地方議会会議録におけるオノマトペの自動抽出手法の提案

Proposal of a Method for Automated Extraction of Onomatopoeia in Regional Assembly Minutes

◯ [1] 木村　泰知, [2] 渋木　英潔, [3] 内田　ゆず,
◯ [1]Yasutomo Kimura, [2]Hideyuki Shibuki, [3]Yuzu Uchida,
[4] 乙武　北斗, [5] 高丸　圭一, [2] 森　辰則
[4]Hokuto Ototake, [5]Keiichi Takamaru, [2]Tatsunori Mori

[1] 小樽商科大学
[1]Otaru University of Commerce
[2] 横浜国立大学
[2]Yokohama National University
[3] 北海学園大学
[3]Hokkai-Gakuen University
[4] 福岡大学
[4]Fukuoka University
[5] 宇都宮共和大学
[5]Utsunomiya Kyowa University

Abstract: An onomatopoeia is an useful linguistic expression to describe sounds, conditions, degrees and so on. Japanese has rich onomatopoeic expressions. They are frequently used in daily conversations. An onomatopoeia in a region may have a different meaning from one in different regions even if it is the same expression. Therefore, we attempt to investigate practical usage of onomatopoeias taken into account for the regional difference. However, general morphological analyzers cannot always recognize onomatopoeias. In this paper, we propose a method for automated extraction of onomatopoeias in regional assembly minutes. Although most previous work treats only four-letter onomatopoeias, our work treats not only four-letter onomatopoeias but also ones shorter than four letters.

1　はじめに

オノマトペ（擬音語および擬態語）は音，雰囲気，程度，様子を効果的に伝える手段であり，日本語の話しことばでは多用されることが知られている．近年，オノマトペの工学的な利活用を目指した取り組みが盛んである [1].

筆者らは現代の日本語におけるオノマトペの諸相を明らかにし，オノマトペを工学的に利活用することを目指して，地方議会会議録コーパスを対象としたオノマトペの分析を進めている [2]．地方議会会議録は都道府県議会または市区町村議会における議員や首長，行政職員などの発言を書き記したものである．発言者の属性（年齢・性別・肩書きなど）が明らかで，かつ，特定の自治体に居住する者の発言が，地域別・年度別に記録されている．地方議会会議録は自然言語処理，言語学，政治学等の様々な分野で利用すべき研究資源である．ただし会議録は自治体ごとに個別に提供されているため，横断的な研究は容易ではない．そこで近年，地方自治体がウェブに公開している地方議会会議録を収集・整形し，関係データベースに登録することにより，コーパスとして学際的に利用することを目指した研究が進められている [3].

筆者らの先行研究 [2] において，全国 402 自治体の 2010 年度の議会会議録（約 3 億語）を対象として，オノマトペの出現傾向を分析した．オノマトペ辞典 [4] に意味分類付きで掲載されている 1,751 語のオノマトペを形態素解析器 JUMAN[1] のユーザ形態素辞書にす

[1]http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 1: 先行研究における文字長と正抽出率の関係

文字長	正抽出数	誤抽出数	正抽出率
2	15	604	2.4%
3	580	1,549	27.2%
4	999	356	73.7%
6	61	0	100.0%

べて副詞として登録し，形態素解析によるオノマトペの抽出を試みた．地方議会会議録には事物の程度を表すオノマトペ，特に「しっかり」「どんどん」「はっきり」など政策の推進や適切な判断などに関わるとみられるオノマトペが高頻度で出現することが分かった．少数の都道府県に高頻度に出現する 61 語のオノマトペ（4,164 用例）の分析から，多義的に用いられ，かつ，地域によって語義の異なるオノマトペが確認された．しかし，この 61 語（4,164 用例）のうち 2,509 例 (60.3%) はオノマトペではない文脈で出現した。誤抽出は，(i) 方言に起因する解析誤り (1,527 例)，(ii) 名称・固有名詞の一部 (720 例)，(iii) 他のオノマトペの一部 (58 例)，(iv) 言い間違い・入力ミス等 (27 例)，(v) 同音異義語 (28 例)，(vi) その他 (149 例) の 6 パタンに大きく分類された．オノマトペの文字長別の正抽出率は表 1 のとおりであった．文字長の短いオノマトペと一致する部分文字列が文中に多数存在するため，3 文字以下のオノマトペでは正抽出率が低い結果となっている．

