Modeling Semantic Imagery Categories of Japanese Onomatopoeia Using Word Embeddings*

Yoshiki Sakurai

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic

Abstract

This paper focuses on the unique property called sound symbolism that Japanese onomatopoeic words have, and investigates whether semantic imagery categories can be computationally modeled using word embeddings. Our proposed method is based on the task of classifying 220 Japanese onomatopoeic words into five categories: Emotion, Movement, Sound, Texture, and Visual. We compare three different models: fastText, which captures subword information; Bidirectional Encoder Representations from Transformers (BERT), which represents text as contextual vectors; and Long Short-Term Memory (LSTM) networks over the phoneticized written form of words, which model phonemic information. The results show that fastText outperforms BERT. This suggests that subword-level morphological features provide a more robust semantic representation for onomatopoeic words, particularly under limited dataset conditions. We also observe that Visual category classification performance is significantly lower across all models, implying that text-based embeddings are insufficient to capture the mapping from visual impressions to sound. Overall, this paper demonstrates the effectiveness of machine-learning approaches for modeling sound symbolism and highlights challenges that go beyond text-only information.

Keywords

fastText, BERT, LSTM, Japanese onomatopoeia, sound symbolism, word embedding

1. Introduction

Onomatopoeic words refer to vocabulary encompassing both sound-imitative words and mimetic words. They differ in representing actual physical sound (sound-imitative words) or a state of silence or an emotion (mimetic words). Their defining characteristic is the strong, non-arbitrary connection between the sound and the word's meaning. This connection is called sound symbolism. It is thought that the motivation for sound symbolism phenomena may involve humans perceiving the properties of sound through their own physicality.

Japanese onomatopoeic words are used frequently in many levels of the language. This implies that learning those vocabularies is essential for language learners. However, this is significantly challenging because various factors may be considered, and particularly noteworthy is the difficulty in grasping the intuitive semantic nuances regarding the sound symbolism of Japanese onomatopoeic words, as discussed by Zhou and Ge [1] . In addition, in the area of natural language processing (NLP), onomatopoeic words are characterized by significant inconsistencies (e.g., the interchangeable use of hiragana and katakana for the same word) and a high degree of ambiguity; in other words, one word can have several meanings depending on context. Thus, many challenges remain regarding the computational processing of these sounds.

The sound symbolism of Japanese onomatopoeic words is understood mainly on the basis of experience and intuition. Japanese onomatopoeic words can be categorized into five categories of semantic imagery (Akita [2], Hamano [3]): Movement, Sound, Emotion, Texture, and Visual. Each onomatopoeic word can be analyzed based on those categories through computational approaches. Word embedding is a technique that converts words into dense numerical vectors in a continuous vector space, which enables computers to consider semantic relationships or context. Thus, it is possible to evaluate how well the differences based on those representations can be captured. Previous studies (Ototake et al. [4], Motomura et al. [5]) analyze the systematic sound symbolism of Japanese onomatopoeic words and verify

ITAT'26: Information Technologies – Applications and Theory, September 25–29, 2026, Biele Karpaty, Slovakia
sakuraiyoshiki682@gmail.com (Y. Sakurai)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-09

CEUR-WS.org/Vol-4323/paper16.pdf

some parts of those theories using word embeddings, while there is little quantitative evaluation of semantic imagery categories in vector space. Especially, there exists the fundamental question of which approach to semantic modeling of onomatopoeic words using computers is more important: capturing the form of words or the context of surrounding words. Thus, we investigate to what extent word embedding representations preserve the semantic imagery categories of Japanese onomatopoeic words. The aim of this paper is to construct a labeled dataset of Japanese onomatopoeic words annotated with semantic imagery categories and to carry out computational analysis using word embeddings to examine the representation of these categories in vector space.

2. What Are Japanese Onomatopoeic Words

The term "Japanese onomatopoeic words" generally refers to vocabulary that verbalizes environmental sounds, the status of things, or people's feelings, in ways similar to some sound.

In the field of Japanese linguistics, onomatopoeic words can be broadly divided into Giongo (onomatopoeic words that imitate sounds) and Gitaigo (mimetic words that describe states). For example, from Giongo,「ワンワン」(wan-wan) is a dog's barking and「ニャーニャー」(nyānyā) is a cat's meow. Additionally, from Gitaigo, 「キラキラ」(kira-kira) represents the sight of stars or gems sparkling, and 「サラサラ」(sara-sara) represents the way something flows lightly, without any dampness or stickiness. They are an important group of words not only in art but also in real life when we describe people's emotions in detail or the status of various situations vividly. As Mikami [6] notes, many onomatopoeic expressions are used in manga, which is a major motivation these days for language learners to study Japanese. To appeal to the visual sense with impact in manga, onomatopoeic words are useful in many scenes. They can create a sense of realism, as if we were watching a movie or TV, through the size of letters and layout.

2.1. Classification

Kindaichi [7] classified Giongo and Gitaigo into five word groups and named them as follows.

First, we divide Giongo into Giseigo (「擬声語」) and Giongo (「擬音語」). Giseigo (animate phonomime) represents human or animal voices. Giongo (inanimate phonomime) represents environmental sounds or noise. We note that the terms Giongo and Gitaigo are used both as subcategories.

Next, we divide Gitaigo into Gitaigo (「擬態語」), Giyogo (「擬容語」), and Gizyogo (「擬情語」). Gitaigo (phenomime) describes an inanimate state, i.e., a state of an inanimate object. Giyogo represents an animate state, i.e., a state of a living or animate creature. Gizyogo (psychomime) refers to a person's mental state or sensations such as pain. Examples of the five groups are as follows:

- Giseigo (「擬声語」): 「ワンワン」(wan-wan), 「ニャーニャー」(nyānyā), 「ゲラゲラ」(gera-gera)[1]
- Giongo (「擬音語」):「ザーザー」(zāzā),「ゴロゴロ」(goro-goro),「チリンチリン」(chirin-chirin)[2]
- Gitaigo (「擬態語」):「キラキラ」(kira-kira),「ツルツル」(tsuru-tsuru),「どんより」(donyori)[3]
- Giyogo (「擬容語」):「ウロウロ」(uro-uro),「のろのろ」(noro-noro),「とぼとぼ」(tobo-tobo)[4]
- Gizyogo (「擬情語」): 「イライラ」(ira-ira), 「ウキウキ」(uki-uki), 「うじうじ」(uji-uji)[5]

We note that it is possible for a word to belong to multiple groups, depending on the context. For example,「ゴロゴロ」(goro-goro) can belong to all five groups. While it can describe a cat purring as Giseigo or thunder rolling as Giongo, it can also be used for a heavy log rolling as Gitaigo. Furthermore, it can describe a person idling at home as Giyogo or a gritty sensation in the eye as Gizyogo. Thus, polysemy in onomatopoeic words is an interesting feature of Japanese linguistics.

