COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12922

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12922

Visual and Affective Multimodal Models of Word Meaning in Language and Mind

Simon De Deyne,[a] Danielle J. Navarro,[b] Guillem Collell,[c] Andrew Perfors[a]

[a]*School of Psychological Sciences, University of Melbourne*
[b]*School of Psychology, University of New South Wales*
[c]*Department of Computer Science, KU Leuven*

Received 10 January 2019; received in revised form 26 October 2020; accepted 10 November 2020

Abstract

One of the main limitations of natural language-based approaches to meaning is that they do not incorporate multimodal representations the way humans do. In this study, we evaluate how well different kinds of models account for people's representations of both concrete and abstract concepts. The models we compare include unimodal distributional linguistic models as well as multimodal models which combine linguistic with perceptual or affective information. There are two types of linguistic models: those based on text corpora and those derived from word association data. We present two new studies and a reanalysis of a series of previous studies. The studies demonstrate that both visual and affective multimodal models better capture behavior that reflects human representations than unimodal linguistic models. The size of the multimodal advantage depends on the nature of semantic representations involved, and it is especially pronounced for basic-level concepts that belong to the same superordinate category. Additional visual and affective features improve the accuracy of linguistic models based on text corpora more than those based on word associations; this suggests systematic qualitative differences between what information is encoded in natural language versus what information is reflected in word associations. Altogether, our work presents new evidence that multimodal information is important for capturing both abstract and concrete words and that fully representing word meaning requires more than purely linguistic information. Implications for both embodied and distributional views of semantic representation are discussed.

Keywords: Multimodal representations; Semantic networks; Distributional semantics; Visual features; Affect

Correspondence should be sent to Simon De Deyne, School of Psychological Sciences, University of Melbourne, Melbourne, 3010 Vic., Australia. E-mail: simon.dedeyne@unimelb.edu.au

1. Introduction

When you look up the word *rose* in the 2012 *Concise Oxford English Dictionary*, it is defined as "a prickly bush or shrub that typically bears red, pink, yellow, or white fragrant flowers, native to north temperate regions and widely grown as an ornamental." How central are each of these aspects to our representation of a rose, and in what form are they represented? Different theories give different answers to this question, particularly with respect to how much linguistic and nonlinguistic sensory representations contribute to meaning. In embodied theories, meaning is based on the relationship between words and internal bodily states corresponding to multiple modalities, such as somatosensation, vision, olfaction, and perhaps even internal affective states. By contrast, lexico-semantic views focus on the contribution of language, suggesting that the meaning of *rose* can be derived in a recursive fashion by considering its relationship to the meaning of words in its linguistic context, such as bush, red, and flower (cf. Firth, 1968). These two views are extremes on a spectrum, and current theories of semantics tend to take an intermediate position that both linguistic and nonlinguistic information contribute to meaning.

How is information from sensory modalities and the language modality combined, and which is more important to understand the meaning of words? One idea is that language is a symbolic system that represents meaning via the relationships between (amodal) symbols but is also capable of capturing sensory representations since these symbols refer to perceptual information. This is consistent with the *symbol interdependency hypothesis*, proposed by Louwerse (2011) and closely related to Clark's (2006) hypothesis that we use language as proxy of the world.

Others have argued for hybrid approaches that combine both symbolic and sensorily grounded representations to varying degrees, depending on task and word characteristics (Andrews, Frank, & Vigliocco, 2014; Paivio, 1971; Riordan & Jones, 2011; Vigliocco, Meteyard, Andrews, & Kousta, 2009). One reason for the importance of nonlinguistic sensory information is that although language has the capacity to capture sensory properties, this capacity is imperfect in making fine-grained distinctions. For example, in Japanese the word *ashi* (あし) does not distinguish between foot and leg. As another example, consider the fact that across languages, commonly used color terms only cover a subset of all the colors that humans can detect. Thus, one would expect that models of lexical semantics should perform better when provided with visual training data in addition to linguistic corpora (e.g., Bruni, Tran, & Baroni, 2014; Johns & Jones, 2012; Silberer & Lapata, 2014). This may be true even for abstract words that lack visual referents: Recent work suggests that nonlinguistic factors such as sensorimotor experience, emotional experience, interoception, and sociality ground the meaning of abstract words (for an overview, see Borghi et al., 2017). This suggests that models incorporating only linguistic information will fare less well in capturing human representations than those that combine linguistic and sensory input. The opposite may also hold—that models incorporating only sensory input will fare less well than those based on both.

