

パターンマッチングによるオノマトペ候補語抽出の試み ーオノマトペ形態変換プログラムを用いてー

廉 沢奇 (神戸大学国際文化学研究所)

An Attempt to Extract All the Possible Onomatopoeic Words By Using a Pattern Matching Technique: Utilizing an Onomatopoeic Morphological Transformation Program

LIAN Zeqi (Graduate School of Intercultural Studies, Kobe University)

要旨

日本語オノマトペは種類が多いが、実際にいくつの形態パターンがあるか、また、その具現形はいくつあるかという点に関して、先行研究の見解は一致していない。そこで本研究では、先行研究で触れられたすべての形態パターンとして全 72 種を特定し、BCCWJ と CEJC に含まれる全テキストデータにおいて、自作の Python プログラムでパターンマッチングを実行した。その結果、現代日本語の書き言葉・話し言葉に実際に出現する形態パターンは最大 60 種であり、その具現形は約 2,200 語となることが明らかになった。今回の調査で得られたオノマトペ形態パターンおよび具現形のリストは、実データから作られた網羅的リストであるという点で価値を有し、今後のオノマトペ研究において基礎資料の一つになるものと考えられる。

1. はじめに

日本語の語彙体系において、オノマトペはきわめて重要な位置を占めている(廉 2022)。オノマトペは、日本語学においても、また、日本語教育学においても、研究が盛んに行われている。しかし、研究対象となるオノマトペについて、その範囲は必ずしも明確ではない。日本語学においては、オノマトペの範囲を明確に規定しないまま、数十語から数千語まで、それぞれ異なる範囲のオノマトペを対象としてその性質が議論されてきた。また、日本語教育においても、目標として習得すべきオノマトペリストが多数提案されているが、それらの語数はまちまちであった。日本語学においても、日本語教育学においても、議論の出発点となるオノマトペの範囲や数が定まっていないという看過できない問題が存在する。

この背景には、オノマトペの語彙としての特殊性と不安定性がある。オノマトペは、日々、新たに生み出されており、擬音語の多くは一回限りのものである。こうしたオノマトペを広く集めていけば、その総数は非常に大きなものとなりうる。たとえば、小野(2007)の辞典には、4500 を超えるオノマトペが収録されている。しかし、これらのすべてが一般的に使用されているわけではない。たとえば、小野の辞典に含まれている「しゅりり」(嫌味なく、爽やかなさま)や「げんべいツツジしろツツジ」(ホオジロという鳥の鳴き声)などを、現代の日本語で耳にすることはほとんどないだろう。こうしたものを含めるかどうかによってオノマトペの数は大きく変わり、結果として、既存のオノマトペ辞典の収録数は 1000 種から 4500 種、教育的オノマトペリストの収録数も 70 種から 1400 種の幅が生じている。

こうした現状をふまえると、オノマトペ研究のさらなる深化にとって必要なことは、現代

日本語で実際に出現しているオノマトペの最大数を正確に推定することである。その際、上述のように、オノマトペの具現形は不安定であるため、具現形そのものではなく、より安定した形態パタン(例:「ABAB」)の側に注目することが重要であると思われる。これにより、「もふもふ」など、小野の辞書にも含まれていない、言わば気付かれずに「眠っている」オノマトペも網羅的に取り出すことができる。

そこで、本研究では、先行研究で触れられたすべての形態パタンを書き出し、書き言葉・話し言葉の大型コーパスに含まれるテキストと機械的に照合することで、網羅的なオノマトペ形態パタンおよび具現形の調査を実施することとした。

2. 先行研究

オノマトペには「きらきら」「どんどん」などの形態もあれば、「きちんと」「しっかり」などの形態もある。こうした形態的多様性はオノマトペの大きな特徴である。それについて、先行研究はオノマトペの取りうる形態を分類・整理してきた。また、オノマトペの抽出に関して、先行研究はそれぞれ異なるアプローチを取り、価値ある知見を示している。以下では、(1)オノマトペ形態パタンに関する研究と、(2)教育的重要なオノマトペの選定に関する研究に分けて概観を行いたい。

まず、(1)について、金田一(1978)はオノマトペの形態を「語基(語根)+その他の要素」と認定している。「その他の要素」について、促音「っ」、撥音「ん」、「り」、「ウ延長」、「い」、「りん」、「りっ」の7種が挙げられている。結果として、単語の長さ(モーラ数)、オノマトペの語基(語根)、その他の要素でオノマトペの形態を18種に分類している。

