DOI: 10.66069/ojspub.1811260811

Iconic-systematicity in Guangzhou Cantonese Reduplicative Ideophones

Danrui Li

School of Humanities, Zhuhai City Polytechnic, Zhuhai, Guangdong, China
danrui.li@outlook.com

Abstract: ***This study investigates whether Guangzhou Cantonese ABB ideophones exhibit iconic-systematicity, the statistical association between speech sounds and sensory meanings that is grounded in perceived resemblance. While such systematic sound-meaning correspondences have been widely documented in Indo-European languages, it remains unknown whether they exist in tonal languages with rich reduplicative morphology. To address this gap, a corpus of 181 Guangzhou Cantonese ABB ideophones was compiled. Phonological features were coded as 82 binary variables, and semantic embeddings were derived from dictionary definitions and example sentences. Global prediction models (Partial Least Squares and Random Forest) did not reliably predict the continuous semantic space, suggesting that phonological features do not reliably predict the overall semantic space. However, theory-driven univariate tests revealed two significant associations after false discovery rate correction. The high front vowel /i/ was strongly linked to physical smallness, replicating a cross-linguistic pattern. The voiceless alveolar stop /t/ showed a position-sensitive effect: only coda /t/ was associated with smallness, while onset /t/ was not. No evidence was found for an /a/-largeness link. These findings provide evidence for iconic-systematicity in Guangzhou Cantonese ideophones, but suggest that it is restricted to specific phonemes and sensitive to syllabic position, highlighting a language-specific constraint on sound-symbolic patterning.***

Keywords: Iconic-systematicity, Sound symbolism, Ideophone, Guangzhou Cantonese**.**

1. Introduction

The relationship between linguistic form and meaning has traditionally been regarded as arbitrary. Nevertheless, a growing body of research has shown that speech sounds may exhibit systematic associations with sensory meaning, a phenomenon commonly discussed under the labels of sound symbolism and iconicity. Experimental studies have repeatedly demonstrated that speakers associate particular phonological patterns with semantic dimensions such as size, shape, texture, and movement. More recently, corpus-based research on iconic-systematicity has examined whether these associations are reflected in the statistical organization of natural-language lexicons, suggesting that semantically similar words may also display recurring phonological tendencies.

Quantitative evidence for iconic-systematicity has expanded considerably in recent years. Studies of English and several Indo-European languages have reported associations between high front vowels and smallness, low vowels and largeness, as well as correspondences between consonantal patterns and sensory properties such as sharpness or roughness. However, the typological scope of this research remains relatively narrow. Much less is known about whether comparable sound-meaning correspondences emerge in tonal languages with extensive reduplicative morphology. This limitation is particularly important because tonal systems and reduplicative structures may influence phonological salience and lexical organization in ways that differ substantially from Indo-European languages.

Guangzhou Cantonese provides a particularly suitable context for investigating this issue. Guangzhou Cantonese contains a large inventory of ABB reduplicative ideophones characterized by vivid sensory depiction and marked phonological structure. In many cases, the reduplicated syllable contributes expressive or affective force while lacking fully independent lexical meaning, potentially increasing the perceptual salience of phonological form. Despite the abundance of such forms in Guangzhou Cantonese, quantitative studies of their phonosemantic organization remain limited, with previous research focusing primarily on descriptive classification or psycholinguistic ratings.

The present study addresses this gap through a quantitative analysis of 181 Guangzhou Cantonese ABB ideophones. The study was designed as an exploratory pilot study. Semantic representations were derived from dictionary definitions and example sentences using multilingual sentence embeddings, while phonological structure was encoded through binary phonological features. The study combines exploratory multivariate modelling using Partial Least Squares (PLS) and Random Forest (RF) regression with hypothesis-driven univariate analyses targeting previously documented sound-symbolic associations involving /i/, /a/, and /t/. By integrating exploratory semantic modelling with targeted phonological hypothesis testing, the study aims to evaluate whether Guangzhou Cantonese ideophones exhibit systematic sound-meaning correspondences at both global and localized levels.

2. Literature Review

2.1 Iconicity, Sound Symbolism and Systematicity

Iconicity refers to a perceived resemblance between the form of a sign and its meaning. Unlike arbitrary signs, iconic signs exhibit some motivated connection between expression and content. Perniss et al. (2010) argue that iconicity is not limited to spoken language but constitutes a widespread semiotic principle found across gesture, sign language, visual communication, and speech. In spoken language, iconicity is most obvious in onomatopoeia, where words imitate environmental sounds directly, such as *buzz* or *bang*.

However, iconicity can also operate cross-modally, involving perceived correspondences between sound and non-auditory sensory properties.

The classic example of cross-modal iconicity is the *maluma/takete* effect first reported by Köhler (1929). Speakers consistently associate rounded pseudowords such as *maluma* with curved shapes and angular pseudowords such as *takete* with sharp shapes. Similar effects have been repeatedly replicated across cultures and languages (Nielsen and Rendall, 2011; Sidhu et al., 2021). Another long-standing pattern concerns size symbolism. Jespersen (1922) observed that high front vowels such as /i/ tend to evoke smallness while low back vowels such as /a/ evoke largeness. Sapir's (1929) experiment using the pseudowords *mil* and *mal* provided one of the earliest empirical demonstrations of this tendency, and Newman (1933) later confirmed these findings through additional experiments. More recent cross-linguistic studies have shown that these associations recur across hundreds or even thousands of languages (Haynie et al., 2014; Blasi et al., 2016; Johansson et al., 2019).

Several theoretical explanations have been proposed for these recurring sound-meaning correspondences. One influential account is Ohala's (1983) frequency code hypothesis, which argues that high-frequency acoustic signals are naturally associated with small bodies while low-frequency signals are associated with large bodies. Because high front vowels such as /i/ have relatively high acoustic frequencies, they are more readily associated with smallness, whereas low vowels such as /a/ are associated with largeness. This explanation provides an important physiological and acoustic basis for cross-modal iconicity.