Not all agree that multimodal information is necessary for the representation of word meaning, at least not for abstract words that lack physical referents (Paivio, 2013). For

instance, it is not clear how the meaning of concepts like "romance" or "purity," which are connotations of the word *rose* that are not directly derived from sensory impressions, might be captured by sensory-based models that learn low-level visual features from images (e.g., Chatfield, Simonyan, Vedaldi, & Zisserman, 2014; He, Zhang, Ren, & Sun, 2016; Lazaridou, Pham, & Baroni, 2015).

In this paper, we revisit the question about the extent to which the language modality encodes sensory properties by comparing linguistic representations derived from text corpora or word associations with multimodal models that encode sensory or affective information as well. We focus on concrete and abstract concepts and the role of visual and affective properties. With affective information we refer more specifically to the emotional valence, or how positive/negative a word is and the arousal, or how calming/exciting a word is.[1] This will allow us to determine whether models of abstract word meaning require multimodal affective representations in the same way that models of concrete word meaning do.

1.1. Multimodal representations of concrete concepts

Concrete concepts are concepts that refer to a perceptible entity. Multimodal models of concrete concepts combine linguistic and sensory representations to determine to what degree linguistic representations capture sensory properties. In theory, these sensory properties could reflect any modality. In a study by Kiela and Clark (2015), for example, a multimodal model was constructed that combines language with auditory input. In practice, most studies focus on visual multimodal models, and this is reflected in the type of concepts that are modeled (typically concrete nouns) and the way concreteness is measured (typically focusing on the visual modality, see Brysbaert, Warriner, & Kuperman, 2014). Early work relied on psycho-experimental methods to capture these sensory properties; a common approach is to derive representations from feature elicitation tasks in which participants are asked to list meaningful properties of the concept in question. These (presumably nonlinguistic) features are then combined in a multimodal model whose linguistic representations are derived from text corpora (Andrews et al., 2014; Johns & Jones, 2012; Steyvers, 2010).

In recent years, multimodal models have begun to use visual representations derived from large image databases instead of feature listing tasks. These range from approaches in which visual features are obtained by human annotators (Silberer, Ferrari, & Lapata, 2013) to *bag-of-visual words* approaches in which a large set of visual descriptors are mapped onto vectors encoding low-level scale-invariant feature transformations (e.g., Bruni et al., 2014). Deep convolutional networks are starting to be used as well, since they typically outperform the low-level representations captured by simpler feature extraction methods (e.g., Lazaridou et al., 2015).

1.2. Multimodal representations of abstract concepts

Because of their reliance on visual information, most previous work on multimodal models has focused on concrete concepts. However, there are good reasons to study

abstract concepts as well. For one, they are extremely prevalent in everyday language: 72% of the noun or verb tokens in the British National Corpus are rated by people as more abstract than the noun *war*, which many would already consider quite abstract (e.g., Lazaridou et al., 2015). Abstract concepts also pose a particular challenge to strong embodiment views, which maintain that meaning primarily relies on nonlinguistic information. According to these theories, concrete and abstract concepts are not substantially different from each other; both are grounded in the same systems that are engaged during perception, action, and emotion (Borghi et al., 2017). The difficulty for such views is that abstract concepts like "opinion" do not have clearly identifiable referents. By contrast, distributional semantic views can easily explain how abstract concepts are represented: in terms of their distributional properties in language (Andrews et al., 2014).

A range of other theories about abstract concepts have been proposed in recent years, going beyond either embodied or distributional semantic accounts (for an overview, see Borghi et al., 2017). For example, according to the conceptual metaphor theory, abstract concepts are represented in terms of metaphors derived from concrete domains; this process provides a perceptual grounding for abstract meaning (Lakoff & Johnson, 2008). We focus here on a different view that highlights the role of affect in particular, known as the *Affective Embodiment Account* (*AEA*). The AEA posits that abstract concepts are grounded in internal affective states (Vigliocco et al., 2009). The grounding is quite broad, covering not just abstract words for emotions but also words that evoke an affective state like *cancer* or *birthday* (Kousta, Vinson, & Vigliocco, 2009). Consistent with this, emotional valence leads to a processing advantage for abstract words in a lexical decision task, even when confounding factors like imageability or context availability are taken into account (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011).

