オノマトペの語義決定に寄与するコロケーションの分析

藤田 実智斗 [1]　内田 ゆず [2]　荒木 健治 [3]
[1] 北海道大学 大学院情報科学院　[2] 北海学園大学 工学部
[3] 北海道大学 大学院情報科学研究院
[1]fujita.michito.u9@elms.hokudai.ac.jp　[2]yuzu@hgu.jp
[3]araki@ist.hokudai.ac.jp

概要

日本語学習者が日本語オノマトペの使い方を理解するためには，定義文や用例だけでなく，そのオノマトペの語義に関係するコロケーションを知っておくことが有効である．そこで本研究では，オノマトペの語義決定に寄与している文節についてアノテーションを行い，その結果に基づいた分析を行った．分析により，いくつかの語義で定義文からだけでは読み取れない特徴がみられた．また，人手で抽出されたコロケーションについて Word2Vec と k-means++によるクラスタリングを行い，クラスタと語義の関係を分析した．

1　はじめに

オノマトペ（擬音語および擬態語）は，音や物事の様子，動作の状態を表現する語であり，微細なニュアンスを表現する目的で日常的によく用いられている．特に日本語のオノマトペは他言語に比べて種類が多く，またその多くは複数の語義を持つ．日本語母語話者は，周辺文脈から感覚的に語義を判別することでこれらの語義曖昧性を解消しているが，日本語学習者にとってはその判別基準を学ぶことが難しい．そこで我々は，オノマトペの語義決定に寄与していると考えられるコロケーション（以下，寄与コロケーションと呼ぶ）を収集し，オノマトペの使い分けが分かりやすくなるようなオノマトペコロケーションデータベースの構築を目指している．

オノマトペコロケーションデータベースを構築するには，多義のオノマトペについて語義ごとに用例を集め，どのコロケーションが寄与コロケーションであるかを判別する作業を要する．こうした作業を人手で行うことは時間的コストが高く，大量にデータを集めるためには自動化することが望ましい．そのためには，多義オノマトペの自動語義分類と寄与コロケーションの自動抽出を実現する必要がある．

これまで自動語義分類については，事前学習済みの BERT[1] から得られる単語分散表現を用いた手法や，辞書中の用法情報から生成されたルールに基づく手法が先行研究で提案されている [2][3]．一方，オノマトペのコロケーションについては，オノマトペの係り先動詞に着目し，係り先動詞および係り先動詞の係り元文節の表層格を対象として分析が行われてきた [4][5]．しかし，これまで分析対象とされてきたコロケーションを日本語母語話者が実際に寄与コロケーションとして認識しているかについては，分析が不足している．そこで本研究では，寄与コロケーション抽出の自動化に向けて，実際に日本語母語話者がどのコロケーションを基に語義を決めているかをアノテーションし，人手で抽出した寄与コロケーションについて分析を行う．また，寄与コロケーションについてクラスタリングを行い，語義との関係を分析する．

2　使用するデータ

2.1　分析対象オノマトペ

本研究では『擬音語擬態語使い方辞典』[6]（以下，オノマトペ辞書と呼ぶ）で語義が 3 つ以上定義されているもので，かつ『現代日本語書き言葉均衡コーパス』[7]（以下，BCCWJ と呼ぶ）上で頻度が高い 10 種類のオノマトペ『あっさり，きっちり，ぎりぎり，くるくる，ごろごろ，さっぱり，ばたばた，ぴったり，ぶつぶつ，ぶらぶら』を分析対象として選定した．各オノマトペについて BCCWJ から各 400 文，合計 4,000 文を取得し，分析対象のテキストとした．

表 1　アノテーション例

文節	評価者 1	評価者 2	評価者 3	合計値
自力だけでは	0	0	0	0
動きそうもない	0	1	0	1
大石が	1	1	0	2
ごろごろ	0	0	0	0
している。	1	1	1	3

