

大規模言語モデルにおける日本語オノマトペの音韻的・意味的特徴の分析

佐々木 稔（茨城大学工学部情報工学科）[†]

小関 竜真（茨城大学工学部情報工学科）

Analysis of Phonological and Semantic Characteristics of Japanese Onomatopoeia in a Large Language Model

Minoru Sasaki (Computer and Information Sciences, College of Engineering, Ibaraki University)

Ryoma Koseki (Computer and Information Sciences, College of Engineering, Ibaraki University)

要旨

本研究は、大規模言語モデル(LLM)が多義的な感覚的な意味を持つ日本語オノマトペを内部でどのように表現しているかを分析する。LLM におけるオノマトペの処理メカニズムを音韻的特徴と意味的特徴の観点から解明するため、埋め込みベクトルを用いてその特徴の可視化を試みた。具体的には、オノマトペを含む文に対して、「オノマトペ単体」、「オノマトペを含む周辺文脈」、「両者の平均」の3条件で LLM を用いて埋め込みベクトルを取得し、次元縮約手法により各文を2次元平面に配置する。その結果、オノマトペ単体では音韻的類似性、文脈中では意味的類似性に基づきクラスが形成された。また両者の平均では、音韻的特徴が優位に現れた。これは LLM が文脈的な特徴に応じてオノマトペの音韻情報と意味情報を選択的に利用していることを示唆しており、こうした分析によって将来的に LLM がオノマトペをより適切に認識できることに繋がると考えられる。

1. はじめに

日本語は、物事の状態や音を感覚的に描写するオノマトペが他言語に比べて非常に豊富な言語である。オノマトペは、その音の響きが特定の意味や感覚を連想させる音と意味の対応関係があり、微細なニュアンスの表現に不可欠な役割を担っている。近年、大規模言語モデル (LLM) は自然言語処理分野で目覚ましい性能向上を達成しているが、その内部でオノマトペのような感覚的・音韻的な情報がどのように表現され、処理されているかは依然として未解明な点が多い。LLM が文脈に応じてオノマトペの「音の響き」と「文中での意味」をどう捉え分けているのかを理解することは、LLM の言語理解能力の解明において重要な課題である。

オノマトペに関する先行研究は数多く存在する。音韻要素から微細な印象を推定するシステムの構築や、辞書情報と意味ネットワークを用いてオノマトペの語義分類を行う手法が提案されている。また、Word2Vec などの分散表現を用いてオノマトペの分布を可視化する研究も行われてきた。しかし、これらの研究の多くは、触覚や食感といった特定の共通点を持つオノマトペ群を分析対象としていた[卒論 5]。そのため、音が似ていて意味が異なるペアや、音は異なるが意味が似ているペアなど、多様な関係性を持つオノマトペ群を横断的に扱った上で、LLM が文脈の有無に応じてその内部表現をどう変化させるかに焦点を当てた研究は十分ではなかった。

[†] minoru.sasaki.01@vc.ibaraki.ac.jp

そこで本研究では、LLM の埋め込み空間における日本語オノマトペの分布を可視化・分析し、文脈情報がオノマトペの内部表現、すなわち音韻的特徴と意味的特徴のどちらを重視するかに与える影響を解明することを目的とする。この目的を達成するため、本研究では、(1) オノマトペ単体、(2) オノマトペを含む周辺文脈、(3) 両者の埋め込みベクトルの平均という 3 つの異なる条件で LLM から埋め込みベクトルを取得し、比較分析を行う。これらのベクトルを主成分分析 (PCA) を用いて 2 次元に削減し、可視化することで、LLM がオノマトペの音韻情報と意味情報をどのように符号化しているかを考察する。

2. 関連研究

オノマトペの計算言語学的研究は、その音韻的特徴と意味的印象の関係性に着目したアプローチと、分散表現を用いてその意味を捉えるアプローチに大別される。

音韻的特徴に着目した研究として、清水らは子音・母音の種類や濁音の有無といった音韻上の要素からオノマトペの微細な印象を推定するシステムを構築した (清水 2016)。また、戸本らは音素特徴に基づきニューラルネットワークを用いることで、オノマトペの印象を 2 次元平面上に可視化する「オノマトペマップ」を作成する手法を提案している (戸本 2012)。これらの研究は、オノマトペの音象徴性を工学的にモデル化する試みとして重要である。

