



OPEN ACCESS

EDITED BY

Christian Huyck,
Middlesex University, United Kingdom

REVIEWED BY

Marianna Marcella Bolognesi,
University of Bologna, Italy
Marit Lobben,
Oslo Metropolitan University, Norway
Sandro Rama Fiorini,
IBM, Brazil

*CORRESPONDENCE

Akira Utsumi
utsumi@uec.ac.jp

SPECIALTY SECTION

This article was submitted to
Cognitive Science,
a section of the journal
Frontiers in Psychology

RECEIVED 28 March 2022

ACCEPTED 14 September 2022

PUBLISHED 04 October 2022

CITATION

Utsumi A (2022) A test of indirect
grounding of abstract concepts using
multimodal distributional semantics.
Front. Psychol. 13:906181.
doi: 10.3389/fpsyg.2022.906181

COPYRIGHT

© 2022 Utsumi. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

A test of indirect grounding of abstract concepts using multimodal distributional semantics

Akira Utsumi*

Department of Informatics, Artificial Intelligence Exploration Research Center, The University of Electro-Communications, Tokyo, Japan

How are abstract concepts grounded in perceptual experiences for shaping human conceptual knowledge? Recent studies on abstract concepts emphasizing the role of language have argued that abstract concepts are grounded indirectly in perceptual experiences and language (or words) functions as a bridge between abstract concepts and perceptual experiences. However, this “indirect grounding” view remains largely speculative and has hardly been supported directly by empirical evidence. In this paper, therefore, we test the indirect grounding view by means of multimodal distributional semantics, in which the meaning of a word (i.e., a concept) is represented as the combination of textual and visual vectors. The newly devised multimodal distributional semantic model incorporates the indirect grounding view by computing the visual vector of an abstract word through the visual vectors of concrete words semantically related to that abstract word. An evaluation experiment is conducted in which conceptual representation is predicted from multimodal vectors using a multilayer feed-forward neural network. The analysis of prediction performance demonstrates that the indirect grounding model achieves significantly better performance in predicting human conceptual representation of abstract words than other models that mimic competing views on abstract concepts, especially than the direct grounding model in which the visual vectors of abstract words are computed directly from the images of abstract concepts. This result lends some plausibility to the indirect grounding view as a cognitive mechanism of grounding abstract concepts.

KEYWORDS

abstract concepts, indirect grounding, embodied cognition, multimodal distributional semantic model, conceptual representation, symbol grounding problem

1. Introduction

1.1. Abstract concepts and embodied cognition

Since Harnad (1990) pointed out the symbol grounding problem, embodied approaches to cognition have emerged as promising solutions of how symbols (or words) acquire their meanings. Embodied cognition theories argue that concepts or word meanings are grounded in our perceptual or sensorimotor experiences.

For example, Barsalou's (1999) theory of perceptual symbol systems states that concepts (and word meanings as well) are inherently modal, perceptual symbols grounded in the real world. Perceptual symbols refer to neural representations, or activation patterns of sensorimotor regions of the brain elicited during direct perceptual experiences (e.g., seeing, touching, and playing with dogs for the concept of dog). Once these representations are encoded in the brain by repeated experiencing, they can be reactivated, that is, experiences are mentally simulated when words are encountered even in the absence of direct experience. Embodied cognition theories have been empirically supported by a considerable number of studies (e.g., Glenberg and Kaschak, 2002; Kaschak et al., 2005; Pecher and Zwaan, 2005; Barsalou, 2008; Pulvermüller, 2013; Scorolli, 2014; Barsalou, 2016; Coello and Fisher, 2016; Fisher and Coello, 2016).

However, abstract concepts pose a serious challenge to the embodied theory of cognition. Because abstract concepts such as *love* and *justice* do not have clearly perceivable referents, it is difficult to see how representations grounded in perceptual experiences can capture the content of abstract concepts. The empirical studies on embodied cognition have previously focused primarily on concrete concepts, such as *dog* and *kick*, which directly refer to perceivable objects or physical motions. Recently, however, the focus of concepts research has recently shifted from the embodied nature of concrete concepts to the complex nature of abstract concepts (Bolognesi and Steen, 2018; Borghi et al., 2018).

Some embodied theories claim a general mechanism of grounding common to both concrete and abstract concepts. Barsalou (1999, 2003) advocates that abstract concepts are represented in the same perceptual symbol systems. Perceptual symbols for abstract concepts are acquired from sensorimotor and introspective experiences in specific situations and abstract words elicit mental simulations of those situations. For example, people visualize and emotionalize two people kissing when seeing the word "love" and the court when seeing the word "justice." A recent view of situated simulation is more radical; Barsalou et al. (2018) proposed that the distinction between concrete and abstract concepts is no longer useful and should be abandoned because all concepts can be explained within the situated simulation view. The situated simulation view is supported by a number of empirical studies. Barsalou and Wiemer-Hastings (2005) and Wiemer-Hastings and Xu (2005) found using a property generation task that, when participants generated properties for abstract concepts, they were likely to describe social and introspective aspects of the situations, whereas for concrete concepts they tended to describe properties of entities in the situations. McRae et al. (2018) demonstrated that pictures of specific situations facilitated lexical decisions to abstract words relevant to the picture primes, and conversely abstract words also facilitated processing of pictures depicting the relevant situations. These empirical findings may suggest

that the situated simulation view is plausible and at least partially resolves the problem of how abstract concepts are grounded in our perceptual or sensorimotor experiences.

The situated simulation view, however, is not essentially sufficient to explain abstract concepts. Imagine that you have to explain abstract concepts using only visual images or videos without language. For example, to explain what is love, people may show a picture of two people kissing and hugging, a picture of wedding ceremony, and/or a picture of specific dating situations. These pictures can convey some conceptual knowledge about love, and more elaborate visual images such as films can convey greater knowledge. We feel nevertheless that only seeing them lacks something to fully understand the concept of love. This difficulty becomes more serious when more abstract concepts (e.g., *justice* and *democracy*) have to be explained; they are more difficult to explain using only visual images. This simple thought experiment suggests that situated simulation is somewhat limited as a thorough theory of abstract concepts. Another limitation is that the situated simulation view seems not to provide a clear explanation of how abstract concepts are linked to the relevant situations in acquiring those concepts. It is much less likely that people think of abstract concepts (e.g., *democracy*), or see or hear abstract words that refer to those concepts, at the same time as experiencing the situations associated with those concepts (e.g., *casting a vote in a polling station*), in contrast to concrete concepts (e.g., *dog*), which are often mentioned verbally in the situations including their referents. Bergelson and Swingley (2013) actually demonstrated through a video-corpora analysis of mother–infant interaction that mothers used abstract words less often in the presence of their referent events than they used concrete words in the presence of their referent objects.

These limitations of the situated simulation view can be largely overcome by taking into account language not only as a source of conceptual knowledge but also as an effective means of grounding abstract concepts in the real world. It has been widely accepted that language is much more important for representing abstract concepts (e.g., Borghi et al., 2017; Dove, 2018). Neuroimaging studies have demonstrated that processing of abstract concepts elicits greater activation of the left-dominant Perisylvian language network (including the left inferior frontal gyrus and the left superior temporal cortex) as compared to processing of concrete concepts (e.g., Binder et al., 2009; Wang et al., 2010). Recent embodied theories of abstract concepts have therefore emphasized the role of language in forming and processing abstract concepts or words (Borghi et al., 2017; Bolognesi and Steen, 2018). One of the important questions to be addressed by these theories is how language and embodied experience contribute to shaping our conceptual knowledge of abstract concepts and meaning representation of abstract words (e.g., Bolognesi and Steen, 2018). This question is what we address in this paper.

1.2. Hybrid theory integrating symbolic and embodied cognition

The dual coding theory (Paivio, 1971, 1986) is an early influential theory that integrates symbolic and embodied cognition. The main claim of this theory is that concepts are represented in two separate systems, that is, a verbal system for linguistic information and a visual system for mentally visual images. Furthermore, the dual coding theory argues that concrete words activate both the verbal and visual systems, but abstract words activate only the verbal system. This argument is consistent with the concreteness effect, whereby concrete words have processing and mnemonic advantages over abstract words. By contrast, it implies that abstract words are represented primarily by linguistic information, and it is not clear whether and how the visual system contributes to the representation of abstract words. To explain the concreteness effect, Schwanenflugel et al. (1988) and Schwanenflugel (1991) also proposed the context availability theory. According to this theory, concrete words are strongly associated with a few contexts, whereas abstract words are weakly associated with many contexts. Therefore, the context availability theory can explain the concreteness effect because abstract words require more effort to activate their contexts. Although contexts in this theory can be both linguistic or embodied, this theory is not devoted to the differences and relations between linguistic and embodied contexts or representations.

Recent hybrid theories, which are collectively referred to as “multiple representation theories” (Borghi et al., 2017), are more committed to how language processing interacts with embodied cognition. The multiple representation theory proposed first is Language And Situated Simulation (LASS) theory (Barsalou et al., 2008). The LASS theory focuses on the temporal interplay between language processing and situated simulation during conceptual processing. According to the LASS theory, when a word is perceived, both linguistic and simulation systems become active initially, but the linguistic system becomes engaged immediately to categorize the word. For the tasks requiring only shallow comprehension (e.g., lexical decision task), language processing would suffice. When deeper conceptual processing (e.g., property generation task) is required, the simulation system is activated later after the activation of the linguistic system peaks. However, the LASS theory is not aimed at explaining how abstract concepts are represented. It claims that both concrete and abstract concepts activate a mixture of linguistic and embodied information, and which information is dominant is determined depending on the task¹, not the concept. The situated simulation view still holds in this framework, and thus the LASS theory also suffers

¹ As explained in section 3, the task used in the experiment of this study is to predict a featural representation of concepts comprising attributes and their salience. This task can be regarded as a kind of deeper

from the limitations of the situated simulation view described in section 1.1.

A more influential theory for the multiple representation views of abstract concepts is the “Words As social Tools” (WAT) theory (Borghi et al., 2013; Borghi and Binkofski, 2014; Borghi et al., 2019). The WAT theory claims that abstract concepts depend more on language than concrete concepts, but the role of language is not limited to word association. It emphasizes the importance of language (or words) as tools to perform social actions, and argues that the situated simulation (i.e., re-enactment) of social experience through language is necessary for representing and acquiring abstract concepts. The WAT theory is supported by a number of empirical findings on language acquisition and brain organization (see, Borghi et al., 2017, 2019), but it does not clearly explain the mechanism of how language shapes the meaning representation of abstract concepts. Words not only are tools for direct social experiencing of language-related actions and events, but also function as a bridge to direct perceptual and sensorimotor experience, which is the main tenet of the indirect grounding view described in the next section.

1.3. Indirect grounding view

The hybrid views mentioned above assume that language provides a separate source of conceptual knowledge independent of embodied experience or the use of language is itself a constituent of embodied experience in which abstract concepts are grounded. Unlike the hybrid views, some recent studies have been devoted specifically to how language is used to relate abstract concepts to embodied experience.

The symbol interdependency hypothesis proposed by Louwerse (2011, 2018) argues for the role of language as a shortcut to the perceptual or embodied system. According to the symbol interdependency hypothesis, language comprehension is symbolic through interdependencies of amodal linguistic symbols, while it is indirectly embodied through the references linguistic symbols make to perceptual representations. Hence, “language has evolved to become a communicative short-cut for language users and encodes relations in the world, including embodied relations (Louwerse, 2011, p. 279).” Dove (2014) also argues that language provides an important means of extending our cognitive capabilities and encoding abstract concepts by enabling access to an embodied representational system that exists independently of language. Thill et al. (2014) suggest a similar view for robotic models of language grounding. Their “division of labor” approach assumes two layers of conceptual processing; a perceptual layer that associates basic, concrete concepts with perceptual features and a relational (i.e., linguistic)

conceptual processing such as a property generation task that requires both linguistic and simulation (i.e., embodied) systems.

layer that grounds more complex and abstract concepts in relation to basic concepts. The relational layer can be formed by the interdependency of linguistic symbols obtained through distributional learning. Lupyán and Lewis's (2019) "words-as-cues" view is consistent with these views; they argue that language provides a cue to meaning that can augment semantic knowledge derived from perceptual experiences or construct semantic knowledge.

In fact, the same line of thought has been suggested earlier in the context of word learning or language acquisition. Howell et al. (2005) proposed a mechanism of "propagation of grounding" in which an abstract word inherits some meaning from the concrete words to which it is related. In their words, "the grounded meaning propagates up through the syntactic links of the co-occurrence meaning network, from the simplest early words to the most abstract (Howell et al., 2005, p. 260)." Gleitman et al. (2005) explained how "hard words" are acquired along the same line. In the early stage of lexical acquisition, the meaning of concrete words is acquired directly from perceptual information via word-to-world pairing. In the later stage, the meaning of hard words, which is not easily accessible through perception, is acquired by a structure-to-world mapping procedure that combines linguistic observations with co-occurring perceptual experience.