Sound symbolism is a more specific concept referring to motivated correspondences between speech sounds and meaning. While iconicity is a broader semiotic phenomenon, sound symbolism focuses specifically on phonological form. Research in psycholinguistics has demonstrated that such correspondences affect language processing (Vinson et al., 2015), lexical decision (Sidhu et al., 2020), and word learning (Lockwood et al., 2016). Akita and Imai (2022) further propose the "iconicity ring" model, arguing that sound symbolism ranges from highly direct imitative forms to weaker forms of secondary iconicity shaped by lexical patterns.

Importantly, modern research increasingly distinguishes iconicity from systematicity. Dingemanse et al. (2015) define systematicity as recurring statistical relationships between linguistic form and meaning within a lexicon. A linguistic pattern may therefore be systematic without necessarily being iconic. For example, nouns and verbs may display different phonological tendencies even when no perceptual resemblance exists between sound and meaning (Farmer et al., 2006). Iconic-systematicity arises when statistically recurring patterns are themselves grounded in perceived resemblance. Sidhu (2024) emphasizes that iconic-systematicity is probabilistic rather than deterministic: not every word associated with smallness contains /i/, but such phonemes occur more frequently than chance would predict.

Recent quantitative studies have demonstrated iconic-systematicity across multiple semantic domains. Winter and Perlman (2021) identified robust size-related phonological patterns in English adjectives, finding that /i/, /ɪ/, and /t/ are associated with smallness whereas /ɑ/ is associated with largeness. Sidhu et al. (2021) demonstrated systematic shape symbolism in English nouns, while Winter et al. (2022) showed that words associated with rough textures disproportionately contain /r/, especially in languages with trilled /r/ realisations. Emotional meaning also exhibits iconic tendencies (Adelman et al., 2018). Yu et al. (2021) reported that words containing /i/ tend to be associated with positive valence, whereas words containing /ʌ/ are associated with negative valence. Similarly, Aryani et al. (2018) found that arousing German words preferentially contain short vowels and voiceless consonants.

At the same time, scholars have emphasized that iconicity is shaped by linguistic and cultural experience. Occhino et al. (2017) argue that iconicity is partly "in the eye of the beholder", meaning that language users' prior experience influences how form-meaning correspondences are perceived. Consequently, sound-symbolic patterns may vary across languages even when certain broad tendencies recur cross-linguistically. Perry et al. (2015), for example, observed cross-linguistic differences in how iconicity is distributed across lexical categories. This raises the important possibility that different languages may exhibit both universal and language-specific forms of iconic-systematicity.

2.2 Ideophone in Guangzhou Cantonese

Ideophones are marked lexical items that vividly depict sensory imagery, movement, texture, sound, emotion, or manner across many languages (Dingemanse, 2012, 2019). Nuckolls (1999) argues that ideophones provide some of the clearest evidence that linguistic form can exhibit motivated relationships with meaning. Unlike ordinary lexical items, ideophones are often phonologically marked, prosodically expressive, and semantically depictive rather than merely descriptive. Their forms are therefore especially suitable for investigating sound symbolism and iconicity.

Cross-linguistically, ideophones frequently employ reduplication, unusual phonotactics, or expressive prosody, and they have been shown to provide a particularly productive domain for examining iconic-systematicity (Dingemanse, 2012, 2019). For example, Japanese mimetics display systematic correspondences between phonological structure and sensory meaning (Hamano, 1998), suggesting that ideophones constitute a productive domain for investigating iconic-systematicity across languages.

Guangzhou Cantonese possesses a large inventory of reduplicative ideophones, especially ABB forms in which a monosyllabic root is followed by a reduplicated syllable. T'sou (1978) observed that ABB constructions in Cantonese frequently convey vivid sensory impressions and often display expressive or evaluative functions. More recently, Van Hoey (2023) proposed that Chinese ABB constructions are best understood as a collocate-ideophone schema in which the reduplicated BB component contributes depictive force rather than independent lexical meaning. This interpretation aligns closely with Dingemanse's (2019) characterization of

canonical ideophones as structurally marked depictive forms. Examples such as 眼甘甘 /ngaan^{5} gam^{1} gam^{1}/ (staring greedily), 滑捽捽 /waat^{6} lyut^{1} lyut^{1}/ (very smooth), and 湿趿趿 /sap^{1} taap^{1} taap^{1}/ (wet and sticky) illustrate the vivid sensory quality of Guangzhou Cantonese ABB ideophones. In these constructions, the reduplicated syllable often lacks a clear referential meaning of its own, yet contributes strongly to the overall perceptual impression. This semantic opacity or semantic vacancy makes the phonological form itself especially salient, thereby increasing the likelihood that sound-meaning correspondences may emerge systematically.

Despite the abundance of Guangzhou Cantonese ideophones, quantitative research on their sound-symbolic structure remains limited. Existing work has focused primarily on descriptive classification or psycholinguistic ratings. Van Hoey et al. (2024), for instance, provided subjective iconicity and imageability ratings for Mandarin ABB words, demonstrating their high depictive potential. However, no large-scale statistical study has yet examined whether particular phonemes in Guangzhou Cantonese ideophones are systematically associated with particular sensory meanings. The present study addresses this gap by combining exploratory semantic modelling with theory-driven statistical testing.

2.3 Limitations of Previous Research

Despite growing evidence for iconic-systematicity across several semantic dimensions, two major gaps remain. First, nearly all quantitative studies have been conducted on English or other Indo-European languages. It is unclear whether similar patterns exist in languages with typologically different phonological and morphological structures, such as tonal languages with rich reduplicative morphology. Guangzhou Cantonese, with its large inventory of ABB ideophones and lexically specified tones, offers an ideal testing ground. Second, most previous work has adopted a theory-driven approach that pre-defines sensory dimensions (e.g., roundness, smallness, roughness) and then tests for over-representation of particular phonemes. While this confirms existing hypotheses, it may miss unforeseen sound-meaning associations that are not captured by pre-established categories. A complementary exploratory approach that first lets semantic categories emerge from the lexicon has rarely been attempted, and never for Guangzhou Cantonese.