[1]English translations:*gera-gera* ('guffawing').
[2]English translations: *zāzā* ('pouring rain'), *goro-goro* ('rumbling'), *chirin-chirin* ('jingling bell').
[3]English translations: *tsuru-tsuru* ('smooth/slippery'), *donyori* ('gloomy/overcast').
[4]English translations: *uro-uro* ('wandering aimlessly'), *noro-noro* ('sluggishly'), *tobo-tobo* ('trudgingly').
[5]English translations: *ira-ira* ('irritated'), *uki-uki* ('excited/cheerful'), *uji-uji* ('hesitant/moping').

2.2. Onomatopoeic Words Usage

Japanese onomatopoeic words are used not only as sound imitation, but also in verbal, adjectival, adverbial, and noun forms, as described by Tamori [8].

Verbal usage. Onomatopoeic words with 「する」(suru) function as verbs that express changes in action or state (such as 「イライラする」(ira-ira suru), meaning "frustrating"). They are more commonly found in Gitaigo or Gizyogo than in Giongo.

Adjectival usage. Onomatopoeic words modify nouns or describe a state at the end of a sentence. It is used with「の」(no) such as「ツルツルの」(tsuru-tsuru no) meaning smooth in English or with「だ」(da) such as 「キラキラだ」(kira-kira da) which is "It's sparkly" in English.

Adverbial usage. This is the most typical use of onomatopoeic words, where they modify a verb to describe the manner of the action in detail. They are used with「と」(to) such as「キラキラと」(kira-kira to) + Verb, but the adverbial use is also possible without the particle. Also, we use onomatopoeic words with 「に」 such as 「ピカピカに」(pika-pika ni) [6] + verb (sparkling).

Noun usage. They function as nouns referring to the phenomenon itself. For example, "I want「ドキドキ」(doki-doki)" means "I want excitement", or「ワンワン」(wan-wan) can refer to a dog. Tamori [8] mentioned that noun-like onomatopoeic words are typically found in baby talk, especially when pointing out animals, while some onomatopoeic words with a two-syllable repetitive form, such as「チカチカ」(chika-chika)[7], are not limited to baby talk but are also used as ordinary nouns.

2.3. Sound Symbolism

Sound symbolism is the phenomenon that causes specific meanings or images to arise directly from the sounds (echoes) of a word. Ferdinand Mongin de Saussure, a Swiss linguist, semiotician, and philosopher, proposed the principle that there is no inherent connection between sound and the meaning of language. However, sound symbolism is an exception to this arbitrariness. In other words, many onomatopoeic words are partially non-arbitrary. Sound symbolism is a phenomenon in which the sound itself carries motivation. Akita [2] mentioned that, given its inherent resistance to arbitrariness, it is expected to possess universal aspects and has long been tested in various languages.

Hamano [9] argued that Japanese onomatopoeic words are not merely a mimicry of sound, but are systematically phonological according to sound symbolism. Hamano [3] reorganized the systematic sound symbolism, according to Tables 1 and 2. Sound symbolism in Japanese onomatopoeic words has many connections with movement, emotions, texture, visual state, and the resonance of sound.

3. Previous Works

We discuss previous works related to Japanese onomatopoeic words and sound symbolism and clarify the academic context of this paper. In particular, we review previous work from the perspectives of basic research in NLP.

[6] English translation: *pika-pika* ('sparkling' or 'shining brightly').
[7] English translation: *chika-chika* ('flickering' or 'flashing').

Table 1
Consonant sound symbolism

Consonant	Sound symbolism	
p	impact with or rupture of a taut surface	lightweight, compact, and detailed
b		heavy, large, rough
t	tap a surface with low surface tension	lightweight, compact, and detailed
d		heavy, large, rough
k	hard surface, firmness, reliability	lightweight, compact, and detailed
g		heavy, large, rough
s	smooth surfaces, liquids, and fluids	lightweight, compact, and detailed
z		heavy, large, rough
n	slippery, hard to grip, sticky, sluggish	
y	shaking, unsteady movements	
h	beauty, vulnerability	
w	excitement, agitation	
m	oppression	
r	fluid movements	

Table 2
Vowel sound symbolism

Vowel	Sound symbolism
i	linear, thin, high temperature, tension
e	vulgar
a	wide, flat, extensive, conspicuous
o	inconspicuous
u	protrude

3.1. Sound Symbolism in Linguistics and Cognitive Science

Sound symbolism, which underlies onomatopoeic words, has historically been a research topic in linguistics and cognitive science. Monaghan et al. [10] discussed that while sound symbolism assists in early childhood language acquisition, the mature adult lexicon becomes increasingly arbitrary to maximize distinctiveness, prevent confusion, and ensure communicative efficiency.

On the other hand, a systematic study of Japanese onomatopoeic words is provided by Hamano [3]. She confirmed that semantic rules of onomatopoeic words are linked to their phonological and morphological characteristics. For example, the initial unvoiced and voiced consonants represent the size or weight of the target item, and the vowels represent the spatial expanse. As a typical example of the cognitive science area, the Bouba-kiki effect reported by Ramachandran and Hubbard [11] indicates that people can perceive a relationship between the resonance of sounds and the roundness or sharpness of shapes.

3.2. Onomatopoeic Words Analysis using NLP

With recent developments in NLP, experiments that treat onomatopoeic words computationally have increased. With the introduction of word embeddings and deep learning, studies that capture the semantic space of onomatopoeic words are progressing.

As an approach considering context, Ototake et al. [4] demonstrated a method to classify word meaning from surrounding context using the pre-trained model Bidirectional Encoder Representations from Transformers (BERT) to address the strong polysemy of Japanese onomatopoeic words. This study shows that LLMs based on Transformers are effective at capturing contextual information surrounding onomatopoeic words. In contrast, as an approach that focuses on morphology, Motomura et al. [5] analyzed representations of sound symbolism through contrasts of voiced and voiceless sounds

in Japanese onomatopoeic words using static word embeddings such as fastText and Word2Vec. The results showed that these static models contain phonetic and morphological features of onomatopoeic words. However, these studies remain at the verification stage from the tasks and perspectives of BERT and fastText, respectively. There are no studies that compare which model is fundamentally superior—a dynamic model specialized in understanding context or a subword-based model emphasizing word form—based on classification of diverse semantic imagery categories of onomatopoeic words.

3.3. Position of this Study

Based on the previous works mentioned above, the question of which model classifies semantic imagery categories of onomatopoeic words more effectively, BERT or fastText, remains unresolved. In addition, modeling the phonemic sequences of onomatopoeic words using LSTM to capture their semantics poses an open question. Therefore, this paper sets up a task to classify 220 Japanese onomatopoeic words into five categories (more details in Section 4.1) and compares fastText, BERT, and Long Short-Term Memory (LSTM) models. As a result, this paper contributes a point of view to quantitatively demonstrate the prevalence of morphological features in onomatopoeic words from the perspective of machine learning.