The basic idea underlying all these views is that abstract concepts or the meaning of abstract words are grounded in sensorimotor or perceptual experiences, but the grounding is indirect, rather than direct in the case of concrete concepts. Language not only provides a means to understand and represent abstract concepts (and the meaning of abstract words as well) through statistical regularities in linguistic surface structure, but also functions as a mediator between abstract concepts and perceptual experiences for a deeper understanding of abstract concepts in the absence of direct experiences with the words referring to the abstract concepts. For example, people can understand the concept of love by associating the word "love" with "kiss" (and many other relevant mediator words) via linguistic interdependency and mentally simulating the situation of kissing, even though they have never encountered the word "love" directly in the situation of kissing. In this paper, we collectively refer to these views as *indirect grounding views*.

Although the indirect grounding view may be able to provide a promising solution for the symbol grounding problem of abstract concepts, it remains largely speculative and has hardly been supported directly by empirical evidence. To empirically justify the symbol interdependency hypothesis, Louwerse and Jeuniaux (2010) demonstrated that the symbolic factor (i.e., frequency of word pairs) predicted error rates and response time in both semantic and iconicity judgments, whereas the embodied factor (i.e., iconic configuration) predicted error rates and response time in iconicity judgment. Malhi and Buchanan (2018) recently extended Louwerse and Jeuniaux's (2010) findings by using both concrete and abstract words as

stimuli and found that, for both concrete and abstract words, the symbolic factor dominated in semantic judgment and embodied factor dominated in iconicity judgment. Although these findings support the general claim that language comprehension is both embodied and symbolic, which can be predicted by the indirect grounding view and even by some of the hybrid views, they do not provide direct evidence for indirect grounding of abstract words.

Recently, Günther et al. (2020) provided more direct evidence for the indirect grounding view using an experimental paradigm (e.g., Zwaan and Yaxley, 2003) in which target words are faster to process when their perceptually embodied meaning (e.g., spatial location) is congruent with perceptual experiences that participants have in the experiment. They applied this paradigm to new concepts for which participants had no direct perceptual experience, but which they learned from language alone referring to vertical (i.e., up or down) concepts. The result was that, after learning new concepts via language, participants were faster at responding to sentences describing those concepts when their implied vertical position matched the direction of their hand movement for responding. This finding indicates that novel (unknown) concepts, even though not grounded directly, can be grounded indirectly by establishing a connection with directly grounded concepts via language network. In this paper, we test further the validity of the indirect grounding view for existing abstract concepts, in particular the role of language in the grounded representation of abstract concepts, using another methodology, that is, by means of multimodal distributional semantics described next.

1.4. Multimodal distributional semantics

Distributional semantics is an effective computational approach to constructing word meaning representations (i.e., word vectors) from the distributional statistics of words in large collections of text (Turney and Pantel, 2010; Lenci, 2018; Pilehvar and Camacho-Collados, 2020). Distributional semantics has been widely used in natural language processing (NLP) as meaning representations for neural networks or deep learning (Goldberg, 2017) and in cognitive science as a cognitive modeling method (Jones et al., 2015; Kumar, 2021). In cognitive research on concepts, in particular on embodied vs. symbolic processing, distributional semantics is regarded as a de facto standard language model (de Vega et al., 2008; Bolognesi and Steen, 2018).

Distributional semantics has been criticized as psychologically implausible because it is based only on linguistic (i.e., symbolic) information and thus suffers from the symbol grounding problem (de Vega et al., 2008; Baroni, 2016). Although it is controversial whether distributional semantics cannot essentially capture human semantic or conceptual knowledge, it is undoubtedly unable to represent the meaning

of some kinds of words, in particular concrete words, just as they are represented in human semantic memory.

An earlier approach to this problem is to integrate feature-based information, which is often produced by humans in property generation tasks, with distributional semantics (Andrews et al., 2009; Johns and Jones, 2012; Silberer and Lapata, 2012; Hill and Korhonen, 2014). For example, Andrews et al. (2009) used perceptual features collected as featural norm to ground a language-based topic model on perceptual experience, and demonstrated that the integrated model outperformed the language-based topic model. However, the grounding ability of these feature-integrated models is not sufficient for modeling embodied cognition. Perceptual features produced by humans are limited to what can be conveyed verbally, and they are often only salient and distinctive. Hence, implicit perceptual features characterizing a concept cannot be taken into account in the models (Bruni et al., 2014). Additionally, the number of concepts (or words) used in experiments with human-generated properties is relatively small.

A more promising and common approach is to directly integrate non-verbal information with (text-based) distributional semantics. Multimodal distributional semantics has been proposed for this purpose (for a review, see Baroni, 2016). In multimodal distributional semantics, linguistic or textual information is integrated with perceptual information computed directly from non-linguistic inputs such as visual (Bruni et al., 2014; Kiela et al., 2014; Silberer et al., 2017), auditory (Kiela and Clark, 2015), or olfactory (Kiela et al., 2015) ones. Furthermore, another type of approach has been proposed which utilizes visual information without directly computing visual vectors from images. Bolognesi's (2016, 2017) Flickr® Distributional TagSpace uses the co-occurrence statistics between user-generated tags that appear in the same image to generate word vectors. In this paper, we use multimodal distributional semantics to model the hybrid view of conceptual representation.

Multimodal distributional semantic models generally compute and utilize perceptual vectors in the same way for all words; they do not take account of the difference between concrete and abstract words in terms of how concepts are grounded in perceptual information, which is claimed by the indirect grounding view. Furthermore, it has been empirically demonstrated that a simple addition of perceptual information is beneficial only for concrete concepts (Bruni et al., 2014; Kiela et al., 2014). Therefore, in section 2, we devise a new multimodal distributional semantic model for indirect grounding by incorporating the indirect grounding view into an algorithm for constructing multimodal word vectors. In the devised model, the perceptual vector of an abstract word is computed from the perceptual representations of concrete words that are semantically related to (or associated with) that abstract word.

It must be noted that technically word vectors in (both unimodal and multimodal) distributional semantics are regarded as representing the meaning (or semantics) of words. In this paper, however, we consider the meaning of words and concepts (or the conceptual knowledge) as interchangeable, as is usually assumed in the cognitive science literature (e.g., Vigliocco and Vinson, 2007; Jackendoff, 2019). Although we do not intend to argue that concepts and word meanings are the same, it is much difficult or impossible to distinguish between concepts and word meanings in most of the cases; empirical studies on embodied cognition have used words for the tasks on conceptual representation in human adults (Borghi et al., 2017). In what follows, therefore, we assume that distributional word vectors also represent the concept referred to by a word, and that multimodal distributional semantics can be applied to modeling embodied conceptual processing.

1.5. Aim of this study

The aim of the present study is to test the indirect grounding view of abstract concepts by using computational modeling based on multimodal distributional semantics. For this purpose, we compare the indirect grounding view and other competing views mentioned above, the basic tenets of which are summarized in Table 1. By examining which of these views can predict the performance difference among distributional semantic models that mimic these views (and other baseline models), we attempt to test the validity of the indirect grounding view. The performance of distributional semantic models is evaluated in terms of the degree to which human conceptual representation can be predicted by the models, using Utsumi's (2020) experimental framework for analyzing and evaluating distributional semantic vectors. In the evaluation experiment, we focus on visual images as a source of perceptual (i.e., non-verbal) information, as used in many other studies on multimodal distributional semantics.

In the rest of this paper, after describing in detail a new multimodal semantic model for simulating indirect grounding in section 2, we explain the method of our evaluation experiment in section 3. We then report the results of the evaluation experiment in section 4 and discuss the implications and limitations of the findings in section 5.

2. Distributional semantic model for indirect grounding

We devise a new model to incorporate the indirect grounding view into multimodal distributional semantics. According to the indirect grounding view arguing that grounding of abstract concepts is mediated by language, different methods are used for computing visually grounded

TABLE 1 Summary of conceptual representation theories.

Theory	Basic tenets	Processing difference between abstract and concrete concepts
Situated simulation view (e.g., Barsalou, 1999; Barsalou et al., 2018)	<ul style="list-style-type: none"> • Concepts are grounded in perceptual experiences via mental simulation. • Language information is not necessary for understanding concepts. 	No
Dual coding theory (Paivio, 1971, 1986)	<ul style="list-style-type: none"> • Concrete concepts are both linguistic and grounded in perceptual experiences. • Abstract concepts are only linguistic. 	Yes
Hybrid view (e.g., Barsalou et al., 2008; Borghi et al., 2019)	<ul style="list-style-type: none"> • Concepts are both linguistic and grounded in perceptual experiences. • The mechanism of grounding does not differ between concrete and abstract concepts; both concepts are grounded directly. 	No
Indirect grounding view (e.g., Howell et al., 2005; Louwerse, 2011)	<ul style="list-style-type: none"> • Concepts are both linguistic and grounded in perceptual experiences. • Abstract concepts are grounded indirectly via language, whereas concrete concepts are grounded directly. 	Yes

vectors depending on whether words are concrete or abstract. For a concrete word, its visually grounded vector is computed directly from the visual images tagged with that word, as shown in Figure 1A. For example, starting from the blind and not knowledgeable assumption that “love” is a concrete word, the visually grounded vector of “love” is computed directly from the images tagged with “love” using deep neural networks (DNNs). In this case, the visually grounded vector $\vec{g}(\text{love})$ is identical to the directly computed visual vector $\vec{v}(\text{love})$, as generally assumed by multimodal distributional semantic models that take no account of the difference between concrete and abstract concepts. By contrast, the visually grounded vector of an abstract word is computed from the visual images of concrete words semantically associated with the abstract word, that is, *semantic neighbors* of the abstract word, assuming that semantic neighbors are good mediator words. As shown in Figure 1B where the word “love” is supposed to be abstract,

its visually grounded vector $\vec{g}(\text{love})$ is not computed directly from the images of “love,” but computed indirectly using the direct visual vectors, e.g., $\vec{v}(\text{kiss})$, $\vec{v}(\text{mother})$, $\vec{v}(\text{wedding})$, derived from the images tagged with semantic neighbors (e.g., “kiss,” “mother,” and “wedding”). This computation is regarded as an implementation of indirect grounding for abstract concepts.

Formally, we define a multimodal distributional semantic model for indirect grounding as follows. We assume that the vocabulary V is divided into a set of concrete words V_C and a set of abstract words V_A . Each word $w_i \in V$ has a textual vector $\vec{t}(w_i) \in DSM_T$ trained from a text corpus and a direct visual vector $\vec{v}(w_i) \in DSM_V$ computed directly from images for the word w_i . We build an indirect grounding model DSM_I in which a word is represented by a pair $[\vec{t}(w_i), \vec{g}(w_i)]$ of a textual vector $\vec{t}(w_i) \in DSM_T$ and visually grounded vector $\vec{g}(w_i) \in DSM_G$. The visually grounded vector $\vec{g}(w_i)$ is defined as follows:

$$\vec{g}(w_i) = \begin{cases} \vec{v}(w_i) & \text{(for a concrete word } w_i \in V_C) \\ \frac{\sum_{w_j \in SN_k(w_i)} \vec{v}(w_j)}{k} & \text{(for an abstract word } w_i \in V_A) \end{cases} \quad (1)$$

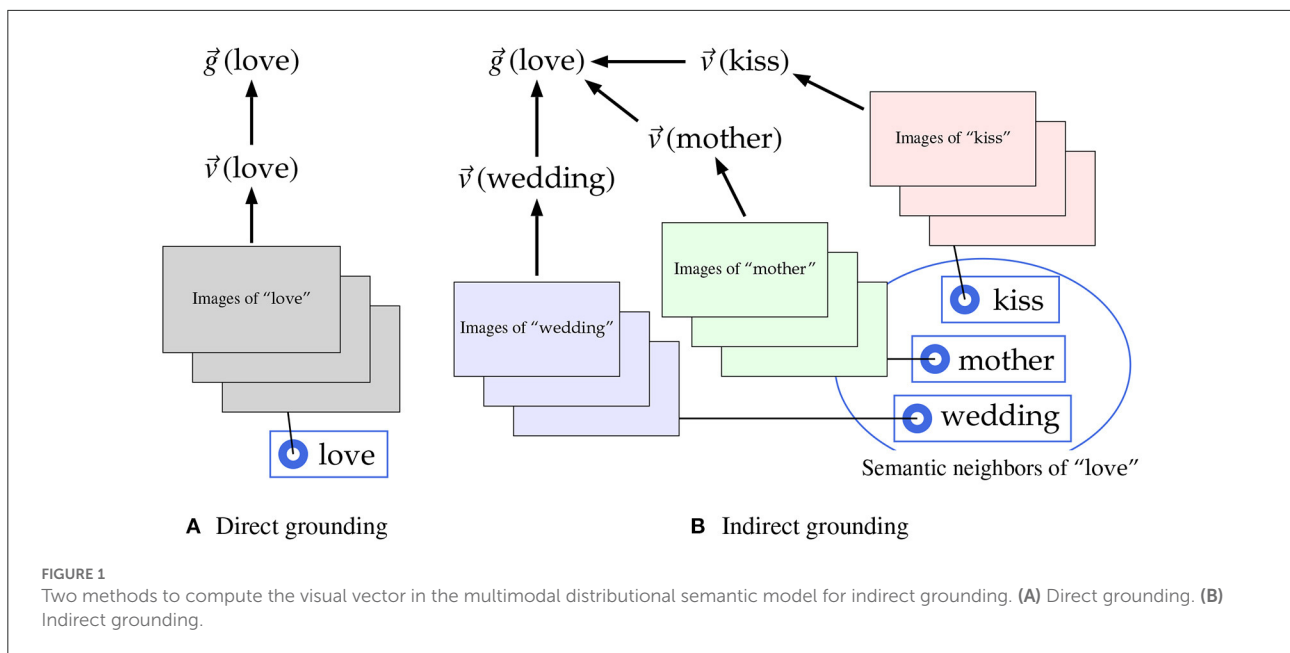
where $SN_k(w_i) \subset V_C$ is a set of k semantic neighbors of w_i , that is, k concrete words most semantically related to w_i . The visually grounded vector for an abstract word is thus obtained by averaging k direct visual vectors of semantic neighbors.

