Investigating Iconicity in Vision-and-Language Models: A Case Study of the *Bouba/Kiki* Effect in Japanese Models

Hinano Iida (iida.hinano.i1@s.mail.nagoya-u.ac.jp)
Department of English Linguistics, Nagoya University, Furo-cho, Chikusa-ku, Nagoya
Aichi 464-8601 Japan

Hayate Funakura (funakura.hayate.28p@st.kyoto-u.ac.jp)
Graduate School of Human and Environmental Studies, Kyoto University, Yoshida-Honmachi, Sakyo-ku
Kyoto 606-8501 Japan

Abstract

Extensive evidence from diverse areas of the cognitive sciences suggests that iconicity—the resemblance between form and its meaning—is pervasive and plays a pivotal role in the processing, memory, and evolution of human language. However, despite its acknowledged importance, iconicity in language models remains notably underexplored. This paper examines whether Japanese language models learn iconic associations between shape and sound, known as the *bouba/kiki* (or *maluma/takete*) effect, which has been widely observed in human language as well as English and multilingual vision-and-language models, including Finnish, Indonesian, Hungarian, and Lithuanian models in previous studies. A comparison between the current results and the previous studies revealed that Japanese models learn language-specific aspects of iconicity, such as the associations between /p/ and roundness, and /g/ and hardness, reflecting the sound symbolic system in Japanese.

Keywords: iconicity; sound symbolism; *bouba/kiki* effect; crossmodal correspondence; language-specificity; vision-and-language model; embodiment

Introduction

Iconicity—traditionally defined as the resemblance between form and meaning (Peirce, 1932)—is pervasive at different levels of human language, both signed and spoken langauge (Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015; Perniss, Thompson, & Vigliocco, 2010). For example, certain phonemes can be associated with certain meanings, a phenomenon known as “sound symbolism”. Examples include shape sound symbolism, known as the *bouba/kiki* (or *maluma/takete*) effect, in which sonorant and bilabial consonants (e.g., /m/, /l/, /b/) and rounded vowels (e.g., /o/, /u/) are associated with roundness, whereas voiceless stops (e.g., /t/, /k/) and high-front vowels (e.g., /i/) are associated with spikiness (Köhler, 1929; McCormick, Kim, List, & Nygaard, 2015; Ramachandran & Hubbard, 2001). This effect has been robustly demonstrated cross-linguistically using pseudowords (Ćwiek, Fuchs, Draxler, Asu, Dediu, Hiovain, Kawahara, Koutalidis, Krifka, Lippus, & Lupyan, et al., 2021), and also observed in object nouns in English (Sidhu, Westbury, Hollis, & Pexman, 2021), as well as among preverbal infants as young as 4 months old (Ozturk, Krehm, & Vouloumanos, 2013).

Another example of iconicity is ideophones, including onomatopoeia (e.g., *bowwow*, *cock-a-doodle-doo*), which abound in many of the world’s languages. They depict sensory information in an imitative fashion (e.g., *kira-kira* ‘twinkling’ in Japanese), illustrating word-level iconicity (Dingemanse, 2012). This pervasive presence of iconicity indicates that it is “a general property of language” (Perniss, Thompson, & Vigliocco, 2010).

Furthermore, compelling evidence from diverse areas of the cognitive sciences suggests that iconicity plays a pivotal role in the processing (Sidhu, Vigliocco, & Pexman, 2020), learning (Imai, Kita, Nagumo, & Okada, 2008; Kantartzis, Imai, & Kita, 2011; Imai & Kita, 2014; Imai, Miyazaki, Yeung, Hidaka, Kantartzis, Okada, & Kita, 2015), memory (Sonier, Poirier, Guitard, & Saint-Aubin, 2020; Sidhu, Khachatoorian, & Vigliocco, 2023), and evolution of human language (Akita & Imai 2020; Imai & Kita 2014). However, despite its recognized significance in human language, iconicity in language models is still one of the least explored areas. An exception can be the experiments conducted by Alper and Averbuch-Elor (2023), who investigated shape sound symbolism (the *bouba/kiki* effect) with English vision-and-language models and multilingual vision-and-language models (Finnish, Indonesian, Hungarian, and Lithuanian). They concluded that vision-and-language models learn the association between sound and shape, such as the associations between voiceless stops (e.g., /p, t, k/) and “sharpness” and between voiced stops and sonorants (e.g., /b, d, m, n, l/) and “roundness”, paralleling with human perception demonstrated by studies conducted in psycholinguistics (McCormick et al., 2015).