The present study addresses these gaps by combining two complementary strategies. First, we use a data-driven exploratory method: continuous semantic embeddings derived from dictionary definitions and example sentences, together with global prediction models (PLS and RF), to assess whether phonological features collectively predict semantic space without imposing categorical boundaries. Second, we conduct hypothesis-driven univariate tests targeting three previously documented sound-meaning associations involving /i/, /t/, and /a/. By applying both approaches to a carefully constructed corpus of 181 Guangzhou Cantonese ABB reduplicative ideophones, we aim to determine whether iconic-systematicity exists in this under-studied language and, if so, which specific associations are robust across languages and which may be language-specific.

3. Method

3.1 Corpus Construction

The data for this study were taken from the Guangzhouhua Cidian (广州话词典), a standard dictionary of the Guangzhou variety of Cantonese. Following Dingemanse's (2019, p.13) characterization of canonical ideophones as "marked words that depict sensory imagery", we first used the structural property of reduplication as a proxy for markedness. This step ensures that the candidate pool consists of words that are formally conspicuous, a prerequisite for ideophone status.

Among reduplicative patterns, we chose to focus on the ABB type, where a monosyllabic root (A) is followed by a reduplicated syllable (BB). The ABB pattern is a common and productive reduplicative form for adjectives and depictive words in Cantonese. Our decision is motivated by two methodological considerations. First, identical reduplication, as in ABB, provides a more transparent test for iconicity than non-identical patterns, because the repeated syllable does not carry independent lexical meaning and thus its phonological form is more likely to contribute directly to sensory depiction. Second, according to Smith (2015), non-identical reduplicatives (e.g., ABAB forms) are typically derived from existing lexical items and show weaker and less transparent sound-meaning mappings. Therefore, limiting the present analysis to fully identical reduplication may yield a more rigorous investigation of systematic sound-symbolic patterns.

From the initial pool of ABB entries, we applied explicit exclusion criteria to retain only genuine ideophones. Lexicalised nouns and verbs that happen to have an ABB shape were removed because their primary function is referential or predicative rather than depictive. Excluded nouns include 乜东东 /met^{1} tʊŋ^{1} tʊŋ^{1}/ (what thing), 伏匿匿 /pɔk^{6} nei^{1} nei^{1}/ (hide-and-seek), 炒粒粒/tʃʰau^{2} nɐp^{1} nɐp^{1}/ (diced meat) and 摸盲盲/mɔ^{2} maŋ^{4} maŋ$^{4\text{-}1}$/ (hide-and-seek); excluded verbs include 跕跛跛/kɐt^{6} pɐi^{1} pɐi^{1}/ (hop on one foot), 洗白白/ʃɐi^{2} pak^{6} pak^{6}/(to bathe) and 屙啡啡/ŋɔ^{1} fɛ^{4} fɛ$^{4\text{-}2}$/ (to defecate).

Among the remaining adjective-like ABB words, we further distinguished depictive from descriptive instances. One illustrative case is 面黄黄/min^{6} wɔŋ^{4} wɔŋ^{4}/, which the dictionary defines as "a pale yellow complexion due to illness or weakness". This meaning evokes a sensory impression of an unhealthy physical state and was therefore retained as a depictive ideophone. By contrast, two other colour-related ABB words were excluded: 面红红/min^{6} hʊŋ^{4} hʊŋ^{4}/ (red-faced) and 面青青/min^{6} cɛŋ^{1} cɛŋ^{1}/(pale-faced). Their dictionary definitions merely describe facial colour without conveying an additional layer of sensory imagery or an evaluative connotation of a specific condition. Accordingly, they were treated as descriptive rather than depictive and removed from the corpus. After applying these criteria, a final corpus of 181 Guangzhou Cantonese ABB ideophones was established for subsequent analysis.

3.2 Phonological Feature Coding

The International Phonetic Alphabet (IPA) transcription for

each word was taken directly from the Guangzhouhua Cidian, which provides IPA forms for all sounds in Guangzhou Cantonese. This dictionary practice reflects the fact that the BB part is semantically vacant and carries no independent lexical meaning, whereas the A root contributes the core descriptive content. Consequently, the BB syllable is the most plausible locus of sound-symbolic patterning, and all subsequent phonological coding was limited to this syllable.

A set of 82 binary features was compiled to capture the phonological properties of the BB syllable. These features were organised into four major categories: onset, nucleus, coda, and tone. For the onset, separate binary variables were coded for the presence or absence of each consonant phoneme attested in the corpus, including plain and aspirated stops (e.g., /p, pʰ, t, tʰ, k, kʰ/), labialised velars (/kʷ, kʷʰ/), affricates (/tʃ, tʃʰ/), fricatives (/f, s, h, ʃ/), nasals (/m, n, ŋ/), laterals (/l/), and approximants (/j, w/). Additional articulatory groupings (e.g., plosive, fricative, nasal, aspirated, bilabial, alveolar, velar) were also included as separate binary columns.

The nucleus and coda were coded in a similar vein. Each monophthong that occurs as a nucleus (/a, ɐ, ɛ, œ, æ, i, y, u, ʊ, ɔ, ø/) received its own binary column. For diphthongs (e.g., /ai, ɐi, au, ɐu, ei, iu, ou/), we followed a componential approach: the diphthong as a whole was marked in a separate set of columns, and its starting and ending points were further decomposed into height, frontness, and roundness features. This componential decomposition was motivated by the possibility that the dynamic formant transition of diphthongs might itself carry iconic meaning, which would be ignored if diphthongs were treated as static monophthongs. Coda consonants (/p, t, k, m, n, ŋ/) were binary-coded individually, and summary columns (plosive coda, nasal coda) were also added.