Semantic neighbors of abstract words are determined using word similarity in the text-based distributional semantic model DSM_T . This implies that linguistic interdependency for indirect grounding is modeled by (text-based) distributional semantics. In the devised model, semantic neighbors $SN_k(w_i) \subset V_C$ of an abstract word w_i are determined by first selecting $N (> k)$ nearest neighbors of w_i from the whole vocabulary V and then selecting k nearest concrete words from the set of N neighbors. Nearest neighbors are computed using cosine similarity in the textual model DSM_T . The reason for limiting N neighbors before selecting k concrete words is that some highly abstract words (e.g., “truth,” “wisdom”) may not have semantically related concrete words, and in this case it is more appropriate not to consider a visual representation. Hence, when no semantic neighbors are selected [i.e., $SN_k(w_i) = \emptyset$], no visual vector is considered and only the textual vector is used for representing the abstract word w_i in DSM_I .

3. Materials and methods

3.1. Experimental design and predictions

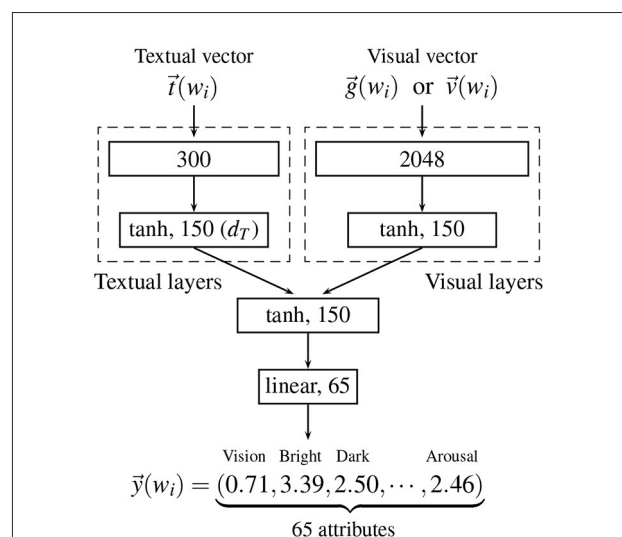
To evaluate the representational ability of a given distributional semantic model, we examined how accurately the model can predict human conceptual representation using Utsumi’s (2020) experimental framework for analyzing and evaluating distributional semantic models. As human conceptual representation, we used a brain-based semantic



representation provided by Binder et al. (2016), which is described in detail in section 3.2. In the evaluation experiment, a function from the bimodal (or unimodal) vector of a given word w_i (section 3.3) to the target human conceptual representation $\vec{y}(w_i)$ (section 3.2) was trained using the feed-forward neural network shown in Figure 2. The conceptual representation of untrained words was predicted by the trained neural network. Prediction performance was evaluated by comparing the estimated conceptual representation with the target representation. The details of training and test procedure is described in section 3.4.

To test whether the indirect grounding view is more plausible than other views on abstract concepts listed in Table 1, we conducted the evaluation experiment described above using the following distributional semantic models, that is, four models that correspond to each of the four views in Table 1 and two additional baselines.

- Indirect grounding model $DSM_I = \{[\vec{t}(w_i), \vec{g}(w_i)] \mid \vec{t}(w_i) \in DSM_T, \vec{g}(w_i) \in DSM_G, w_i \in V\}$: A bimodal model for the indirect grounding view described in section 2.
- Hybrid model $DSM_H = \{[\vec{t}(w_i), \vec{v}(w_i)] \mid \vec{t}(w_i) \in DSM_T, \vec{v}(w_i) \in DSM_V, w_i \in V\}$: A standard bimodal model in which all visual vectors are computed directly from images for a word w_i . This model is assumed to simulate the hybrid view.
- Dual coding model $DSM_D = \{[\vec{t}(w_i), \vec{v}(w_i)] \mid \vec{t}(w_i) \in DSM_T, \vec{v}(w_i) \in DSM_V, w_i \in V_C\} \cup \{[\vec{t}(w_i) \mid \vec{t}(w_i) \in DSM_T, w_i \in V_A\}$: A partially bimodal model in which a pair of textual and direct visual vectors is used for representing a concrete word, whereas only a textual vector



is used for an abstract word. When abstract words are given in the training and test procedure, the visual layers in Figure 2 are not either trained or used. This model corresponds to the dual coding theory.

- Visual model $DSM_V = \{\vec{v}(w_i) \mid w_i \in V\}$: A unimodal model in which only direct visual vectors are used for

TABLE 2 Predictions of evaluation performance by conceptual representation theories.

Theory	Best (highest performance) model predicted by each theory	
	Abstract concepts	Concrete concepts
Situated simulation view (e.g., Barsalou, 1999; Barsalou et al., 2018)	Visual	Visual
Dual coding theory (Paivio, 1971, 1986)	Dual coding, textual	Dual coding, hybrid, indirect grounding
Hybrid view (e.g., Barsalou et al., 2008; Borghi et al., 2019)	Hybrid	Dual coding, hybrid, indirect grounding
Indirect grounding view (e.g., Howell et al., 2005; Louwerse, 2011)	Indirect grounding	Dual coding, hybrid, indirect grounding

representing words. Hence the textual layers in Figure 2 are not used in the training and test procedure. This model corresponds to the situated simulation view.

- Textual model $DSM_T = \{\vec{t}(w_i) | w_i \in V\}$: A unimodal baseline model in which only textual vectors are used for representing words. Hence the visual layers in Figure 2 are not used in the training and test procedure.
- Indirect visual model $DSM_G = \{\vec{g}(w_i) | w_i \in V\}$: A unimodal baseline model in which only indirect visual vectors (defined in Equation 1) are used for representing words. The textual layers in Figure 2 are ignored in the training and test procedure.

By comparing the prediction performance of the indirect grounding model with those of other models, we test the validity of indirect grounding as a plausible mechanism for embodied representation of abstract concepts. Different views of abstract concepts summarized in Table 1 make different predictions about the performance of the six models, as shown in Table 2. Basically, each of the four views predicts that its corresponding model would outperform other models. For abstract concepts, the dual coding theory predicts that the textual model also achieves the best performance because it argues that abstract concepts are represented only by linguistic information. For concrete concepts, the indirect grounding, hybrid, and dual coding views do not differ and thus make the same prediction that the indirect grounding, hybrid, and dual coding models do not significantly differ in performance and outperform the remaining models.

TABLE 3 Example of words included in Binder et al.'s (2016) dataset, which are selected mainly from abstract words.

POS	Category	Word examples
Noun	Abstract construct	analogy, irony, truth, verb, worth
	Cognitive entity	belief, hope, knowledge, sympathy, wit
	Emotion	gratitude, joy, love, shame, woe
	Social event	advice, deceit, matinee, snub, tribute
	Time period	day, era, evening, semester, summer
Verb	Locative action	approach, deliver, go, leave, walk
	Social action	arrest, celebrate, help, play, write
Adjective	Visual property	black, dark, new, red, shiny
	Emotional property	angry, dangerous, happy, lonely, peaceful

3.2. Human conceptual representation

As a target human conceptual representation $\vec{y}(w_i)$, we used Binder et al.'s (2016) brain-inspired featural representation². They provided 65-dimensional real-valued vectors of 535 words, some of which are listed in Table 3. These words comprise 434 nouns, 62 verbs, and 39 adjectives and are classified into 47 categories that reflect grammatical classes and semantic classes. Note that, for the reason explained later in section 3.3.3, two nouns were excluded from the experimental materials and thus the remaining 533 words were used in the evaluation experiment.

The dimensions of the vectors correspond to neurobiologically plausible attributes whose neural correlates have been well described. Table 4 lists all 65 attributes in 14 domains used in Binder et al.'s (2016) vectors. Binder et al. (2016) selected these attributes according to two fundamental principles; they correspond to distinguishable neural processors that can be identified by an extensive body of evidence from brain imaging and neurological studies, and they can contribute to concept acquisition and composition. Each value of the conceptual vectors represents the degree of salience of the corresponding attribute for the target word. Binder et al. (2016) collected these values using Amazon Mechanical Turk. The participants of the experiment were given a single word and questions such as "To what degree do you think of this thing as a characteristic or defining color (for the attribute *Color*)" with some examples, and asked to rate the degree on a 7-point scale ranging from 0 to 6. Collected ratings were averaged for each word and attribute after data screening, and these mean ratings were used in conceptual vectors.

² Binder et al.'s (2016) data is available at <http://www.neuro.mcgw.edu/semanticrepresentations.html>.

TABLE 4 Sixty-five attributes used in Binder et al.'s (2016) conceptual representation.

Domain	Attribute
Vision	Vision, bright, dark, color, pattern, large, small, motion, biomotion, fast, slow, shape, complexity, face, body
Somatic	Touch, temperature, texture, weight, pain
Audition	Audition, loud, low, high, sound, music, speech
Gustation	Taste
Olfaction	Smell
Motor	Head, upper-limb, lower-limb, practice
Spatial	Landmark, path, scene, near, toward, away, number
Temporal	Time, duration, long, short
Causal	Caused, consequential
Social	Social, human, communication, self
cognition	Cognition
Emotion	Benefit, harm, pleasant, unpleasant, happy, sad, angry, disgusted, fearful, surprised
Drive	Drive, needs
Attention	Attention, arousal

3.3. Distributional semantic model

3.3.1. Textual vector

Textual vectors $\vec{t}(w_i) \in DSM_T$ were trained on the Corpus of Contemporary American English (COCA), which included 0.56G word tokens. Words that occurred less than 30 times in the corpus were ignored, resulting in the training vocabulary of 108,230 words. As a distributional semantic model for training textual vectors, we used skip-gram with negative sampling (SGNS), which is one of two algorithms in word2vec model (Mikolov et al., 2013). In SGNS, a feed-forward neural network with one hidden layer of d units is trained to predict co-occurring words of an input word (i.e., w words appeared on either side of the input word in the corpus), and d -dimensional activation vectors in the hidden layer of the trained network are used as textual vectors. We set the vector dimension $d = 300$ and the window size $w = 10$. The choice of corpus, distributional semantic model, and parameter values was determined considering the result of the similar experiment (Utsumi, 2020).

3.3.2. Visual vector

To compute direct visual vectors $\vec{v}(w_i) \in DSM_V$, we collected 20 images using Flickr image retrieval for each of the words in the vocabulary. The image retrieval was performed using the API `flickr.photos.search` with the argument `sort=relevance` and the top 20 most relevant images were downloaded for each word. Note that these relevant images

are often tagged with other words, but we did not use the information of these tags.

To compute the feature vector of each downloaded image, we utilized the ResNet152-hybrid1365 model (Zhou et al., 2018)³. This model is the Residual Network (ResNet), which is a recent high-performance version of the deep convolutional neural networks, trained on both ImageNet1000 dataset for object recognition and Places365-standard dataset for scene recognition. Each image was entered into this model and a 2,048-dimensional activation vector was extracted from the last hidden layer. The activations in the last hidden layer are deemed to be appropriate for a visual vector in distributional semantics, because they are generally assumed to represent visually intrinsic features of a concept. Finally, the visual vector $\vec{v}(w_i)$ was computed as the centroid (i.e., average) of the activation vectors of 20 images.

3.3.3. Indirect visual vector

To compute indirect visual vectors $\vec{g}(w_i) \in DSM_G$, we must determine how to split the whole vocabulary V into concrete words V_C and abstract words V_A . For this purpose, we used Brysbaert et al.'s (2014) concreteness ratings for 39,354 English words including 37,058 single words and 2,896 two-word expressions. These words were rated on a 5-point scale ranging from 1 (abstract) to 5 (concrete) and the collected ratings were averaged per each word. In the instructions given to raters, Brysbaert et al. (2014) stressed that the assessment of word concreteness would be based on perceptual experiences involving all senses and motor responses. Specifically, the following instruction was used:

Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words concrete words. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are abstract words. (Brysbaert et al., 2014, p. 906)

This definition of concrete words as experience-based and abstract words as language-based is consistent with our view of abstract concepts, and thus the use of their concreteness ratings is appropriate for the indirect grounding model⁴.

