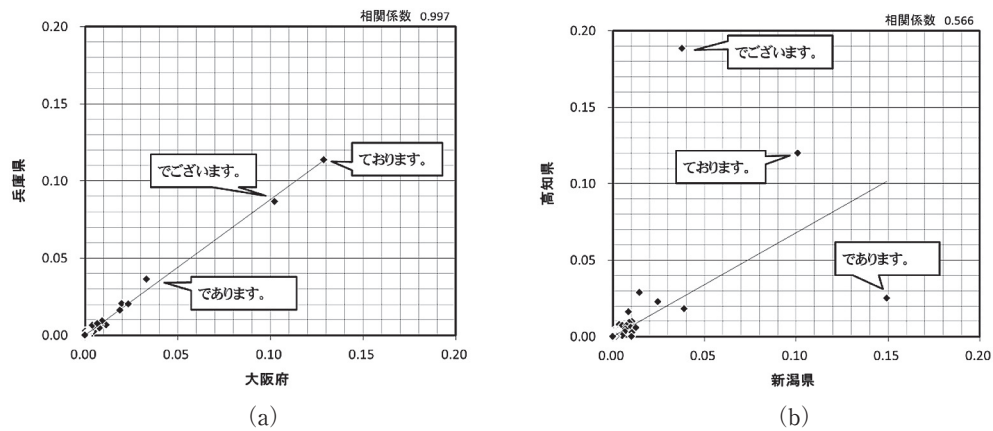


図2 都道府県間の出現確率の散布図

相関係数の低い図2(b)と比べると図2(a)は多くのフレーズが回帰直線の付近に分布しており、大阪府と兵庫県では、文末表現の使用傾向がかなり類似していることが分かる。しかし、(a)(b)ともに、大半のフレーズが原点付近に集中していることから、都道府県間の相関係数は、全国的に出現頻度の高い「でございます。」「ております。」「であります。」などのフレーズの出現のばらつきによって決定づけられているといえる。

方言差などの“ある地域には出現し、他の地域にはほとんど見られない”という特徴をもつフレーズはこれらに比べて出現頻度が低く、原点付近に存在していると考えられるため、相関係数に微細な差異を与えるに過ぎない。したがって、都道府県間の相関係数は、出現頻度の高いフレーズの使用傾向を説明できるものの、いわゆる地域差をよく表現する値とはいえない。

3.3 フレーズ間の相関

次に、行と列を入れ替えフレーズ間の相関係数について考える。1,331フレーズ間の組み合わせ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3に示

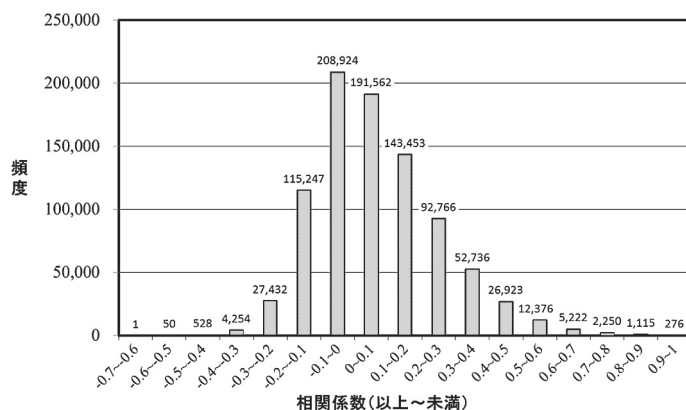


図3 フレーズ間の相関係数の分布

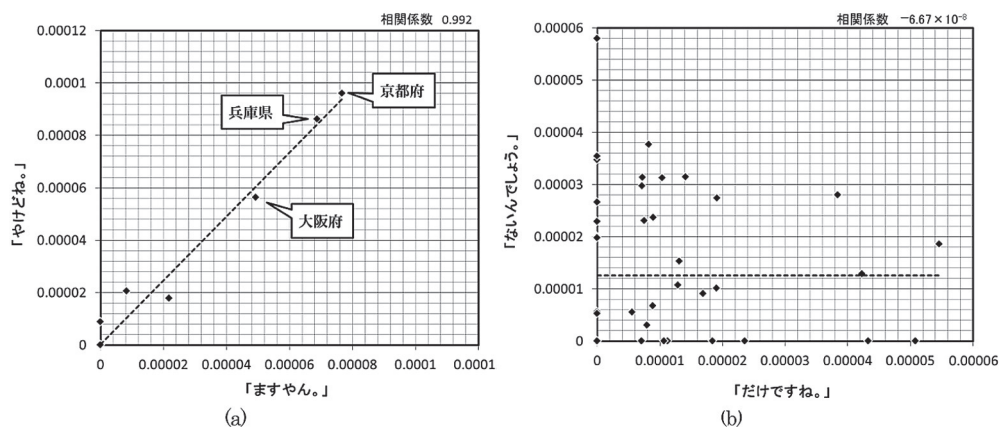


図4 フレーズ間の出現確率の散布図

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4(a)(b)に示す。

図4(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、その他の都道府県では出現頻度が0である。このように出現傾向の地域特性が共通している2つのフレーズの間では、相関係数が高くなる。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4 相関係数の高いフレーズ（上位10組）

	フレーズ1	フレーズ2	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

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。



図5 ネットワーク表現の考え方

3.4 フレーズ間の相関係数に基づくネットワーク表現

3.3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。図5に示すように、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、フレーズのパター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⁽⁴⁾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）とその相関係数を同ソフトウェアに入力した。生成されたクラスタの概要を表5に示す。