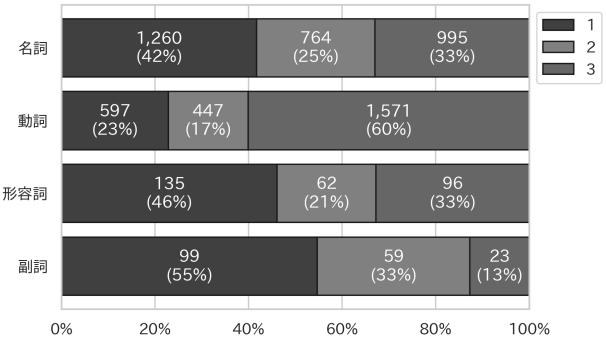

図 1　寄与コロケーションの一致数（品詞別）

2.2　寄与コロケーションを含む文節のアノテーション

本研究では文中のオノマトペの語義を決定する語について分析を行う．そこでまず 3 名の評価者による寄与コロケーションを含む文節のアノテーションを行う．アノテーションの具体例を表 1 に示す．まず，KNP[1)] を用いて分析対象のデータを文節単位に分割する．そして，評価者は語義を決める際に手がかりとなる文節に 1 をマークする．表 1 の例では「している。」という文節に対しては評価者 3 名が共通して寄与語を含む文節としてマークしているのに対し，「大石が」という文節では評価者 2 名がマークしている．このように評価者間で意見が分かれるケースについても，3 章で分析を行う．

3　寄与コロケーションの特徴分析

3.1　アノテーションの一致

アノテーションした結果について，品詞別に評価者間でどの程度寄与コロケーションが一致するかを確認する．1 名以上が語義決定に寄与する単語が含まれるとした文節について，文節先頭の形態素を寄与コロケーションとして集計し，品詞別に一致数を調べた結果を図 1 に示す．動詞の寄与コロケーションは全体総数が 2,615 個と多く，そのうち評価者 3 名全員が一致する割合は 60%であった．評価者 2 名

1）　https://nlp.ist.i.kyoto-u.ac.jp/?KNP

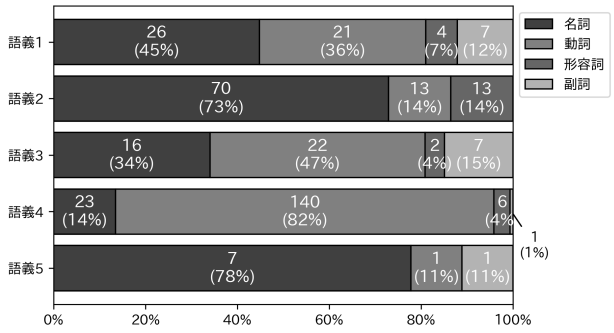

図 2　「さっぱり」の語義別品詞比率

が一致した割合も含めると 77%であり，動詞はアノテーションにばらつきが少ないことがわかった．一方，名詞の寄与コロケーションは全体総数 3,019 個のうち，評価者 1 名のみがアノテーションした寄与コロケーションが 42%と高く，ばらつきが多かった．要因としては，一般に文の構成要素として動詞よりも名詞の割合が高いため，選択肢が増えたことが挙げられる．またオノマトペは副詞として用いられることが多く [8]，直後の動詞の文節に係ることが多いため，評価者間で動詞の選択が一致したことも考えられる．

