3G4-OS-05b-5

BCCWJ コアデータにおけるオノマトペ出現実態の分析

現代オノマトペ実例辞書アプリ構築に向けて

内田ゆず[*1]
Yuzu Uchida

高丸圭一[*2]
Keiichi Takamaru

[*1] 北海学園大学
Hokkai-Gakuen University

[*2] 宇都宮共和大学
Utsunomiya Kyowa University

乙武北斗[*3]
Hokuto Ototake

木村泰知[*4]
Yasutomo Kimura

[*3] 福岡大学
Fukuoka University

[*4] 小樽商科大学
Otaru University of Commerce

An onomatopoeia is a useful linguistic expression to describe sounds, conditions, degrees and so on. It is said Japanese is rich in onomatopoeic expressions. They are frequently used in daily conversations. We aim to develop a collocation dictionary of Japanese onomatopoeia with abundance of examples. This paper reports the detailed analysis of onomatopoeias in BCCWJ and the onomatopoeia extraction method.

1. はじめに

オノマトペ（擬音語・擬態語）は音や程度，状態を効果的に伝達する手段であり，豊かな日本語表現には欠かすことができないものである．近年，オノマトペを様々な分野で利活用することを目指した研究が進められている[小松 15]．

オノマトペには多様な語義をもつという特徴がある．例えば日本語オノマトペ辞典[小野 07]の「ごろごろ」の項目には 6 つの語義が掲載されている（「雷の響く音」「猫がのどを鳴らす音」等の擬音の語義と「無造作に転がっているさま」「仕事をせずに無駄に暮らしているさま」等の擬態の語義）．[高丸 15]の地方議会会議録コーパスにおける「ごろごろ」を含む文の分析では，辞典中の語義に加えて「たくさんある（いる）さま」，「変わりゆくさま」などの語義が見られた．このように 1 つのオノマトペは擬音，擬態の語義を持ち，さらにそれらから派生した語義や新たな語義が追加されることがある．また，語義が類似したオノマトペが多数あるという特徴もある．例えば「ごろごろ」に対して，「ころころ」「ごろんごろん」「ごろっ」は類似の語義をもつものの，それらが表現する様子や修飾できる語はやや異なると予想される．これらのことは日本語母語話者にとっては直感的に理解可能であるが，日本語学習者にとっては理解が容易ではない．また，対話システムにおける文生成処理においてもオノマトペを適切に利用することは容易ではない．オノマトペの語義については，文の係り受け関係を利用して，あるオノマトペの擬音的用法と擬態的用法を区別する研究[Fukushima 14]や，SD 法によってオノマトペの語義を定量的に表現する研究[清水 14]が進められている．

日本語非母語話者がオノマトペを適切に使用するためには，あるオノマトペがどのような場面で使用可能であるかという実例を示すことが重要であるし，対話システムの文生成処理においても，前方および後方の文脈に基づいて，適切なオノマトペを選択する必要があると考えられる．そこで，筆者らはオノマトペの実際の用例に着目した研究を進めている．現代の日本語における最新の用例を収集するために，ウェブ上の文書からオノマトペを抽出する．オノマトペを含む用例文から，例えば「ごろごろ」＋「寝る」，「ごろごろ」＋「転がる」という係り先のコロケーションや，「石が」＋「ごろごろ」，「雷が」＋「ごろごろ」という係り元のコロケーションを抽出し，そのオノマトペが使用できる文脈を明らかにする．人間がオノマトペを学習する際には，さらに各コロケーションの具体的な例文を提示することで，語義を計り知ることが可能であろう．また，「A が」＋「ごろごろ」＋「転がる」と，「B が」＋「ころころ」＋「転がる」という共起を考えたときに，「ごろごろ」と共起する単語集合 A と，「ころころ」と共起する単語集合 B の差異を見れば，2 つのオノマトペの意味の違いを理解することにつながると考えられる．このような観点から，本研究では現代のオノマトペの最新の用法を提示できるウェブ上の実例に基づく辞書の構築を目指す．ユーザの利便性を考慮して，携帯端末上で動作するアプリケーションの構築も視野に入れている．