Further evidence that supports the AEA comes from a study in which brain activation was predicted by linguistic representations derived from word embedding models as well as experiential (e.g., visual, affective) representations derived from a feature rating task (Wang et al., 2017). Both types of representations were only weakly correlated with each other. When Wang et al. (2017) related the neural activation during a word familiarity judgment task with the linguistic and affective experiential representations using features like valence, arousal, emotion, or social interaction for the same words, they found dissociable neural correlates for each. Notably, affective experiential features were sensitive to areas involved in emotion processing. A subsequent principal component analysis of the whole-brain activity pattern showed that the first neural principal component, which captured most of the variance in abstract concepts, was correlated significantly with experiential information but not linguistic information. Moreover, the correlation was stronger for valence than other factors such as social interaction, which have been proposed as other factors in which abstract concepts could be embodied.

1.3. Current study

This study aims to investigate how and to what extent distributional models based on word associations or word co-occurrences derived from natural language are able to capture the

meaning of concrete and abstract words. In the word association model the meaning of a word is measured as a distribution of respectively weighted associative links encoded in a large semantic network, whereas the distributional linguistic model derives word meaning from word co-occurrences derived from large text corpora. Both models are considered to be primarily linguistic in nature. However, for word associations, we expect that associations reflect access to not only the language modality but nonlinguistic sensory modalities as well.

Our second aim is to measure the extent to which distributional models based on word association and word co-occurrence models can be improved by adding nonlinguistic visual and affective information. Our research is motivated by several observations. First, recent performance improvements by corpus-based distributional semantic models (e.g., Mikolov, Chen, Corrado, & Dean, 2013) have led to suggestions that these models might learn representations like humans do (Mandera, Keuleers, & Brysbaert, 2017). Our work evaluates to what extent this is the case. Second, models that learn visual categories have recently shown a similarly striking improvement in their ability to correctly identify a wide range of concrete objects (Chatfield et al., 2014; He et al., 2016; Lazaridou et al., 2015).

Our study investigates new multimodal models that integrate visual and linguistic information. Our goals is to evaluate the extent to which such hybrid models capture human performance and to explore what insights they can provide about how humans use environmental cues (linguistic or visual) to build and represent semantic concepts.

Our final aim is motivated by the observation that previously reported gains of multimodal models over purely linguistic models in predicting human performance have been fairly modest (e.g., Bruni et al., 2014). This leaves some uncertainty about whether language in itself is sufficiently rich to encode detailed perceptual information (see Louwerse, 2011).

This study goes further by looking beyond concepts consisting of concrete nouns to determine whether multimodal representations for abstract words provide better predictions of behavioral measures of word meaning. In particular, we computationally test a corollary from the AEA hypothesis, namely that internal affective states provide the foundation for the meaning of abstract concepts.

We achieve these ends by extending existing multimodal models to incorporate both affective and visual multimodal representations. Our performance evaluation focuses on basic-level concepts within a superordinate category, and it compares text-based linguistic distributional models to models derived from experimental word association data that might reflect access to both linguistic and experiential representations. As described below, these aspects of our approach enable us to better interpret the performance of the models and what they mean about human cognition.

1.3.1. Comparisons between basic-level concepts within a superordinate category

In this work, we evaluate models by evaluating their performance on basic-level concepts (*apple*, *guitar*, etc.) belonging to natural language superordinate categories (*fruit*, *musical instruments*, etc.). The basic level is the most inclusive one within which the attributes are common to all or most members of the category (Rosch, 1978). Basic-level concepts are not only the most informative; they are also acquired early and tend to be easy to form an image of (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Tversky

& Hemenway, 1984). They are also the default label: For instance, in picture naming studies, subordinate items are usually named at the basic level, even when subordinate names are known (Rosch, 1978).

Given these factors as well as the fact that visual features are shared to some extent between basic-level objects, one might expect comparisons between them (like *apples* to *pears*) to provide a sensitive test of unimodal distributional linguistic models. Despite this, the human similarity benchmarks used to evaluate such models in previous work focus more on wider comparisons (like *apples* to *food* or *tree*). These non-basic-level and more abstract comparisons might be more readily encoded through language. This could explain the high correlations between model predictions and human similarity judgments found in such studies (e.g., Mandera et al., 2017), as well as the low correlations found in studies involving more concrete and basic-level objects (De Deyne, Peirsman, & Storms, 2009). In this work, by focusing on basic-level concepts, we aim to better control the nature of the semantic relationships and thereby clarify the utility of the information encoded in both linguistic models *and* multimodal models.