3.2　品詞比率

収集した寄与コロケーションについて，品詞の比率を確認する．3.1 節の結果から，評価者によって寄与コロケーションにばらつきがあることが確認されたため，本研究における以降の分析では，2 名以上が寄与コロケーションだと判断した単語を対象とする．また，品詞は名詞，動詞，形容詞，副詞の 4 種類を対象とする．寄与コロケーションとして抽出された単語の品詞比率は名詞:44%，動詞:50%，形容詞:4%，副詞:2%であり，名詞と動詞が大半を占めていた．そのため，今後寄与コロケーションの分析は名詞と動詞を中心に行う．具体例として「さっぱり」について語義ごとに比較した結果を図 2 に示す．「さっぱり」では「味が濃厚でなく、口当たりがさわやかであるようす。」と定義される語義 2 において，名詞の比率が 73%と高くなっている．これは，語義 2 の用法では食べ物について表現する際に「さっぱり」が用いられることが多く，寄与コロケーションとして食べ物に関連した語が多く抽出されたためであると考えられる．実際に抽出された語には「スープ」「味わい」「サラダ」「梅干し」などがある．

表 2　「ごろごろ」の寄与コロケーション（一部抜粋）

語義番号	定義	寄与コロケーション	
		名詞	動詞
語義 1	雷が鳴る音。また、雷が鳴るような音。	音, 雷, 雨, おなか, 轟き, 声, 喉, 天気, 猫, 腹, お腹, 雨粒	鳴らす, 鳴る, する, 聞こえる, いう, ひびく
語義 2	かなり重量のある物体や肉体が連続してころがるようす。	台車, 蒲団, 地面, 上, 身, 回転, 斜面, 石	転がる, 転がす, 寝返る, 横たえる
語義 3	働かないで時を過ごすようす。何もしないで時を過ごすようす。	家, 部屋, 毎日, 元旦, リビング, 和室, ベッド, 日曜日, こたつ, お家, 休み	する, 過ごす, できる, やる, 寝る, 休む, くつろぐ
語義 4	たくさんありすぎて珍しくも貴重でもないようす。	男, 石, 連中, 岩, 人, 肉, 砂利, かたまり, じゃがいも, にんじん	する, いる, 転がる, ある, 出る, あふれる
語義 5	かたまりや異物がはいり込んでいて違和感を感じるようす。	目, ヤニ, 糸, 右眼, 眼, 口, コンタクト	する, 痛む

表 3　寄与コロケーションの文節位置

文節	マーク数	単語	距離
彼女は	0	彼女	3
バッグを	2	バッグ	4
拾いあげると、	0	拾う	3
それを	1	それ	2
ぶらぶら	0	ぶらぶら	0
させながら、	2	する	1
ぼくを	0	ぼく	3
見おろした。	0	見おろす	2

3.3　寄与コロケーションの具体例

アノテーションにより収集した寄与コロケーションの特徴を分析する．3.1 節で述べたように，寄与コロケーションの大半は名詞と動詞であったため，本節では名詞と動詞に限定して分析を行う．例として「ごろごろ」の寄与コロケーションの一部を表 2 に示す．語義 1 は擬音用法であるため，名詞および動詞共に音に関わる語が多く出現した．また，「猫の声」や「雷」を表現する場面で多く用いられていたため，「猫」や「天候」に関する名詞が多く出現したと考えられる．これらのことはオノマトペ辞書の定義や用法・用例から読み取ることができるが，一方で語義 3 ではオノマトペ辞書の情報からだけでは読み取れない「休日」や「家」に関する名詞が散見された．これは語義 3 の定義で示されている「働かないで時を過ごすようす。何もしないで時を過ごすようす。」という状態が，基本的に「休日」や「家の中」のシチュエーションで起こり得ることを示唆している．このように，定義文などの辞書情報からだけではオノマトペの使い方が十分に読み取れない場合でも，大量のテキストからコロケーションを抽出して分析することで，日本語学習者にとってより理解しやすいデータベースを構築できると考えられる．