オノマトペ実例辞書構築のためには，まずウェブ上の文書からオノマトペ抽出処理を行い，「オノマトペ用例データベース」を構築する．オノマトペは文字長の短いひらがな／カタカナの文字列であり，特殊拍（促音・撥音・長音）が挿入により変形が可能であるため，文書中からオノマトペを正確に抽出することは難しい．ブログ[内田 12]や議会会議録[木村 14][池田 15]からオノマトペを自動抽出する手法が検討されているが，更なる検討が必要な点である．

「オノマトペ用例データベース」内の文に対して，係り受け解析や共起する単語の纏め上げを行うことで，オノマトペ実例辞書に必要なコロケーションデータを得る．大規模言語資源とコロケーションに関する先行研究には[田野村 10]，[郜 06]などがある．[田野村 10]では，ウェブコーパスから得られるコロケーション情報からのコロケーション辞典作成の手法について幾つかの具体例を元に考察している．[郜 06]では，「しんみり」「しみじみ」の 2 語を対象に新聞コーパスにおけるコロケーション（共起する動詞）を調査し，アンケート調査によって得た人間が想起する係り先の動詞と比較している．

本稿では，オノマトペコロケーション抽出の出発点として，「現代日本語書き言葉均衡コーパス（BCCWJ）」のコアデータに含まれる全てのオノマトペの表層形態を分析する（3 章）．さらに，

連絡先：内田ゆず，北海学園大学工学部電子情報工学科，札幌市中央区南 26 条西 11 丁目 1 番 1 号，yuzu@hgu.jp

この結果に基づき，品詞情報を利用してオノマトペの抽出を行い(4 章)，抽出手法の拡張を試みる(5 章)．最後に，コアデータから得られるコロケーションの例について触れつつ結論を述べる(6 章)．

2. 対象データ

本研究で使用するデータについて説明する．

2.1 コーパス

本研究で分析対象とするコーパスは，大学共同利用機関法人人間文化研究機構国立国語研究所と文部科学省科学研究費特定領域研究「日本語コーパス」プロジェクトが共同で開発した『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese，以降 BCCWJ)である．BCCWJ には，現代の日本語の書き言葉の全体像を把握できるように集められたサンプルが書籍全般，雑誌全般，新聞，白書，ブログ，ネット掲示板，教科書，法律などのジャンルにまたがって1億430 万語収録されている．

なお，BCCWJ には人手で形態素解析結果を修正したサブセットであるコアデータが含まれている．コアデータは約 9 万短単位のデータである．

2.2 オノマトペ辞典

ある単語がオノマトペであるかを判断する際に，日本語オノマトペ辞典[小野 07]を基準として用いる．この辞典には古事記などの古典から現代に至るまでのオノマトペが掲載されており，見出し語は 4,564 語となっている．

この辞典には 2 種類の索引がある．

一つ目の「意味分類別さくいん」は，辞典に収録されている見出し語のうち，延べ 2,470 語(異なり 1,751 語)を採り上げ，自然・人間・事物に三分類し，それぞれに簡略な解説を付してあるものである．一般性の高い語が厳選されたオノマトペ集合と考えられる．

二つ目の「五十音順さくいん」には，辞典の本編ならびにコラム，付録(漢語オノマトペ，鳴き声オノマトペ)に収録されている全 4,506 語が掲載されている．漢語オノマトペを含む表現(例：焔焔に滅っせずんば炎炎を若何せん)や，オノマトペではないがコラムで言及されている語(例：あいまい)も対象であるため，語数は多いがオノマトペとして不適切なものも含まれている．

3. コアデータの全オノマトペ分析

我々は本研究に取り組むにあたり，BCCWJ に出現するオノマトペの傾向を概観するため，意味分類別さくいんに掲載されたオノマトペ(1,751 語)のコアデータにおける出現傾向を分析した．紙面の都合上詳細は割愛するが，392 語のオノマトペが延べ 1,370 回出現すること，2・3 文字の短いオノマトペを抽出するためには品詞等の情報が必要であることが明らかになった．一方で，意味分類別さくいんに掲載されていないオノマトペの使用実態は明らかになっていない．したがって，本章ではコアデータ中の全てのオノマトペを抽出し，それらの品詞を分析する．