Waida(1984)は、2モーラ語はそれだけではオノマトペとして不適格で、そこに「促音」「撥音」「り」「反復」という4種の「オノマトペ標識」が伴っている必要があると指摘した。これにより、従来の研究において「その他の要素」と呼ばれていた存在が、「オノマトペ標識」という名称を得て、広く認知され研究対象となるようになった。田守・スコウラップ(1999)はこれを受け、4種の「オノマトペ標識」に「母音の長音化」を加えた。そしてオノマトペの語基に限定し、語基+オノマトペ標識で分類を行った。結果の整理にあたり、まず、小分類だけでなく、語基のモーラ数で1モーラ/2モーラ/その他という3つの大分類を設定し、その語基の音を文字で表記するのではなく、子音と母音のパタンで整理した。これによりオノマトペの形態的性質がより明確に示されることとなった。結果としてオノマトペの形態を3つの大分類と22種の下位分類に分類している。

日向・笹目(1999)は、まず(1)オノマトペの形態分類は語の長さに基づいて行うべきか語基の長さに基づいて行うべきか、(2)語基に付属する要素として取り上げられるべきものはなんであるのか、という2点については先行研究においても共通理解がなされていないと指摘した。(1)の解決策として、語の長さや語基の長さではなく、語の形成パタンに注目することにした。(2)の解決策として、先行研究の多くに注目されている「なし」「っ」「り」「ん」「反復」「長音」というオノマトペ標識のほか、金田一(1978)の研究に挙げられた「い」と「ら」も枠組みに取り込んだ。その後、浅野鶴子(1978)編の『擬音語・擬態語辞典』に取り上げられた1647語を対象にオノマトペの形態分類を行った結果、オノマトペの形態を49種に分類した。

上記で見たように、オノマトペ形態パタンに関する先行研究は、「促音」「撥音」などいわゆる「オノマトペ標識」を意識して取り込んでいる。日向・笹目は「オノマトペ標識」とされてきていない「い」と「ら」を気づき、さらなる精緻な分類ができた。しかしながら、そ

それぞれの分類枠組みにおいて、分類できないものも存在し、さらなる精緻化する余地が残されている。また、全体的に小規模の研究が多く、各形態パタンの頻度情報を調査する研究が少ない。

次に、(2)について、三上(2007)は、辞書の収録語数 32600 語のうち、オノマトペは 791 語で全体の約 2%を占めることを明らかにした。このことを根拠として、全体で 4000 語を学ぶ日本語中級学習者であれば、その 2%、つまり 80 語程度のオノマトペを学ぶべきだと考えた。この発想をもとに、三上は 8 種の基本語彙リストを調査し、3 種以上に含まれていた 87 語のオノマトペを抽出し、内省で取捨選択を行うことで、教育用オノマトペとして 70 語を選定した。

獅々見(2016)は、まず 4 冊のオノマトペ辞典を調べ、2 冊以上のオノマトペ辞典に出現したオノマトペ 273 語を抽出した。その後、日本語母語話者を対象とするアンケートで個々の語の親密度（見聞き、使用）を調査し、親密度の低い 21 語を除外した。最後に「BTS による多言語話し言葉コーパス」で頻度調査を行った。こうして得られた出現頻度データ、親密度（見聞き）データ、親密度（使用）データを主成分分析で合成し、全体を得られた主成分得点順に並べかえ、オノマトペの順位を決定した。こうして、順位付きの重要オノマトペ 252 語を提案している。

張(2020)は 6 種のオノマトペ辞典を調べ、最低 1 冊以上に含まれるオノマトペ (4,189 語) と 2 冊以上に含まれるオノマトペ (1,740 語) を抽出した。その後、BCCWJ で 4189 語の使用頻度を調査し、頻度 1 のものを排除したところ、4,189 語は 2,006 語になった。この 2,006 語と 1,740 語を比較したところ、共通するのは 1,463 語となった。最終的に、出現頻度に辞書における収録率をかけて調整した頻度で全体を並び替え、1,463 種の重要オノマトペとした。

上記で見たように、オノマトペの抽出に関する先行研究は、使用するデータの選定、重要語抽出の手法、最終的に得られたリストの各点においてそれぞれ価値ある知見を示している。しかしながら、全体として辞書に依存する研究が多く、結果的に書き言葉に偏っている。また、直接的にコーパスから抽出した研究が極めて少ないという課題が残されている。

以上を踏まえ、先行研究の課題を 3 点にまとめた。 (1) 話し言葉データが少ない。 (2) 実在のデータの使用が少ない。 (3) 頻度情報の調査が少ない。そこで、本研究は (1') 書き言葉と話し言葉の統合調査を行い、 (2') 大型コーパス 2 種を組み合わせ、 (3') 形態パターン・具現形の両者に対して悉皆的頻度調査の実施し、現代日本語オノマトペを悉皆的抽出することを目指す。

3. リサーチデザインと手法

3.1 研究目的と RQ

すでに述べたように、本論文の目的は、先行研究で触れられたすべてのオノマトペ形態パターンを特定した上で、コーパスデータとの網羅的なパターンマッチングにより、現代日本語の書き言葉・話し言葉に実際に出現しているオノマトペ具現形の最大数を特定することにある。この目的に沿って、以下の 3 つの研究設問を設定した。