³ The ResNet152-hybrid1365 model is publicly available at <https://github.com/CSAILVision/places365>.

⁴ Furthermore, although Brysbaert et al. (2014) instructed that word concreteness would be judged according to experiences involving all five senses, the analysis on the obtained concreteness ratings revealed that participants largely focused on visual and haptic experiences. This result further justifies the appropriateness of their data for using to determine abstract words for our model focusing only on visual experience.

For the vocabulary V in the indirect grounding model, we chose 28,437 words from Brysbaert et al.'s (2014) word concreteness dataset that were also included in the training vocabulary of COCA corpus and associated with at least 20 images. As a result, two words "ire" and "oration" in Binder et al.'s (2016) dataset were not included in the chosen vocabulary because they are not contained in Brysbaert et al.'s (2014) dataset. Hence, these two words were not used in the entire experiment.

Each word in the vocabulary V was judged as abstract if its concreteness rating was less than a given threshold θ_c , and otherwise as concrete. We performed the same experiment with different thresholds ranging from 1.2^5 to 5.0 with a step size of 0.1. In section 4, we report the overall result with a representative threshold $\theta_c = 3.0$ at which any words whose concreteness rating is toward the language-based side of the continuum are classified as abstract. Additionally, we use the threshold $\theta_c = 4.0$, at which words are treated as abstract unless they are rated as highly experience-based. Note that 116 and 214 out of 533 words in Binder et al.'s (2016) dataset (14,305 and 21,471 out of 28,437 words in the whole vocabulary V), respectively, were judged as abstract when $\theta_c = 3.0$ and $\theta_c = 4.0$.

After concrete words V_C and abstract words V_A are determined, indirect visual vectors $\vec{g}(w_i)$ are computed according to Equation (1). When a word w_i is concrete (i.e., $w_i \in V_C$), its direct visual vector $\vec{v}(w_i)$ defined in section 3.3.2 is used as a visual vector $\vec{g}(w_i)$. When a word w_i is abstract (i.e., $w_i \in V_A$), its semantic neighbors are determined using the textual vectors in section 3.3.1 as follows. First, $N(> k)$ nearest neighbors of w_i are selected from the whole vocabulary V by computing cosine similarity between $t(w_i)$ and $t(w_j)$ for all words $w_j \in V(j \neq i)$ and selecting words with top N highest cosine. Then, k nearest (i.e., highest cosine) concrete words are selected from the set of N neighbors. Finally, the direct visual vectors $\vec{v}(w_j)$ of k nearest neighbors are averaged and the resulting vector is used as $\vec{g}(w_i)$ for the abstract word w_i .

3.4. Training and prediction

To train the mapping (i.e., prediction function) from bimodal (or unimodal) word vectors to Binder et al.'s (2016) conceptual representation, we used a feed-forward neural network shown in Figure 2. The activation function and the number of units are shown in each of the layers denoted by solid rectangles. The training was performed by minimizing the mean squared error (MSE), and gradient descent with Adam was used as an optimization method. The learning rate for Adam was

fixed at 0.001. The weights (and biases) were initialized by the normalized initialization heuristic (Glorot and Bengio, 2010).

As an overall framework for evaluation (i.e., the procedure for training and prediction), we used a "leave-one-cluster-out" cross-validation procedure (Utsumi, 2020). This procedure is a variant of n -fold cross-validation in which semantic clusters for all words are used instead of randomly and equally partitioned groups. The reason for using leave-one-cluster-out cross-validation instead of n -fold cross-validation (and other methods with random sampling) is that words in Binder et al.'s (2016) dataset are not equally distributed in the semantic space. Some groups of words are semantically rich and they are very close to one another, whereas some other groups of words have only a small number of semantically less similar words. If we apply n -fold cross-validation to this dataset, semantically rich words with many close neighbors are likely to be better predicted independent of the representation ability of distributional semantic models, because their neighbors have more chance of being included in the training set.

To obtain word clusters for this procedure, we used Utsumi's (2020) method in which all 533 words were classified into 20 clusters using the k-means algorithm. Given textual vectors, we repeated k-means clustering 100 times and selected the best clustering result according to the Dunn index, which is a metric for evaluating clustering quality. Furthermore, to ensure the generality of the experimental results, we repeated this clustering procedure 10 times, and as a result, 10 different sets of 20 clusters were generated.

In the leave-one-cluster-out cross-validation procedure, for each cluster, the neural network in Figure 2 (i.e., prediction function) is trained using all words in the other clusters, and the conceptual vectors of words in the target cluster were predicted using the trained neural network. By repeating this procedure using each word cluster as a target, we obtained estimated conceptual vectors $\hat{y}(w_i)$ for all 533 words. Prediction performance was measured by Pearson's correlation between the estimated vector $\hat{y}(w_i)$ and the original vector $y(w_i)$. Spearman's rank correlation ρ and MSE were also used as secondary measures. For each set of 20 clusters, this experimental run was carried out three times under the same condition (i.e., hyperparameters), and the result of the run with the highest mean correlation across all words was retained. Finally, the results obtained using 10 sets of clusters were averaged and used for the analysis reported in section 4.

Hyperparameters other than the concreteness threshold were determined using a grid search. First, we determined the number of epochs for training from hybrid, dual coding, visual, and textual models. Using the leave-one-cluster-out cross-validation, we computed MSE across all words with the number of epochs ranging from 1 to 50. The lowest MSE was obtained at 19 epochs for the hybrid model, 17 epochs for the dual coding and textual models, and 10 epochs for the visual model. For the indirect grounding model, two parameters N and k for

⁵ We did not use $\theta_c = 1.1$ because the minimum concreteness rating in Binder et al.'s (2016) dataset was 1.19 (for the word "belief") and no words were judged as abstract when $\theta_c = 1.1$.

computing semantic neighbors were optimized together with the number of epochs using grid search of $N = 100, 200, 300$ and $k = 1, \dots, 10$. In this grid search, we used the concreteness threshold $\theta_c = 4.0$. Mean squared error was computed over all words and the lowest MSE was obtained at $N = 300, k = 10$, and 20 epochs. For the indirect visual model, we determined the number of epochs using indirect visual vectors at $N = 300$ and $k = 10$, and as a result, 16 epochs achieved the lowest MSE. These hyperparameters were used in all experimental runs for evaluation.

4. Results

4.1. Performance difference among models

Table 5 lists mean correlations between the original conceptual vector and the vectors estimated by the indirect grounding model and other models. Figure 3 shows the variation in correlation coefficients over abstract, concrete, and all words. Note that in this section we report the results of Pearson's correlation when used as performance measure, but we also analyzed the performance with two additional measures (i.e., Spearman's rank correlation and MSE). Because these results do not significantly differ from those of Pearson's correlation, the detailed results of the additional analysis are provided in Appendix S1 of the Supplementary material.

For abstract words, the indirect grounding model DSM_I achieved the highest mean correlation in both concreteness thresholds. The Friedman test conducted on abstract words revealed a significant difference between word correlations of six models, $\chi^2(5, N = 116) = 280.20, p < 0.001$ for $\theta_c = 3.0$ and $\chi^2(5, N = 214) = 583.23, p < 0.001$ for $\theta_c = 4.0$. Multiple pairwise comparisons using the Wilcoxon signed-rank test with Ryan's procedure ($p < 0.05$) showed that the correlation of the indirect grounding model was significantly higher than those of all other models at $\theta_c = 3.0$ and than those of other four models except the dual coding model at $\theta_c = 4.0$. For other pairwise differences, only the difference between the dual coding and hybrid models was not significant for either thresholds. This result is consistent with the prediction of the indirect grounding view in Table 2. Additionally, the result that the indirect visual model DSM_G predicted the conceptual representation better than the simple visual model DSM_V also indicates the effectiveness of computing visual vectors using semantic neighbors. These results clearly support the indirect grounding view of abstract concepts.

Although it is not the main concern of this paper whose focus lies in abstract concepts, the Friedman test conducted on correlations of concrete words also indicated a significant difference among six models, $\chi^2(5, N = 417) = 1181.66, p < 0.001$ for $\theta_c = 3.0$ and $\chi^2(5, N = 319) = 895.97, p < 0.001$

for $\theta_c = 4.0$. The highest mean correlation was achieved by the hybrid model for both concreteness thresholds, but multiple pairwise comparison revealed that pairwise differences among the hybrid, indirect grounding, and textual models were not significant. For $\theta_c = 4.0$, the difference between the dual coding and textual models and between the dual coding and indirect grounding models also did not reach the significance level. All the other pairwise differences were significant. The absence of significant difference among the indirect grounding, hybrid, and dual coding models is a predictable result, as shown in Table 2. What is somewhat surprising is that bimodal models for abstract concepts (i.e., the indirect grounding and hybrid models) did not achieve significantly higher performance than the textual (i.e., unimodal) model, given that a number of studies on multimodal distributional semantics have shown the superiority over text-based unimodal models for concrete words (e.g., Bruni et al., 2014; Baroni, 2016). This result is not consistent with the prediction of Table 2 made by the indirect grounding, hybrid, and dual coding views.

One possible reason would be that the textual layer may cover most of the information needed to predict the target conceptual representation of concrete words; the textual layer compresses 300-dimensional input textual vectors to 50% (= 150/300) of their original dimension, but the visual layer compresses 2,048-dimensional input visual vectors to a much lower percent, 7.3% (= 150/2,048). To test this possibility, we conducted an additional experiment with the same experimental procedure by decreasing the dimension d_T of the textual (hidden) layer. The detailed result of this additional experiment is provided in Appendix S2 of the Supplementary material. The result is supportive of this possibility; when the dimension d_T was 30 (whose compression rate 10.0% is nearly equal to that of the visual layer) or lower, the bimodal models (i.e., indirect grounding and hybrid models) achieved significantly higher correlations than the unimodal textual model. Note also that decreasing the dimension d_T of the textual layer did not affect the result of abstract words; the same result of performance differences were obtained regardless of d_T . From these results, it follows that the visual layer actually contributes to model performance in an expected way and our bimodal distributional semantic model achieves a result fully consistent with the prediction of the indirect grounding view when the impact of the textual and visual layers is equalized.

In addition, the difference among six models was also significant for all words, $\chi^2(5, N = 533) = 1438.50, p < 0.001$ for $\theta_c = 3.0$ and $\chi^2(5, N = 533) = 1444.01, p < 0.001$ for $\theta_c = 4.0$. The difference between the indirect grounding and hybrid models at $\theta_c = 3.0$ and the differences among the indirect grounding, hybrid, and dual coding models at $\theta_c = 4.0$ were not significant, but all the other pairwise comparisons were significant.

Summarizing, the obtained results are most consistent with the predictions of the indirect grounding view shown in

TABLE 5 Mean correlations for the indirect grounding model and other models.

Model	$\theta_c = 3.0$			$\theta_c = 4.0$		
	Abstract	Concrete	All	Abstract	Concrete	All
Bimodal						
Indirect grounding (DSM_I)	0.772	0.742	0.749	0.731	0.761	0.749
Hybrid (DSM_H)	0.764	0.744	0.748	0.724	0.764	0.748
Dual coding (DSM_D)	0.762	0.734	0.740	0.729	0.756	0.745
Unimodal						
Visual (DSM_V)	0.536	0.475	0.488	0.480	0.494	0.488
Textual (DSM_T)	0.755	0.740	0.744	0.716	0.762	0.744
Indirect visual (DSM_G)	0.626	0.490	0.520	0.529	0.513	0.519

Boldfaced numbers indicate the highest correlations (i.e., the best performance) among the models.