4. Methodology

The source code and datasets used in this study are publicly available on GitHub.[8]

4.1. Dataset

For a supervised classifier, we prepared an annotated dataset. The onomatopoeic-words dataset includes each word, its category, and an example sentence. We call this dataset the onomatopoeia dataset. Prior work motivates the semantic domains used in our labeling scheme. For example, Hayakawa et al. [12] discussed texture expressions focusing on Japanese onomatopoeic sound symbolism; Akita [2] summarized broader semantic classes (e.g., movement, visual, tactile experience, and emotions); and Hamano [3] argued that sound symbolism systematically corresponds to domains such as emotions, movement, sound, texture, and visual states.

Therefore, the onomatopoeia dataset is constructed so that each Japanese onomatopoeic word is assigned to one primary semantic imagery category, resulting in a single-label classification setup consisting of five classes. The five classes are the following:

- **Movement:** Words describing physical motion or manner of movement.
- **Sound:** Words directly imitating auditory phenomena.
- **Emotion:** Words describing psychological states, feelings, or internal conditions.
- **Texture:** Words describing tactile sensations, surface qualities, or material properties.
- **Visual:** Words describing visual states, appearances, or light phenomena.

We collect a subset of 220 onomatopoeic words from the dictionary by Yamaguchi [13]. According to the category definition above, each onomatopoeic word is manually categorized by the author. To ensure the selected words are actively used and observable in natural text contexts, we cross-reference them with the Wikipedia Japanese Corpus.[9] We note that many onomatopoeic words change categories depending on context. For example, 「ドキドキ」(doki-doki) can be Sound (heart beating sound) and Emotion (excitement). We restrict this subset to words that are prototypical and straightforward to categorize, and each word is annotated based on its primary dictionary entry. Additionally, onomatopoeic words can be written in either hiragana or katakana. To resolve the orthographic form (hiragana vs. katakana), a corpus-based selection method is used. The sentences are obtained from

[8] https://github.com/yoshiyan1001/embeddings-NLP
[9] https://pumpup.sakura.ne.jp/wikicorpus/japanese/en/index.html

the Wikipedia Japanese Corpus. The orthographic form that appears more frequently in contexts that match the annotated category is manually selected by the author as the standard representation for the dataset.

As a result, each entry is structured as follows: the word, its semantic category, and a representative sentence derived from the Wikipedia Japanese Corpus. It includes 55 Movement words, 45 Sound words, 45 Texture words, 41 Emotion words, and 34 Visual words.

Additionally, Maekawa et al. [14] provided a Short Unit Word list from BCCWJ (Balanced Corpus of Contemporary Written Japanese) as a text file that includes 133,489 katakana words (one of the basic Japanese spellings). This list of those words in addition to 220 Japanese onomatopoeic words obtained by the onomatopoeia dataset is called the Large Japanese Word dataset. This dataset is used with the LSTM model in Task B (which will be mentioned below).

4.2. fastText

While Word2Vec (as the alternative method) treats words as single atomic entities, fastText is specifically selected because its subword mechanism (character n-grams) is relatively suited to capture the morphological reduplication patterns inherent in sound symbolism.

4.2.1. Preprocessing

Since fastText processes only words, there is no additional preprocessing.

4.2.2. Task Design and Classification

We use a pre-trained model on a large-scale corpus for Japanese, downloaded from the fastText website (cc.ja.300.bin).[10] The fastText treats a Japanese word as a set of character *n*-grams (subwords), rather than as a single unit. For example, 「ドキドキ」(doki-doki) is calculated as the sum of vectors of the subword pieces: < ドキ, キド, ドキ >. fastText is useful for onomatopoeic-word analysis because Japanese onomatopoeic words have many derivatives, such as 「ドキドキ」(doki-doki), 「ドキリ」(dokiri), and 「ドキン」(dokin), which share the root 「ドキ」(doki). The subword design is also appropriate for capturing reduplication patterns and inferring representations for unseen words from their character structure. In contrast, Japanese onomatopoeic words have spelling variations in hiragana (e.g.,「どきどき」; doki-doki) and katakana (e.g.,「ドキドキ」; doki-doki). They share the same pronunciation and can be written in romanization (e.g.,「どきどき」and「ドキドキ」→ "doki-doki"). However, their Unicode code points differ (e.g., 「ど」→ $U + 3069$ vs. 「ド」→ $U + 30C9$, romanized to "do").

Therefore, because the fastText subword mechanism depends on string collation at the Unicode level, fastText cannot recognize morphological similarity between hiragana and katakana spellings of the same word. To address this Japanese script-variation issue, future work could incorporate romanization (converting Japanese spellings to Latin spellings) as an additional representation so the model can capture the similarity between hiragana and katakana forms.

4.3. BERT

4.3.1. Preprocessing

Because Japanese text is not whitespace-delimited, tokenization and morphological segmentation are required before contextual embedding. In particular, BERT may split an onomatopoeic word into multiple subword tokens, which motivates the matching procedure described below. We use a pre-trained Japanese BERT model (cl-tohoku/bert-base-japanese-v3)[11] with `BertJapaneseTokenizer`, which performs tokenization, adds `[CLS]` and `[SEP]`, converts tokens to token IDs, and truncates sequences

[10] https://fasttext.cc/docs/en/crawl-vectors.html

[11] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

longer than 128 tokens. The tokenizer performs morphological analysis with MeCab introduced by Kudo [15] and subword segmentation with WordPiece. For example, after MeCab processing:
「私は擬音語が大好きです。」is segmented as「私 (noun)/ は (particle)/ 擬音語 (noun)/ が (particle)/ 大好き (adjectival noun)/ です (auxiliary verb)/ 。(symbol)」(watashi / wa / giongo / ga / daisuki / desu).

If a MeCab-segmented word is not in the BERT vocabulary, it is further split by WordPiece. For example,「ドキドキ」(doki-doki) may be split into `[' ドキ', '## ドキ']` (`'##'` means it is joined to the preceding piece). This mitigates the out-of-vocabulary (OOV) problem and prevents the input from becoming entirely `[UNK]`.

Additionally, Japanese onomatopoeic words have multiple related forms based on the same root. For example,「キラ」(kira) is the root and can derive「キラキラ」(kira-kira),「キラリ」(kirari), and「キラッ」(kira'). These can be processed as subwords, which may help the model learn associations between similar-sounding forms even with limited data.

4.3.2. Task Design and Classification

To extract features, we try to obtain contextualized embeddings to resolve polysemy of Japanese onomatopoeic words. We extract the vectors corresponding to the target onomatopoeic words from the last hidden state obtained by inputting the full text into BERT.

We implement sliding window search to process the case of multiple tokens in one onomatopoeic word due to BERT subword segmentation. Sliding window search is an algorithmic technique that efficiently finds a solution by shifting a contiguous sub-range (window) of an array to the right, thereby avoiding unnecessary recalculations. We calculate a single vector representing a target word by identifying the sequence of tokens of the target word in the sequence of tokens of a whole sentence, and then we average the vectors of all matching tokens (average pooling).