Tone was coded using the six lexical tones of Cantonese. The citation tone of the character was used as the base tone, following the dictionary's indication. In cases where a tone change notation (e.g., xud$^{3\text{-}1}$) appears, the surface tone (the tone after the arrow) was taken, because it reflects the actual pronunciation in the reduplicated form. When the two B syllables were transcribed with different tones (e.g., lou^{1}-lou^{6}), we consulted the character's citation tone; if one of the transcribed tones matched it, that tone was selected; otherwise the tone of the first B was used.

3.3 Semantic Representation for Exploratory Modelling

For each ideophone, the semantic input was taken from the dictionary definition concatenated with a typical example sentence. To convert this text into a numerical vector, we used a pretrained multilingual sentence embedding model, paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019). The model maps any input sentence to a 384-dimensional dense vector. All 181 textual inputs were fed into the model independently, resulting in a 181 × 384 matrix.

Because the original 384 dimensions are highly redundant and may contain noise, we applied principal component analysis (PCA) to reduce the dimensionality while preserving the main semantic variation. PCA was performed on the centered and scaled 181 × 384 matrix. The first 20 principal components were retained, which together explained 66.5% of the variance in the original embeddings. This 20-dimensional representation was used as the semantic outcome variable Y in subsequent regression analyses (PLS and RF). Because the study was designed as an exploratory analysis rather than an out-of-sample prediction study, PCA was fitted on the full dataset prior to the cross-validation analyses.

3.4 Global Sound-meaning Association: PLS and RF

Building on the semantic representations established in the previous section, this stage of the analysis examined whether systematic relationships could be identified between semantic structure and phonological form in the Guangzhou Cantonese ABB ideophone lexicon. Following recent work on iconic-systematicity (Winter and Perlman, 2021), multivariate statistical modelling was employed to evaluate whether semantic distinctions are reflected in phonological patterning across the lexicon.

To examine whether the full set of phonological features collectively predicts semantic representations, two complementary regression methods were employed: Partial Least Squares (PLS) and Random Forest (RF). Both models used the 82 binary phonological features as predictors (X) and the 20-dimensional PCA-reduced semantic vectors derived from Section 3.3 as the outcome variables (Y). The phonological predictors included onset consonants, vowel qualities, coda consonants, and tone represented as binary presence-absence variables. Both models were implemented in Python 3.14.

PLS analysis was used to model the covariance structure between phonological and semantic representations. PLS is particularly suitable for high-dimensional linguistic datasets because it can accommodate multiple correlated predictor and response variables simultaneously while reducing the data into latent components. The method therefore provides an interpretable representation of broader phonosemantic structure within the lexicon. The optimal number of PLS components was selected using five-fold cross-validation based on mean cross-validated R^2. Predictive performance was evaluated using mean five-fold cross-validated R^2 as the primary metric.

To complement the linear latent-structure modelling provided by PLS, RF regression was additionally implemented as a non-parametric ensemble learning approach capable of capturing non-linear interactions among phonological predictors. RF models are also comparatively robust to multicollinearity, which is particularly relevant for phonological datasets involving naturally interrelated features. The RF models were run with 100 trees and otherwise default settings. Predictive performance was evaluated using mean five-fold cross-validated R^2. Variable importance measures were additionally examined to identify phonological predictors contributing most strongly to semantic prediction.

However, previous methodological research has shown that RF importance measures may become biased when predictors differ substantially in frequency or correlation structure (Strobl et al., 2007, 2009). Since many phonological variables in the present dataset were naturally interdependent, the RF

analysis was treated primarily as exploratory rather than confirmatory. Accordingly, the purpose of this stage was not to establish deterministic relationships between individual phonemes and meanings, but rather to evaluate whether global phonological patterning provides predictive information about semantic structure.

3.5 Hypothesis-driven Univariate Tests

In addition to the global multivariate analyses, this study conducted targeted hypothesis-driven statistical tests to examine specific sound-symbolic associations that have been repeatedly documented in previous research. Whereas the PLS and RF analyses evaluated whether semantic representations could be predicted from phonological structure at the global lexical level, the present analyses focused on localized phonosemantic effects involving theoretically motivated phonological features.

Particular attention was given to size symbolism, one of the most consistently reported phenomena in the sound symbolism literature. Since the early work of Jespersen (1922), Sapir (1929), and Newman (1933), high front vowels such as /i/ have repeatedly been associated with smallness, while low vowels such as /a/ are commonly associated with largeness. More recent large-scale quantitative studies have replicated these tendencies across natural-language lexicons and across languages (Blasi et al., 2016; Winter and Perlman, 2021). In addition, Winter and Perlman (2021) identified significant associations between voiceless stop consonants and smallness-related meanings in English adjectives. Based on these findings, the present study examined the sound-symbolic behaviour of /i/, /a/, and /t/ in Guangzhou Cantonese ABB ideophones.

To identify the semantic categories relevant to these hypotheses, lexical items were manually annotated on the basis of dictionary definitions and example sentences. Smallness was defined as reduced physical size in terms of length, height, width, thickness, or overall volume. A lexical item was classified as expressing smallness if its core meaning unambiguously referred to shortness, lowness, narrowness, slenderness, thinness, flatness, or diminutive overall size. Conversely, largeness was defined as increased physical size in one or more of these dimensions, including length, tallness, width, thickness, bulkiness, or swollen volume. Metaphorical extensions, such as quantity, degree, loudness, intensity, or age, were excluded from both categories in order to restrict the analysis to physical size symbolism. Two coders, both trained linguists blind to the research hypotheses, independently classified each lexical item according to the operational definitions of smallness and largeness. Cohen's κ was calculated to quantify inter-rater agreement. Disagreements were resolved through discussion, and the final binary labels were used for subsequent statistical analyses.