The purpose of this paper is to replicate the findings of Alper and Averbuch-Elor (2023) using Japanese vision-and-language models, employing the same probing methodology as theirs. We investigate whether models trained on general image generation tasks are aware of the associations between sound and shape, without any additional training specifically for these associations, as conducted by Alper and Averbuch-Elor (2023). Our study reveals significant divergences from their reported findings, shedding light on the language-specific, systematic dimensions of iconicity originating from the sound symbolic system of imitative words (i.e. ideophones) in Japanese. Our data and codes will be made

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

publicly available at: https://github.com/hfunakura/vlm-ssw-ja.

Related Work

Universality and Language-Specificity of Iconicity

Iconic form-meaning associations are assumed to be identifiable across languages because iconicity relies on perceptuomotor analogies (i.e., biological bases) (Akita & Imai, 2022; Dingemanse et al., 2015; Imai & Kita, 2014). For example, shape sound symbolism has been observed among speakers of 25 languages from nine unrelated language families, including English and Japanese (Ćwiek et al, 2022). Words containing bilabials (e.g., /m, b/), sonorants (e.g., /m, n, l/), and rounded back vowels (e.g., /u/, /o/) such as *bouba* and *maluma*, sound round, and words containing non-bilabial obstruents (e.g., /t, k, s, d, g, z/) and unrounded front vowels (e.g., /i, e/) such as *takete* or *kiki* sound sharp. These sound-shape associations are based on articulatory (e.g. rounded vowels corresponding to round shapes) (Akita & Imai, 2022; Ramachandran & Hubbard, 2001; Sapir, 1929; Sidhu & Pexman, 2018) and/or acoustic features of the sounds (e.g., the gradual amplitude changes of sonorants corresponding to rounded shapes) (D'Onofrio, 2014).

However, recent studies have revealed that not all types of iconic form-meaning associations are necessarily universal or biologically grounded, indicating that iconicity can be language-specific. For example, Japanese exhibits somewhat unique associations between voicing and meaning contrasts, which speakers of other languages may not perceive. Iwasaki, Vinson, and Vigliocco (2017) asked native Japanese speakers and English speakers to rate Japanese ideophones for laughing (e.g., *gera-gera* 'laughing loudly', *kusu-kusu* 'giggling') and walking (e.g., *peta-peta* 'slapping', *yochiyochi* 'toddling') on 14 semantic differential scales, such as Loud–Soft (in volume), Graceful–Vulgar for laughing, and Big strides–Small strides, Feminine–Masculine for walking. Their findings revealed that English-speaking participants failed to accurately guess the semantic contrast between voiced and voiceless obstruents, especially for ideophones for manner of walking (e.g., *bura-bura* 'wandering' vs. *furafura* 'walking unsteadily'). Saji, Akita, Kantartzis, Kita, and Imai (2019) used a production-elicitation task and demonstrated that the systematic associations between voiceless consonants and 'small' and 'light' and between voiced and 'big' and 'heavy' were identified by Japanese speakers but not by English speakers. Uno, Kobayashi, Shinohara, and Odake (2016) also report on the Japanese unique associations between voicing and hardness.

Thus, it may be possible that form-meaning associations in language models are not as robust as those in humans because they should not be aware of the biological or bodily bases which iconicity relies on. Additionally, it may also be possible that we may observe different form-meaning associations across language models, reflecting the language-specific aspects of iconicity observed in individual human languages. This paper examines these possibilities in an exploratory fashion.