議会会議録におけるオノマトペの研究をさらに発展させるためには，より精度の高い抽出手法が必要である．そこで，本研究では形態素解析に加えて構文解析を利用することで精度の向上を試みる．

2 関連研究

形態素解析精度の向上や日本語学習の支援を目的として，オノマトペの用例を自動抽出するための研究が行われている．

奥村らは，Web からオノマトペの用例を収集し，オノマトペ概念辞書の自動構築を行った [5]．この研究では，オノマトペによく見られる音韻パタンを用いてオノマトペの候補語を生成し，それらを含む文を Web から抽出している．既存の辞書に掲載されていないオノマトペも抽出対象になるが，ノイズの除去に工夫が必要である．

香林らは，オノマトペの用例を日本語，英語，中国語，韓国語で表示するオンライン多言語辞書を開発した [6]．この研究では，小説から得た大量の用例を人手で分析し，オノマトペの用例を抽出しているため，質の高い辞書を実現している．

Asaga らは，オノマトペが用いられている文章を Web コーパスから自動抽出し，オノマトペ用例辞典を開発した [7]．この辞典は，単純な用例抽出手法を用いてオノマトペ用例文を Web から収集し，オノマトペを意味によって分類した結果をユーザに提示する．また，この研究の成果に基づき，80 語余りのオノマトペについて，用例文や共起する単語が一般に公開されている．

内田らは，ブログ記事を対象としたオノマトペ用例文の自動抽出手法を構築した [8]．この手法は，オノマトペの後続要素と係り受け関係を利用したものである．ドメインをブログ記事に限定し，オノマトペの係り先を制限することで，高い適合度での抽出を実現している．

3 手法

前章で述べたように，先行研究において幾つかのオノマトペ抽出 (収集) 手法が提案されているが，汎用的な抽出手法ではないため，これらを地方議会会議録にそのまま適用し，高い精度でオノマトペを抽出することは困難であると予想される．

そこで，本稿では形態素解析器と構文解析器を用い，以下に示す基本的な文法的特徴を利用した 2 つの規則のみによってオノマトペの抽出を試みる．

まず，会議録において，抽出対象のオノマトペと一致する部分文字列を「オノマトペ候補」とする．このうち，形態素解析に基づく規則 (図 1) と構文解析に基づく規則 (図 2) のいずれか（または両方）を満たしているものをオノマトペとして抽出する．

基本的な考え方としては，規則 A では，形態素解析によって 1 形態素として解析された副詞のオノマトペと名詞化して用いられているオノマトペを抽出する．また，規則 B でオノマトペ候補の文字列が文節の先頭であると解析されたものをオノマトペとして抽出する．規則 B は構文解析における文節まとめ上げ処理を利用するものであり，形態素解析が適切に行われなかった場合に有効な規則となる．文節まとめ上げ処理では，文として不自然な並びの形態素列は一つの文節としてまとめられる傾向にあるため，不自然な位置で解析されたオノマトペ候補を排除することができると考えた．

また，先に述べたように，3 文字以下のオノマトペは誤抽出が生じやすい．そこで本手法では，3 文字以

形態素解析によって 1 つの形態素に分割されており，以下の品詞に解析されているオノマトペ候補
(A-1) 4 文字以上の「副詞」「名詞」
(A-2) 3 文字以下の「副詞」

図 1: 形態素解析結果に基づく規則（規則 A）

文節の先頭にあり，以下の 3 条件のいずれかにあてはまるオノマトペ候補
(B-1) 4 文字以上である
(B-2) 3 文字で最後が「り」である
(B-3) 3 文字以下で直後に「と」「っと」「に」のいずれかが続く