Based on these annotations, the study tested five specific hypotheses corresponding to the phonological features and semantic categories listed in Table 1. First, ideophones associated with physical smallness were predicted to contain the vowel /i/ more frequently than ideophones belonging to other semantic domains (H1). Second, because Guangzhou Cantonese permits unreleased stop codas, onset /t/ and coda /t/ were analysed separately to examine whether syllabic position influences sound-symbolic interpretation: H2a predicted an association between onset /t/ and smallness, H2b predicted an association between coda /t/ and smallness, and H2c predicted an association between /t/ in any position and smallness. Third, ideophones associated with physical largeness were predicted to contain the vowel /a/ more frequently than ideophones in other semantic domains (H3).

To test these hypotheses, five univariate analyses were conducted (Table 1), each based on a 2×2 contingency table comparing the presence or absence of a target phonological feature across semantic categories. Phonological features were converted into binary categorical variables indicating their presence or absence within each lexical item, and ideophones belonging to the target semantic categories were compared against all remaining items in the dataset. Because several contingency tables contained relatively small expected cell frequencies, statistical significance was assessed using Fisher's exact test. To control the false discovery rate associated with multiple hypothesis testing, p-values were adjusted using the Benjamini-Hochberg procedure with an adjusted α level of .05. Effect sizes were reported as odds ratios with 95% confidence intervals in order to evaluate the strength and direction of observed associations. All univariate analyses were conducted in SPSS 32.

Table 1: Summary of independent tests in the present study

Hypothesis	Phonological feature (present in BB syllable)	Predicted semantic dimension
H1	/i/	Smallness
H2a	onset /t/	Smallness
H2b	coda /t/	Smallness
H2c	/t/ in any position	Smallness
H3	/a/	Largeness

Unlike the global multivariate models, which assessed distributed phonosemantic structure across the lexicon as a whole, the univariate analyses were designed to detect localized sound-symbolic correspondences involving individual phonological features. Given the probabilistic nature of iconic-systematicity (Dingemanse et al., 2015; Sidhu, 2024), statistically significant results were interpreted as increased tendencies rather than deterministic phonosemantic rules.

4. Results

4.1 Overall Prediction: PLS and RF

We first examined whether the 82 binary phonological features could predict the 20-dimensional PCA-reduced semantic embeddings. A Partial Least Squares (PLS) regression was fitted with five-fold cross-validation to select the optimal number of latent components. The cross-validated R^2 values for component numbers 1 through 15 were consistently negative; the best mean cross-validated R^2 was -0.0985 (SD = 0.0177) at 1 component, and the training set R^2 was only 0.0129. These figures indicate that the PLS model performed worse than a simple mean baseline and captured virtually no systematic variance. Therefore, no evidence of a linear predictive relationship between phonological features and semantic representations was identified in the present analysis. As a complementary non-linear approach, a Random

Forest (RF) regression with 100 trees was run under the same cross-validation scheme. The mean cross-validated R^2 was -0.2864 (SD = 0.0304), similarly failing to detect any predictive relationship.

Taken together, these results suggest that, under the current analysis, the phonological form of Guangzhou Cantonese ABB ideophones does not exhibit a statistically reliable linear or non-linear association with the continuous semantic representations derived from dictionary definitions and example sentences.

4.2 Univariate Tests of Specific Sound-meaning Pairs

Of the 181 words, 9 were classified as denoting physical smallness and 11 as physical largeness. Inter-rater agreement was nearly perfect for both semantic categories (Cohen's κ = 0.883 for smallness and 0.903 for largeness). Table 2 reports the results of the five Fisher's exact tests. After Benjamini-Hochberg correction for multiple comparisons (α= 0.05), two associations remained significant. The high front vowel /i/ was strongly associated with physical smallness ($p < 0.01$). The voiceless alveolar stop /t/ showed a position-dependent effect: it was significantly associated with smallness when occurring as a coda ($p = 0.004$), but not as an onset ($p = 1.000$). The combined measure indicating /t/ in any position showed a trend towards association ($p = 0.052$), although it did not reach significance after correction. No evidence was found for the predicted association between /a/ and physical largeness ($p = 0.692$). These findings indicate that iconic-systematicity in Guangzhou Cantonese ABB ideophones is present for certain sound-meaning pairs, particularly /i/ and coda /t/ with smallness, while the size-sound pattern for largeness was not replicated in this dataset.

Table 2: Fisher's exact tests for sound-meaning associations

Hypothesis	Phonological feature	Semantic category	P-value	FDR-adjusted	Odds ratio	95% CI
H1	/i/	Small	<0.01	YES	13.083	[3.17, 53.99]
H2a	onset /t/	Small	1.000	NO	.671	[.081, 5.587]
H2b	coda /t/	Small	0.004	YES	8.523	[2.125, 34.178]
H2c	/t/ in any position	Small	.052	NO	3.869	[.993, 15.075]
H3	/a/	Large	.692	NO	.415	[.051, 3.358]

*Note. All p-values are from two-tailed Fisher's exact tests. FDR correction was applied using the Benjamini-Hochberg procedure (α=0.05) across the five tests.

5. Discussion

The present study combined a global, data-driven approach (PLS and RF) that examined the entire semantic space derived from dictionary definitions, with theory-driven univariate tests focusing specifically on physical size. The global models failed to predict the continuous semantic embeddings, indicating that at a coarse level, the phonological form of Guangzhou Cantonese ABB ideophones does not reliably map onto the overall semantic space. In contrast, the focused univariate tests revealed two significant associations: the high front vowel /i/ was strongly linked to physical smallness, and the voiceless alveolar stop /t/ showed a position-dependent effect, being significantly associated with smallness only when occurring as a coda. No evidence was found for the predicted association between /a/ and physical largeness. These findings demonstrate that iconic-systematicity may exist in this under-studied language, but is restricted to specific sound-meaning pairs and sensitive to syllabic position, supporting the view that size sound symbolism is not a general property of the entire lexicon but rather a domain-specific phenomenon (Winter and Perlman, 2021; Sidhu, 2024).