Iconicity in Computational Models

Iconicity in computational models is less explored compared to iconicity in human language. Yamagata, Kwon, Kawashima, Shimoda, and Sakamoto (2021) investigated associations between sound and tactile sensations using Japanese sound symbolic words. They developed a computer vision method to generate the phonemes and structure comprising sound-symbolic words that probabilistically correspond to the input images. Their evaluation indicated that the sound-symbolic words output by their system had an accuracy rate of about 80%. While these findings suggest that computer vision systems can learn certain associations between sound and visual information, the study does not specify which aspects of form are associated with particular meanings. Additionally, since the study focused solely on Japanese sound symbolic words, it does not provide insights into whether these associations transcend different languages.

Alper and Averbuch-Elor (2023) more directly examined the extent to which vision-and-language models capture sound symbolism, particularly the *bouba/kiki* effect, by investigating whether multilingual vision-and-language models (English, Finnish, Indonesian, Hungarian, and Lithuanian) encode a relationship between sounds and sharp or round shapes. They employed vector representations obtained through Stable Diffusion and CLIP, focusing on English adjectives, nouns, and pseudowords corresponding to sharpness and roundness. Additionally, they also conducted a user study to validate that the pseudowords have a similar effect on human perception. They concluded that multilingual vision-and-language models learn the associations between sound and shape sharpness and roundness even though the models should not be aware of embodied motivations (i.e., perceptuomotor analogies) for the iconic form-meaning mappings as humans are. In the current study, we will try to replicate Alper and Averbuch Elor's (2023) findings using models of Japanese, which is phylogenetically unrelated to the five languages they investigated.

Methodology

In this section, we describe the methodology we adopted from Alper and Averbuch-Elor (2023). Due to computational resource constraints, we focused on adjectives and pseudo-words, not including nouns. We first describe the models used—Stable Diffusion and CLIP—and then the scores for predicting the association between sound and meaning, namely the geometric score and phonetic score. Finally, we introduce evaluation metrics to assess the classification ability of the models.

Models

Alper and Averbuch-Elor (2023) chose Stable Diffusion and CLIP as representative models in the field of vision-and-language models. We will describe these models below.

Stable Diffusion Stable Diffusion is a generative model that takes textual inputs and outputs corresponding to images. Utilizing a process of iterative refinement, it transforms a random pixel pattern into a coherent image that corresponds with the provided text. In the subsequent sections, we denote the output image from Stable Diffusion when given a prompt $\boldsymbol{i}$ as the following:

$$\mathrm{SD}(i)$$

CLIP CLIP (Contrastive Language–Image Pre-training) is a model accepting both text and images as input. It is trained to output similar vector representations for similar text-image pairs. Alper and Averbuch-Elor (2023) utilized CLIP for the specific purpose of vectorizing images generated by Stable Diffusion, and for evaluating the performance of CLIP itself. Hereafter, the vector output by CLIP when given a certain image or text i as input will be represented as follows:

$$\mathrm{CLIP}(i)$$

Obtaining Vector Representations

Here, A_r and A_s are defined as sets of adjectives associated with roundness and sharpness, respectively, and P_r and P_s as sets of pseudowords associated with roundness and sharpness, respectively. The elements of each set will be described in the following section. Alper and Averbuch-Elor (2023) create prompts for the primary input to the model following the template below.[1]

"a 3D rendering of a X object"

By substituting a word for X, a corresponding prompt i_w (= "a 3D rendering of a object") can be obtained. For each pseudoword $p = P_r \cup P_s$, the vector representing it can be obtained through two different pipelines.

$$v_p^{\mathrm{SD}} = \mathrm{CLIP}\left(\mathrm{SD}(i_p)\right)$$
$$v_p^{\mathrm{CLIP}} = \mathrm{CLIP}(i_p)$$

v_p^{SD} represents the vector created by CLIP from an image generated by Stable Diffusion using a prompt i_p. v_p^{CLIP} is the vector output by CLIP when it directly receives the prompt. The former is used for evaluating Stable Diffusion, and the latter is used for evaluating CLIP.