RQ1 先行研究で触れられたオノマトペ形態の総数はいくつ？ (オノマトペ形態の最大数の特定)

RQ2 二種の大型コーパスにおいて、上記のオノマトペ形態はいくつ出現しているか？また、それらに基づく具現形の最大数はいくつか？（実データにおけるオノマトペ形態・具現形の最大数の特定）

RQ3 先行研究の主張とコーパスから得られた言語事実にどのようなズレが認められるか？（先行研究と実データの乖離の解明）

3.2 データ

本研究で使用したコーパスは『日本語日常会話コーパス』（Corpus of Everyday Japanese Conversation, CEJC）と『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese, BCCWJ）である。CEJC は日常場面の中で当事者たち自身の動機や目的によって自然に生じた会話を収集することを目的としたコーパスである。性別・年齢などのバランスを考慮して抽出された40名の調査協力者を集め、機材機器を約3か月間貸与し、協力者の日常生活で自然に生じる会話を記録させた。結果として異なり856名発話者による約240万語を収録した。1億語という大規模コーパスの目安に比べて大きいサイズとは言えないが、集めにくい日常会話を200万語以上集めたということは、世界的にも珍しいことから、十分な分量のデータが収録されていると考えられる。

BCCWJ は国立国語研究所が開発した日本国内ではじめての大規模均衡コーパスである。BCCWJ は2011年に公開され、データ量はおよそ1億450万語である。BCCWJ は現代日本語の統計的縮図になるように、現代日本語書き言葉という母集団に対してその分布を過不足なく表現できるように設計した。従来の単一種類で構成されたコーパスに比べて、BCCWJ は書籍、新聞、雑誌、白書、ブログ等異なるレジスターのテキストの集まりであり、書き言葉の多様な実態を捉えるデータとなっており、日本語研究や日本語教育研究にとって非常に有益で有用なコーパスであると言える。

3.3 手法

RQ1 は2つのアプローチがある。1つ目は金田一(1978)、田守・スコウラップ(1999)、日向・笹目(1999)という3つの日本語学のオノマトペ分類研究から形態を抽出し、本研究基準に合わせて再整理することである。2つ目は三上(2007)、獅々見(2016)、張(2020)という3つの日本語教育学のオノマトペリストを自作のオノマトペ形態変換プログラムで処理し、個々の具現形を形態パターンに変換し統計することである。この2つのアプローチの結果を合わせ、重複を除き、これにより、形態パターン数の最大値を特定することができる。

オノマトペ形態変換プログラムについて、本研究では個々のオノマトペ具現形から形態パターンへの変換が必要とすることから、日向・笹目(1999)の分類手法に参照し、語の形成パターンに注目し、加えてオノマトペ標識で分類する手法を決めた。具体的に、1つの語に対して、語頭である左から1文字ずつ「A」からアルファベット順に変換する。また、下表のようなオノマトペ標識に遭遇するとき、その処理方法を準じて働き、最後は結果を出力する。

表1 オノマトペ標識と処理のアルゴリズム

オノマトペ標識	処理方法	変換例
促音	すべてが変換せず、そのまま出力する	ぱっ→Aっ
撥音	一回だけ出現の場合はせず、そのまま出力する	うんざり→AんBり

り	一回だけを出現し、かつ、語尾に出現する場合は変換せず、そのまま出力する	しっかり→AっBり
反復	同じ文字を同じアルファベットに変換する	どんどん→ABAB
母音の長音化	長音記号をアルファベットに変換する	きゃーっ→ABっ
と	一回だけを出現し、かつ、語尾に出現する場合は変換せず、そのまま出力する	ちょこっ→ABっ と

上表が示されている通り、本研究のアルゴリズムは5種のオノマトペ標識に加え、語尾の「と」もオノマトペ標識と同様に処理することにした。「と」は従来オノマトペ標識とされていないが、本研究は一部のオノマトペは「と」がないと不自然であることに気付き（例：「きちんと」と「きちん」(?), 「ほっと」と「ほっ」(?)), 語尾の「と」もオノマトペ標識と認定する。このオノマトペ形態変換プログラムを使用することにより、データの高速処理だけでなく、先行研究のような分類できないものの出現を回避し、すべての語はそれに対応する形態パターンになるという利点を持つ。

RQ2 では、オノマトペは主に副詞として機能することから、まず調査範囲を副詞に限定する。BCCWJとCEJCの語彙表を「品詞=副詞/名詞-普通名詞-副詞可能、語種=和語」という条件でフィルターし、得られた単語リストをRQ1のプログラムで形態パターンに変換する。次にRQ1で得られた形態パターンを条件にし、フィルターすることで、コーパスからオノマトペ形態パターンに合致するすべての語を抽出する。そして得られた結果からオノマトペでないもの（例：ぜんぜん、まだまだ、いろいろ・・・）を除去し、オノマトペリストとなる。最後はリストを調査対象にし、形態パターンと具現形の出現度を考察する。