To clarify this process and show how we avoid improperly merging different onomatopoeic words in the same sentence, we use the following example:
「キラキラ光るプレゼントにドキドキした。」(Kirakira hikaru purezento ni doki-doki shita.) means "My heart raced at the sight of the sparkling gift.". This sentence includes two onomatopoeic words, which are「キラキラ」(kira-kira) and「ドキドキ」(doki-doki). When processed by the BERT tokenizer, this sentence can be split into the following subword tokens:
`['[CLS]', ' キラ', '## キラ', ' 光る', ' プレゼント', ' に', ' ドキ', '## ドキ', ' し', ' た', '。', '[SEP]']`
If the target word is「キラキラ」(kira-kira), the sliding window picks up only the subwords `[' キラ', '## キラ']` and performs pooling. It strictly avoids merging the「ドキドキ」(doki-doki) token into this calculation. This is because, if we merged it, the independent semantic features would be mixed up, and the semantic representation in the vector space would be distorted.

In addition, when we are unable to identify the target word due to inconsistencies in morphological analysis, we ensure the robustness of the extraction process by using the vector of the `[CLS]` token, which holds a semantic summary of the whole sentence, as a fallback. Also, if any word or character doesn't exist in the model's predefined vocabulary, then unknown token `[UNK]` is allocated. The task design of BERT pre-training is Masked Language Modeling (MLM). Since Japanese onomatopoeic words are highly polysemous (e.g.,「ふわふわ」(fuwa-fuwa) can describe a soft texture or a floating/-light feeling), their meaning cannot be determined from the isolated word alone. Therefore, BERT's bidirectional context modeling is essential for capturing the semantics of onomatopoeic words from surrounding context.

4.4. Phoneme-Based LSTM Model: Task A and Task B

In the computational methods covered in the previous sections, we focus mainly on character strings of onomatopoeic words and classification of semantic imagery categories based on morphological and contextual features. However, as theoretic background of sound symbolism indicates, the essence of

onomatopoeic words is the non-arbitrary connection between sound itself and its semantic image. In addition to analysis of textual information, it is also important to consider the analysis of onomatopoeic words as phonemic series.

Therefore, we use the LSTM network and design a model to capture phonemic sequences or phonological rules unique to Japanese onomatopoeic words. While standard feed-forward networks or Transformer models could be considered, Transformers require significantly more data and compute, whereas an LSTM provides a highly efficient and well-suited baseline for modeling the short, character-level sequential dependencies of isolated onomatopoeic words. We set up two tasks that have different compositions of the training data to validate to what extent the extracted hidden vector represents the semantic imagery categories in vector space using the LSTM model. The dataset in Task A consists solely of the onomatopoeia dataset. In addition to the onomatopoeia dataset, Task B is trained on the Large Japanese Word dataset.

Many general Japanese words have an arbitrary form–meaning relationship, whereas many onomatopoeic words do not. Therefore, by comparing Tasks A and B, we can verify how adding general Japanese words to the onomatopoeia dataset affects semantic representations in the vector space.

4.4.1. Preprocessing

For the phoneme-based LSTM model, we apply text normalization. Using `jaconv`, we standardize katakana converted from all hiragana and katakana since the phoneme map (`kanaMap`) is based on katakana. The `kanaMap` is constructed according to the Phonetic Symbol Chart for Syllables in Model Japanese according to Dual Daijirin.[12]

We apply the following four processes to help the model learn the phonological features of Japanese:

1. Long vowel 「ー」 is processed by repeating the preceding vowel, e.g., 「オー」 → [o,o]. This represents sound duration in sound symbolism.
2. For the geminate consonant 「ッ」, we process it according to its phonetic function of doubling the following consonant. Once 「ッ」 is found, we anticipate the next letter's initial consonant (e.g., 「パ」 → p) and copy it, such as 「パッパ」(p a p p a). Kubozono [16] notes that Japanese geminate consonants characterize Japanese rhythm. Hamano [3] argued that gemination used in infixes (e.g.,「パッチリ」(pacchiri)) can express a stronger force or severity in sound symbolism.
3. We process the assimilation of the place of articulation for the nasal sound 「ン」. Following Kuribayashi [17], we distinguish three types of phonemes (m, n, N) depending on the following sound:
   - If the next consonant is `b`, p, or `m` (labial), use `m` (e.g., 「サンポ」(sampo)).
   - Else if the next consonant is `k` or `g` (velar), use `N` (e.g., 「マンガ」 [maNga]).
   - Otherwise, use `n`.

   Hamano [3] notes that 「ン」 at the end of a word can express a lingering sound (e.g., 「パタン」(patan)).
4. For palatalized consonants (e.g.,「シャ」(sya) and「チャ」(cha)), although they are written with two characters (Kana + small character), they behave as a single unit of sound that determines syllable weight. We therefore implement a greedy matching algorithm that captures palatalized consonants by dictionary lookup on the next two characters.

After preprocessing, we insert tokens <SOS> and <EOS> at the beginning and end of each word, respectively, to train phonological rules.

4.4.2. Task Design and Classification

The model structure for the phoneme-based LSTM is an embedding layer plus a Long Short-Term Memory (LSTM) network. We construct a phoneme language model and train it not on the original

[12] http://daijirin.dual-d.net/extra/nihongoon.html

Model (SVM)	Macro-F_1	Accuracy	F_1-Emotion	F_1-Movement	F_1-Sound	F_1-Texture	F_1-Visual
fastText	0.84	0.85	0.89	0.86	0.78	0.88	0.79
BERT	0.81	0.81	0.86	0.80	0.84	0.83	0.69
LSTM Task A (Ono-only)	0.30	0.31	0.18	0.41	0.42	0.14	0.37
LSTM Task B (General)	0.28	0.29	0.29	0.44	0.21	0.26	0.21
Majority Class Baseline	0.08	0.25	0.00	0.40	0.00	0.00	0.00

Table 3
Overall classification performance for each model with a linear SVM classifier.

written form but on its conversion into phonemes. For example, 「ザーザー」(zāzā) is converted into `['z', 'a', 'a', 'z', 'a', 'a']`, which is used as an input sequence for the model. This model is composed of three layers: (i) an embedding layer that converts input phoneme IDs to continuous vectors, (ii) an LSTM layer that maintains and updates time-series information, and (iii) a fully connected (linear) layer for next-phoneme prediction. This architecture is designed to retain early phonetic information through the end of the sequence via the LSTM cell state.

Its task design is autoregressive language modeling. Given an input sequence prefixed with <SOS>, we train the model to predict the next phoneme step-by-step until <EOS> with cross-entropy loss.

The primary objective of this task design is to extract the last hidden state of the model. By solving the autoregressive next-sound prediction task (rather than merely modeling phoneme addition), the hidden representation is intended to encode phonotactic rules specific to Japanese onomatopoeic words and the structural characteristics of sound symbolism.