図 2: 構文解析結果に基づく規則（規則 B）

下のオノマトペについて品詞 (A-2)，表層形態 (B-2)，助詞の接続 (B-3) の観点から誤抽出抑制規則を設ける．

なお，本稿の実験では形態素解析に MeCab(IPA 辞書)[2] を，構文解析に CaboCha(IPA 辞書)[3] をそれぞれ用いる．

4 実験データ

地方議会会議録コーパスの中で整形済みの文書数がもっとも多い，2010 年度の会議録を研究対象とする．2010 年度のコーパスには，すべての都道府県を網羅した 402 自治体（19 道県，323 市，13 特別区，42 町，8 村）の地方議会会議録が収録されており，データ数は 13,192,936 文（約 3 億語）である．膨大な文に含まれるオノマトペをすべて人手で確認し正解データを作成することは困難であるため，まず，JUMAN のユーザ辞書にオノマトペを登録した上で形態素解析を行い，オノマトペの予備抽出を行った．このうち，都道府県別の出現確率の和が 50×10^{-7} 以上 500×10^{-7} 未満の 177 語を本稿における抽出対象オノマトペとする．これは文書中に一定の出現頻度があり，かつ，全国に分布する―多様な方言や文脈の下で出現する―オノマトペを対象として実験を行うためである．

正解ラベルを作成するために，予備抽出したオノマトペ (177 語，全 18,792 発言) を人手によって確認した．共著者相互の合意に基づきオノマトペか否かを判定し，正解ラベルを付与した．「オノマトペ」は 12,261 発言，「非オノマトペ」は 6,437 発言であった．

[2]https://code.google.com/p/mecab/
[3]https://code.google.com/p/cabocha/

表 2: 実験結果

		実験結果	
正解ラベル	総数	抽出	非抽出
オノマトペ	12,261	**11,484**	777
非オノマトペ	6,437	1,390	**5,047**

5 結果と考察

5.1 抽出精度

18,792 発言に対して，提案手法を適用した結果を表 2 に示す．全体の精度は 88.1%であった．

予備抽出の手法（オノマトペをユーザ辞書に登録した形態素解析器のみを用いた手法）では，抽出精度が 67.5%(12,261/18,792) であった．これと比較すると本手法では 20.6 ポイントの精度の向上が見られた．

このうち，文字長が 4 文字以上のオノマトペ (例えば「ぴったり」「ごちゃごちゃ」) における精度は予備抽出で 91.12%，提案手法で 91.09%であった．一方，3 文字以下のオノマトペ (例えば「にやり」「じん」) では，予備抽出で 39.5%，提案手法で 85.4%であった．このことから，提案手法は 3 文字以下のオノマトペの抽出精度向上に有効であるといえる．

本手法は非オノマトペラベルのついた 6,437 例のうち，21.6%(1,390 例) を，誤抽出した．また，オノマトペラベルのついた 12,261 発言のうち，6.3%(777 例) は，提案手法で抽出することができなかった（未抽出）．これらの要因を形態素解析，同音異義語，その他に分けて以下に述べる．

5.2 形態素解析の問題

誤抽出例の多かった上位 15 語[4] のオノマトペのうち，「しとり」「かったん」「かんから」は方言文法（「～**しとり**ます。」「～言わへん**かったん**で～」「～せない**かんから**～」）を誤抽出したものである．形態素解析器の口語表現や方言文法への対応，または，解析誤りへ対応する誤抽出抑制ルールの追加が必要であると考えられる．

また，形態素解析誤りによって生じた未抽出には，「ばさっと」（ば [助詞] ＋さっと [副詞]），「ごちゃ」（ご [接

[4]たった，ごくごく，おいおい，しとり，さらさら，くすり，さんさん，かったん，かんから，とことこ，なんなん，たっ，ぽっ，とくとく

頭詞] ＋ちゃ[名詞]),「ずらずらと」(ずら [動詞] ＋ず [助動詞] ＋ら [名詞] ＋と [助詞]) などがある．これらは形態素解析の精度向上によって解消する可能性がある．

5.3 同音異義語の問題

オノマトペと表記が同一の副詞や名詞を誤抽出する例が見られた．誤抽出の上位 15 語を見ると,「たった」(「唯」の転化),「ごくごく」(極々)「おいおい」(追々), さらさら (更々) はオノマトペと表層的に一致する副詞である．また,「くすり」は「薬」,「なんなん」は「垂ん」のひらがな表記である.「さんさん」「とことこ」は, オノマトペに起因せずに制度や施策等の名称に使用される例が存在した．同音異義語については形態素情報や文節情報だけからオノマトペか否かを判断することは困難であり，この問題を解決するには，語の意味を考慮して，構文解析における係り先情報等を利用する必要があると考えられる．