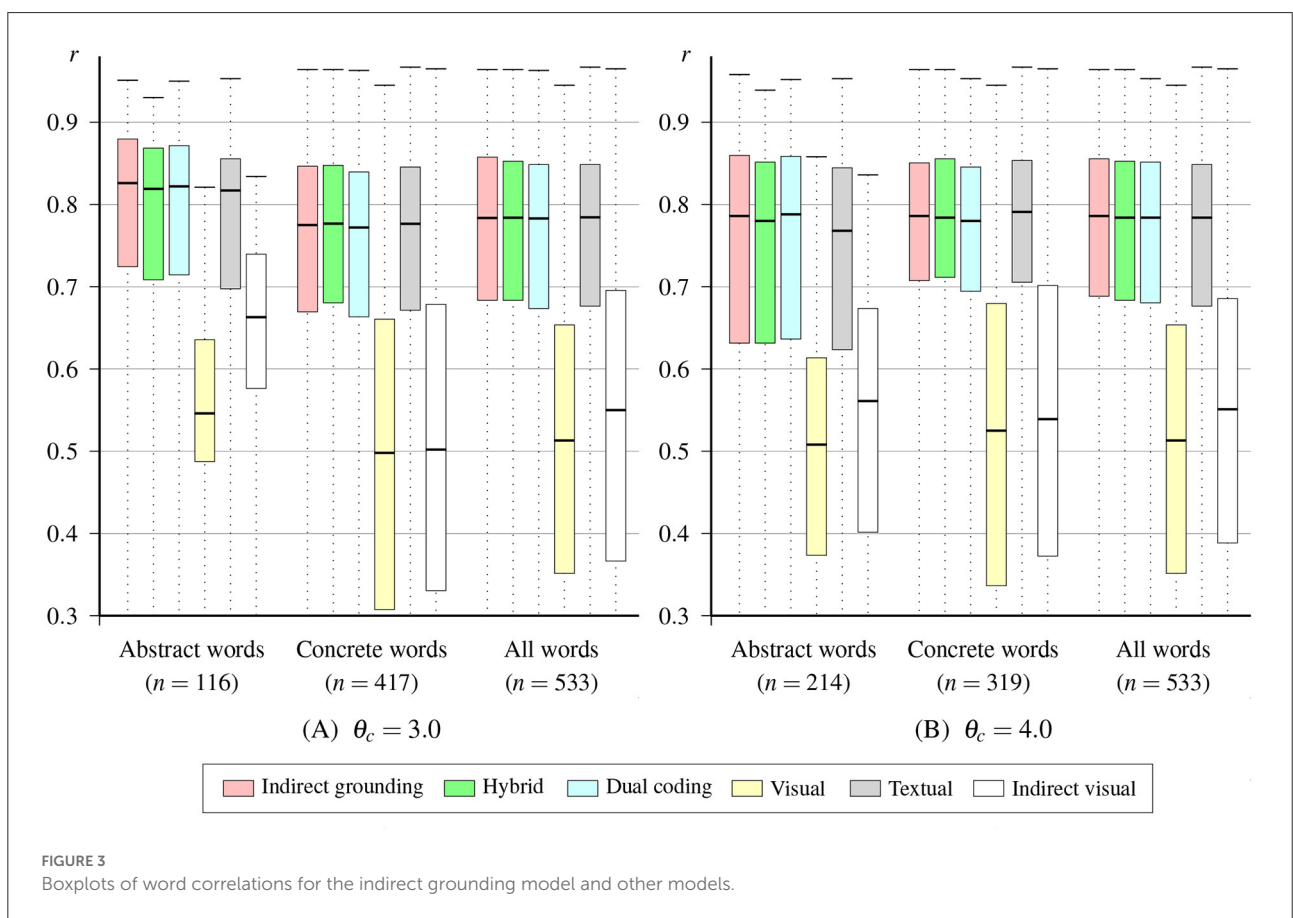


FIGURE 3 Boxplots of word correlations for the indirect grounding model and other models.

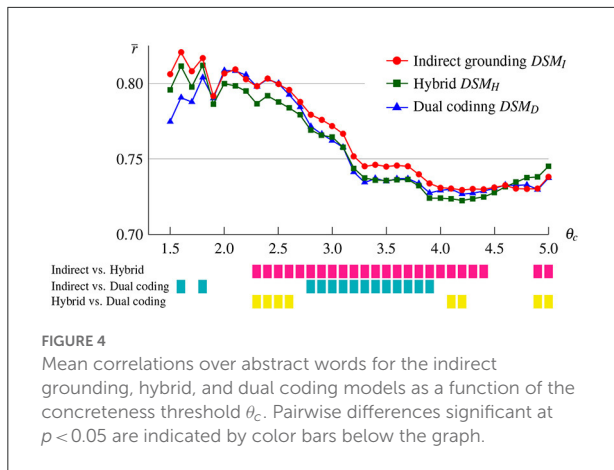
Table 2. It is therefore concluded that the indirect grounding view is plausible as a conceptual representation theory of abstract concepts.

4.2. Effect of the concreteness threshold

To test whether the superiority of the indirect grounding model for abstract concepts reported in the last section holds

for other concreteness thresholds, we conducted the same experiment at different thresholds θ_c ranging from 1.5 to 5.0 with a step size of 0.1. Figure 4 shows mean correlations over abstract words for the indirect grounding, hybrid, and dual coding models. Color bars shown below the graph denote whether pairwise differences were statistically significant by multiple pairwise comparison ($p < 0.05$).

The indirect grounding model yielded a higher correlation than other two models when the concreteness threshold was



between 1.5 and 4.5, although at some lower threshold the correlation of the indirect grounding model was slightly lower than that of the dual coding model. Specifically, the correlation of the indirect grounding model was significantly higher than that of the hybrid model between $\theta_c = 2.3$ and 4.4, and than that of the dual coding model between $\theta_c = 2.8$ and 3.9. This confirms the finding reported in section 4.1 and indicates that the superiority of the indirect grounding model was not accidentally observed in some concreteness thresholds. At $\theta_c = 4.6$ or higher, the mean correlation of the indirect grounding model was lower than the hybrid model. In these cases, highly concrete words were selected as abstract and their visual vectors were computed via semantic neighbors, even though it is appropriate that they are grounded directly through their own visual images. This “less plausible” grounding may generate a harmful effect on the prediction performance of the indirect grounding model. This behavior of the model is also consistent with the indirect grounding view.

4.3. Relation between word concreteness and improvement by indirect grounding

In this section, we examine whether the performance improvement of the indirect grounding model compared to the hybrid and dual coding models depends on word concreteness. To quantify the degree of improvement, we considered the difference of correlation computed by subtracting the correlation coefficient of a baseline model from the correlation coefficient of the indirect grounding model.

We computed a correlation between the difference of correlation and word concreteness. The difference of correlation was not correlated with word concreteness when the hybrid model is a baseline, $r = -0.104$ ($\theta_c = 3.0$) and $r = 0.075$ ($\theta_c = 4.0$). In the case of the dual coding model used as a baseline, the difference of correlation was not correlated with

word concreteness for $\theta_c = 4.0$, $r = 0.071$, but they were weakly correlated for the threshold $\theta_c = 3.0$, $r = 0.222$ ($p < 0.05$). These results indicate that there was generally no monotonic relationship between the degree of improvement and word concreteness, but appending indirect visual vectors to textual vectors may be somewhat more effective for less abstract concepts.

To examine more closely the relation between performance improvement and word concreteness, we also computed the mean difference of correlation per each of the intervals into which the entire range of concreteness values was equally divided, whose results are shown in Figure 5. Overall, the indirect grounding model improved the prediction performance regardless of word concreteness, but two concreteness thresholds showed different patterns of improvement. For $\theta_c = 3.0$, the indirect grounding model improved the performance for highly abstract words (i.e., words with concreteness rating is less than 1.75) against both competing models. This result suggests that the indirect grounding model is effective in representing purely abstract words. Furthermore, only when comparing with the dual coding model, the degree of improvement was higher for less abstract words (i.e., the ones within the range of $2.50 \leq$ concreteness rating < 3.00) than for more abstract words. This may suggest that these words, which may be difficult to judge concreteness or have both concrete and abstract senses, benefit from adding visual information whether directly or indirectly.

For $\theta_c = 4.0$, however, the performance of highly abstract words (whose concreteness rating was less than 2.0) were not improved and the difference of correlation peaked at higher concreteness range [3.50, 3.75]. These different results for highly abstract words between both concreteness thresholds can be attributed primarily to the difference of semantic neighbors, because θ_c determines a word pool (i.e., a set of concrete words) from which semantic neighbors are chosen. Hence the effectiveness of indirect grounding depends on not only the concreteness of a word to be grounded, but also the choice of semantic neighbors for that word.

4.4. Word-level analysis on the impact of indirect grounding

Abstract (and concrete) concepts have been considered as a unitary whole, but recent research has argued that abstract concepts should be treated as a heterogeneous category including various different types of abstract concepts (Ghio et al., 2016; Troche et al., 2017; Borghi et al., 2018; Villani et al., 2019). Therefore, to examine the types of abstract concepts for which indirect grounding, in particular, visually indirect grounding, is effective, we analyzed the degree of improvement in terms of semantic categories of abstract words.

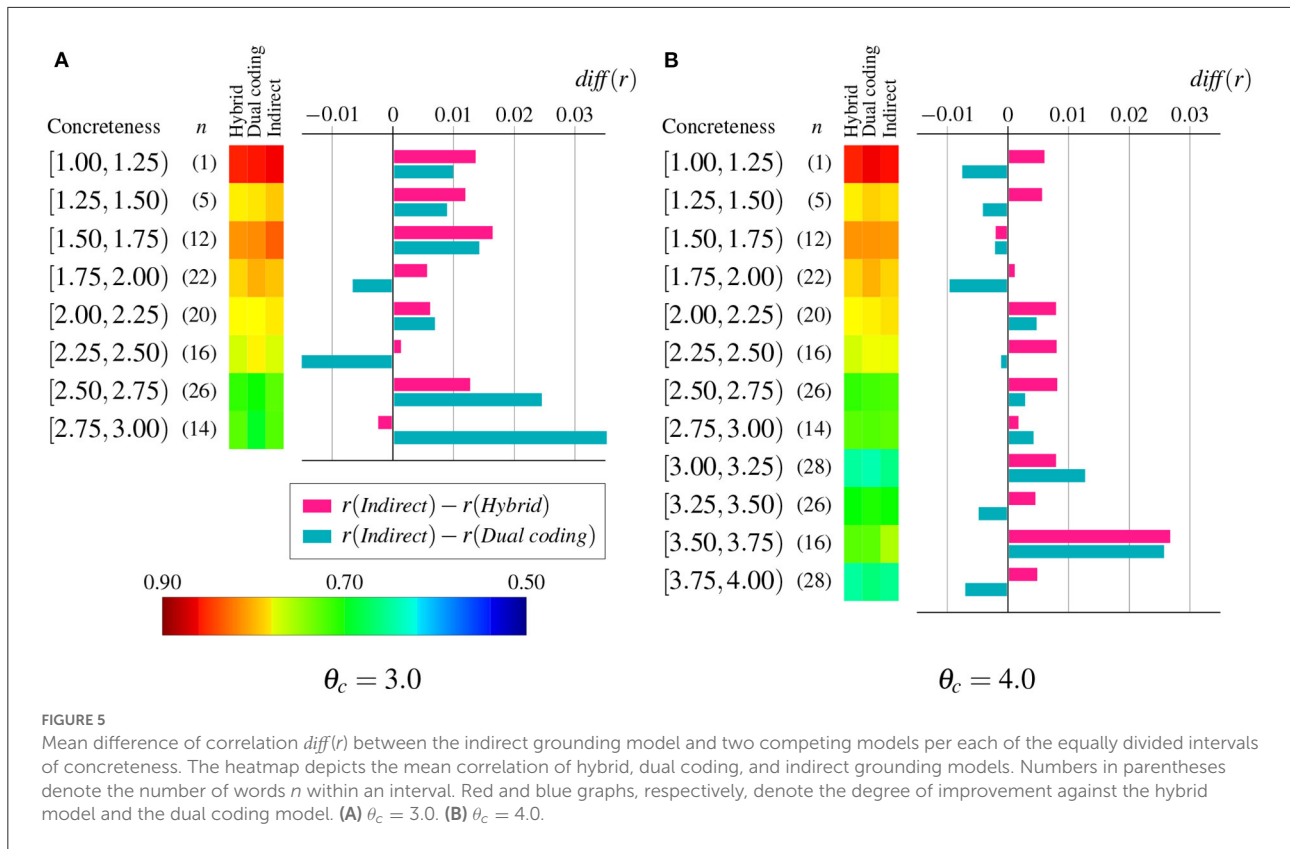
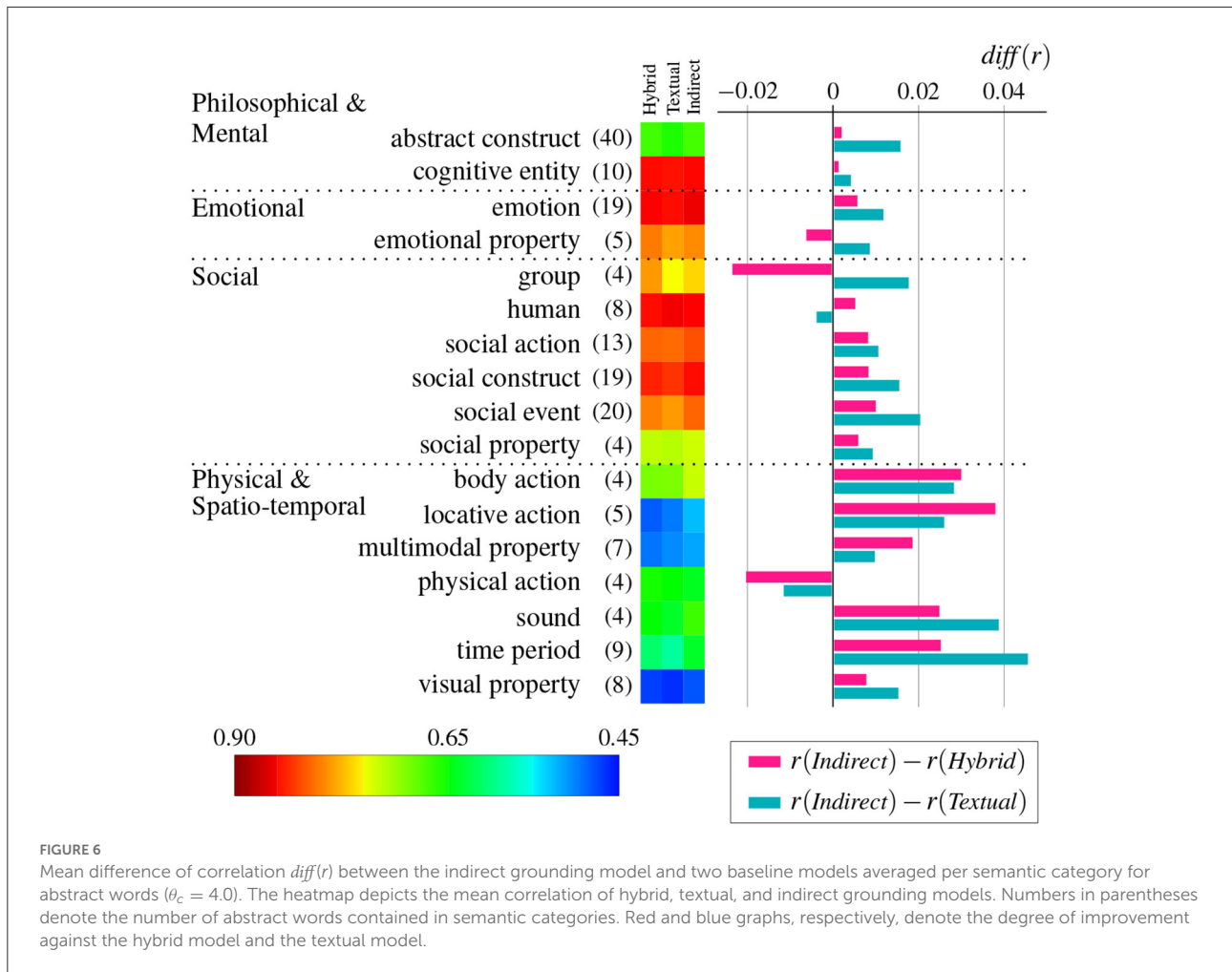