The vector representation for each adjective a is obtained using CLIP as per the following equation:

$$v_a^{\mathrm{CLIP}} = \mathrm{CLIP}(i_a)$$

Henceforth, the normalized version of any vector will be referred to as $\hat{v}$. Next, we will explain the metrics used to measure the extent to which these vector representations distinctly express the association between sound and shape in the model.

Scores

To classify the vectors into sharpness or roundness, Alper and Averbuch-Elor (2023) defined two scores (numerical criteria for prediction): the geometric score and phonetic score. Informally speaking, the former is a score for classifying pseudowords using abstract adjective vectors and individual pseudoword vectors, while the latter uses abstract pseudoword vectors and individual adjective vectors for classifying adjectives. Although the geometric score and phonetic score defined in their paper differ from those defined in their implementation,[2] this difference does not affect the evaluation metrics. The definitions provided below follow their implementation.

Geometric Score The geometric score γ_p for a pseudoword p is defined by the following equation:

$$\gamma_p \coloneqq \widehat{v_p} \cdot \widehat{v_{adj}}$$

Here, $v_{adj} = \Sigma_{a \in A_r} \widehat{v_a} - \Sigma_{a \in A_s} \widehat{v_a}$. The vector v_{adj} represents the difference between the sum of adjective vectors with a "round" meaning and the sum of adjective vectors with a "sharp" meaning. We refer to this as the abstract adjective vector. γ_p represents how much the pseudoword p is biased towards either the "round" end or the "sharp" end in the semantic space. A larger γ_p indicates a bias towards the "round" end, while a smaller γ_p suggests a bias towards the "sharp" end. Assuming that vector v_{adj} represents the difference in meaning between "round" and "sharp", γ_p indicates whether the pseudoword p leans towards "round" or "sharp" in the semantic space. In this sense, γ_p is referred to as the "geometric" score.

Phonetic Score The phonetic score ϕ_a for an adjective a is defined by the following equation:

$$\phi_a \coloneqq \widehat{v_a} \cdot \widehat{v_{pse}}$$

Here, $v_{pse} = \Sigma_{p \in P_r} \widehat{v_p} - \Sigma_{p \in P_s} \widehat{v_p}$. The vector v_{pse} represents the difference between the sum of pseudoword vectors with a "round" meaning and the sum of pseudoword vectors with a "sharp" meaning. We refer to this as the abstract pseudoword vector. ϕ_a represents how much the pseudoword a is biased towards either the "round" end or the "sharp" end in the semantic space. A larger ϕ_a indicates a bias towards the "round" end, while a smaller ϕ_a suggests a bias towards the "sharp" end. As previously mentioned, since ϕ_a is obtained by the text-to-vector model CLIP, it reflects only the features of the adjective a as a string of characters. Since strings at least partially reflect their corresponding phonetic features, it is called the "phonetic" score. Therefore, it should be noted that "phonetic" here does not mean that this metric is based on phonetics, but rather that it does not include the visual features of adjectives.

Evaluation Metrics

Words are classified as "round" if their score is high and "sharp" if it is low, but the threshold value that defines this classification boundary is not obvious. Therefore, the AUC and Kendall's correlation are used as threshold-agnostic evaluation metrics. Both metrics are calculated from the score sequence of each word and the gold label sequence, where "round" corresponds to 1 and "sharp" to 0.

Experimental Settings

Pseudowords and Adjectives

Here, we describe the set of pseudowords and adjectives used in this experiment. Although each word is represented in the Latin alphabet below, they were written using kanji (Sino-Japanese character) and hiragana (Japanese phonogram) in our implementation. We define the adjective set as follows:
A_r = {*marui, en-kei-no, marumi-o obita, yawarakai, futotta, debu-no, marumaru-to shita, pocchari shita, fukkura shita, magatta*}
A_s = {*eeri-na, togatta, kakubatta, kado-no aru, supaiku-joo-no, toge-darake-no, chiku-chiku-shita, kifuku-no aru*}