4.5. Evaluation and Visualization

We adopt a unified evaluation and validation framework to compare three models (fastText, BERT, and LSTM) that use different features. We perform all evaluations using frozen representations extracted from the respective models. While fastText and BERT utilize embeddings extracted from the pre-trained language models, the LSTM representations are extracted from the hidden states of our custom-trained phonological model. The resulting vector representations have varying dimensionalities (fastText: 300, BERT: 768, LSTM: 64). We use vectors of these different dimensionalities as inputs to a simple machine learning method and see to what extent it can identify the class of the onomatopoeic word in question. This method is called linear probing, and it estimates how accurately a model can perform classification. Importantly, this classification task is performed without further training of the representation model (fastText, BERT, or LSTM). Only the classifier is trained on the fixed representations because we would like to study the models' internal representations, without additional training. To verify the linear separability of each semantic imagery category in the feature space, we use a linear support vector machine (SVM) classifier.

For metrics, considering the unbalanced onomatopoeia dataset ($N = 220$) across categories, we perform stratified 5-fold cross-validation and report the Macro-F_1 score. To complement quantitative evaluations, we perform a qualitative analysis using the visualization of the high-dimensional vectors. Using the Uniform Manifold Approximation and Projection (UMAP) method, we project higher feature space dimensions created by each model onto two-dimensional space.

5. Results

5.1. Overall Classification Performance

Table 3 reports the Macro-F_1 score and accuracy for each model with a linear SVM classifier. To provide a lower bound for our results, we establish a majority class baseline of 0.25, representing a classifier that constantly predicts the most frequent category (Movement). According to the experimental results, the highest score across all models is achieved by fastText (Macro-F_1: 0.84), which is 0.03 points higher

than BERT (0.81). On the other hand, the LSTM models perform significantly worse than the other two models (Task A: 0.30; Task B: 0.28).

5.2. Category-Specific Performance

Next, we compare the classification performance for each of the five semantic categories (Emotion, Movement, Sound, Texture, Visual). We confirm that each model shows different strengths and weaknesses across categories (Table 3). While fastText records the highest score overall, it demonstrates particularly high performance for Texture (SVM: 0.88) and Emotion (SVM: 0.89). BERT shows high scores for Emotion (SVM: 0.86) and Sound (SVM: 0.84). The Sound score for BERT is higher than that of fastText. Across all models, it is observed that the Visual category tends to have relatively low scores. For fastText, Visual and Sound are the only categories with scores below 0.8, and for BERT, the Visual category score is also relatively low (SVM: 0.69).

5.3. The Result of Tokenization in BERT

Table 4 shows the result of BERT tokenization across the semantic categories. Within all 220 words, there are 161 words (73.2%) that are fragmented into 3 or more tokens. The fragmentation rate by category is highest for Movement at 83.6%, followed by Texture at 82.2%. In contrast, Sound, Visual, and Emotion have lower rates: 66.7%, 64.7%, and 63.4%, respectively. Across all categories, there are zero words that include the unknown word (`[UNK]`) token. In addition, a fallback to `[CLS]` occurs in 5 examples across all categories, which is about 2.3%. By category, the Visual category has the highest fallback rate at 5% (2 cases), followed by Emotion at 5% (2 cases) and Movement at 2% (1 case).

Table 4
BERT Tokenization Result across Semantic Categories

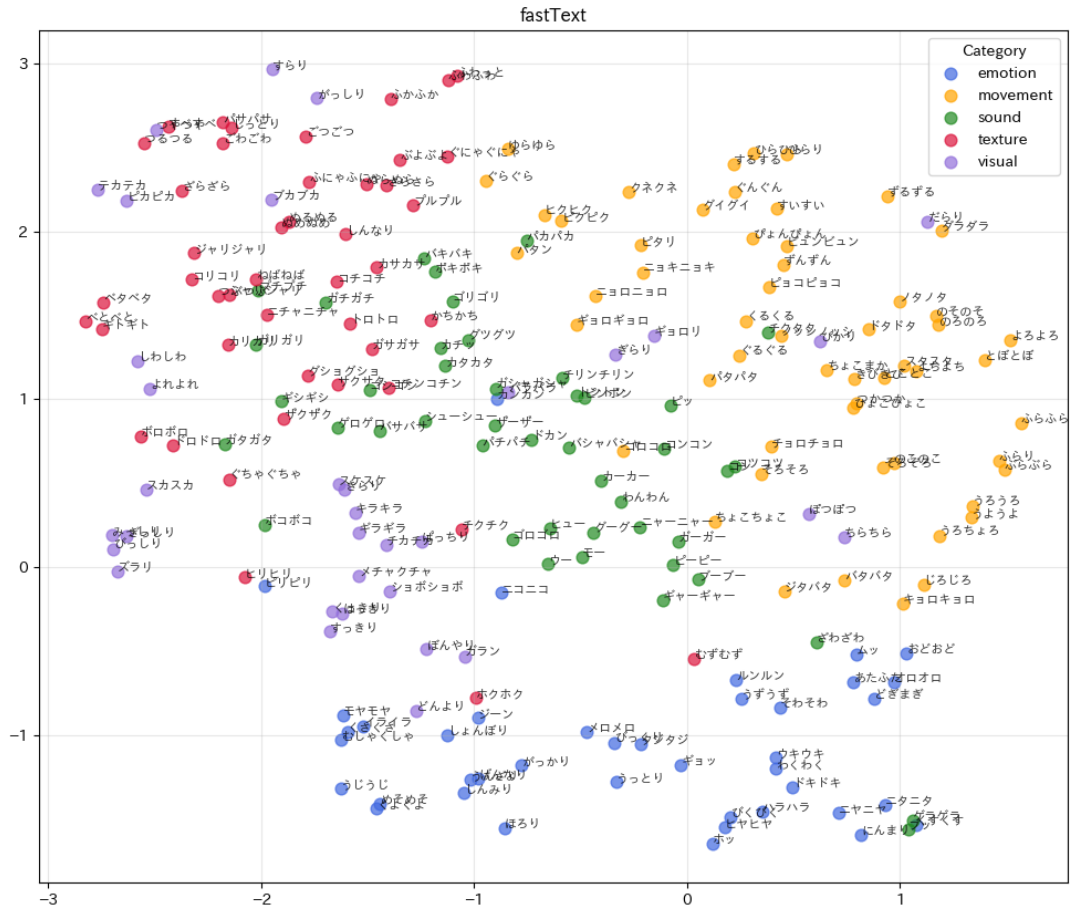
Category	Total (N)	Fallback to [CLS]	Contains [UNK]	Fragmented (≥ 3)
Sound	45	0 (0.0%)	0 (0.0%)	30 (66.7%)
Movement	55	1 (1.8%)	0 (0.0%)	46 (83.6%)
Texture	45	0 (0.0%)	0 (0.0%)	37 (82.2%)
Visual	34	2 (5%)	0 (0.0%)	22 (64.7%)
Emotion	41	2 (4.9%)	0 (0.0%)	26 (63.4%)
Total	220	5 (2.3%)	0 (0.0%)	161 (73.2%)