Figure 6 shows how the degree of improvement is related to semantic categories for abstract words with concreteness rating less than 4.0 (i.e., $\theta_c = 4.0$). We used semantic categories provided by Binder et al. (2016) to classify abstract words. They classified all 535 words into 47 categories that reflect semantic and grammatical classes. Among them, we selected 17 categories for analysis that included four or more abstract words. In addition, these 17 categories were grouped according to the following four clusters revealed by Villani et al. (2019): physical, spatio-temporal, and quantitative (*Physical and Spatio-temporal*) concepts, self and sociality (*Social*) concepts, philosophical/spiritual (*Philosophical and Mental*) concepts, and emotional/inner states (*Emotional*) concepts.

As shown in Figure 6, *Social* and *Physical and Spatio-temporal* clusters are likely to show higher improvement than *Philosophical and Mental* and *Emotional* clusters. In particular, *body action*, *locative action*, *sound*, and *time period* categories in *Physical and Spatio-temporal* cluster achieved relatively higher improvement by the indirect grounding model. The high improvement of physical and spatio-temporal concepts is consistent with the indirect grounding view, because these abstract concepts are likely to be associated (via language) with specific visual images. For example, the concept *evening* easily evokes visual experiences associated with the concepts of *dinner* and *sunset*. In the experiment of this paper, “night,” “dinner,”

“supper,” “dusk,” “dawn,” “sundown,” “twilight,” “candlelight,” “sunset,” and “sunrise” were selected as semantic neighbors of the word “evening,” and the indirect visual vectors computed from these images improved the baselines as shown in Figure 7. The relatively high improvement of social categories can be explained along the same line. A number of social concepts are associated with perceptually grounded concepts. For example, the verb *play* can be captured by relevant concepts such as *game*, *soccer* and *football* (objects to be played) and *sandlot* (place to play), which were selected as semantic neighbors. Meanwhile, some other social concepts (e.g., *business* and *joke*) are more complex and difficult to capture by grounded concrete concepts.

By contrast, emotional categories were not improved by the indirect grounding model. Emotional information is highly likely to be encoded in textual vectors (e.g., Recchia and Louwerse, 2015; Utsumi, 2020), and thus indirect grounding may not be necessary for emotional concepts. A more plausible explanation would be that emotional concepts are directly grounded in emotional experiences, and thus relatively less dependent on indirect grounding in perceptual (i.e., visual) experiences. The lesser degree of improvement for the cluster of *Philosophical and Mental* is not surprising. These abstract concepts are generally thought of as “highly disembodied” concepts (Dove, 2016), which are divorced from experiential (at



least visual) grounding regardless of whether the grounding is direct or indirect.

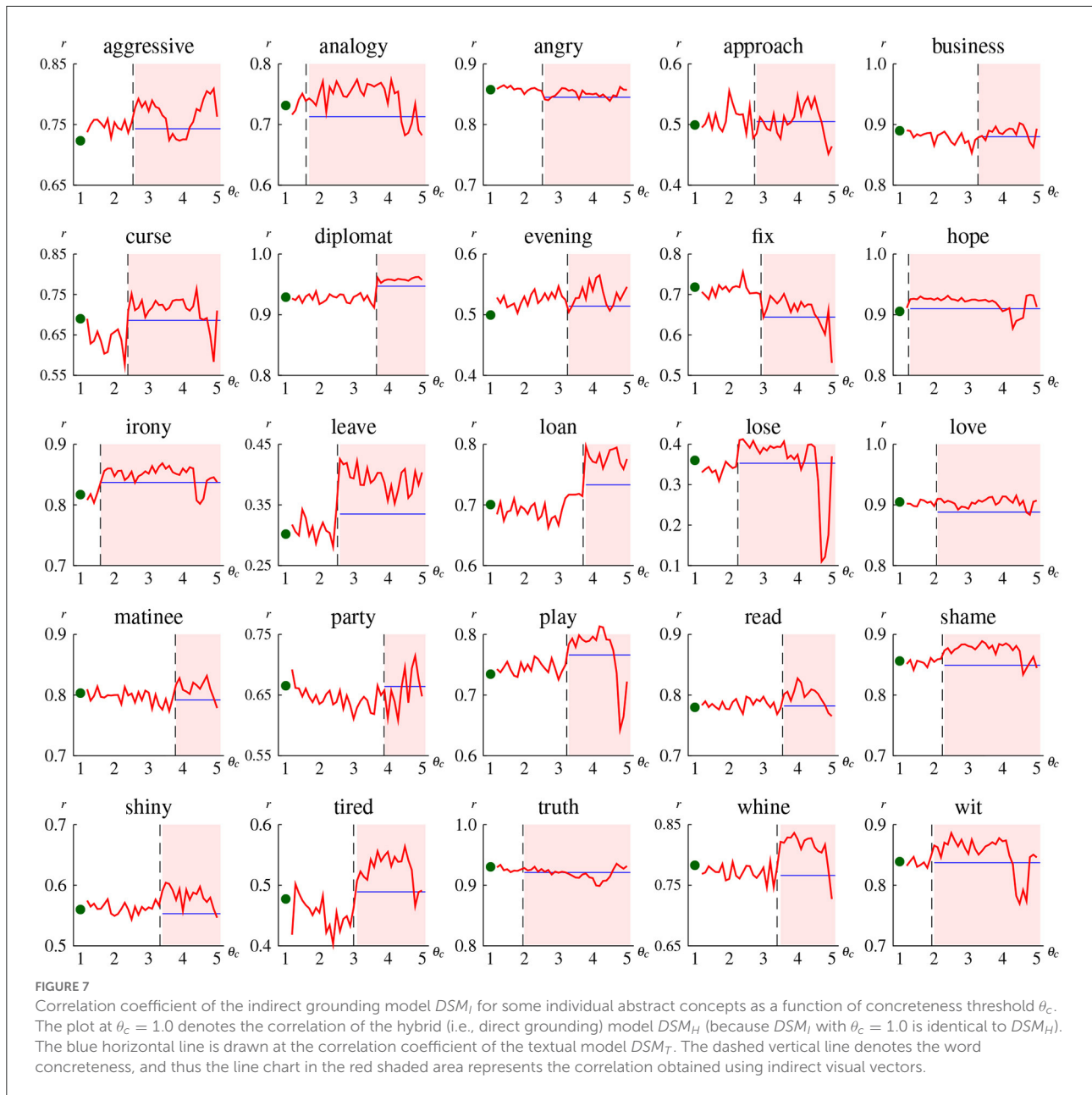
Figure 7 shows the change of correlation coefficient in terms of concreteness threshold for some abstract words. The result of all abstract words with concreteness rating less than 4.0 is provided in Appendix S3 of the Supplementary material. The important point to note is the difference of correlation between non-shaded and shaded areas. Correlation plotted in the red shaded area (i.e., the region right of the dashed vertical line) shows the prediction performance of the indirect grounding model when a word is regarded as abstract and thus indirect visual vectors are used, whereas correlation in the non-shaded area shows the performance when that word is regarded as concrete and direct visual vectors are used. Hence, the indirect grounding model is found to be effective for words whose correlations in the shaded area are higher than those in the non-shaded area. A typical pattern in this effective case is that correlations in the shaded area are higher than those in the non-shaded area as well as higher than those of textual and hybrid models (e.g., “evening,” “leave,” “play”). This pattern is

also marked by the decrease of correlation at higher thresholds (i.e., $\theta_c \geq 4.5$), because good mediator words for an abstract word are erroneously judged as abstract and thus no longer selected as semantic neighbors at higher thresholds. By contrast, some words (e.g., “angry,” “business,” “truth”) show a different pattern that correlation does not largely differ between the shaded and non-shaded area, which indicates no performance improvement by the indirect grounding model. For some other words (e.g., “fix”), correlation decreases in the shaded area; this pattern indicates that the indirect grounding model is harmful to predicting conceptual representation.

5. Discussion

5.1. Contribution to the research on abstract concepts

The present study makes an original contribution to research on abstract concepts. As mentioned in section 1, very



few empirical studies have demonstrated direct evidence in favor of the indirect grounding view, although a number of studies have empirically shown that both symbolic/linguistic and perceptual/embodied representations are required for shaping and processing abstract concepts. Given the current lack of direct evidence, the present study provides empirical support specific to the role of language posited by the indirect grounding view. The higher prediction performance of the indirect grounding model compared to the hybrid (i.e., direct grounding) model suggests that a mere combination of symbolic and perceptual representation (e.g., Barsalou et al., 2008) is less adequate for explaining abstract concepts (or at least those that

can be grounded in visual experiences); abstract concepts are more likely to be indirectly grounded through their linguistic relations to the concepts directly grounded in the world. In other words, language functions as a bridge between abstract concepts and perceptual experiences.

The plausibility of indirect grounding is also supported by the result of comparing two visual models. Even when textual vectors were not used for prediction, the indirect visual model outperformed the simple visual model, as reported in section 4.1. This implies that the use of visual vectors derived from concrete words strongly associated with an abstract word stands on its own merit.

Furthermore, the superiority of the indirect grounding model over two unimodal models (i.e., textual and visual models) and the dual coding model suggests that abstract concepts are both linguistic and grounded in perceptual experience. This result is consistent with the recent empirical findings (Louwerse and Jeuniaux, 2010; Malhi and Buchanan, 2018) and thus lends further support to the hybrid views of abstract concepts (Louwerse, 2011; Dove, 2016; Borghi et al., 2019).

5.2. Related work on computational approaches to indirect grounding

Some existing studies have proposed a computational model for semantic processing that is based on similar views to indirect grounding. Howell et al. (2005) attempted to simulate the mechanism of propagation of grounding (mentioned in section 1.3) using a simple recurrent network. The network was trained to predict, from the current input word, both what the next word would be and the featural (i.e., sensorimotor) representation of the current word. They demonstrated that the network trained with featural representation achieved better performance in next word prediction than the network trained without featural representation, and argued that this result supports the propagation of grounding. However, their model does not directly simulate the process in which the sensorimotor features of concrete concepts are propagated to abstract concepts. Hoffman et al. (2018) proposed a more sophisticated neural network model that assimilates a very similar idea to indirect grounding; “Knowledge of abstract words is acquired through (a) their patterns of co-occurrence with other words and (b) acquired embodiment, whereby they become indirectly associated with the perceptual features of co-occurring concrete words (Hoffman et al., 2018, p. 293).” Their model is based on a hub-and-spoke architecture in which the information of an input word, its sensorimotor properties and past states (as context) is integrated into a hidden “hub” layer. They trained the model to predict the next word in a word sequence and showed that the trained model could represent the semantic knowledge of concrete and abstract words in a hub layer and accounted for behavioral patterns consistent with normal and impaired semantic cognition. However, they did not quantitatively test whether indirect grounding is more plausible than other competing views, such as one that abstract concepts are also grounded directly in sensorimotor experience.

These previous studies differ from the present study in some important respects. First of all, they use, as perceptual or sensorimotor representation, only verbally expressed featural information, which is essentially symbolic and discrete. Second, they do not directly test the plausibility of indirect grounding for representing abstract concepts; their proposed models are

not compared with other competing models to be considered. Furthermore, their studies are limited in their coverage of the vocabulary of words and features; only a relatively small set of words and features are used in the experiments. The training corpus is also small in size and generated artificially. By contrast, the present study directly uses non-verbal (i.e., visual) information as perceptual representation and quantitatively tests the indirect grounding view by comparing other competing models including the direct grounding model. The vocabulary and corpus used in the experiment of the present study are relatively large.