一方で、コーパスに基づく分散表現を用いたアプローチも広く研究されている。古宮らは、コーパスから Word2Vec を用いてオノマトペの分散表現を作成し、クラスタリングによって可視化する手法を提案している (古宮 2016)。また、渡邊らはこの種の分布図を利用して、触素材の快・不快評価といった感性評価に応用する研究を行っている (渡邊 2014)。これらのアプローチは、実際の言語使用データからオノマトペの意味的類似性を捉える点で有効である。このほか、藤田らはオノマトペ辞書と日本語 WordNet を用いて、多義的なオノマトペを語義ごとに分類する手法を提案しており、オノマトペの意味解析の精緻化に貢献している (藤田 2022)。

しかし、これらの先行研究の多くは、触覚や食感といった特定の共通点を持つオノマトペ群を対象とした分析が中心であった。そのため、音が似ていて意味が異なるペアや、音も意味も異なるペアといった、より多様な関係性を持つオノマトペ群を体系的に扱った上で、特に近年の大規模言語モデル (LLM) がそれらをどのように表現するかを調査した研究は十分ではなかった。

本研究は、これらの背景を踏まえ、多様な関係性を持つオノマトペペアを対象に、文脈の有無という条件が LLM の内部表現に与える影響に焦点を当てる点で新規性を持つ。これにより、先行研究では捉えきれなかった、LLM によるオノマトペの動的な表現メカニズムの解明を目指す。

3. 大規模言語モデルを用いたオノマトペの特徴分析手法

本章では、LLM がオノマトペの音韻的特徴と意味的特徴をどのように表現するかを分析するための手法について詳述する。分析の基本方針は、3 つの異なる入力条件でオノマトペの埋め込みベクトルを取得し、それらに対して主成分分析 (PCA) を用いて 2 次元空間に射影・可視化することで、その分布を比較・考察するものである。

3.1 埋め込みベクトルの取得

本研究では、オノマトペの音韻的特徴と意味的特徴を分離して分析するため、以下の 3 つの条件で埋め込みベクトルを取得する。

3.1.1 オノマトペ単体の埋め込み

文脈から独立したオノマトペ自体の基底的な表現、すなわち音韻的特徴が強く反映されると想定される表現を捉えることを目的とする。具体的には、分析対象とするオノマトペの

文字列そのものを、OpenAI が提供する埋め込みモデル `text-embedding-ada-002` に入力する。このモデルは、テキストの類似性評価タスクにおいて高い性能を示すことが報告されており、応答として得られる 1536 次元のベクトルを当該オノマトペの埋め込みベクトルとして使用する。

3.1.2 オノマトペを含む周辺文脈の埋め込み

文脈中でのオノマトペの意味的な役割を反映した表現を捉えることを目的とする。『現代日本語書き言葉均衡コーパス (BCCWJ)』から抽出したオノマトペを含む前後 n 単語の単語列を、同様に `text-embedding-ada-002` に入力し、文全体の 1536 次元の埋め込みベクトルを取得する。これにより、オノマトペが特定の文脈でどのように解釈されるかを分析する。

3.1.3 平均埋め込み

オノマトペ自体の特徴と文脈中の意味的特徴の両方を統合した表現を分析することを目的とする。この条件では、文脈全体の埋め込みと、文中のオノマトペ部分の埋め込みを統合する。しかし、`text-embedding-ada-002` はテキストの一部分のみの埋め込みを取得できないため、オノマトペ部分の埋め込みは、トークンレベルの表現が取得可能な日本語 BERT モデル (`cl-tohoku/bert-base-japanese`) を用いて別途取得する。

具体的な手順として、対象となる用例集合の各文に対し、以下の 2 つの 2 次元ベクトルを求める。

- 3.1.2 節で取得した文全体の埋め込みベクトルを PCA で 2 次元に次元削減したベクトル
- BERT で取得したオノマトペ部分の埋め込みベクトルを PCA で 2 次元に次元削減したベクトル