Of course, while the notion of a basic level applies clearly to concrete concepts, it is not obvious whether abstract concepts have a basic level. Nevertheless, we can make use of the fact that abstract concepts can be described as a part of a taxonomic hierarchy. Linguistic resources such as WordNet (Fellbaum, 1998) distinguish between superordinate, coordinate, and subordinate abstract concepts. For example, a hyponym (subordinate) listed in WordNet for the term *envy* is *covetousness*, whereas a hypernym (superordinate) is *resentment*. This suggests that abstract concepts, like concrete ones, can be described at a kind of "basic" level that has high cue validity and many shared attributes.[2] In this work we rely on WordNet to ensure that all of our basic-level (concrete and abstract) concepts belong to superordinate concrete categories of similar size.

1.3.2. Comparison of distributional linguistic models to a word association baseline

To better understand to what degree linguistic information can predict human semantic judgments, we will compare distributional (corpus-based) semantic models to a baseline model designed to capture subjective meaning. A variety of methods have been proposed to measure meaning, including semantic differentials (Osgood, Suci, & Tannenbaum, 1957), feature elicitation (De Deyne et al., 2008; McRae, Cree, Seidenberg, & McNorgan, 2005), and word association responses (De Deyne, Navarro, & Storms, 2013; Kiss, Armstrong, Milroy, & Piper, 1973; Nelson, McEvoy, & Schreiber, 2004). Distributional semantic representations derived from all of these methods appear to reflect both linguistic and nonlinguistic semantic information (Taylor, 2012). For example, modality-specific brain regions related to imagery are also activated when participants generate semantic features or word associations (Simmons, Hamann, Harenski, Hu, & Barsalou, 2008). In the remainder of this article we will use the term "distributional linguistic model" to refer to a distributional semantic model that uses language corpora.

In this study we derive our baseline model of semantic meaning using word association data. This approach makes fewer assumptions than feature elicitation tasks and is relevant for abstract words as well. If, as hypothesized, this model incorporates both linguistic and

nonlinguistic information, we would expect that adding visual or affective features would have little impact on the performance of the model. If, as the AEA account suggests, visual and affective information cannot be derived from linguistic information alone, we might expect that distributional corpus-based models would be improved by adding that information.

Alternatively, if it is indeed the case that natural language does encode perceptual and affective features in a sufficiently rich way, then adding visual or affective features to these language models would have little impact on their performance. This possibility is not unreasonable, given recent improvements in the ability of distributional semantic models to predict a variety of human judgments (see, e.g., Mandera et al., 2017). These improved models, which rely on word co-occurrence predictions (e.g., word2vec, Mikolov et al., 2013) instead of co-occurrence counts (e.g., latent semantic analysis; Landauer & Dumais, 1997), may thus capture more of the information inherent in linguistic data and provide a better proxy of the world (Clark, 2006).

The structure of this paper is as follows. We first describe how the distributional linguistic, word association, and experiential (visual or affective) models are constructed. We then evaluate how the experiential information (visual or affective) augments the performance of the distributional linguistic model and the word association baseline model in two similarity judgment tasks. Study 1 required participants to pick the most related pair out of three words (e.g., ***rose* vs. *tulip*** vs. *daffodil*) while Study 2 had participants rate the similarity of pairs like *rose* and *tulip* on an ordinal scale.

2. Constructing unimodal and multimodal models

2.1. Distributional linguistic model

The linguistic model captures semantic representation from the distributional properties in the language environment as is, without integrating it with information from other modalities.

2.1.1. Corpus

The model was trained on a combination of different text corpora described in detail in De Deyne, Perfors, and Navarro (2016). The corpora were constructed to provide a reasonably balanced set of texts that is representative of the type and amount of language a person experiences during a lifetime: formal and informal, spoken and written. It contains four parts: (a) a corpus of English movies subtitles; (b) written fiction and nonfiction taken from the Corpus of Contemporary American English (COCA; Davies, 2009); (c) informal language derived from online blogs and websites available through COCA; and (d) the Simple English Wikipedia, a version of Wikipedia with usually shorter articles aimed at students, children, and people who are learning English (SimpleWiki, accessed February 3, 2016). The resulting corpus consisted of 2.16 billion word tokens and 4.17 million word types. It thus encompasses knowledge that is likely available to the

average person but is sufficiently generous in terms of the quality and quantity of data to ensure that our models perform similarly to the existing state of the art.