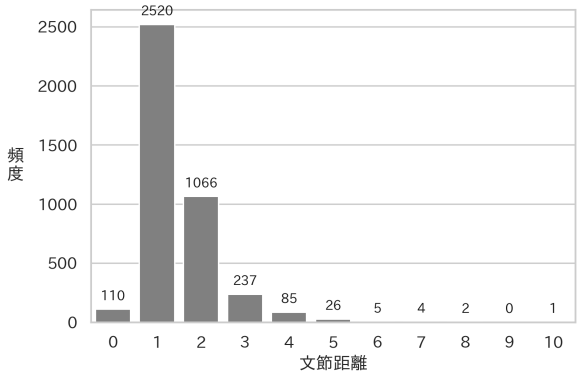

図 3　文節位置

3.4　寄与コロケーションの文節位置

先行研究では，オノマトペの係り先動詞を含む文節を基準として，その文節に係る文節を対象としたコロケーションの分析を行っている [4][5]．しかし，文の構造が複雑である場合や，文の長さが長い場合には，オノマトペを含む文節から離れた位置にある単語から語義決定に必要な文脈を読み取る可能性がある．そこで，寄与コロケーションの出現位置についての実態を知るために，寄与コロケーションを含む文節について，係り受け関係に基づく距離を調べる．分析結果を図 3 に示す．多くの場合は文節距離が 1 または 2 であった．一方で，文節距離 3 以上の例もいくつか見受けられた．例えば表 3 の例では，「バッグ」が「それ」という指示代名詞の先行詞であることが原因で，文節距離が 4 となっている．このほか，主語が省略されることで寄与コロケーションが離れて出現するケースや，係り受け解析のミスにより距離が離れるケースがある．これらのことから，寄与コロケーションの抽出において，係り受け解析による文節距離は寄与コロケーションの探索条

表 4 寄与コロケーション例（あっさり，名詞）

語義	寄与コロケーション
語義 1	味，醤油，スープ，味わい，ハンバーグ
語義 2	顔，メール，供述，切れ，表現
語義 3	別れ，承諾，同意，放棄，スルー
語義 4	解決，敗退，確定，実現，接続

表 5 クラスタリング結果（あっさり，名詞）

クラスタ	寄与コロケーション
クラスタ 1	表現，スルー，ＯＫ，形状，風景
クラスタ 2	敗退，切れ，実現，断念，決着
クラスタ 3	態度，別れ，供述，返事，承諾
クラスタ 4	味，醤油，スープ，味わい，ハンバーグ

件の重みづけとしては適しているが，探索範囲を限定することには適していないと考えられる．

4 寄与コロケーションのクラスタリング

「ぶつぶつ」の語義 2「蒸気やガスが液体の表面に連続して噴出したりわき立ったりする音。」のように，オノマトペ辞書で定義された語義にはドメインを限定するものがあり，こうしたドメインの特徴が寄与コロケーションに反映されると想定される．オノマトペコロケーションデータベースの構築にあたっては，各語義ごとの特徴が解釈できるように寄与コロケーションをまとめあげることが理想的である．そこで本章では，各オノマトペの寄与コロケーションについて，Word2Vec を用いて単語分散表現を取得し，k-means++によるクラスタリングを行うことで，クラスタと語義の関係を分析する．

4.1 単語分散表現

3.1 節と同様に，オノマトペの語義決定に寄与しているとされた文節の先頭の形態素を寄与コロケーションとする．これらの寄与コロケーションについて，Word2Vec の学習済みモデルである日本語 Wikipedia エンティティベクトル [9] を用いて単語分散表現を取得する．

4.2 k-means++によるクラスタリング結果

各オノマトペの寄与コロケーションについて，k を語義数として k-means++を用いてクラスタリングを行い分析する．ここでは具体例として「あっさり」の名詞の寄与コロケーションおよび「ごろごろ」の動詞の寄与コロケーションについて，分析結果の一部を表 4, 5, 6, 7 に示す．

表 4 と表 5 を比較すると，語義 1「色、味などの濃度が薄いようす。」の寄与コロケーション群とクラスタ 4 が対応していることがわかる．語義 1 およびクラスタ 4 には，抜粋した語以外にも食べ物や味に関する語が多く含まれているため，こうした特徴に基づいたクラスタが形成できたと考えられる．一方，そのほかのクラスタは各語義と対応せず，複数の語義の語が混在していた．