5.4. Comparison of LSTM Tasks

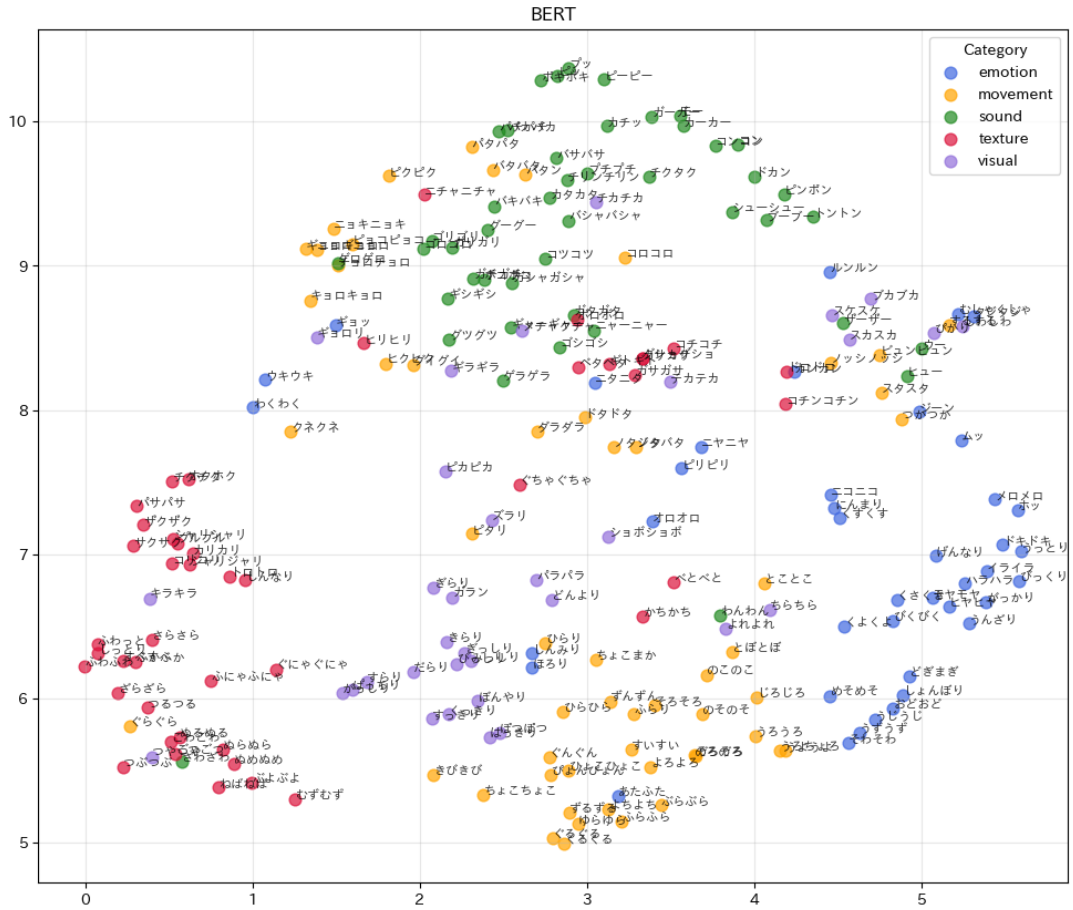
We compare the effect of training data composition (Task A vs. Task B) for the LSTM model. Task A, trained only on onomatopoeic words, records higher classification scores (SVM Macro-F_1: 0.30 vs. 0.28) than Task B, trained additionally on general Japanese words.

5.5. Qualitative Visualization of Embedding Spaces

To support the quantitative evaluation scores, we compress and visualize the high-dimensional vectors extracted from each model to two-dimensional space using UMAP (fig. 1). FastText's visualization shows the formation of independent clusters quite clearly, particularly Texture (red) and Emotion (blue). BERT's visualization demonstrates a tendency for clusters to form; however, compared to fastText, the boundaries between categories are more ambiguous, and the distribution shows a mixture of multiple categories, particularly near the center of the space. LSTM visualizations form many dense micro-clusters in both Task A and Task B. However, individual clusters often include multiple mixed semantic categories, and they do not correspond to the classification based on the five defined semantic categories. Additionally, Task A extends more widely throughout the entire space than Task B.