表5 ネットワーク図の概要

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

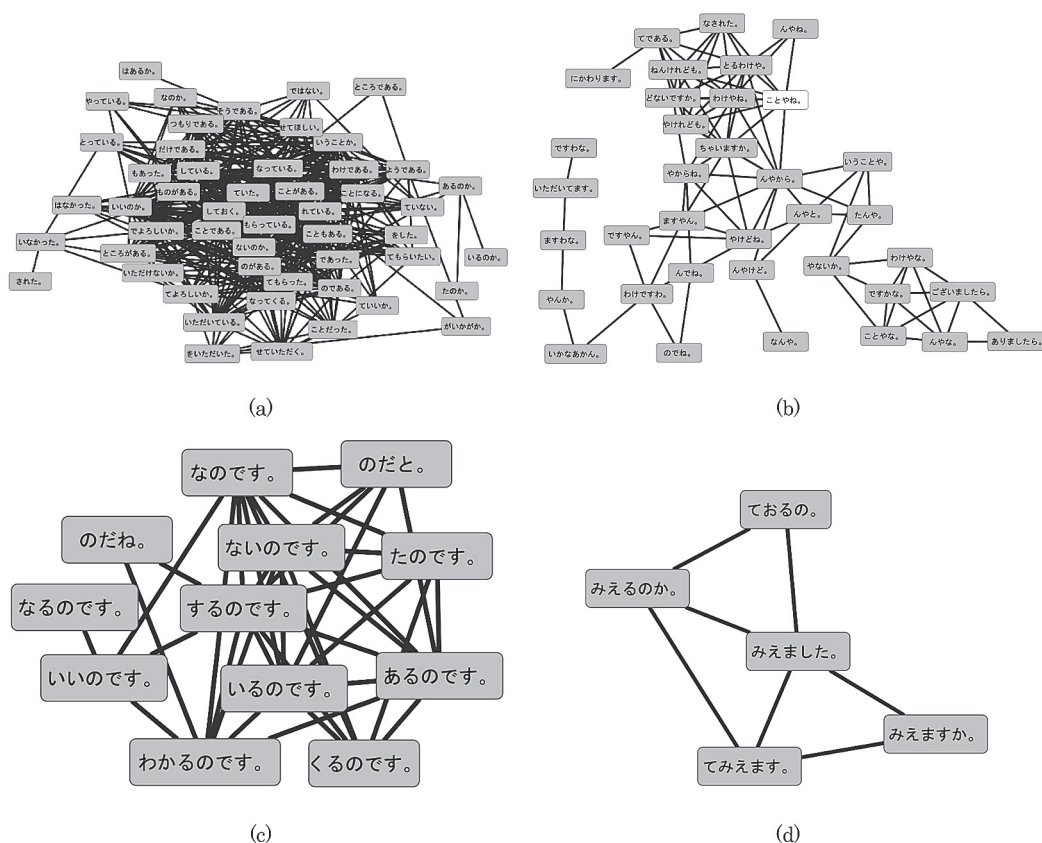


図6 相関係数が高いフレーズを結んだネットワーク図の例

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、3.3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（図6(a)）。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（図6(b)）。また、宮城県に多い「～のです。」はノード数12のク

ラストを形成した(図6(c))。このほか、東海地方で使われる尊敬表現の「～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、クラスタを形成している(図6(d))。

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、クラスター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。都道府県間のフレーズ出現確率の相関係数では、出現確率が全体的に高いフレーズのみが相関係数に強く影響するため、それ以外の地域差を見つけることが困難であった。一方、フレーズ間の相関係数は出現地域の類似性にし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。

関西方言が全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。東海地方で使われる尊敬表現「～してみえる」が全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全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。

謝 辞

本研究の一部は、科研費 No:23652102 および No:22300086 によるものです。

〈注〉

- (1) <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html>
- (2) 例えば札幌市議会では、最終形態素が句点であるN-gramのうち先行部分がすべてひらがなである割合は、2-gramでは99.5%、3-gramでは78.4%、同様に、4～7-gramでは、それぞれ67.6%、45.8%、32.1%、23.3%である。3-gram(文末2形態素のみを分析)では終助詞の連続を捉えきれない可能性があり、また5-gram以上(文末の4形態素以上を分析)では、ひらがなの連続といっても、ひらがなの内容語が含まれる割合が高くなると予想されるため、本研究では4-gramを用いることとした。
- (3) 計算機の処理能力の観点から、4-gramの生成時には足りきりを行ったため明らかではないが、例えば札幌市議会の3-gramでは、254,724パタンのうち約73%(185,722パターン)の出現頻度は1である。4-gram以上では出現頻度1のパターンはさらに増大する。
- (4) <http://www.cytoscape.org/>

参考文献

- 井上史雄 (2011)「Google 言語地理学入門」『明海日本語』 Vol. 16, pp.43-52
- 井上史雄 (2012)「[ことばの散歩道] 171 去った〇日」『日本語学』 Vol. 31-10
- 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
- 木村泰知, 渋谷英潔, 高丸圭一, 乙武北斗, 森辰則 (2012)「地方議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
- 齋藤誠, 大城卓, 菅原晃平, 永井隆広, 渋谷英潔, 木村泰知, 森辰則 (2011)「地方議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』, p.2-21
- 高丸圭一, 木村泰知 (2010)「栃木県の地方議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
- 高丸圭一 (2011)「規模の異なる自治体における地方議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
- 伝康晴・小木曾智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (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
- 野村稔・鶴沼信二 (1996)『地方議会実務講座 第3巻』ぎょうせい
- 松田謙次郎・編 (2010)『国会議録を使った日本語研究』ひつじ書房
- 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/>)