To the best of our knowledge, no prior studies on multimodal distributional semantics or other computational models using non-verbal information have tested indirect grounding of abstract concepts (or words), but a noteworthy observation was reported. Lazaridou et al. (2015) proposed a multimodal skip-gram model by extending the objective function of the original skip-gram (Mikolov et al., 2013) so as to take into account visual similarity computed using visual vectors. Using the trained multimodal word vectors, they showed that some abstract words had nearest neighbors in the trained multimodal space whose visual images depict relevant concrete situations (e.g., the nearest neighbor picture of the word “theory” depicts a bookshelf with many books), although the nearest neighbors of many other abstract words were not visually relevant. The extended skip-gram model does not directly simulate the process of indirect grounding, but this result suggests that the multimodal skip-gram may be a useful model for exploring the grounding mechanism of abstract concepts.

5.3. Limitation of this study and future direction

Our distributional semantics-based approach to embodied cognition of abstract concepts has its limitations. One important limitation is that the method for modeling perceptual experiences used in this paper does not deal with perceptual information other than visual one. Although people are supposed to acquire a large percentage of information from visual perception, conceptual knowledge is also grounded in other types of perceptual experiences, such as auditory, somatosensory, gustatory, and olfactory ones, as well as from emotional and social experiences (e.g., Borghi et al., 2017). Our finding in favor of indirect grounding is thus confined to visual grounding; the detailed analysis based on semantic categories reported in section 4.4 revealed that abstract concepts in only some categories (i.e., *Social* and *Physical and Spatio-temporal* categories) benefit from visually indirect grounding. Future research is required to investigate whether and how other types of abstract concepts are grounded directly or indirectly.

A more noteworthy limitation is that the DNN by which visual vectors are computed may diverge from human visual perception. The progress of deep learning technique has demonstrated that DNNs surpass human-level performance on some specific image classification tasks (Zhang et al., 2021) and it has been shown that their internal representations match coarsely with the brain (Cichy et al., 2016; Serre, 2019). By contrast, recent studies have also revealed that DNNs show behavioral deviations from human visual perception, for example, in terms of the sensitivity to global shape (Baker et al., 2018) and the visual representational structure in the human brain (Xu and Vaziri-Pashkam, 2021). Jacob et al. (2021) demonstrated that some phenomena (e.g., surface invariance, sensitivity to 3D shape) seen in human visual perception were not observed in ResNet-152, which is used for extracting visual vectors in this study, as well as in other DNNs. These deviations suggest that DNNs and its visual vectors are limited as a cognitive model of visual grounding. Given these potential limitations, we must be cautious about interpreting the obtained results, in particular the lower performance of the visual model DSM_V as evidence against the situated simulation view.

Our choice of text-based distributional semantic model (i.e., SGNS) by which textual vectors are computed is unlikely to greatly affect the obtained results, given the current technical possibilities. Utsumi (2020) compared three distributional semantic models, that is, SGNS, GloVe (Pennington et al., 2014), and PPMI+SVD (Bullinaria and Levy, 2007), in terms of performance in predicting Binder et al.'s (2016) conceptual representation only from textual vectors. He demonstrated that SGNS achieved the highest prediction performance, and more importantly, the relative performance differences among words and attributes were quite similar among three models. Chersoni et al. (2021) extended this result by comparing a wider variety of textual models using the same prediction task. These models include BERT (Devlin et al., 2018), which is a deep neural model for contextual embeddings that achieves state-of-the-art performance in many NLP tasks. They reported that BERT did not show any significant differences from the distributional semantic models analyzed by Utsumi (2020), and word and attribute correlation of BERT is equivalent to that of SGNS. These results imply that the obtained findings in this paper have certain generality with respect to textual models. Obviously, however, it does not mean that current language models sufficiently capture semantic (or conceptual) knowledge people can acquire from language (for a review, see Rogers et al., 2020; Lake and Murphy, 2021). If a psychologically more plausible language model is developed in the future, it would be interesting to explore whether the indirect grounding model still yields the same result.

The present study also suffers from methodological limitations. Binder et al.'s (2016) dataset used as a target conceptual representation for the evaluation experiment is the most comprehensive and fine-grained featural representation publicly available at present, but yet not sufficient to capture

the richness of human conceptual knowledge. For example, concepts (particularly abstract concepts) generally involve the knowledge of binary and multiary relations among concepts and higher-order relations that cannot be expressed by feature-based representation. Hence the obtained result does not reflect the representational ability of relational knowledge. Furthermore, the training procedure for predicting Binder et al.'s (2016) representation poses an additional concern about whether it can precisely capture processing differences among competing models. This concern is particularly salient for the dual coding model. According to its definition in section 3.1, the dual coding model should approximate the performance of the textual model for abstract words and that of the hybrid model for concrete words. However, the result of section 3.1 (Appendix S2) diverged from these expectations; the dual coding model outperformed the textual model in predicting abstract words and showed lower performance for concrete words than the hybrid model. This discrepancy between expectations and performance may be caused by simultaneous training of concrete and abstract words. Specifically, network parameters (i.e., weights and biases) between the output layer (i.e., the bottom layer of Figure 2) and the hidden layer just above reflect the visual information of concrete words, even when visual vectors are not given. Because of this, prediction of an abstract word may indirectly reflect visual information of concrete concepts whose textual vectors are similar to that of the abstract word, and thus the dual coding model would perform better than the textual model. For concrete words, the dual coding model does not benefit from visual information for abstract words and this may cause lower performance of the dual coding model than the hybrid model. Although this discrepancy does not affect the findings on the superiority of the indirect grounding model, more valid training procedures should be pursued.

The use of concreteness rating may be controversial because of its limitations. Most concreteness ratings including Brysbaert et al.'s (2014) are collected by presenting words in isolation, thereby including an ambiguity in judgment for polysemous words (e.g., Reijnierse et al., 2019). This is particularly problematic for words with both concrete and abstract senses, which tend to be rated around the middle of the scale. In other words, words with low concreteness ratings are likely to be unambiguous and regarded as definitely abstract. The analysis reported in section 4.3 showed that the indirect grounding model improved the performance for words with low concreteness rating (when appropriate semantic neighbors were given), thus suggesting that our finding in favor of the indirect grounding view does not severely affected by this problem of concreteness rating.

Despite the positive result, there is another problem with the use of concreteness rating that must be addressed in future work. Recent studies on abstract concepts have argued that abstract concepts are not a unitary whole and should be treated as a heterogeneous category including various different types of abstract concepts (Troche et al., 2017; Borghi et al., 2018;

Villani et al., 2019). This argument implies that concreteness rating is not sufficient for determining words (or concepts) to be grounded indirectly. As reported in section 4.4, the impact of indirect grounding differs among various types of abstract concepts. Exploring this issue may provide an interesting avenue for future investigation.

The choice of semantic neighbors as mediator concepts, whose visual vectors define the indirectly grounded representation of abstract words, is an important process of the indirect grounding model. Appropriate mediator concepts need to have perceivable referents that also have perceptually clear-cut boundaries. For example, basic-level concepts such as *desk* and *chair* have specific referents that are characterized by perceptual features such as shapes, and thus they become good mediators. Superordinate concepts such as *furniture* also have perceivable referents but are difficult to distinguish from other concepts by perceptual features, and thus they are less likely to be good mediators. Highly underspecified concepts such as *artifact* no longer function as mediators because they are very generic and their defining features are not based on perceptual or other bodily experiences. However, our simple method of generating a pool of candidate words from which semantic neighbors are chosen has a potential problem in that words with higher concreteness rating do not necessarily have such specific referents. For example, the word “furniture” has a higher concreteness rating of 4.89 than “desk” (4.87) and “chair” (4.58), and the word “artifact” also has a very high rating of 4.50 in Brysbaert et al.’s (2014) dataset. Recently Bolognesi et al. (2020) empirically examined the same line of argument and demonstrated that categorical specificity should be considered as a distinct dimension from concreteness to characterize concepts. Categorical specificity is therefore an important property for choosing appropriate mediator words, although an appropriate level of specificity depends on the concept to be indirectly grounded.

Some other methods for determining mediator words can be considered to refine the indirect grounding model. Age-of-acquisition ratings (Kuperman et al., 2012) may be used to limit the vocabulary of candidate words for semantic neighbors, because basic words learned at the early stage of lexical acquisition are represented primarily perceptually, while other words learned at the later stage are acquired through the knowledge of basic words (Gleitman et al., 2005; Thill et al., 2014). A more promising approach is to use minimal grounding sets (Vincent-Lamarre et al., 2016) as a candidate set of semantic neighbors. A minimal grounding set is the smallest set of words (i.e., a subset of a vocabulary) from which all the other words in a vocabulary can be defined. Vincent-Lamarre et al. (2016) proposed a method for computing the minimal grounding sets from dictionary definitions. Although the minimal grounding set is not uniquely determined, it can be a theoretically more motivated pool of potential mediator words than a set of concrete words simply selected based on concreteness rating or other human-rating-based measures. Image tags may also

be a useful source of information for mediator words because tag co-occurrence statistics reflect visually motivated semantic knowledge. For example, words that are similar to a target abstract word in Flickr distributional tagspace (Bolognesi, 2017) are expected to be good mediators.

6. Conclusion

To test the indirect grounding view, we devised a new multimodal distributional semantic model in which a visual vector of an abstract word (i.e., embodied representation of perceptual experiences) is computed from the visual images of concrete words semantically related to the abstract word. Through the evaluation experiment, we have demonstrated that the indirect grounding model outperformed the hybrid (i.e., direct grounding) model, the dual coding model, and unimodal models. Despite the limitations described above, this finding lends some plausibility to the indirect grounding view and the present study is regarded as a first step toward empirically exploring the grounding mechanism of abstract concepts. For future work, we would like to explore a further mechanism of what and how linguistic processes come into play for grounding abstract concepts, together with to test indirect grounding via language using psychological experiments.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

AU conceptualized the study, performed the evaluation experiment and statistical analysis, and wrote the draft of the manuscript.

Funding

This research was supported by JSPS KAKENHI Grant Numbers JP15H02713 and JP20H04488.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those