同音異義語の影響による未抽出も見られた．例えば,「きらら」は「明るくまぶしく輝き続けているさま」[4] を表すオノマトペであるが，一般の辞書に掲載される意味は「雲母」の別称であり，名詞として解析されるため未検出となった．

5.4 その他

3 文字以下のオノマトペは，促音や長音を伴いながら連続して出現することがある．例えば,「ぽん」は「ぽん」「ぽんぽん」「ぽんぽんぽん」などの形で出現する．このうち「ぽん」と「ぽんぽん」はオノマトペ辞典 [4] に見出し語として掲載されている．本手法では,「ぽんぽん」のようなオノマトペを「ぽん [副詞] ＋ぽん [副詞]」と解析し，オノマトペ「ぽん」を 2 回抽出する．評価においては，オノマトペ「ぽんぽん」が未抽出であったと分類されるが，オノマトペの検出には成功していると考えることができるため，精度評価の方法については検討が必要である．

また，2 文字のオノマトペ「たっ」「ぽっ」は「**たっ**とい」(尊い),「まち**ぽっ**と」(組織名称) などが規則 (B-3) により誤抽出された．例えば「たっ」は対象の会議録中に 87,534 回出現しているが，そのうちオノマトペであるものの割合はそれほど高くないことが予想される．文字長の短いオノマトペの抽出規則については，出現事例を分析し，さらなる検討が必要である．

6 まとめ

本稿では，形態素解析と構文解析を利用した地方議会会議録からのオノマトペ抽出手法について検討した．177 語（12,261 例）のオノマトペを含む地方議会会議録から，88.1%の精度でオノマトペを抽出することができた．オノマトペの文字長を考慮した規則の適用により，短いオノマトペにおいて誤抽出が大幅に削減された．オノマトペをユーザ辞書に追加した形態素解析による手法に比べ，3 文字以下のオノマトペの抽出精度が 45.9 ポイント向上した．

誤抽出および未抽出の分析から，形態素解析において口語表現や方言文法への対応が必要であることを指摘した．また，同音異義語に対応するために，構文解析によって得られる係り先情報の利用を検討する必要がある．係り先の動詞の情報を利用してオノマトペを抽出する手法 [8] などを参考に，高精度な抽出を目指す．さらに，未知の（あらかじめ辞書に登録されていない）オノマトペの探索手法，複数の語義を持つオノマトペにおける語義の曖昧性の解消手法などの検討を進め，日本語オノマトペの使用の地域差，語義の地域差などの分析にも取り組んでいく予定である．

謝辞

本研究の一部は科学研究費 No.25370524 および No.26370498 による．

参考文献

[1] 小松孝徳，中村聡史: OS-08「オノマトペの利活用：オノマトペ研究の分野横断連携を目指して」，人工知能学会誌 27(6), pp.653-654 (2012)

[2] 高丸圭一，内田ゆず，乙武北斗，木村泰知: 地方議会会議録におけるオノマトペの出現傾向に関する基礎的検討，言語処理学会第 20 回年次大会，pp.566-569 (2014)

[3] 木村泰知，渋木英潔，高丸圭一，乙武北斗，森辰則: 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 (2012)

[4] 小野正弘編: 日本語オノマトペ辞典，小学館 (2007)

[5] 奥村敦史, 齋藤豪, 奥村学: Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築, 情報処理学会研究報告, Vol. 2003, No. NL-154, pp.63-70, 2003.

[6] 香林隆子, 増永良文: オノマトペのオンライン多言語辞書の構築, DEWS2002 論文集, A4-4, 2002.

[7] C. Asaga, M. Yusuf and C. Watanabe: Onomatopedia: Onomatopoeia Online Example Dictionary System Extracted from Data on the Web, The 10th Asia Pacific Web Conference, 2008.

[8] 内田ゆず, 荒木健治, 米山淳: ブログ記事からのオノマトペ用例文の自動抽出手法, 知能と情報（日本知能情報ファジィ学会誌）, Vol. 24, No.3, pp.811-820, 2012.

連絡先

高丸圭一

E-mail: takamaru@kyowa-u.ac.jp