最後に、RQ3では、まず形態パターのずれを考察するために、RQ2で得られた形態パターンリストを調査対象にし、金田一(1978)、田守・スコウラップ(1999)、日向・笹目(1999)という3つの日本語学のオノマトペ分類研究それぞれに出現したかを検証する。また、日向・笹目(1999)の研究には頻度情報があるため、本研究との頻度差を計算し比較する。得られた結果で過剰の注目と過少の注目の形態パターンを考察する。次に具現形のずれを考察するために、RQ2で得られた具現形オノマトペリストを調査対象にし、三上(2007)、獅々見(2016)、張(2020)という3つの日本語教育学のオノマトペリストそれぞれに出現したかを検証し、過剰の注目と過少の注目のオノマトペを考察する

4. 結果と考察

4.1 RQ1 オノマトペ形態の最大数の特定

4.1.1 言語学的研究から形態パターンを抽出

すでに述べたように、RQ1では(1)既存の分類研究から整理、(2)オノマトペリストから変換、という2つのアプローチで形態パターの最大値を確保する。まず、(1)について、3つの日本語学のオノマトペ分類研究を本研究基準に合わせて再整理したところ、以下の結果を得た。

表2 金田一(1978)と本研究基準の形態パターン分類

金田一(1978)分類	本研究基準
(1) 1 拍語のもの	A と

(2) 1 拍の語根+「い」「ん」「っ」	AB、Aん、Aっ
(3) 2 拍の語根のもの	AB
(4) 1 拍の語根に+「い」「う」「ん」「っ」のうちものが2箇	ABC、ABん、ABっ
(5) 2 拍の語根+「っ」	ABっ
(6) 2 拍の語根+「ん」	ABん
(7) 2 拍の語根+「り」	ABり
(8) (7) の1種「り」でないもの。古風な語	ABC
(9) 2 拍の語根の中間に、つめ、はねの入ったもの	AんB、AっB
(10) (7) の形の第1拍と第2拍の間に、「っ」、「ん」の入ったもの	AっBり、AんBり
(11) 2 拍の語根の繰り返し	ABAB
(12) 前項に似て類音のものを重ねるもの	ABCB、ABCD
(13) 全く似ていない2拍を重ねたもの	ABCD
(14) 2 拍語+「りん」「りっ」の形	ABCん、ABCっ
(15) 5 拍のもの	-
(16) (7) (8) (9) の繰り返し	ABCABC、AっBA っB
(17) (16) に似てあとのものは、多少形のちがうもの	ABCDBC
(18) その他の6拍もの	-

表3 田守(1999)と本研究基準の形態パターン分類

田守(1999)分類	本研究基準	田守(1999)分類	本研究基準
CV	Aと	CVCVri	ABり
CVQ	Aっ	CVCVN	ABん
CVN	Aん	CVQCV	AっB
CVV	AB	CVNCV	AんB
CVVQ	ABっ	CVQCVri	AっBり
CVVN	ABん	CVNCVri	AんBり
CVQ-CVQ	AっAっ	2 モーラの反復	ABAB
CVN-CVN	ABAB	2 モーラの反復の変種	ABCB
CVV-CVV	ABAB	CVCVri 反復	ABCABC
CVCV	AB	CVCVN 反復	ABCABC
CVCVQ	ABっ		

表4 日向・笹目(1999)と本研究基準の形態パターン分類

日向・笹目	本研究	日向・笹目	本研究
Aい	AB	AーAー	ABAB
Aいっ	ABっ	ABAB	ABAB
Aいーっ	ABCっ	ABんABん	ABCABC
Aん	Aん	ABりABり	ABCABC

Aっ	Aっ	ABら ABら	ABCABC
Aー	AB	AっBAっB	AっBAっB
Aーん	ABん	AっBんAっBん	AっBCAっBC
Aーっ	ABっ	AっBりAっBり	AっBCAっBC
AB	AB	AーんAーん	ABCABC
ABん	ABん	ABっABっ	ABっABっ
ABっ	ABっ	AっBらAっBら	AっBCAっBC
ABり	ABり	AーっAーっ	ABっABっ
AんB	AんB	AっAっA	AっAっA
AっB	AっB	AんBAんB	ABCABC
AんBり	AんBり	ABB	ABB
AっBり	AっBり	ABBっ	ABBっ
AっBら	AっBC	ABCB	ABCB
ABーん	ABCん	ABCD	ABCD
AっBん	AっBん	AらりBらり	ABCDBC
AっBーん	AっBCん	ABらCDら	ABCDEC
ABーっ	ABCっ	ABABっ	ABABっ
AんAん	ABAB	ABんCDん	ABCDEC
AっAっ	AっAっ		