オノマトペを含む文ごとにこれら 2 つの 2 次元ベクトルを平均し、それにより得られたベクトルを対象文の最終的な表現ベクトルとする。

3.2 可視化手法

上記の埋め込みモデルである `text-embedding-ada-002` や日本語 BERT モデルで取得した埋め込みベクトルは高次元であり、直接的な解釈が困難である。そこで、本研究では次元削減手法として主成分分析 (PCA) を用いる (Hyvarinen 2005)。PCA は、データの分散が最大となる方向を新たな軸 (主成分) として見出す手法であり、高次元データの本質的な構造を保持したまま低次元空間へ射影するために広く用いられる。

本研究では、各埋め込みベクトル (または 3.1.3 節で取得した 2 つの埋め込みベクトル) の集合に対し PCA を適用し、第 1 主成分と第 2 主成分を軸とする 2 次元平面上に各ベクトルをプロットする。これにより、元の高次元空間におけるベクトル間の類似度を、2 次元平面上のユークリッド距離として視覚的に捉え、オノマトペの表現が入力条件によってどのように変化するかを分析することが可能となる。

4. オノマトペ特徴分析実験

第 3 節で述べた分析手法に基づき、LLM がオノマトペの音韻的・意味的特徴を文脈の有無によってどのように埋め込みベクトルを表現するのかを検証するための実験について説明する。

4.1 分析対象データ

オノマトペの音韻的な特徴と意味的な特徴の組み合わせによって、LLM の内部表現がどのように変化するかを体系的に分析するため、以下の 4 つのカテゴリに分類したオノマト

ペアを人手により選定した。各カテゴリには5つのオノマトペのペアで構成された合計20ペア（40語）のオノマトペを分析対象とする。選定したオノマトペペアを表1に示す。

- A. 音韻・意味ともに類似: 音の響きが似ており、意味・用法も近い関係にあるペア。（例:「あっさり」と「さっぱり」）
- B. 音韻類似・意味非類似: 音の響きは似ているが、意味・用法が異なるペア。（例:「ひらひら」と「ふらふら」）
- C. 音韻非類似・意味類似: 音の響きは異なるが、意味・用法が近い関係にあるペア。（例:「ぎらぎら」と「じりじり」）
- D. 音韻・意味ともに非類似: 音も意味も関連性が低いペア。（例:「くんくん」と「だらだら」）

表1: 分析対象とした4カテゴリのオノマトペペア

	ペア1	ペア2	ペア3	ペア4	ペア5
A	あっさり, さっぱり	あつあつ, ぐつぐつ	びしょびしょ, べちゃべちゃ	ぼかぼか, ほくほく	ふわふわ, ふんわり
B	ひらひら, ふらふら	ぼたぼた, ばたばた	ころころ, ごろごろ	さらさら, ざらざら	ぼろぼろ, ぼろぼろ
C	ぎらぎら, じりじり	ぶくぶく, ぼこぼこ	てくてく, とぼとぼ	ぐっすり, すやすや	がたがた, ぶるぶる
D	くんくん, だらだら	ぽかん, ゲロゲロ	ぺらぺら, ぐつぐつ	ぽちゃん, ごほごほ	ばらばら, ぶんぶん

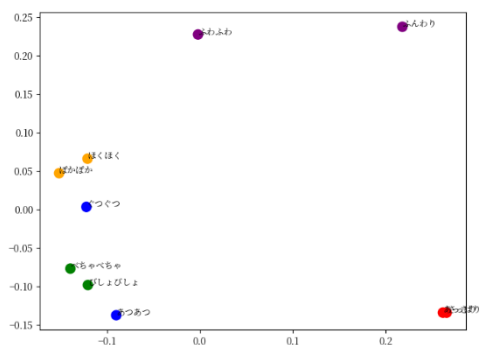
4.2 用例文データと実験条件

「周辺文脈の埋め込み」および「平均埋め込み」の分析に用いる文脈データは、『現代日本語書き言葉均衡コーパス (BCCWJ)』から抽出した。さらに、文脈の長さがオノマトペの表現に与える影響を調査するため、各オノマトペが出現する文において、その前後に含まれる語数 n を 10, 20, 100 の3パターンに設定した。