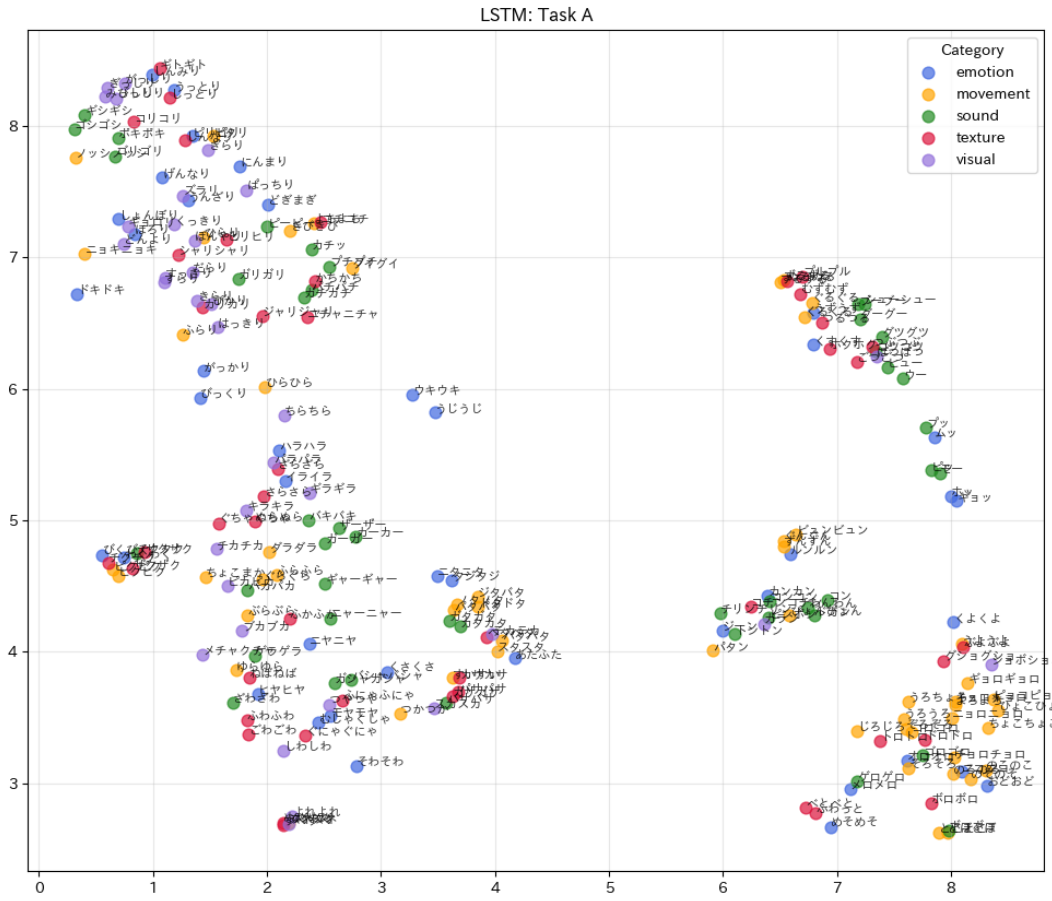


(a) fastText

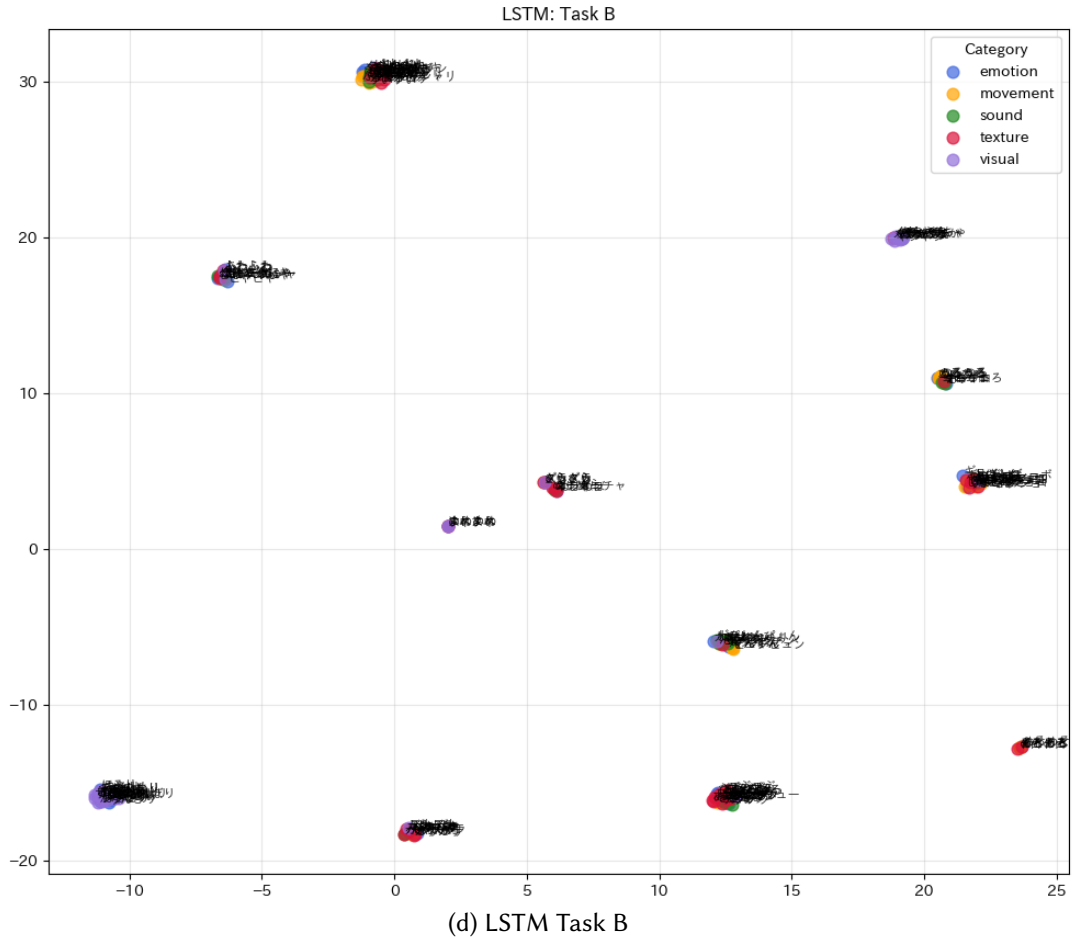


(b) BERT

Figure 1: UMAP visualization of embedding spaces for each model. Colors indicate semantic categories (Emotion, Movement, Sound, Texture, Visual).



(c) LSTM Task A



(d) LSTM Task B

Figure 1: UMAP visualization of embedding spaces for each model (continued).

6. Analysis and Discussion

6.1. Why fastText is Better Than BERT

According to this classification experiment, we obtain the very interesting result that fastText (SVM: 0.84), a static embedding method, produces clearly higher scores than BERT (SVM: 0.81), which typically holds state-of-the-art performance in common natural language processing tasks. This reversal phenomenon can be explained by special linguistic properties of Japanese onomatopoeic words. Three perspectives are stated below.

6.1.1. Capturing Sound Symbolism via Character *n*-grams

We assume that fastText's key advantage is its mechanism of representing words as character *n*-grams (subwords). Japanese onomatopoeic words have strong semantic and sensory associations (sound symbolism) that are closely tied to morphological patterns, as well as markers like 「っ」(geminate consonant) and「ん」(nasal consonant). Because fastText breaks words down into character-level components, it can capture morphological patterns that are central to Japanese onomatopoeic words. In contrast, automatic tokenization of subwords in BERT (e.g., WordPiece) can split words into fragments (such as ' ドキ' (doki) and '## ドキ'), which can weaken the visual and auditory imagery conveyed by the character sequence. As a result, fastText, which preserves this morphological information, can better represent essential characteristics of onomatopoeic words.

6.1.2. Static vs. Contextual Representations

Another key difference is whether the vector representation is static or contextual. Ototake et al. [4] reported that classifying onomatopoeic words with BERT relies heavily on contextual information. Even when the onomatopoeic form is the same, BERT is susceptible to differences in syntactic usage. Additionally, if words from different categories appear in similar surrounding contexts (e.g., expressions with 「とする」 (to suru)), the classifier may be more prone to confusion.

Each onomatopoeic word in the onomatopoeia dataset is assigned only one golden semantic imagery category (i.e., the annotation does not allow polysemy, rather than homonymy). Because fastText provides one fixed vector for each word regardless of context, its output vector aligns naturally with this single label annotation.

In contrast, BERT's contextual architecture is designed to capture polysemy based on the surrounding context. However, this strength becomes a disadvantage under this annotation rule. The context-dependent vector variations can work as noise, so this causes a factor that reduces the accuracy of the classification. As a result, fastText is less affected by contextual variation and can better capture the semantic concept associated with the word form, making the semantic classes easier to distinguish using linear decision boundaries (fig. 1a).

6.1.3. Generalization Ability of Subword Information

One notable difference between fastText and BERT is the structural approach for the subword expression and how this approach affects morphological normalization. Since many onomatopoeic words in the onomatopoeia dataset consist of four characters, BERT may split them into multiple subword tokens. In such cases, BERT can struggle to infer a word's meaning from those small components. Table 5 shows that comparing the categories of Sound and Texture (both 45 words), the fragmentation rate for Texture (82.2%) is higher than for Sound (66.7%); nevertheless, the SVM F_1 score for Sound (0.84) is higher than for Texture (0.83), while Table 4 shows that no unknown word is found in all categories. A high fragmentation rate does not automatically degrade performance; instead, it may force BERT to rely entirely on surrounding context.