5.1 Cross-linguistic Evidence for the /i/-smallness Link

The present study provides corpus-based evidence that the high front vowel /i/ is significantly associated with physical smallness in Guangzhou Cantonese ABB ideophones. This finding extends previous cross-linguistic observations to a tonal Sinitic language with productive reduplicative morphology, suggesting that the /i/-smallness association is not restricted to Indo-European lexicons.

Rather than merely replicating earlier findings, the present study shows that this association remains detectable in naturally occurring Guangzhou Cantonese ideophones despite substantial typological differences in phonology and morphology. This supports the view that certain iconic sound-meaning correspondences are sufficiently robust to emerge across typologically diverse languages, while still allowing language-specific phonological realisations.

The observed pattern is also compatible with Ohala's (1983) Frequency Code, which proposes an association between higher-frequency acoustic signals and smaller sound sources. Although the Frequency Code does not directly predict a vowel-specific association between /i/ and smallness, the present finding is consistent with the broader possibility that recurrent perceptual biases contribute to sound-meaning associations across languages.

5.2 The Positional Effect of /t/

The present study found that the association between the voiceless alveolar stop /t/ and physical smallness in Guangzhou Cantonese ABB ideophones is restricted to the coda position; when /t/ occurred as an onset, no such association was observed, and the combined measure collapsed across positions did not reach significance. This pattern suggests that the iconic potential of /t/ is not a general property of the phoneme but is specifically tied to its occurrence as a coda in Guangzhou Cantonese. Winter and Perlman (2021) reported a /t/-smallness link in English size adjectives, but their analysis did not distinguish syllabic position. The present result thus adds a new dimension to our understanding of this sound-symbolic pattern: position may matter in some languages.

The special status of coda /t/ in Guangzhou Cantonese can be explained by its phonetic properties. Coda stops in Guangzhou Cantonese are unreleased and lack an audible burst, making them perceptually abrupt and brief. While Ohala's frequency code (1983) primarily links high acoustic

frequency to smallness, brief acoustic events can also evoke smallness, possibly through an ecological correlation between small sound sources and short durations. The onset /t/, by contrast, is a fully released plosive, which may not share the same perceptual affinity. This phonetic difference offers a plausible grounding for the observed positional asymmetry, though direct perceptual testing is needed to confirm the link between unreleased stops and the concept of smallness.

The pattern may be understood as a language-specific realisation of a broader iconic tendency. Cross-linguistic evidence indicates that sound-symbolic associations can vary with phonetic realisation. For example, the association of /r/ with roughness is particularly strong in languages that have a trilled /r/, a phonetic variant that enhances the tactile sensation (Winter et al., 2022, Ćwiek et al., 2024). By analogy, the unreleased coda /t/ in Guangzhou Cantonese may provide an especially effective perceptual match for smallness, whereas onset /t/ does not. In English, Winter and Perlman (2021) detected a robust /t/-smallness effect using size adjectives where /t/ predominantly occurs in onset position; there, different acoustic cues, such as the high-frequency burst of released /t/, may support the same iconicity. Thus, the same phoneme can be iconically linked to smallness through different phonetic realisations in different languages, a finding consistent with the growing recognition that iconicity is not monolithic but is shaped by language-specific phonetic and phonological structures.

The positional effect of /t/ identified in the present study raises several questions for future research. The present study focused exclusively on ABB reduplicative ideophones, which represent a productive but delimited subset of Guangzhou Cantonese depictive vocabulary. Extending the analysis to other reduplicative patterns, such as AAB, ABAB, or ABAC forms, would help determine whether the observed positional effect of /t/ generalises to a broader range of ideophones and would also increase the overall sample size, allowing for more stable estimates of the effects. Future research should also incorporate perceptual experiments to directly test whether unreleased stop codas are perceived as smaller or more abrupt. Despite these limitations, the present findings provide a first indication that in Guangzhou Cantonese, the sound-symbolic association of /t/ with smallness is selectively tied to the coda position, offering a novel example of language-specific iconicity.

5.3 Methodological Implications of the Global Prediction Failure

The global models (PLS and RF) did not capture any systematic relationship between the full set of phonological features and the continuous semantic embeddings derived from dictionary definitions and example sentences. This negative result, however, does not imply the absence of sound-symbolic associations in Guangzhou Cantonese ABB ideophones, as the subsequent univariate tests revealed robust effects for specific phonemes. Instead, it points to an important methodological issue.

One likely reason for the failure is the low signal-to-noise ratio inherent in the continuous semantic space. The embeddings were generated from complete dictionary definitions and example sentences, which contain a large amount of semantic information unrelated to sound symbolism, such as grammatical functions, abstract concepts, concrete object names, and emotional overtones. Sound-symbolic associations are typically confined to specific sensory dimensions such as size, shape, or texture. These dimensions constitute only a small portion of the total semantic variance. Predicting such a high-dimensional, noisy space from a limited set of binary phonological features is inherently difficult, and even genuine iconic signals can be drowned out by irrelevant variation.

The contrast between the present approach and that of Winter and Perlman (2021) is instructive. They adopted a categorical approach, restricting their analysis to a carefully curated set of size adjectives. This allowed them to detect robust size-symbolic patterns. The present study, by contrast, employed a continuous approach without pre-selecting semantic categories. While continuous methods can cover a broader range of vocabulary and are useful for exploratory analysis, they are less sensitive to local, dimension-specific iconicity. As Sidhu (2024) notes in his review, the categorical approach is better suited for testing specific form-meaning contrasts, whereas the continuous approach may dilute the signal when the semantic space is too heterogeneous. This methodological complementarity echoes the call by Van Hoey et al. (2024) to not take rating data at face value but to contrast it with other data types; in the present study, we similarly argue that global prediction models and focused univariate tests should be seen as complementary rather than competitive tools in the detection of iconic-systematicity.