To construct the pseudoword set, we define the following character set based on Alper and Averbuch-Elor (2023).[1]

$$V_r = \{o, u\} \quad V_s = \{e, i\} \quad V = \{a\}$$
$$C_r = \{b, d, g, m, n\} \quad C_s = \{p, t, k, s, h\}$$

V_r and C_r are vowels and consonants associated with roundness, and V_s and C_s with sharpness. V_contains the neutral vowel a. Based on these, we define the pseudoword set as follows:

$$P_r = \{C_1V_1C_2V_2C_1V_1 | C_1, C_2 \in C_r \ \& \ V_1, V_2 \in V_r \cup V\}$$
$$P_s = \{C_1V_1C_2V_2C_1V_1 | C_1, C_2 \in C_s \ \& \ V_1, V_2 \in V_s \cup V\}$$

For example, Pr includes pseudowords such as *bobobo*, *gunagu*, while Ps contains words such as *pipipi*, *tekate*, etc.

Pre-trained Models

Similar to Alper and Averbuch-Elor (2023), we also used Stable Diffusion and CLIP architectures. The difference lies in our use of Japanese-specific model. More specifically, we adopted the following checkpoints:

- Stable Diffusion: stabilityai/japanese-stable-diffusion-xl[2]
- CLIP: sonoisa/clip-vit-b-32-japanese-v1[3]

Each model checkpoint is specialized for Japanese, and we use them without any additional training.

Image Generation

When generating images, we set the hyperparameters as follows, consistent with Alper and Averbuch-Elor (2023):

- Guidance scale: 9
- Inference steps: 20

The guidance scale represents the degree to which the model adheres to the input prompt, and the inference steps denote the number of inference iterations from the input noise image to the output image. The images generated by Stable Diffusion are influenced by the fluctuations in random noise that serve as the starting point for inference. Therefore, like prior research, we generate 50 images per prompt and consider the average of the corresponding vectors for each image as the vector associated with that prompt.

Results

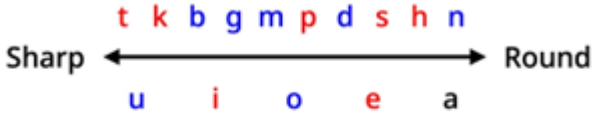
We present the results of both our experiment and that of Alper and Averbuch-Elor (2023) in Table 1.

Table 1: Evaluation metrics for each model. (en) and (ja) indicate the English and Japanese models, respectively. The results for the English models were reported by Alper and Averbuch-Elor (2023). Our experimental results are shown in bold.

Model	$\gamma_{\langle w \rangle}$		$\phi_{\langle w \rangle}$	
	AUC	τ	AUC	τ
Stable Diffusion (en)	0.74	0.34	0.97	0.68
CLIP (en)	0.77	0.39	0.98	0.70
Stable Diffusion (ja)	**0.51**	**0.01**	**0.37**	**-0.19**
CLIP (ja)	**0.43**	**-0.10**	**0.69**	**0.28**
Random	0.50	0.00	0.50	0.00

As the table indicates, all metrics in the Japanese model considerably underperform compared to those in the English model.

We grouped the pseudowords by each consonant and vowel that compose their first two letters and calculated the average geometric score. The results are shown in Figure-1 and 2. Alper and Averbuch-Elor (2023) have also reported the average scores per character with English models, where a significant correlation between the scores and sharpness-roundness was observed. However, our results clearly differ from that. Particularly intriguing differences were observed in consonants, such as /p/ and /s/, which were expected to be round in Alper and Averbuch-Elor's study, is closer to the end associated with roundness in our data from both Japanese Stable Diffusion and CLIP. Conversely, the voiced consonant /g/, expected to be round, being closer to the sharp end. Additionally, the vowel /u/, typically associated with roundness, was closer to the sharp end, whereas /e/ was closer to the round end. We will discuss the potential sources of these differences in the next section.



[1] They include x and l in their character set, but as it is difficult to map these to any Japanese sounds, we do not use them in this study.