Furthermore, according to Table 5, the fragmentation rate of Visual (64.7%) is close to that of Sound (66.7%) and Emotion (63.4%). However, BERT keeps high F_1 score in Emotion (0.86) and Sound (0.84), on

Table 5
BERT results by category.

Category	Total	Fragmented rate (%)	F_1 (SVM)
Sound	45	66.7	0.84
Movement	55	83.6	0.80
Texture	45	82.2	0.83
Visual	34	64.7	0.69
Emotion	41	63.4	0.86

the other hand, the performance for Visual category is relatively low. In addition, the fallback rate for the Visual category is the highest (5.88%), and the size of dataset for Visual (34 words) is the smallest across all categories. This suggests that the small size of the data leads to a loss of precision in the classification task in BERT.

Meanwhile, fastText can still infer a word's meaning well if the character *n*-grams overlap, even when onomatopoeic words have a low frequency in the corpus. For example, if the model learns 「サラサラ」(sara-sara), it can infer a reasonable vector for 「サラッ」(sara') as an unseen word. We can therefore estimate that, although each category is small (only a few dozen words), fastText achieves more stable performance across categories, including Visual, by compensating for limited data via subword generalization.

6.2. Separating Phonology and Semantics

Our results demonstrate that the isolated phonemic sequence of an onomatopoeic word does not carry enough sound-symbolic information to classify its semantic meaning.

The LSTM model scores significantly lower than the contextual models. In the UMAP visualization (fig. 1, panels [c] and [d]), the data points from the five semantic categories are thoroughly mixed. This failure stems directly from the fact that the LSTM is trained strictly on isolated words, completely deprived of surrounding sentence context.

While BERT and fastText capture meaning through a word's co-occurrence with neighboring words in large sentences, the LSTM has absolutely no exposure to semantic usage in a natural environment. Combined with a much smaller training dataset, this lack of context forces the LSTM to rely entirely on formal phonology.

This outcome highlights an important linguistic property: the separation between phonological structure and semantic imagery of onomatopoeic words, as pointed out by Nasu [18]. Because the LSTM only sees isolated phonemes, it groups words based on prosodic rules rather than meaning.

7. Conclusion

This paper empirically evaluates three different approaches (fastText, BERT, and LSTM) on their ability to classify 220 Japanese onomatopoeic words into five semantic imagery categories.

We show that choosing an appropriate model that matches the morphological characteristics of the target linguistic phenomenon is important. In data-scarce conditions, fastText achieves the highest score, demonstrating the robustness of subword-level morphological features. While BERT may overfit differences in syntactic usage, fastText can reduce contextual noise and better capture lexical semantics. However, we note that our single-label annotation scheme inherently disadvantages BERT's contextual strengths, presenting an avenue for future refinement. In contrast, the phoneme-based LSTM model achieves low scores on the classification task. We see this as a sign that Japanese onomatopoeic words are not as phonologically unique as one might think, since they follow common Japanese phonological and morphological patterns.

The possible future work focuses on expanding the dataset size and introducing a multi-modal approach that integrates visual and audio data to more accurately capture human synesthetic mapping.

Acknowledgments

This research was partially supported by SVV.

Declaration on Generative AI

During the preparation of this work, the author has used the Overleaf AI assistant as an auxiliary tool to improve the stylistic and linguistic quality of written content. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] C. Zhou, Q. Ge, A study on Chinese Japanese learners' acquisition of onomatopoeia: Taking phonetic symbolism as an example, 2022.
[2] K. Akita, Onomatopoe・onshōchō no kenkyūshi [a history of research on onomatopoeia and sound symbolism], in: K. Shinohara, R. Uno (Eds.), Onomatopoe kenkyū no shatei: Chikazuku oto to imi, Hituzi Syobo, 2013, pp. 333–364.
[3] S. Hamano, Nihongo no onomatope: Onshōchō to kōzō [Japanese onomatopoeia: Sound symbolism and structure], Kurosio Publishers, 2014.
[4] H. Ototake, Y. Uchida, K. Takamaru, Y. Kimura, A method for classifying onomatopoeia's senses using BERT's contextual information 32 (2020) 518–522.
[5] S. Motomura, Y. Kubo, Y. Nozaki, M. Sakamoto, An analysis of sound symbolism by voiced-voiceless contrast in japanese onomatopoeia using word embeddings, in: Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2024), Hamamatsu, Japan, 2024.
[6] K. Mikami, Selection and teaching of the basic Japanese onomatopoeias, ICU Studies in Japanese Language Education (2007) 49–63.
[7] H. Kindaichi, Giongo gitai-go gaisetsu [An overview of onomatopoeia and mimetics], Kadokawa Shoten, 1978.
[8] I. Tamori, Characteristics of Japanese onomatopoeic words, Journal of the Acoustical Society of Japan 54 (1998) 215–222. doi:10.20697/jasj.54.3_215.
[9] S. Hamano, The Sound-Symbolic System of Japanese, volume 10 of *Studies in Japanese Linguistics*, illustrated ed., Cambridge University Press, 1998.
[10] P. Monaghan, R. C. Shillcock, M. H. Christiansen, S. Kirby, How arbitrary is language?, Philosophical Transactions of the Royal Society B: Biological Sciences 369 (2014). doi:10.1098/rstb.2013.0299.
[11] V. S. Ramachandran, E. M. Hubbard, Synaesthesia—a window into perception, thought and language 8 (2001) 3–34.
[12] T. Hayakawa, S. Matsui, J. Watanabe, Classification method of tactile textures using onomatopoeias, Transactions of the Virtual Reality Society of Japan 16 (2010) 327–330.
[13] N. Yamaguchi (Ed.), Giongo gitaigo jiten [Dictionary of Onomatopoeia], Kodansha Gakujutsu Bunko, 2015.
[14] K. Maekawa, M. Yamazaki, T. Ogiso, T. Maruyama, H. Ogura, W. Kashino, H. Koiso, M. Yamaguchi, M. Tanaka, Y. Den, Balanced corpus of contemporary written Japanese, Language Resources and Evaluation 48 (2014) 345–371. doi:10.1007/s10579-013-9261-0.
[15] T. Kudo, Mecab: Yet another part-of-speech and morphological analyzer, http://mecab.sourceforge.net/ (2005). URL: https://cir.nii.ac.jp/crid/1572543025344508032.
[16] H. Kubozono, Nihongo no sokuon to akusento [geminate obstruents and accent in japanese] 13 (2017) 3–15.
[17] H. Kuribayashi, Nihongo no hatsuon to sokuon no onsei tokuchou [Phonetic characteristics of the Japanese moraic nasal and geminate consonant], Nihon Daigaku Kokubungakkai, Tokyo, 1991.
[18] A. Nasu, Nihongo onomatope no gokeisei to inritsu kozo [word formation and prosodic structure of japanese onomatopoeia], 2002.