上記より、3つの先行研究は本研究の分類基準にすると、金田一(1978)は18から20種、田守(1999)は22種から15種、日向・笹目(1999)は49種から29種となることがわかった。以上を統合し、重複を除くと、全体として31種に整理され、既存の分類研究で触れられてきた形態パタンの最大値を確定することができた。

ここで先行研究の分類手法と比較すると、金田一(1978)による分類はより簡潔で、田守(1999)と日向・笹目(1999)による分類はより精緻的であることがわかる。前者の分類は複数のオノマトペ標識をまとめて1分類にするため、本研究基準に合わせると複数の小分類になることがある。また、(15)と(18)のような語長だけで分類するときがあり、その組み合わせは極めて多いため、本研究基準に合わせるのには難しいと考えられる。一方、後者についてはオノマトペ標識を意識して分けて議論するため、1つの小分類は本研究基準に合わせると唯一の結果が得られる。また、オノマトペ標識に対する処理手法は異なることにより、本研究基準に合わせると形態パターン数が減少した。特に日向・笹目(1999)はさらに「い」と「ら」なども処理したため、本研究より精緻な分類ができたと言えるが、それぞれ分類できないものが残されたのは課題である。

4.1.2 教育的リストから形態パターンを抽出

以上で、日本語学のオノマトペ分類研究を考察し、約90種の分類から31種に整理した。しかし、千語以上のオノマトペリストの中、これまでの分類研究が触れていない語が存在する可能性があるため、オノマトペリストから形態パターンに変換し考察する必要があると思われる。そこで、(2)について、3つの日本語教育学のオノマトペリストをオノマトペ形態変換プログラムで処理しところ、以下の結果を得た。

表5 オノマトペリストから形態パターンへの変換結果（上位10種）

形態パターン	頻度	比率	形態パターン	頻度	比率
ABAB	547	36.59%	ABっ	214	14.31%
ABり	126	8.43%	ABん	114	7.63%
AっBり	113	7.56%	Aっ	44	2.94%
ABCABC	43	2.88%	ABCB	37	2.47%
AB	30	2.01%	Aん	27	1.81%

三上（2007）、獅々見（2016）、張（2020）の3つの教育用リストをプログラムで処理したところ、三上（2007）の70語から11種、獅々見（2016）の252語から22種、張（2020）の1463語から61種の形態パターンを抽出することができた。張（2020）のリストは三上（2007）の20倍以上にオノマトペの具現形を収録しているのに対し、形態パターン数は6倍未満にとどまり、オノマトペは主に上位形態に集中していると考えられる。3つの結果を合わせ、重複を除き、合計69種の形態パターンを発見した。

以上、2つのアプローチでそれぞれ31種と69種の形態パターンを抽出した。本研究の調査枠組みとして形態パターン数の最大値の特定するため、2つのアプローチを合わせて72種となった。分類研究から整理した31種に比べて、今まで発見されていない、または分類できないとされる語は本研究のプログラムの使用により、40種以上に新しく発見することができた。

以上、RQ1に関しては、オノマトペ形態パターン数を特定するため、(1) 分類研究から整理、(2) オノマトペリストから変換、という2つのアプローチを取り、それぞれ31種と69種の形態パターンが得られた。重複を除き、合わせて72種となり、形態パターン数の最大値を特定することができた。この72種の形態を調査することで、より悉皆的にコーパスから現代日本語のオノマトペを抽出することが期待できる。

4.2 RQ2 実データにおけるオノマトペ形態・具現形の最大数の特定

以上で本研究の調査枠組みが決定し、コーパスから72種の形態パターンに合致するオノマトペを抽出し、その出現度を調査したところ、以下の結果を得た。

表6 実際の形態パターンの出現度（上位10種）

形態パターン	頻度	比率	形態パターン	頻度	比率
ABAB	688	30.76%	ABん	239	10.68%
ABC	170	7.60%	ABり	165	7.38%
AっBり	137	6.12%	ABCん	127	5.68%
Aっ	52	2.32%	ABCB	50	2.24%
Aん	44	1.97%	AっBん	42	1.88%
未出現12種：AAっ、ABABっ、AっBCAっBC、ABっABっ、AっBCBC、AんBCり、AっBんCDEC、ABABCっDり、AっBCBん、ABCDBD、ABCBABCB、ABっCり、					