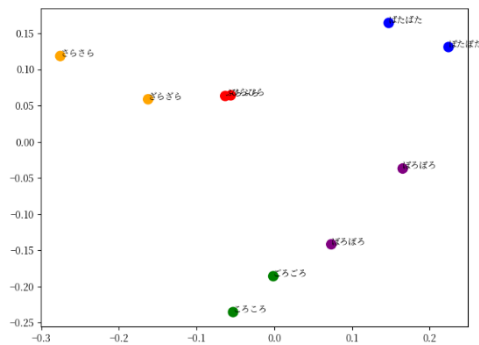
本実験では、3.1節で述べた3つの埋め込み取得条件（オノマトペ単体、周辺文脈、平均埋め込み）と、上記4カテゴリのオノマトペペアを用いて分析を行う。「オノマトペ単体」の条件では、表1の40語のオノマトペを直接入力する。「周辺文脈」と「平均埋め込み」の条件では、表1の各オノマトペを含む文脈データを、文脈長 X, Y, Z の3パターンでそれぞれ入力し、その分布を比較・分析する。

5. オノマトペ特徴分析実験結果

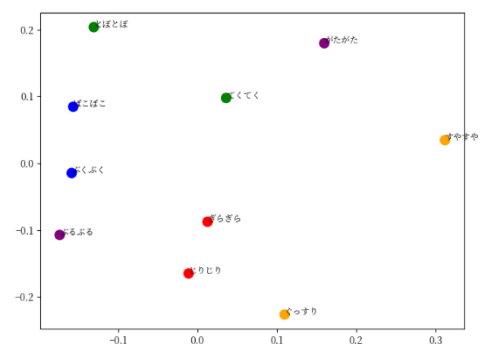
第4節で設定した実験に基づき、3つの条件下で得られたオノマトペの埋め込みベクトルの分布を可視化した結果を示す。



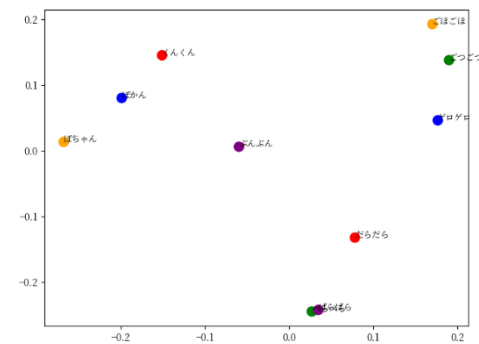
(a) カテゴリ A



(b) カテゴリ B



(a) カテゴリ C



(b) カテゴリ D

図 1：オノマトペ単体の埋め込みベクトル分布．各色のペアが対象オノマトペを示す．

5.1 オノマトペ単体の埋め込みベクトル

文脈情報を含まないオノマトペ単体を text-embedding-ada-002 に入力した場合，その分布は音韻的特徴に強く支配されることが確認された．図 1 に 4 カテゴリ（A: 音・意味類似，B: 音類似・意味非類似，C: 音非類似・意味類似，D: 音・意味非類似）のオノマトペペアを 2 次元平面にプロットした結果を示す．

図 1(a)と(b)が示すように，カテゴリ A（例：「あっさり」と「さっぱり」）およびカテゴリ B（例：「ひらひら」と「ふらふら」）では，ペアとなるオノマトペが意味の類似性に関わらず，非常に近い位置に配置された．これは，モデルがこれらのペアの音韻的・表記的な類似性を捉えていることを示唆する．

対照的に，図 1(c)と(d)が示すように，音韻的に類似していないカテゴリ C（例：「ぎらぎら」と「じりじり」）およびカテゴリ D（例：「くんくん」と「だだだ」）では，ペアとなるオノマトペが意味の類似性に関わらず，互いに遠く離れた位置に配置された．これらの結果から，文脈が与えられない場合，LLM の埋め込み表現はオノマトペの意味内容よりも，その表層的な文字列，すなわち音韻的特徴を優位に符号化していると考えられる．