of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychol. Rev.* 116, 463–498. doi: 10.1037/a0016261
- Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* 14:e1006613. doi: 10.1371/journal.pcbi.1006613
- Baroni, M. (2016). Grounding distributional semantics in the visual world. *Linguist. Issues Lang. Technol.* 10, 3–13. doi: 10.1111/lnc3.12170
- Barsalou, L. W. (1999). Perceptual symbol systems. *Behav. Brain Sci.* 22, 577–660. doi: 10.1017/S0140525X99002149
- Barsalou, L. W. (2003). Situated simulation in the human conceptual system. *Lang. Cogn. Process.* 18, 513–562. doi: 10.1080/01690960344000026
- Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.* 59, 617–645. doi: 10.1146/annurev.psych.59.103006.093639
- Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychon. Bull. Rev.* 23, 1122–1142. doi: 10.3758/s13423-016-1028-3
- Barsalou, L. W., Dutriaux, L., and Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philos. Trans. Roy. Soc. B* 373:20170144. doi: 10.1098/rstb.2017.0144
- Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). “Language and simulation in conceptual processing,” *Symbols and Embodiment: Debates on Meaning and Cognition*, eds M. de Vega, A. Glenberg, and A. Graesser (New York, NY: Oxford University Press), 245–283.
- Barsalou, L. W., and Wiemer-Hastings, K. (2005). “Situating abstract concepts,” in *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, eds D. Pecher and R. Zwaan (New York, NY: Cambridge University Press), 129–163.
- Bergelson, E., and Swingle, D. (2013). The acquisition of abstract words by young infants. *Cognition* 127, 391–397. doi: 10.1016/j.cognition.2013.02.011
- Binder, J. R., Conant, L. L., Humphries, C. J., Fernandez, L., Simons, S. B., Aguilar, M., et al. (2016). Toward a brain-based componential semantic representation. *Cogn. Neuropsychol.* 33, 130–174. doi: 10.1080/02643294.2016.1147426
- Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cereb. Cortex* 19, 2767–2796. doi: 10.1093/cercor/bhp055
- Bolognesi, M. (2016). Modeling semantic similarity between metaphor terms of visual vs. linguistic metaphors through Flickr tag distributions. *Front. Commun.* 1:9. doi: 10.3389/fcomm.2016.00009
- Bolognesi, M. (2017). “Flickr@distributional tagspace: evaluating the semantic spaces emerging from Flickr@tag distributions,” in *Big Data in Cognitive Science*, ed M. Jones (Routledge: Taylor and Francis Group), 144–173.
- Bolognesi, M., Burgers, C., and Caselli, T. (2020). On abstraction: decoupling conceptual concreteness and categorical specificity. *Cogn. Process.* 21, 365–381. doi: 10.1007/s10339-020-00965-9
- Bolognesi, M., and Steen, G. (2018). Abstract concepts: structure, processing, and modeling. *Top. Cogn. Sci.* 10, 490–500. doi: 10.1111/tops.12354
- Borghgi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., and Tummolini, L. (2019). Words as social tools: language, sociality and inner grounding in abstract concepts. *Phys. Life Rev.* 29, 120–153. doi: 10.1016/j.plrev.2018.12.001
- Borghgi, A. M., Barca, L., Binkofski, F., and Tummolini, L. (2018). Varieties of abstract concepts: development, use and representation in the brain. *Philos. Trans. Roy. Soc. B* 373:20170121. doi: 10.1098/rstb.2017.0121
- Borghgi, A. M., and Binkofski, F. (2014). *Words as Social Tools: An Embodied View on Abstract Concepts*. New York, NY: Springer.
- Borghgi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., and Tummolini, L. (2017). The challenge of abstract concepts. *Psychol. Bull.* 143, 263–292. doi: 10.1037/bul0000089
- Borghgi, A. M., Scorolli, C., Caligiore, D., Baldassarre, G., and Tummolini, L. (2013). The embodied mind extended: using words as social tools. *Front. Psychol.* 4:214. doi: 10.3389/fpsyg.2013.00214
- Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.* 49, 1–47. doi: 10.1613/jair.4135
- Brysbaert, M., Warriner, A. B., and Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* 46, 904–911. doi: 10.3758/s13428-013-0403-5
- Bullinaria, J. A., and Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: a computational study. *Behav. Res. Methods* 39, 510–526. doi: 10.3758/BF03193020
- Chersoni, E., Santus, E., Huang, C.-R., and Lenci, A. (2021). Decoding word embeddings with brain-based semantic features. *Comput. Linguist.* 47, 663–698. doi: 10.1162/COLL_a_00412
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatiotemporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755. doi: 10.1038/srep27755
- Coello, Y., and Fisher, M. H. (2016). *Foundations of Embodied Cognition: Vol. 1. Perceptual and Emotional Embodiment*. Oxford UK: Routledge.
- de Vega, M., Glenberg, A., and Graesser, A. (2008). *Symbols and Embodiment: Debates on Meaning and Cognition*. New York, NY: Oxford University Press.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805[cs.CL]*. doi: 10.48550/arXiv.1810.04805
- Dove, G. (2014). Thinking in words: language as an embodied medium of thought. *Top. Cogn. Sci.* 6, 371–389. doi: 10.1111/tops.12102
- Dove, G. (2016). Three symbol ungrounding problems: abstract concepts and the future of embodied cognition. *Psychon. Bull. Rev.* 23, 1109–1121. doi: 10.3758/s13423-015-0825-4
- Dove, G. (2018). Language as a disruptive technology: abstract concepts, embodiment and the flexible mind. *Philos. Trans. Roy. Soc. B* 373:20170135. doi: 10.1098/rstb.2017.0135
- Fisher, M. H., and Coello, Y. (2016). *Foundations of Embodied Cognition: Vol. 2. Conceptual and Interactive Embodiment*. Oxford, UK: Routledge.
- Ghio, M., Vaghi, M. M. S., Perani, D., and Tettamanti, M. (2016). Decoding the neural representation of fine-grained conceptual categories. *Neuroimage* 132, 93–103. doi: 10.1016/j.neuroimage.2016.02.009
- Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., and Trueswell, J. C. (2005). Hard words. *Lang. Learn. Dev.* 1, 23–64. doi: 10.1207/s15473341l1d0101_4
- Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565. doi: 10.3758/BF03196313
- Glorot, X., and Bengio, Y. (2010). “Understanding the difficulty of training deep feedforward neural networks,” in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Sardinia)*, 249–256.
- Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. San Rafael, CA: Morgan and Claypool Publishers. doi: 10.2200/S00762ED1V01Y201703HLT037

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.906181/full#supplementary-material>

- Günther, F., Nguyen, T., Chen, L., Dudschig, C., Kaup, B., and Glenberg, A. M. (2020). Immediate sensorimotor grounding of novel concepts learned from language alone. *J. Mem. Lang.* 115:104172. doi: 10.1016/j.jml.2020.104172
- Harnad, S. (1990). The symbol grounding problem. *Phys. D* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6
- Hill, F., and Korhonen, A. (2014). “Learning abstract concept embeddings from multi-modal data: Since you probably can’t see what I mean,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP2014)* (Doha), 255–265.
- Hoffman, P., McClelland, J. L., and Lambon Ralph, M. A. (2018). Concepts, control, and context: a connectionist account of normal and disordered semantic cognition. *Psychol. Rev.* 125, 293–328. doi: 10.1037/rev0000094
- Howell, S. R., Jankowicz, D., and Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *J. Mem. Lang.* 53, 258–276. doi: 10.1016/j.jml.2005.03.002
- Jackendoff, R. (2019). “Mental representations for language,” in *Human Language: From Genes and Brains to Behavior*, ed P. Hagoort (Cambridge, MA: MIT Press), 7–20.
- Jacob, G., Pramod, R., Katti, H., and Arun, S. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun.* 12:1872. doi: 10.1038/s41467-021-22078-3
- Johns, B. T., and Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Top. Cogn. Sci.* 4, 103–120. doi: 10.1111/j.1756-8765.2011.01176.x
- Jones, M. N., Willits, J., and Dennis, S. (2015). “Models of semantic memory,” in *Oxford Handbook of Mathematical and Computational Psychology*, eds J. R. Busemeyer, Z. Wang, J. T. Townsend, and A. Eidels (New York, NY: Oxford University Press), 232–254.
- Kaschak, M. P., Madden, C. J., Theriault, D. J., Yaxley, R. H., Aveyard, M., Blanchard, A. A., et al. (2005). Perception of motion affects language processing. *Cognition* 94, 79–89. doi: 10.1016/j.cognition.2004.06.005
- Kiela, D., Bulat, L., and Clark, S. (2015). “Grounding semantics in olfactory perception,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Beijing), 231–236.
- Kiela, D., and Clark, S. (2015). “Multi- and cross-modal semantics beyond vision: Grounding in auditory perception,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon), 2461–2470.
- Kiela, D., Hill, F., Korhonen, A., and Clark, S. (2014). “Improving multi-modal representations using image dispersion: why less is sometimes more,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, MD), 835–841.
- Kumar, A. A. (2021). Semantic memory: a review of methods, models, and current challenges. *Psychon. Bull. Rev.* 28, 40–80. doi: 10.3758/s13423-020-01792-x
- Kuperman, V., Stadthagen-Gonzalez, H., and Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behav. Res. Methods* 44, 978–990. doi: 10.3758/s13428-012-0210-4
- Lake, B. M., and Murphy, G. L. (2021). Word meaning in minds and machines. *Psychol. Rev.* doi: 10.1037/rev0000297. [Epub ahead of print].
- Lazaridou, A., Pham, N. T., and Baroni, M. (2015). “Combining language and vision with a multimodal skip-gram model,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO), 153–163.
- Lenci, A. (2018). Distributional models of word meaning. *Annu. Rev. Linguist.* 4, 151–171. doi: 10.1146/annurev-linguistics-030514-125254
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Top. Cogn. Sci.* 3, 273–302. doi: 10.1111/j.1756-8765.2010.01106.x
- Louwerse, M. M. (2018). Knowing the meaning of a word by the linguistic and perceptual company it keeps. *Top. Cogn. Sci.* 10, 573–589. doi: 10.1111/tops.12349
- Louwerse, M. M., and Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition* 114, 96–104. doi: 10.1016/j.cognition.2009.09.002
- Lupyan, G., and Lewis, M. (2019). From words-as-mappings to words-as-cues: the role of language in semantic knowledge. *Lang. Cogn. Neurosci.* 34, 1319–1337. doi: 10.1080/23273798.2017.1404114
- Malhi, S. K., and Buchanan, L. (2018). A test of the symbol interdependency hypothesis with both concrete and abstract stimuli. *PLoS ONE* 13:e0192719. doi: 10.1371/journal.pone.0192719
- McRae, K., Nedjradrasul, D., Raymond Pau, B. P.-H. L., and King, L. (2018). Abstract concepts and pictures of real-world situations activate one another. *Top. Cogn. Sci.* 10, 518–532. doi: 10.1111/tops.12328
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). “Efficient estimation of word representations in vector space,” in *Proceedings of Workshop at the International Conference on Learning Representation (ICLR)* (Scottsdale, AZ).
- Paivio, A. (1971). *Imagery and Verbal Processes*. New York, NY: Holt, Rinehart, and Winston.
- Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York, NY: Oxford University Press.
- Pecher, D., and Zwaan, R. (2005). *Grounding Cognition: The Role of Perception and Action in Memory, Language and Thinking*. Cambridge: Cambridge University Press.
- Pennington, J., Socher, R., and Manning, C. D. (2014). “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Doha), 1532–1543.
- Pilehvar, M. T., and Camacho-Collados, J. (2020). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Williston, VT: Morgan and Claypool Publishers. doi: 10.2200/S01057ED1V01Y202009HLT047
- Pulvermüller, F. (2013). How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends Cogn. Sci.* 17, 458–470. doi: 10.1016/j.tics.2013.06.004
- Recchia, G., and Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: predicting valence, arousal, and dominance. *Q. J. Exp. Psychol. (Hove)*. 68, 1584–1598. doi: 10.1080/17470218.2014.941296
- Reijmiers, W. G., Burgers, C., Bolognesi, M., and Krennmayr, T. (2019). How polysemy affects concreteness ratings: the case of metaphor. *Cogn. Sci.* 43:e12779. doi: 10.1111/cogs.12779
- Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in BERTology: what we know about how BERT works. *Trans. Assoc. Comput. Linguist.* 8, 842–866. doi: 10.1162/tacl_a_00349
- Schwanenflugel, P. J. (1991). “Why are abstract concepts hard to understand?,” in *The Psychology of Word Meanings*, ed P. J. Schwanenflugel (Hillsdale, NJ: Lawrence Erlbaum Associates), 223–250.
- Schwanenflugel, P. J., Harnishfeger, K. K., and Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *J. Mem. Lang.* 27, 499–520. doi: 10.1016/0749-596X(88)90022-8
- Scorolli, C. (2014). “Embodiment and language,” in *The Routledge Handbook of Embodied Cognition*, ed L. Shapiro (New York, NY: Routledge), 127–138.
- Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* 5, 399–426. doi: 10.1146/annurev-vision-091718-014951
- Silberer, C., Ferrari, V., and Lapata, M. (2017). Visually grounded meaning representations. *IEEE Trans. Pattern Recogn. Mach. Intell.* 39, 2284–2297. doi: 10.1109/TPAMI.2016.2635138
- Silberer, C., and Lapata, M. (2012). “Grounded models of semantic representation,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (Jeju Island), 1423–1433.
- Thill, S., Padó, S., and Ziemke, T. (2014). On the importance of a rich embodiment in the grounding of concepts: perspectives from embodied cognitive science and computational linguistics. *Top. Cogn. Sci.* 6, 545–558. doi: 10.1111/tops.12093
- Troche, J., Crutch, S. J., and Reilly, J. (2017). Defining a conceptual topography of word concreteness: clustering properties of emotion, sensation, and magnitude among 750 english words. *Front. Psychol.* 8:1787. doi: 10.3389/fpsyg.2017.01787
- Turney, P. D., and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Intell. Res.* 37, 141–188. doi: 10.1613/jair.2934
- Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. *Cogn. Sci.* 44:e12844. doi: 10.1111/cogs.12844
- Vigliocco, G., and Vinson, D. P. (2007). “Semantic representation,” in *The Oxford Handbook of Psycholinguistics*, ed M. G. Gaskell (Oxford, UK: Oxford University Press), 217–234.
- Villani, C., Lugli, L., Liuzza, M. T., and Borghi, A. M. (2019). Varieties of abstract concepts and their multiple dimensions. *Lang. Cogn.* 11, 403–430. doi: 10.1017/langcog.2019.23
- Vincent-Lamarre, P., Massé, A. B., Lopes, M., Lord, M., Marcotte, O., and Harnada, S. (2016). The latent structure of dictionaries. *Top. Cogn. Sci.* 8, 625–659. doi: 10.1111/tops.12211
- Wang, J., Conder, J. A., Blitzer, D. N., and Shinkareva, S. V. (2010). Neural representation of abstract and concrete concepts: A meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* 31, 1459–1468. doi: 10.1002/hbm.20950

Wiemer-Hastings, K., and Xu, X. (2005). Content differences for abstract and concrete concepts. *Cogn. Sci.* 29, 719–736. doi: 10.1207/s15516709cog0000_33

Xu, Y., and Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat. Commun.* 12:2065. doi: 10.1038/s41467-021-22244-7

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., et al. (2021). *The AI Index 2021 Annual Report*. Stanford,

CA: AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 1452–1464. doi: 10.1109/TPAMI.2017.2723009

Zwaan, R. and Yaxley, R. (2003). Spatial iconicity affects semantic relatedness judgments. *Psychon. Bull. Rev.* 10, 954–958. doi: 10.3758/BF03196557