[2] https://huggingface.co/stabilityai/japanese-stable-diffusion-xl

[3] https://huggingface.co/sonoisa/clip-vit-b-32-japanese-v1

Figure 1: Consonants and vowels sorted by the average geometric score of Stable Diffusion. Blue corresponds to round characters, while red corresponds to sharp characters according to Alper and Averbuch-Elor (2023).

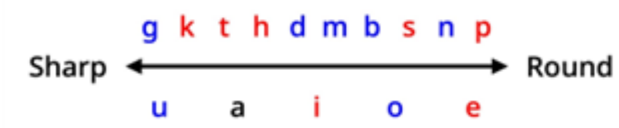


Figure 2: Consonants and vowels sorted by the average geometric score of CLIP.

Discussion

Our results obtained from Japanese vision-and-language models considerably differ from those from English vision-and-language models, not showing the shape-sound associations as demonstrated by Alper and Averbuch-Elor (2023). Where do these differences arise from?

Possible explanations or interpretations include the language-specific, systematic aspect of iconicity. As for consonants, /g/ was placed closer to the sharp end (Figures 1 and 2). In the Japanese sound symbolic system, velar stops (/k, g/) are associated with 'hardness of surface', 'abrupt manner' and 'harshness' according to Hamano (1998, 2014). More significantly, the voiced velar stop /g/ is associated with 'roughness,' along with other voiced obstruents such as /d/ and /z/. These associations are illustrated by Japanese ideophones, such as *giza-giza* 'serrated', *gata-gata* 'uneven', *gari-gari* 'hard', *gotsu-gotsu* 'rugged', and *gowa-gowa* 'rough', and *toge-toge* 'edgy'.

Furthermore, /p/ can be associated with 'roundness' in various respects. Firstly, /p/ is classified as a labial consonant, as well as /b/ and /m/, associated with roundness as in *bouba* and *maluma* cross-linguistically (D'Onofrio, 2014), despite also being categorized as voiceless stops which include /t/ and /k/, which may instead be associated with sharpness. Another factor contributing to the roundness of /p/ is again language-specificity. According to Hamano (1998, 2014), /p/ is associated with 'fatness' as evidenced by ideophones such as *pocha-pocha* 'chubby', *puyo-puyo* 'fat', and *puku-puku* 'puffing up'. In Japanese, there is a unique association between voicing and hardness, with Japanese speakers associating voiceless consonants (e.g., /p/) with softness and voiced consonants (e.g., /b/) with hardness. The crossmodal correspondence between softness and roundness may also have contributed to the roundness of /p/ (Sakamoto & Watanabe, 2018).

The differences in score which come from different places of articulations are effectively captured by the geometric scores of CLIP, as illustrated in Figure 3. A linear model that predicts geometric scores from place of articulation revealed that velars (/k, g/) had lower geometric scores (i.e., sharper) than other types of consonants, including glottal /h/ ($b = 0.01$, $SE < 0.01$, $t = -3.74$, $p < .001$, $R^2 = .24$). Additionally, labials (/b, m, p/) demonstrated higher geometric scores (i.e., rounder) than other consonants, such as alveolars (/t, d, n/) ($b = 0.01$, $SE < 0.01$, $t = -3.74$, $p < .001$, $R^2 = .24$).

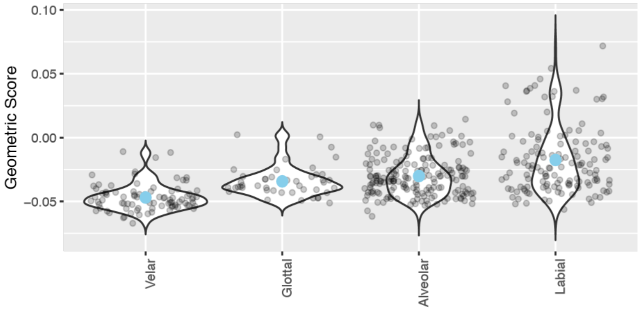


Figure 3: Geometric scores of CLIP by place of articulation (the skyblue bots represent the average).