The negative outcome of the global models thus does not undermine the positive findings of the univariate tests. Rather, it highlights a complementarity: global, hypothesis-free methods can reveal whether phonological features predict overall semantic variation, but they are not designed to detect isolated, dimension-specific associations. For the latter, targeted categorical tests remain indispensable. The present results therefore suggest that in investigations of iconic-systematicity, the choice of methodological approach should be guided by the specificity of the research question. Exploratory continuous analyses may help generate hypotheses, but confirmatory tests of specific sound-meaning pairs require a focused categorical design.

Future studies using continuous semantic spaces could improve sensitivity by filtering the embeddings to retain only dimensions relevant to the sensory domain of interest, or by using sensory-norm databases to weight the vectors. Alternatively, a two-stage strategy can be adopted: first use a categorical approach to establish an effect, then test its generalisation with continuous methods. The present study, with its combination of a global exploratory model and theory-driven univariate tests, illustrates how the two approaches can complement each other.

6. Conclusion

The present study provides the first quantitative investigation of iconic-systematicity in Guangzhou Cantonese ABB reduplicative ideophones. By combining exploratory semantic modelling with theory-driven hypothesis testing, the

study demonstrates that sound-meaning correspondences exist in Guangzhou Cantonese, although they are restricted to specific phonological features and semantic domains.

Most importantly, the robust association between /i/ and physical smallness replicates a long tradition of sound symbolism research extending from Jespersen (1922) and Sapir (1929) to contemporary cross-linguistic studies (Blasi et al., 2016; Winter and Perlman, 2021). The findings therefore suggest that size symbolism persists even in a typologically distinct tonal language with extensive reduplicative morphology. The position-sensitive behaviour of /t/, however, points to a more language-specific form of iconicity. Whereas onset /t/ showed no significant effect, coda /t/ was reliably associated with smallness, indicating that syllabic position and phonetic realization may influence sound-symbolic interpretation.

The failure of the global predictive models is equally informative. Similar to Winter and Perlman's (2021) findings for English, the results suggest that iconic-systematicity is not a general organizing principle of the entire lexicon but instead emerges in semantically concentrated domains where sensory meaning is especially salient. Guangzhou Cantonese ideophones therefore appear to preserve localized pockets of iconicity rather than globally predictable phonological semantics.

More broadly, the study contributes to ongoing debates concerning the role of iconicity in language structure and evolution. Scholars such as Imai and Kita (2014) and Nielsen and Dingemanse (2020) argue that iconicity facilitates communication, lexical learning, and language acquisition. The vivid depictive quality of Guangzhou Cantonese ABB ideophones may help explain why such forms remain productive and cognitively salient despite pressures toward arbitrariness in language change.

Several limitations should nevertheless be acknowledged. The corpus was restricted to ABB reduplicatives, and the number of words associated with smallness remained relatively modest. Future studies should extend the analysis to other reduplicative patterns such as AAB, ABAB, and ABAC constructions, as well as incorporate perceptual experiments to test whether unreleased stop codas are indeed perceived as smaller by Guangzhou Cantonese speakers. Cross-linguistic comparisons with other Sinitic varieties and tonal languages would also help clarify which patterns reflect universal perceptual biases and which arise from language-specific phonological structures.

A further limitation concerns the analysis of lexical tones. While tone was coded as part of the phonological feature set and included in the global prediction models, the present study did not conduct theory-driven tests targeting specific tonal contrasts. This reflects the absence of clear predictions in the existing literature regarding tonal iconicity in Guangzhou Cantonese, as well as the small sample size, which would render any tone-specific analysis severely underpowered. Future research with larger corpora and dedicated experimental paradigms is needed to determine whether lexical tones in Guangzhou Cantonese contribute to iconic-systematicity independently of segmental features.

Despite these limitations, the present study demonstrates that Guangzhou Cantonese ideophones exhibit measurable iconic-systematicity and that these patterns reflect both cross-linguistic regularities and local phonological constraints. The findings therefore contribute to a growing body of evidence suggesting that iconicity is not merely a marginal feature of language but a persistent probabilistic force shaping lexical structure.

Competing Interests

The authors have no competing interests to declare.

Acknowledgements

This work was supported by the Guangzhou Philosophy and Social Science "14th Five-Year Plan" Research Project (广州市哲学社科"十四五"规划课题), "Sound Symbolism and Cultural Embeddedness of Guangzhou Cantonese Reduplicative Ideophones" (Grant No. 2025GZGJ389).

References

[1] Adelman, J.S., Estes, Z. and Cossu, M. 2018. Emotional Sound Symbolism: Languages Rapidly Signal Valence via Phonemes. *Cognition*. **175**, pp.122–130.

[2] Akita, K. and Imai, M. 2022. The Iconicity Ring Model for Sound Symbolism. *Iconicity in Language and Literature*. pp.27–46.

[3] Aryani, A., Conrad, M., Schmidtke, D. and Jacobs, A. 2018. Why "Piss" is Ruder than "Pee?" the Role of Sound in Affective Meaning Making. *PLoS ONE*. **13**(6).

[4] Blasi, D.E., Wichmann, S., Hammarström, H., Stadler, P.F. and Christiansen, M.H. 2016. Sound–meaning Association Biases Evidenced across Thousands of Languages. *Proceedings of the National Academy of Sciences*. **113**(39), pp.10818–10823.

[5] Dingemanse, M. 2012. Advances in the Cross-Linguistic Study of Ideophones. *Language and Linguistics Compass*. **6**(10), pp.654–672.

[6] Dingemanse, M. 2019. Chapter 1. 'Ideophone' as a Comparative Concept In: A. Kimi and P. Pardeshi, eds. *Ideophones, Mimetics and Expressives*. John Benjamins Publishing Company, pp.13–34.

[7] Dingemanse, M., Blasi, D.E., Lupyan, G., Christiansen, M.H. and Monaghan, P. 2015. Arbitrariness, Iconicity, and Systematicity in Language. *Trends in Cognitive Sciences*. **19**(10), pp.603–615.