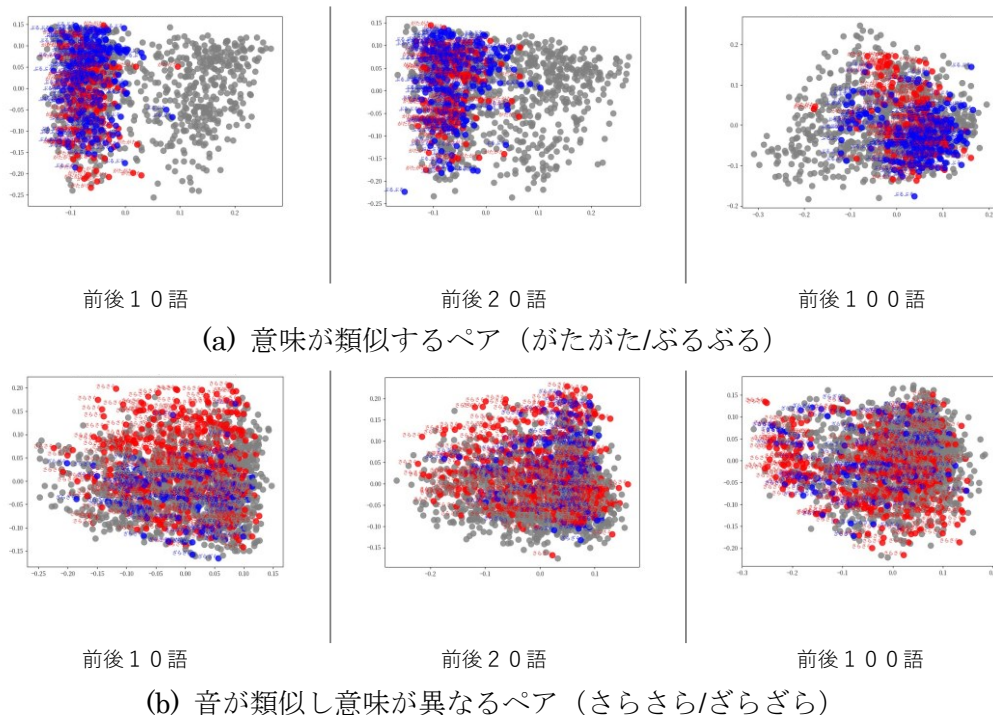


図 2：周辺文脈（前後 10 語，20 語，100 語）の埋め込みベクトル分布

5.2 周辺文脈の埋め込みベクトル

次に、オノマトペを含む周辺文脈を入力した場合、埋め込みベクトルの分布は音韻的特徴から意味的特徴を強く反映するものへと大きく変化した。

図 2 に、意味が類似しているペア（カテゴリ C：「がたがた」と「ぶるぶる」）と、音が似ていて意味が異なるペア（カテゴリ B：「さらさら」と「ざらざら」）の文脈埋め込み（前後 100 語）の分布を例示する。図 2(a)では、音韻的には異なる「がたがた」（赤）と「ぶるぶる」（青）の分布が広範囲にわたって混じり合っている。これは、文脈中において震える様子を表すという共通の意味・用法をモデルが捉え、両者を類似したものとして表現していることを示している。一方、図 2(b)では、音韻的には類似している「さらさら」（赤）と「ざらざら」（青）が、明確に分離された 2 つのクラスターを形成した。これは、文脈情報を利用することで、モデルが両者の質感に関する意味的な違い（滑らかさと粗さ）を区別できていることを示唆する。また、文脈長を長くする（前後 10 語→100 語）と、同一オノマトペ内でも複数のサブクラスターが形成される傾向が見られた。これは、オノマトペが持つ複数の語義が、文脈に応じて分離・表現されている可能性を示唆している。