表 6 寄与コロケーション例（ごろごろ，動詞）

語義	寄与コロケーション
語義 1	鳴る，ひびく，聞こえる，鳴り響く
語義 2	転がる，ころがる，転がす，あてる，寝返る
語義 3	する，過ごす，寝る，休む，くつろぐ
語義 4	する，いる，転がる，ある，出る
語義 5	する，痛む

表 7 クラスタリング結果（ごろごろ，動詞）

クラスタ	寄与コロケーション
クラスタ 1	する，いる，あてる，入れる，横たえる
クラスタ 2	鳴る，ひびく，聞こえる，鳴り響く
クラスタ 3	転がる，ころがる，転がす，浮く
クラスタ 4	休む，くつろぐ，過ごす，眠る，寝る
クラスタ 5	寝返る

表 6 と表 7 を比較すると，語義 1「雷が鳴る音。また、雷が鳴るような音。」の寄与コロケーション群とクラスタ 2 が対応していることがわかる．語義 1 は擬音語用法であり，音を伴う動詞が多く用いられるため，単語分散表現によりクラスタを形成できたと考えられる．一方，そのほかのクラスタは各語義と対応せず，複数の語義の語が混在していた．特に語義 3, 4, 5 においては「ごろごろ」が「する」を伴って動詞として機能する点で共通しているため，周囲の文脈を考慮できない静的な単語分散表現ではクラスタリングが難しい．

これらの結果から，一部の語義については，本手法によって寄与コロケーションをまとめあげることで語義の特徴が解釈できることが明らかになった．今後は，文脈を考慮した単語分散表現を用いた手法を導入し，有効性を検証したい．

5 おわりに

本研究では，オノマトペの語義決定に寄与するコロケーションについてのアノテーションを行い，アノテーション結果について分析を行った．また，Word2Vec と k-means++によるクラスタリングを行い，クラスタと語義の関係を分析した．今後は文脈を考慮した単語分散表現の活用を目指す．また，こうした各語義の特徴に基づいた寄与コロケーションのクラスタリングは，自動語義分類にも有用であると考えられるため，適用方法を検討する．

謝辞

本研究は JSPS 科研費 21K12584 および 17K12791 の助成を受けたものである.

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186. Association for Computational Linguistics, 2019.

[2] 乙武北斗, 内田ゆず, 高丸圭一, 木村泰知. BERT による周辺文脈を考慮したオノマトペの語義分類手法の提案. 知能と情報, Vol. 32, No. 1, pp. 518–522, 2020.

[3] 藤田実智斗, 内田ゆず, 荒木健治. オノマトペ辞書に基づいたルールによるオノマトペ語義分類手法の提案. 日本知能情報ファジィ学会 ファジィ システム シンポジウム 講演論文集, Vol. 38, pp. 580–584, 2022.

[4] 乙武北斗, 内田ゆず, 高丸圭一, 木村泰知. 表層格に着目したオノマトペ共起語の抽出と分析. 言語処理学会第 22 回年次大会予稿集, pp. 195–198, 2016.

[5] 高丸圭一, 内田ゆず, 乙武北斗, 木村泰知. 係り先動詞に着目したオノマトペの語義分類に関する検討. 知能と情報, Vol. 28, No. 4, pp. 693–699, 2016.

[6] 阿刀田稔子, 星野和子. 擬音語・擬態語使い方辞典. 創拓社, 第 2 版, 1996.

[7] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language resources and evaluation**, Vol. 48, No. 2, pp. 345–371, 2014.

[8] 内田ゆず. 現代日本語書き言葉均衡コーパスコアデータにおけるオノマトペ出現実態に基づくオノマトペ自動抽出手法. 工学研究（北海学園大学大学院工学研究科紀要）, Vol. 17, pp. 15–20, 2017.

[9] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会発表論文集, pp. 797–800, 2016.