コアデータ中の 2 文字以上のひらがな・カタカナからなる短単位形態素を全て抽出し，それらがオノマトペであるかを人手で判断する．この分析によって，コアデータ中の全てのオノマトペ(つまり，正解データ)を得ることを意図している．

分析の結果，198,829 個の短単位が抽出され，そのうち 2,048 個がオノマトペであると判断された．意味分類別さくいんに掲載されていないオノマトペが 182 語，延べ 678 回出現している．

意味分類別さくいんには掲載されていないが五十音順さくいんに掲載されているものは以下の 101 語である．これらのオノマトペは，五十音順さくいんを導入することで抽出が可能になる．

あつあつ，あっさり，いちゃいちゃ，がが，かくかく，かっちり，がっつり，がらっ，がらん，ぎざぎざ，ぎっくり，きっちり，きりっ，きりり，ぐいっ，くっきり，くったり，くるくる，ぐるぐる，くるっ，ぐるっ，ぐるり，くんくん，ぐんなり，ごうごう，こじんまり，こぢんまり，こっそり，こてんぱん，ささっ，しっくり，じっと，しゃなり，しゅわしゅわ，ずしり，すっかり，すっきり，ずっしり，ずっと，すっぽり，すぱすぱ，すべすべ，すぽっ，するっ，すれすれ，そっくり，そろり，だらり，ちぐはぐ，ちゃんと，ちゅんちゅん，ちょい，ちょくちょく，ちょこまか，ちょんちょん，つるっ，でこでこ，てっきり，でれでれ，でん，どきん，とことん，どっかり，どっしり，とんかち，どんぴしゃ，にょろにょろ，ぱかっ，ぱっくり，ばったり，ばっちり，ばばば，ばん，びしばし，ひしひし，ひっそり，ぴぴ，ひょっこり，ぴょんぴょん，ぷすん，ぷちぷち，ふつふつ，ふらっ，ふらり，ぺこり，ぺしゃぺしゃ，ぺしゃんこ，べちゃり，ぼうぼう，ぽか，ぽちっ，ぼちぼち，ほっこり，ぽっぽ，ぽつりぽつり，まったり，まんまん，むちゃくちゃ，もちもち，もっちり，もんもん

意味分類別さくいんにも五十音順さくいんにも掲載されていない語は以下の 81 語である．「きちんと」や「くりくりっ」はそれぞれ索引に掲載された「きちん」，「くりくり」に助詞「と」，促音「っ」を付与することで対応できる．このように，一部のオノマトペは単純なルールで抽出が可能になる．一方，「ごふっ」や「ぷんすか」などは比較的新しい表現だと考えられ，このような新出オノマトペを抽出する手法の確立が求められる．

うがうが，うんと，かぁん，がたがたがた，がちゃ，かっかっかっ，かっと，きちっと，きちんと，ぎゃあぎゃあ，きゃっきゃ，ぎゅ，ぎょっと，ぐらぐらぐら，くりくりっ，ぐるぐるぐるっ，ぐんと，こつ，ごふっ，こりこりっ，ささささささっ，さっさと，しんと，じんと，すうっ，ずうっと，ずず，すつかり，ずらずらっ，せっせと，そっと，ぞっと，たたったった，ちびり，ちょいと，ちょこっと，ちら，つるんつるん，てれん，てんかん，とっとと，どよどよ，とんとんとん，のうのう，はたと，ぱっきり，ばっさ，ばっちし，はっと，ばばっ，ぱらぱらっ，ぱんっ，びいい，ひいひい，ぴか，びくと，ぴぽぴぽ，ひょっと，ぶぅぶぅ，ぶふぉっ，ぷぷっ，ぷらっと，ふらふらっ，ぶろろろ，ぷんすか，ぺこ，ぺたりんちょ，べろりっ，ぼうっと，ぼそ，ほたり，ぽち，ほっと，ぽにょ，ぽぽん，ほわり，むっちゃ，めちゃ，めっちゃ，めっちゃめちゃ，よよと