Additionally, /s/ was associated with roundness in our data, contrasting with the findings of Alper and Averbuch-Elor (2023). This discrepancy may arise from the association of /s/ with concepts such as 'surface without resistance' and 'smoothness', as observed in ideophones referring to "smoothness", such as *sarasara*, *surusuru*, and *subesube*, as proposed by Hamano (1998). Consequently, it is conceivable that /s/ was linked with roundness rather than sharpness.

As for vowels, there are two potential explanations for why /u/ had a lower geometric score compared to other vowels. One possibility is that /u/ ([ɯ]) is normally not rounded in the Japanese phonological system. When we categorize the five vowels into two groups—rounded /o/ and unrounded /a, e, i, u/—the rounded vowel /o/ exhibited a higher geometric score (i.e., rounder) than the unrounded vowels, aligning with universal associations between rounded vowels and round shapes, as depicted in Figure 4. A linear model that predicts geometric score from roundness of vowels revealed that rounded vowels had higher geometric scores (i.e., rounder) than unrounded vowels ($b = 0.01$, $SE < 0.01$, $t = -2.10$, $p = .04$, $R^2 = .01$).

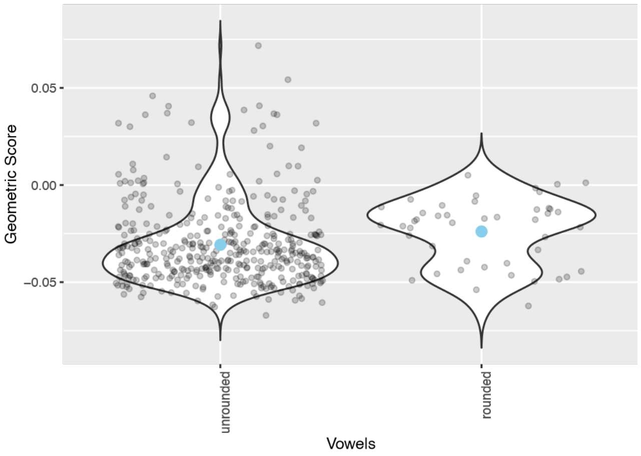


Figure 4: Geometric scores of CLIP by roundness of vowels (the skyblue bots represent the average).

Another possibility is that /u/ was combined with voiced obstruents (/g/, /b/, /d), since they were expected to be round, as investigated by Alper and Averbuch-Elor (2023). As mentioned above, voiced obstruents are generally associated with ‘roughness’ in the Japanese sound symbolic system (Hamano, 1998). This aspect requires further refinement in future research, particularly in the reconsideration of pseudoword construction.

Other possible factors contributing to our results are diverse, including the expressiveness of Stable Diffusion and CLIP and the selection of adjectives: the abstract vectors do not have sufficient representation power. What we have revealed thus far is that the cosine similarity between $\Sigma_{a \in A_r} \widehat{v_a}$ and $\Sigma_{a \in A_s} \widehat{v_a}$ in prior research is approximately 0.93, whereas in this study, it is approximately 0.98. The abstract adjective vector is defined as the difference between these vectors, with the assumption that this represents the difference in meaning between “round” and “sharp”. However, due to the lack of significant difference between these two vectors, there is doubt about whether our created abstract adjective vector is suitable for use as the vector for calculating geometric scores. For a more comprehensive analysis, measuring the quality of the output at each step is one of the future challenges.

The current results encourage further investigations into iconic form-meaning mappings in models of other languages to support the universality that Alper and Averbuch-Elor’s findings may suggest and language-specificity that the current study revealed. Comparing the experimental results from psycholinguistics with those from computational approaches will help us to understand the nature of iconicity in terms of embodied cognition and its possible role in human language and language models.

Acknowledgments

We thank Kimi Akita and the three anonymous reviewers of CogSci 2024 for their thought-provoking comments on earlier versions of this paper. Any remaining inadequacies are our own.

References

Akita, K., & Imai, M. (2022). The iconicity ring model for sound symbolism. *Iconicity in cognition and across semiotic systems, 18*, 27.