[8] Farmer, T.A., Christiansen, M.H. and Monaghan, P. 2006. Phonological Typicality Influences On-line Sentence Comprehension. *Proceedings of the National Academy of Sciences*. **103**(32), pp.12203–12208.

[9] Hamano, S.S. 1998. *The Sound-symbolic System of Japanese*. Stanford, CA: CSLI.

[10] Haynie, H., Bowern, C. and LaPalombara, H. 2014. Sound Symbolism in the Languages of Australia. *PLoS ONE*. **9**(4).

[11] Imai, M. and Kita, S. 2014. The Sound Symbolism Bootstrapping Hypothesis for Language Acquisition and Language Evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **369**(1651).

[12] Jespersen, O. 1922. *Language: Its Nature and Development*. New York: Henry Holt and Company.
[13] Johansson, N.E., Anikin, A., Carling, G. and Holmer, A. 2019. The Typology of Sound Symbolism: Defining Macro-concepts via their Semantic and Phonetic Features. *Linguistic Typology*. **24**(2).
[14] Köhler, W. 1929. *Gestalt Psychology*. New York: Liveright.
[15] Lockwood, G., Dingemanse, M. and Hagoort, P. 2016. Sound-symbolism Boosts Novel Word Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. **42**(8), pp.1274–1281.
[16] Newman, S.S. 1933. Further Experiments in Phonetic Symbolism. *The American Journal of Psychology*. **45**(1), p.53.
[17] Nielsen, A. and Rendall, D. 2011. The Sound of Round: Evaluating the Sound-symbolic Role of Consonants in the Classic Takete-Maluma Phenomenon. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*. **65**(2), pp.115–124.
[18] Nielsen, A.K. and Dingemanse, M. 2020. Iconicity in Word Learning and Beyond: A Critical Review. *Language and Speech*. **64**(1), pp.52–72.
[19] Nuckolls, J.B. 1999. The Case for Sound Symbolism. *Annual Review of Anthropology*. **28**(1), pp.225–252.
[20] Occhino, C., Anible, B., Wilkinson, E. and Morford, J.P. 2017. Iconicity is in the Eye of the Beholder. *Gesture*. **16**(1), pp.100–126.
[21] Ohala, J.J. 1983. Cross-Language Use of Pitch: An Ethological View. *Phonetica*. **40**(1), pp.1–18.
[22] Perniss, P., Thompson, R.L. and Vigliocco, G. 2010. Iconicity as a General Property of Language: Evidence from Spoken and Signed Languages. *Frontiers in Psychology*. **1**.
[23] Perry, L.K., Perlman, M. and Lupyan, G. 2015. Iconicity in English and Spanish and Its Relation to Lexical Category and Age of Acquisition. *PLoS ONE*. **10**(9).
[24] Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *ArXiv*.
[25] Sapir, E. 1929. A Study in Phonetic Symbolism. *Journal of Experimental Psychology*. **12**(3).
[26] Sidhu, D.M. 2024. Sound Symbolism in the Lexicon: A Review of Iconic-Systematicity. *Language and Linguistics Compass*. **19**(1).
[27] Sidhu, D.M., Vigliocco, G. and Pexman, P.M. 2020. Effects of Iconicity in Lexical Decision. *Language and Cognition*. **12**(1), pp.164–181.
[28] Sidhu, D.M., Westbury, C., Hollis, G. and Pexman, P.M. 2021. Sound Symbolism Shapes the English Language: The Maluma/Takete Effect in English Nouns. *Psychonomic Bulletin & Review*. **28**(4), pp.1390–1398.
[29] Smith, J. 2015. Sound Symbolism in the Reduplicative Vocabulary of the *Shijing*. *Journal of Chinese Literature and Culture*. **2**(2), pp.258–285.
[30] Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. 2007. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*. **8**(1).
[31] Strobl, C., Malley, J. and Tutz, G. 2009. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. *Psychological Methods*. **14**(4), pp.323–348.
[32] T'sou, B.K. 1978. Sounds Symbolism and Some Socio-and Historical Linguistic Implications of Linguistic Diversity in Sino-Tibetan Languages. *Cahiers De Linguistique. Asie Orientale Paris*. **3**(1), pp.67–76.
[33] Van Hoey, T. 2023. ABB, a Salient Prototype of Collocate–ideophone Constructions in Mandarin Chinese. *Cognitive Linguistics*. **34**(1), pp. 133–163.
[34] Van Hoey, T., Yu, X., Pan, T.- Le and Do, Y. 2024. What Ratings and Corpus Data Reveal about the Vividness of Mandarin ABB Words. *Language and Cognition*. **16**(4), pp.1674–1696.
[35] Vinson, D., Thompson, R.L., Skinner, R. and Vigliocco, G. 2015. A Faster Path between Meaning and form? Iconicity Facilitates Sign Recognition and Production in British Sign Language. *Journal of Memory and Language*. **82**, pp.56–85.
[36] Winter, B. and Perlman, M. 2021. Size Sound Symbolism in the English Lexicon. *Glossa: a Journal of General Linguistics*. **6**(1).
[37] Winter, B., Sóskuthy, M., Perlman, M. and Dingemanse, M. 2022. Trilled /r/ Is Associated with Roughness, Linking Sound and Touch across Spoken Languages. *Scientific Reports*. **12**(1035).
[38] Yu, C.S., McBeath, M.K. and Glenberg, A.M. 2021. The Gleam-Glum Effect: /i:/ versus /ʌ/ Phonemes Generically Carry Emotional Valence. *Journal of Experimental Psychology: Learning, Memory and Cognition*. **47**(7), pp.1173–1185.
[39] Ćwiek, A., Anselme, R., Dediu, D., Fuchs, S., Kawahara, S., Oh, G.E., Paul, J., Perlman, M., Petrone, C., Reiter, S., Rachid Ridouane, Zeller, J. and Winter, B. 2024. The Alveolar Trill Is Perceived as Jagged/Rough by Speakers of Different Languages. *The Journal of the Acoustical Society of America*. **156**(5), pp.3468–3479.