図 1 にオノマトペであると判断された短単位の品詞の割合を示す．すべての短単位が副詞，形状詞，名詞のいずれかに分類され，88.8%は副詞である．品詞を抽出の条件に加えることで，短いオノマトペの抽出精度を向上させることが期待できる．

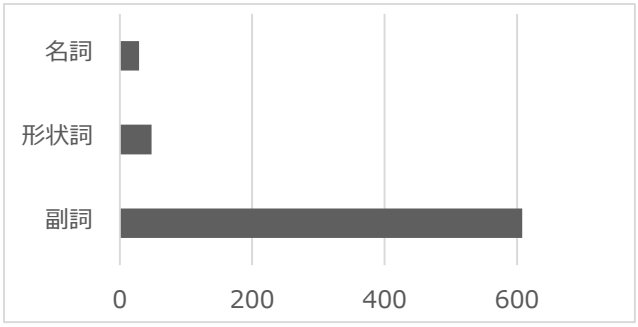

図 1　オノマトペの品詞

4.　品詞情報を利用したコアデータからのオノマトペ抽出（ベースライン）

3.の結果に基づき，品詞情報を利用したオノマトペの抽出実験を行う．コアデータに MeCab[Kudo 04]（Unidic 辞書）で形態素解析を施し，五十音順さくいんに掲載されているオノマトペと字面が一致し，かつ副詞か形状詞になった短単位を人手で分析する．この手法をベースラインとする．

図 2 に抽出結果を示す．2,076 個の短単位が抽出され，1,778 個（85.6%）がオノマトペであった．品詞情報を用いることでオノマトペを高い精度で抽出できることが明らかになった．しかし，3.では考慮していなかった形態素解析誤りや対象オノマトペの拡充に起因するエラーが発生した．

人手で非オノマトペと判断された例を以下に示す．（下線部が該当箇所）

① 2 文字 / 長音 / カタカナ（形態素解析誤り）
- 育ち盛りの高校生，こーゆー添加物のこと…
- …おともだちがサッかーのしあいがありました．
- 一番目立っていたグレートデン．
- どーでもイイ．
- …限定販売する「ビープラスDT」（16万円）だ．

② 助詞とオノマトペ（形態素解析誤り）
- 挽き出すときに，目がちゃっとひっかかるわけですわ．
- 病気のペット（たとえばワンちゃんとしましょう）は…

③ コラム掲載語
- 責任もあいまいだった．
- こわごわ組んだローンだけど…
- わたしは，みにくい姿の魔物がすきだ．
- フルに使いこなすには取説が必要かも．

④ 同音異義語
- 私にはたった一つだけ望みがあった．
- …おうおうにして東洋趣味に走るのよね．
- 「かくかくしかじか？」で…
- これが一般人のごくごく健全な感覚でしょう．
- 二十年も放置され，とうとう空家が一千戸に達した．
- 若い人たちの話をよくよく聞いてみると…

判断不能とされたのは，「しばしば」，「だんだん」，「まだまだ」，「みすみす」，「みるみる」など，一般の副詞として認識されつつあるオノマトペである．

この実験の結果から，本手法の改善には，五十音順さくいんから一部の語を除くことや，同音異義語の問題を回避するためにストップワード（オノマトペと品詞の組）を設けることが有効だと考えられる．

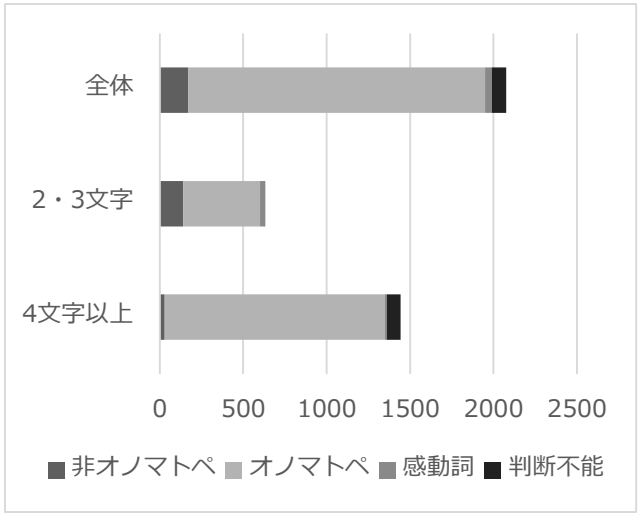

図 2　オノマトペ抽出結果（ベースライン）

5.　BCCWJ のブログデータからのオノマトペ抽出

4.で述べた改善策を導入したオノマトペの抽出手法を構築する．解析誤りが特に起こりやすいカジュアルな文体での本手法のオノマトペ抽出精度を確認するため，BCCWJ に含まれる Yahoo!ブログのデータを対象として抽出実験を行う．