Alper, M., & Averbuch-Elor, H. (2023). Kiki or Bouba? Sound Symbolism in Vision-and-Language Models. *arXiv preprint arXiv:2310.16781.*

Ćwiek, A., Fuchs, S., Draxler, C., Asu, E. L., Dediu, D., Hiovain, K., Kawahara, S., Koutalidis, S., Krifka, M., Lippus, P., Lupyan, G., Oh, G. E., Paul, J., Petrone, C., Ridouane, R., Reiter, S., Schümchen, N., Szalontai, Á., Ünal-Logacev, Ö., Zeller, J., Perlman, M., & Winter, B. (2022). The bouba/kiki effect is robust across cultures and writing systems. *Philosophical Transactions of the Royal Society B*, 377(1841), 20200390.

Dingemanse, M. (2012). *Advances in the cross-linguistic study of ideophones. Language and Linguistics compass, 6*(10), 654-672.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences, 19*(10), 603-615.

D’Onofrio, A. (2014). Phonetic detail and dimensionality in sound-shape correspondences: Refining the bouba-kiki paradigm. *Language and speech, 57*(3), 367-393.

Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language*, 515-540.

Hamano, S. (1998). *The Sound Symbolic System of Japanese*. Stanford, CA: CSLI Publications.

Hamano, S. (2014). Nihongo no onomatope: Onsyootyoo to koozoo [Japanese mimetics: Sound symbolism and structure].

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*(1), 54-65.

Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound symbolism facilitates word learning in 14-month-olds. *PLoS One, 10*(2), e0116494.

Iwasaki, N., Vinson, D. P., & Vigliocco, G. (2007). What do English speakers know about *gera-gera* and *yota-yota*?: A cross-linguistic investigation of mimetic words for laughing and walking. *Japanese-language education around the globe, 17*, 53-78.

Kantartzis, K., Imai, M., & Kita, S. (2011). Japanese soundsymbolism facilitates word learning in English-speaking children. *Cognitive science, 35*(3), 575-586.
Köhler, W. (1929). *Gestalt Psychology*. New York: Liveright.

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: Evidence for sound–shape crossmodal correspondences in 4-month-olds. *Journal of experimental child psychology, 114*(2), 173-186.

Peirce, C. S. (1932). *Collected papers of Charles Sanders Peirce, Vol. 2*, Cambridge, MA: Harvard University Press.

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in psychology, 1*, 227.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International conference on Machine*.

Proceedings of the 38th International conference on Machine Learning, PMLR 139, 8748-8763.
Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia: A window into perception, thought and language. *Journal of consciousness studies, 8*(12), 3–34.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works.

Transactions of the Association for Computational Linguistics, *8*, 842-866.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684-10695.

Sakamoto, M., & Watanabe, J. (2018). Bouba/Kiki in touch: Associations between tactile perceptual qualities and Japanese phonemes. *Frontiers in psychology, 9*, 295.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of experimental psychology, 12*(3), 225.

Sidhu, D. M., Vigliocco, G., & Pexman, P. M. (2020). Effects of iconicity in lexical decision. *Language and Cognition, 12*(1), 164–181.

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (2021). Sound symbolism shapes the English language: The maluma/takete effect in English nouns. *Psychonomic Bulletin & Review, 28*, 1390-1398.

Sidhu, D. M., Khachatoorian, N., & Vigliocco, G. (2023). Effects of Iconicity in Recognition Memory. *Cognitive Science*, 47(11), e13382.

Sonier, R. P., Poirier, M., Guitard, D., & Saint-Aubin, J. (2020). A round Bouba is easier to remember than a curved Kiki: Sound-symbolism can support associative memory. *Psychonomic Bulletin & Review, 27*, 776-782.

Yamagata, K., Kwon, J., Kawashima, T., Shimoda, W., & Sakamoto, M. (2021). Computer Vision System for Expressing Texture Using Sound-Symbolic Words. *Frontiers in Psychology, 12*, 654779.