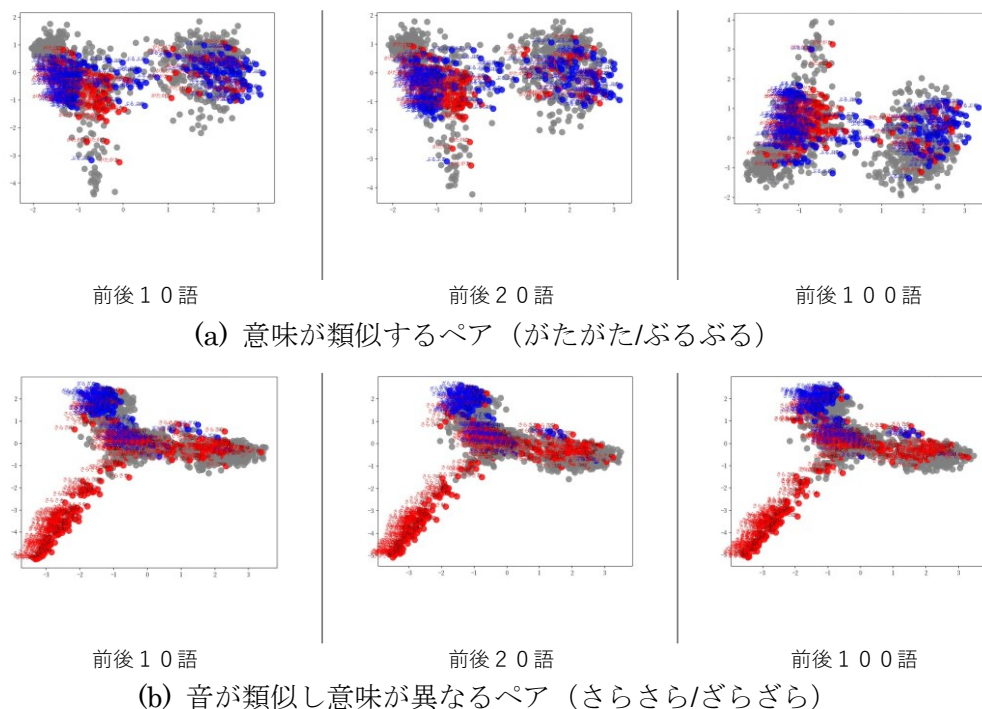


図3：平均埋め込みベクトルの分布

5.3 平均埋め込みベクトル

用例文におけるオノマトペの埋め込み (BERT) と周辺文脈の埋め込み (text-embedding-ada-002) を次元削減後に平均したベクトルでは、音韻的特徴と意味的特徴が混在した分布が観測された。図3に、図2と同じペアの平均埋め込みの分布を示す。図3(a)の意味が類似するペア (がたがた/ぶるぶる) では、5.2節と同様に分布が混じり合う傾向が見られ、意味的類似性が保持されていることがわかる。しかし、図3(b)の音が類似し意味が異なるペア (さらさら/ざらざら) に注目すると、5.2節 (図2(b)) では明確に分離していた2つのクラスが、互いに引き寄せられ、一部が重なる連続的な分布へと変化している。これは、文脈によって分離された意味的表現が、オノマトペ単体の音韻的に類似した表現ベクトルの影響を受け、再び近接したことを示している。この結果から、本手法における平均埋め込みでは、意味的特徴を捉えつつも、オノマトペ単体の持つ音韻的特徴が依然として強く影響を及ぼしていることが確認された。

6. 考察

本節では、前節で示した実験結果を統合的に解釈し、LLMの内部におけるオノマトペの表現メカニズムについて考察する。

6.1 文脈による音韻的・意味的特徴の動的な切り替え

本研究で得られた最も重要な知見は、LLMが文脈の有無に応じて、オノマトペを表現する際に重視する特徴を音韻情報と意味情報の間で動的に切り替えていることである。5.1節で示したように、オノマトペ単体の埋め込みベクトルは、その意味内容とは無関係に、音韻的・表記的な類似性によって分布が決定された。これは、モデルが文脈情報を持たない状態では、単語の表層的な特徴を捉えることを示している。一方で、5.2節で示したように、ひ

とたび周辺文脈が与えられると、その分布構造は一変し、意味的な類似性に基づいてクラスタが形成された。音韻的に類似していても意味が異なるペア（例：「さらさら」と「ざらざら」）が明確に分離されたことは、LLM が文脈を手がかりに、オノマトペが持つ微妙な意味の違いを区別していることを示唆している。さらに、5.3 節の平均埋め込みの分析では、音韻的特徴と意味的特徴が統合された結果、意味的に分離されたクラスタが音韻的類似性によって引き寄せられる現象が観測された。これは、オノマトペの内部表現が、音韻と意味という二つの異なる情報源から構成されており、文脈という条件に応じてその優位性が変化することを示している。この動的な表現の切り替え能力こそが、LLM が多様な言語現象に柔軟に対応できる一因であると考えられる。