具体的な手順は以下の通りである．

I. 3 種類のリストを作成する
- オノマトペリスト：五十音順さくいんから不適切な語を除いたリスト
- 品詞例外リスト：これまでの分析で明らかになった，副詞・形状詞以外に分類されるオノマトペとその品詞をペアのリスト
- ストップワードリスト：これまでの分析で明らかになった，オノマトペとの同音異義語のリスト

II. MeCab（Unidic 辞書）で形態素解析を行う

III. オノマトペリスト中の語と字面が一致する短単位を抽出する

IV. Ⅲで抽出された短単位のうち，以下の条件を満たすものをそれぞれオノマトペと判断する（抽出ルール）
- a) 品詞が副詞，形状詞以外で，品詞例外オノマトペリストに存在する
- b) 品詞が副詞か形状詞で，オノマトペリスト中の語と完全一致し，ストップワードリストに存在しない
- c) 品詞が副詞か形状詞で，オノマトペリスト中の語から最終促音を削除したものと一致し，ストップワードリストに存在しない
- d) 品詞が副詞か形状詞で，オノマトペリスト中の語に最終促音を付加したものと一致する
- e) 品詞が副詞か形状詞で，オノマトペリスト中の語に助詞「と」を付加したものと一致する
- f) 品詞が副詞か形状詞で，長音母音を長音記号に変換，あるいは繰り返しの縮約を行うとオノマトペリスト中の語と一致する

上記の手順で，49,492 個の短単位がオノマトペとして抽出された．これまでの分析で，2・3 文字のオノマトペの抽出精度が特に低いということが明らかになっている．したがって，ここでは

表 1　各ルールのオノマトペ抽出結果

	ルールａ		ルールｂ		ルールｃ		ルールｄ		ルールｅ		ルールｆ		計	
	2 文字	3 文字	2 文字	3 文字	2 文字	3 文字	2 文字	3 文字	2 文字	3 文字	2 文字	3 文字	2 文字	3 文字
非オノマトペ	0	19	2356	942	873	72	0	8	0	122	152	24	3381	1187
オノマトペ	0	13	1689	5312	583	21	0	40	0	1921	5	165	2277	7472
感動詞	0	0	0	1	14	0	0	0	0	0	0	0	14	1
判断不能	0	0	296	11	10	0	0	0	0	16	41	0	347	27
計	0	32	4341	6266	1480	93	0	48	0	2059	198	189	6019	8687

抽出された短単位のうち，2･3 文字のものを全て人手で確認し，オノマトペであるかを判断する．

2･3 文字の短単位は 14,706 個抽出され，9,749 個（66.3%）がオノマトペであった．2 文字の短単位のみでは 37.8%，3 文字のみでは 86.0%の精度である．表 1 に各ルールにおける抽出結果を示す．

2 文字のオノマトペの抽出エラーが全体の抽出精度に悪影響を及ぼしていることがわかる．中でも，2 文字の短単位にルール b を適用すると，非オノマトペをオノマトペとして抽出するエラーが多い．これは，形態素解析誤りによって別の単語の一部や長いオノマトペの一部が切り出されることが主な原因である．具体例を以下に挙げる．例中の下線部はすべて，形態素解析によって副詞と判断されている．

- あなたがた自身も，あらゆる行いにおいて…
- よし，少しベンキョーすっかな･･･まず，初めに･･･
- ぷりんっとしたやつね．

オノマトペによっては，定型句のような表現でしか使われないものもある．形態素解析誤りは避けられないので，個別のオノマトペに抽出ルールをカスタマイズするなど，品詞情報だけに頼らない抽出手法を検討したい．