まず、形態パターンの出現度について、72種形態パターンのうち、実際に出現したのは60種

であることが確認された。ここで高頻度オノマトペ形態パターンと未出現 12 種の形態パターンと比較すると、顕著な語長（モーラ数）の差異が観察された。上位 10 種の形態パターンは 2 モーラから 4 モーラの語が多く、平均語長が 3.3 であるのに対し、未出現 12 種の形態パターンには「AA っ」を除き、全ては 5 モーラ以上であり、計算すると平均語長は 6.17 になり、上位 10 種の平均語長と約 2 倍の差が認められた。この現象の背景には、オノマトペが主に話し言葉で使用されるという特性が関与していると考えられる。話し言葉においては、状況や感覚を即時的かつ効率的に伝達する必要性が高く、そのため発音時間が短い、つまり語長の短いオノマトペが優先的に選択されると考えられる。したがって、短い語長を持つオノマトペが頻繁に使用される一方で、語長が長いオノマトペは使用頻度が低いか、全く使用されないことが示唆される。

次に、オノマトペの具現形の出現度を調査したところ、以下の結果を得た。

表 7 実際の具現形の出現度（上位 10 種）

CEJC			BCCWJ			総合		
オノマトペ	頻度	形態	オノマトペ	頻度	形態	オノマトペ	頻度	形態
ちゃんと	438.58	A んと	ずっと	117.96	A っと	ずっと	277.86	A っと
ずっと	437.75	A っと	はっきり	99.91	A っ B り	ちゃんと	255.11	A ん
どんだん	141.78	ABAB	しっかり	93.68	A っ B り	どんだん	94.26	ABAB
だんだん	99.21	ABAB	ゆっくり	78.54	A っ B り	しっかり	79.08	A っ B り
ゆっくり	74.41	A っ B り	ちゃんと	71.65	A んと	ゆっくり	76.47	A っ B り
しっかり	64.48	A っ B り	きちんと	62.16	AB んと	はっきり	68.14	A っ B り
ぱっ	52.08	A っ	どんだん	46.74	ABAB	だんだん	63.3	ABAB
めちゃくちゃ	45.88	ABCB	すっきり	45.44	A っ B り	きちんと	44.52	AB んと
そろそろ	42.99	ABAB	じっと	40.51	A っと	そろそろ	35.76	ABAB
はっきり	36.38	A っ B り	たっぷり	34.06	A っ B り	ぱっ	33.52	A っ

60 種の形態パターンに合致するオノマトペ具現形を調査した結果、CEJC から 1094 語、BCCWJ から 1886 語、重複を除いた異なり語数として合計 2237 語のオノマトペが抽出された。この数値は、既存のオノマトペ辞書（1000～4000 語収録）と比較すると中間的な位置に相当する。さらに、教育的オノマトペリストとして最大規模とされる張（2020）の 1463 語と比較しても、本研究で得られた語数は 53%増となっている。

以上、RQ2 ではコーパスで 72 種の形態パターンを調査した結果、60 種の形態パターンとそれ

に合致する 2237 語の具現形が得られた。形態パターンに関しては、先行研究から整理した 72 種の中、12 種が検出できず、出現しなかった形態パターンは語長が長く、人は優先的に語長が短い語を選択するため、実際に使われていないと推測される。具現形に関しては、最大 4500 語の辞書には及ばず、教育用オノマトペリストより多く抽出したことがわかった。従来の先行研究は辞書ベースの手法を採用しており、これにより幅広いオノマトペの収録が可能となっている。このアプローチは、オノマトペの形態的多様性を網羅的に理解する上で非常に価値があるものである。しかし、本研究では実際の使用頻度に焦点を当て、コーパスを用いた悉皆的な調査を通じて、実際に使用されているオノマトペ形態パターンと具現形を特定することができ、今後の研究においても一定の参考となることが期待される。

4.3 RQ3 先行研究と実データの乖離の解明

以上、RQ1 では既存の研究を整理し、RQ2 ではコーパスで実際に使われているオノマトペ形態パターンと具現形を調査した。次に、RQ3 では、先行研究と本研究の差異を概観し考察する。まず、形態パタンのずれについて、先行研究と比較した結果、以下の結果を得た。

表 8 形態パタンのずれ

形態パターン	比率	田守 (1999)	金田一 (1978)	日向・笹目 (1999)	日向・笹目(1999) との差
ABC	7.60%	出現	未出現	0.00%	7.60%
ABC ん	5.68%	出現	未出現	1.75%	3.93%
AB ん	10.68%	出現	出現	8.18%	2.51%
ABAB	30.76%	出現	出現	30.90%	0.14%
ABC っ	0.09%	出現	未出現	1.31%	-1.22%
AB り	7.38%	出現	出現	8.80%	-1.43%
AB っ	0.36%	出現	出現	16.17%	-15.81%