6.2 文脈埋め込み空間における多義性の表現

周辺文脈を与えた際に、単一のオノマトペの分布が複数のサブクラスタを形成する傾向が観測された。これは、オノマトペが持つ多義性が埋め込み空間内で表現されていることを示唆している。例えば、「ふわふわ」というオノマトペは、辞書によれば、(1)「やわらかくふくらんでいるさま」、(2)「あてどもなく浮き漂うさま」、(3)「心や動作が落ち着かないさま」といった複数の語義を持つ（小野 2007）。文脈長を長くした場合に観測された複数のクラスタは、これらの異なる語義に対応する用例がそれぞれグループを形成した結果であると解釈できる。この解釈は、「ぺらぺら」や「ぐつぐつ」といったオノマトペの具体的な用例分析によっても裏付けられる。実際に、各クラスタの重心から近い用例と遠い用例を比較したところ、そこには明確な意味・用法の違いが確認された。例えば、「ぺらぺら」のクラスタでは、重心近くに「流暢に話す」という意味の用例が集中する一方、遠方には「薄い様子」を表す用例が分布していた。これは、埋め込み空間内の距離が意味的な差異と相関していることや、LLM が文脈からオノマトペの語義を特定し、それを空間的に配置していることを示していると考えられる。

7. おわりに

本研究では、大規模言語モデル（LLM）が日本語のオノマトペを内部でどのように表現しているかを、文脈の有無という観点から分析した。オノマトペ単体、オノマトペを含む周辺文脈、そして両者の平均という 3 つの条件下で埋め込みベクトルを取得し、その分布を可視化・比較することで、LLM の表現メカニズムの解明を試みた。

分析の結果、LLM は文脈が与えられない状況ではオノマトペの音韻的特徴を優位に捉え、周辺文脈が与えられると意味的特徴を重視する表現へと動的に切り替えていることを明らかにした。この知見は、LLM が単語の表層的な情報と文脈的な意味内容を、状況に応じて柔軟に使い分けていることを示す経験的な証拠であり、モデルの解釈性向上に寄与するものである。

今後の課題として、本研究で用いた分析手法をさらに発展させることが考えられる。具体的には、多義的なオノマトペの用例をあらかじめ語義ごとに分類した上で分析を行い、より精緻に意味表現の分離を検証することや、分析対象とするオノマトペを客観的な指標に基づき拡張することが挙げられる。また、本稿で対象としたモデル以外の多様な LLM に本手法を適用し、その表現戦略を比較分析することも、言語モデルの普遍的な特性と個別性の解明に繋がる課題である。

文 献

清水祐一郎・土斐崎龍一・坂本真樹（2014）「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 巻 1 号 SP1-E.
戸本裕太郎・中村剛士・加納政芳・小松孝徳（2012）「音素特徴に基づくオノマトペの可視

化」, 日本感性工学会論文誌, Vol.11 No.4 pp.545-552.

古宮嘉那子・佐々木稔・新納浩幸 (2016) 「分散表現と文脈ベクトルによるオノマトペの分類の比較」, 言語処理学会 第 22 回年次大会 発表論文集.

渡邊淳司・加納有梨紗・坂本真樹 (2014) 「オノマトペ分布図を利用した触素材完成評価傾向の可視化」, 日本感性工学会論文誌, Vol.13 No.2 pp.353-359.

藤田 実智斗・内田 ゆず・荒木 健治 (2022) 「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」, 第 38 回ファジィシステムシンポジウム講演論文集, TE3-2.

Aapo Hyvarinen, Juha Karhunen, Erkki Oja (根本幾・川勝真喜 訳) (2005) 「独立成分分析 信号解析の新しい世界」, 東京電機大学出版局.

小野正弘(編) (2007) 「擬音語・擬態語 4500 日本語オノマトペ辞書」, 小学館.

関連 URL

『現代日本語書き言葉均衡コーパス (BCCWJ)』 <https://bccwj.ninjal.ac.jp/>

OpenAI text-embedding-ada-002

<https://platform.openai.com/docs/guides/embeddings>

Hugging Face "cl-tohoku/bert-base-japanese" <https://github.com/cl-tohoku/bert-japanese>