また，ルール c～f は，辞典に掲載されていない新しいオノマトペや，既存のオノマトペが変形したものを抽出する役割を果たすが，過剰に適用されることでエラーの原因にもなる．具体例を以下に挙げる．

- どなたかコツ教えて下さい．
  （ルール c 適用：「こつっ」の促音削除）
- …するまでほっとこうと思って今になりました．
  （ルール e 適用：「ほっ」に「と」を付加）
- 友達は私といるだけで巻き添えくうし…
  （ルール f 適用：「くー」に変換）
- 気付けばあっと言う間の12月．
  （ルール f 適用：「ばーっ」に変換）
- しかも土曜日にウチに来て←パパン居なかったから
  （ルール f 適用：「パン」に縮約）

これらのルールの精度向上には，係り受け関係などの利用や，ルールの適用範囲を限定する工夫が必要だろう．

6. おわりに

本稿では，BCCWJ を対象としたオノマトペの抽出及び分析を行った．分析を進める中で，オノマトペに関するいくつかの興味深いコロケーションが見受けられた．たとえば，「ぐんぐん」は「伸びる」や「大きくなる」などの成長に関わる動詞に係ることが多い，「くるくる」や「ぐるぐる」は回転に関わる動詞に係る点で共通しているが，「ぐるぐる」は「さまよう」などの動詞に係ることもある，「しゅわしゅわ」は炭酸入りの飲料水とともに用いられることが多い，などである．しかし，現状では，オノマトペと共起する表現を纏め上げてコロケーションを得るには実例が不足している．

本稿で述べたように，典型的な表層形態をもつオノマトペは高い精度で抽出可能である．したがって，本研究が目標とするオノマトペ実例辞書の構築に向け，地方議会会議録コーパス，ブログコーパスを利用して，大規模なコロケーション抽出を行う予定である．その際，多様な派生的なパターンや新出オノマトペを抽出する手法について，さらなる検討が必要である．

謝辞

本研究は科研費（No. 26370498）の助成を受けたものである．

参考文献

[Fukushima 14] Hironori Fukushima, Kenji Araki, and Yuzu Uchida: Disambiguation of Japanese Onomatopoeias Using Nouns and Verbs, TSD2014, LNAI 8655, pp. 141-149, 2014.

[池田 15] 池田祐一，阪本浩太郎，渋木英潔，森辰則: 国際音声記号を素性とした 3 文字以下の未知のオノマトペ自動抽出手法の提案，言語処理学会第 21 回年次大会論文集，P1-12，2015.

[木村 14] 木村泰知，渋木英潔，内田ゆず，乙武北斗，高丸圭一，森辰則: 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，第 30 回ファジィシステムシンポジウム講演論文集，pp. 638-641，2014.

[郜 06] 郜楓: コーパスを利用した類義語のコロケーション分析―擬態語「しんみり，しみじみ」と動詞の共起から―，ことばの科学，19，pp. 129-140，2006.

[小松 15] 小松孝徳: 論文特集「オノマトペの利活用」にあたって，人工知能学会誌 30(1) , p. 134，2015.

[Kudo 04] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.

[小野 07] 小野正弘編: 日本語オノマトペ辞典，小学館，2007.

[清水 14] 清水祐一郎，土斐崎龍一，坂本真樹: オノマトペごとの微細な印象を推定するシステム，人工知能学会論文誌 29(1)，pp. 41-52，2014.

[高丸 15] 高丸圭一，内田ゆず，乙武北斗，木村泰知: 地方議会会議録コーパスにおけるオノマトペ―出現傾向と語義の分析―，人工知能学会論文誌，30(1)，pp. 306-318，2015.

[田野村 10] 田野村忠温: 日本語コーパスとコロケーション―辞書記述への応用の可能性―コーパスからのコロケーション情報抽出―分析手法の検討とコロケーション辞典項目の試作，阪大日本語研究，21，pp. 21-41，2009.

[内田 12] 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，Journal of Japan Society for Fuzzy Theory and Intelligent Informatics，24(3) , pp.811-820，2012.