先行研究と比較した結果、「ABC」(例：ぶわー)や「ABC ん」(例：ドカーン)などの形態パターンが、実際の使用頻度に比して過少に注目されていた可能性が示唆された。調べたところ、これらのパターンは CEJC における頻度が高く、口語表現において頻繁に使用されていることが明らかになり、先行研究の多くが話し言葉のデータを十分に参照していなかったことがその要因として考えられる。一方、従来の研究で頻繁に言及されてきた「AB っ」(例：ちらっ)のような形態パターンは、実際の使用頻度が 0.36%と極めて低いことが判明した。この乖離の背景には、先行研究において「AB っ」が ABAB 型オノマトペの変形として扱われてきたことが挙げられる。例えば、「にこにこ」や「きらきら」といった ABAB 型に対して、それぞれ「にこっ」「きらっ」といった「AB っ」型の存在が想定されてきた。しかし、実際には ABAB 型が十分に口語的表現であるのに対し、「AB っ」型はさらに使用制限が厳しいと考えられるため、実際の使用頻度が低くなっていると推察される。

次に、具現形のずれについて、先行研究と比較した結果、以下の結果を得た。

表9 具現形のずれ

オノマトペ	ランク	頻度	形態パタン	三上 (2007)	獅々見 (2016)	張 (2020)
もふもふ	29	11.19	ABAB	未出現	未出現	未出現
うんと	49	7.8	Aんと	未出現	未出現	未出現
がつつり	58	6.7	AつBり	未出現	未出現	未出現
もったり	68	5.91	AつBり	未出現	未出現	未出現
があん	78	4.57	ABん	未出現	未出現	未出現
...						
ぴかぴか	254	1.45	ABAB	出現	出現	出現
ばらばら	341	1.07	ABAB	出現	出現	出現
ぺらぺら	367	1.03	ABAB	出現	出現	出現
すらすら	469	0.77	ABAB	出現	出現	出現
がやがや	922	0.25	ABAB	出現	出現	出現

具現形に関しては、本研究の上位 100 語に含まれるにもかかわらず、先行研究で言及されていない語として「もふもふ」、「がつつり」、「うんと」、「もったり」などが挙げられる。以下、これらの語の用例を示しておきたい。

- (1) マイクの先っちょについているもふもふしたスポンジは別売りしていますか？
(OC02_03687)
- (2) 蒸し風呂でうんと汗を搔いて、水で洗い流してくる。(PB49_00014)
- (3) きのががつつり雨に濡れて洗濯しちゃったんですね。(W002_001)
- (4) 全卵と卵黄に和三盆、黒蜜、水飴を加え、ビーターの低速でもったりするまで混ぜる。
(PB4n_00146)
- (5) 頭が一んてぶつけるなどか。(K001_004)

上記の用例において、「もふもふ」は表面の柔らかい触感を表す語で、「うんと」と「がつつり」は量の多さを表す語で、「もったり」は料理の粘りが強い状態を表す語で、「があん」は物を強く打った音を表す語であると考えられる。これらは人間の五感や日常的な体験に密接に関連する語であり、実際の使用頻度が高いにもかかわらず、先行研究では看過されてきた。より身近な体験を描写するオノマトペが過少に注目されていた可能性が示唆された。

一方、「ぴかぴか」、「ばらばら」、「ぺらぺら」、などの語は先行研究に重要オノマトペとして選定されたが、実際の使用頻度は低いことが判明した。以下、これらの語の用例を示しておきたい。

- (6) いい鶏は、羽につやがあってぴかぴか光っている。(LBm5_00015)
- (7) ライブの時はなんとなく椅子がばらばら置いてあった (K007_019)
- (8) お姉さんは、金髪のお兄さんを見あげ、なにかぺらぺら英語で話している。(LBkn_00014)
- (9) 原稿がなくても、一時間くらいの演説をすらすらとやってみせたことがあったね。
(LB13_00074)

(10) がやがや大勢で楽しみたい人も、当然いらっしゃるでしょう。(PB55_00025)

上記の用例において、「ぴかぴか」は表面が光っている様子を描写する語で、「ばらばら」は散らかっている様子を描写する語で、「ぺらぺら」は流暢に外国語を話す様子を描写する語で、「すらすら」はなめらかに物事が進行する様子を描写する語で、「がやがや」は大勢の人が騒ぐ様子を描写する語であると考えられる。これらの語は、物事の様子を詳細に描写する際に用いられるものの、日常的な体験とは距離があり、また重複型(ABAB)であるため書き言葉での使用が制限される可能性があるため、実際の使用頻度が低いと推測される。

以上、RQ3 ではコーパス調査の結果と先行研究を比較した結果、形態パタンについて、「ABC」や「ABCん」は過少に注目され、「ABっ」は過剰に注目されていた可能性が示唆された。その原因について、先行研究が使用頻度データを参照しなかったことが考えられる。具現形について、「もふもふ」や「がっつり」など高頻度かつ日常的な体験を描写する語は先行研究に選定されず、「ぴかぴか」、「ばらばら」など物事の詳細な様子を描写する語は先行研究に選定されたが、その使用頻度が低いと確認された。その原因として、先行研究の多くが母語話者の内省に基づいて選定されたものであることが考えられる。このような日常的に使用されるオノマトペは、母語話者にとって無意識のうちに使われるため、内省の際に見落とされやすい傾向がある。コーパスの使用実態と比較した結果、これまでの学習者は使うオノマトペを教えられず、使わないオノマトペを教えられていた可能性を示唆しており、今後、コーパスに基づいた使用実態を反映した教材の開発が、学習者のより自然な日本語オノマトペの習得に寄与すると期待される。

5. おわりに

日本語で一般的に使用されるオノマトペの数について、先行研究において一致が見られないという現状を出発点として、本研究は、先行研究で触れられたオノマトペ形態の最大数を特定し、パタンマッチングの手法を取ることで、2種の大型コーパスにおいて実際に出現しているオノマトペ形態パタンならびに具現形の最大数の特定を行った。

3つの研究設問につき、以下の知見が得られた。まず、RQ1(オノマトペ形態の最大数の特定)については、まず、過去の3種のオノマトペ形態分類研究から31種を取り出し、次に、3種のオノマトペ具現形リストを形態パタンに戻すアプローチにより69種を取り出した。以上よりオノマトペ形態の最大数を72種と特定した。

RQ2(実データにおけるオノマトペ形態・具現形の最大数の特定)では、BCCWJとCEJCからRQ1で得られた72種の形態パタンに合致する語を抽出した。この結果、72種の形態パタンのうち、60種のみ出現することが確認された。この点に関して、出現パタンは未出現パタンより語長が短く、過度に長い形態パタンは実際には使われないことが示唆された。また、これらに基づく具現形として、BCCWJから1,886語、CEJCから1,094語、重複を除いて合計2,237語を特定した。この数は、各種のオノマトペ辞書の収録語数のおよそ中間的な数となる。

RQ3(先行研究と実データの乖離の解明)では、RQ2で得られた結果を先行研究と比較した結果、いくつかの乖離が示唆された。傾向として、先行研究は、実際には多用される「ABC」と「ABCん」という2つの形態パタンに対する扱いが少ない一方、実際にはあまり使用されない「ABっ」パタンは過度に強調して扱われていると言える。この背景には、先行研究の多くがコーパスデータ、とくに話し言葉データを十分に参照していなかったことがある

と思われる。今回の分析により、過去の研究ではまったく触れられていなかった具現形として、「もふもふ」「がつつり」「うんと」「もったり」「があん」の5語を新たに検出した。こうした語が見落とされていた原因も、先ほどと同様、内省への過度の依存と、コーパスデータの軽視によるものではないかと考えられる。

本研究で得られたリストは、実データから取り出されたものであるという点で、従来の主観ベースのオノマトペリストに比べ、一定の信頼性を持つものになっていると思われる。こうしたリストは、言語学におけるオノマトペ調査の基準リストとして、また、日本語教育における学習者の習得モデルのリストとして、一定の有用性を持つだろう。

以上で見たように、本研究は一定の成果を挙げたが、もちろん、課題も多い。ここでは3点に絞って研究する。1点目は、本研究では既存研究で触れられたオノマトペ形態の最大数を調査の起点としたため、仮に、これまで気付かれていないオノマトペ形態が存在しているとするならば、その形態に基づく具現形は取り出せていないということである。2点目は、コーパス調査の際に副詞に限ったため、そのほかの品詞に分類されているものは取り出せていないということである。3点目は、今回使用した BCCWJ は 2004 年までのデータしか入っておらず、その後に出現したオノマトペは取り出せていないということである。これらの点については、今後の研究で合理的な対応法を考えていきたい。

引用文献

- 小野正弘(2007). 『擬音語・擬態語 4500 日本語オノマトペ辞典』小学館.
- 金田一春彦(1978). 『擬音語・擬態語辞典』角川書店.
- 獅々見真由香(2016). 「日本語の会話におけるオノマトペの基本語彙選定」日本語教育, 165, pp. 73-88.
- 田守育啓・ローレンス＝スコウラップ(1999). 『オノマトペ:形態と意味』くろしお出版.
- 張晶鑫(2020). 「現代日本語におけるオノマトペの用法解明と中国人日本語学習者のためのオノマトペ指導に対する提言 —コーパス言語学の教育的応用の可能性をめぐる—」神戸大学国際文化学研究科博士論文.
- 日向茂男・笹目実(1999). 「語形からみた擬音語・擬態語 2」東京学芸大学紀要第 2 部門人文科学, 50, pp. 189-209.
- 三上京子(2007). 「日本語教育のための基本オノマトペの選定とその教材化」ICU 日本語教育研究, pp. 49-63.
- 廉沢奇(2022). 「話し言葉 10 種の言語データのコーパス調査をふまえた日本語学習者のための ABAB 型基本口語オノマトペの選定」統計数理研究所共同研究レポート, 456, pp. 117-129.