2.1.2. Word2vec embeddings

Word embedding models have recently been proposed as an alternative to count-based distribution models such as latent semantic analysis (Landauer & Dumais, 1997) or word co-occurrence count models (e.g., HAL; Burgess, Livesay, & Lund, 1998). One of the most popular word embedding algorithms consists of a simple neural network that predicts word co-occurrence (Mikolov et al., 2013). In contrast to count-based models, networks like `word2vec` are used to predict words from context (the CBOW or "continuous bag-of-words" approach) or context from words (the skip-gram approach), which more closely resembles human learning. Compared to count-based approaches, these embedding models often lead to better accounts of lexical and semantic processing (Baroni, Dinu, & Kruszewski, 2014; De Deyne, Perfors, et al., 2016; Mandera et al., 2017).[3]

To train `word2vec` on this corpus, we used a CBOW architecture where the learning objective was to predict a word based on the context in which it occurs. A range of parameters determines the efficiency and predictive performance. In this study, parameter values were taken from previous research in which the optimal model was a network trained to predict the context words using a window size of 7 and a 400-dimensional hidden layer from which word vectors are derived (De Deyne, Perfors, et al., 2016). This was the best model on a wide range of similarity judgment studies, and it performs comparably with other published embeddings.

2.2. Word association model

Word associations have long been used as an experimental method to assess the semantic knowledge a person holds about a word (e.g., Deese, 1965). In contrast to a controlled task like feature listing, the *free* association procedure does not censor the type of response. This makes it suitable for capturing the representations of all kinds of concepts (including abstract ones) and all kinds of semantic relations (including thematic and affective ones). It also avoids dissociating the lexicon in two different types of entities (concepts and features), which allows us to represent all concepts in a single graph. The resulting representation is thought to capture broad aspects of meaning, not solely those reflecting our linguistic experiences (Mollin, 2009), since nonlinguistic (i.e., experiential or affective) information is accessed when participants generate associates (Simmons et al., 2008). If this is correct, then word associations can be best characterized as a multimodal model. However, for the purpose of this exposition, unimodal models will refer to either a distributional linguistic or a word association model, and multimodal models will refer to a combination of these models with either visual or affective information.

2.2.1. Word association data

The current data were collected as part of the *Small World of Words* project,[4] an ongoing crowd-sourced project to map the mental lexicon in various languages. The SWOW-

EN2018 data are those reported in De Deyne, Perfors, et al. (2016) and consist of associates given by 88,722 fluent English speakers. Each speaker gave three different responses to between 14 and 18 cue words. For example, a person shown the cue word *miracle* might respond *magic*, *religion*, and *Jesus*. The dataset contains a total of 12,292 cue words for which at least 300 responses were collected for every cue.

2.2.2. Semantic network

In line with previous work, we constructed a semantic weighted graph from these data in which each edge corresponds to the association frequency between a cue and a target word. The graph was constructed by only including responses that also occurred as a cue word and keeping only those nodes that are part of the largest connected component (i.e., nodes that have both in- and out-going edges). The resulting graph consists of 12,217 nodes, which retains 87% of the original data.

Following De Deyne, Navarro, Perfors, and Storms (2016) and De Deyne, Navarro, Perfors, Brysbaert, and Storms (2019), we first transformed the raw association frequencies using positive point-wise mutual information (PPMI). Next, a mechanism of spreading activation through random walks was used to allow indirect paths of varying length connecting any two nodes to contribute to their meaning. The random walks can be thought of as a vector with the same dimensionality as the number of nodes in the graph where each element corresponds to a weighted sum of direct and indirect paths, with longer paths receiving a lower weight. The random walks implement the idea of spreading activation over a semantic network. To limit the contribution of long paths, a decay parameter α was set to 0.75, in line with De Deyne, Navarro, et al. (2016). This algorithm is similar to other approaches (Austerweil, Abbott, & Griffiths, 2012), but differs by taking an additional PPMI weighting of the graph with indirect paths to avoid a frequency or degree bias and to reduce spurious links (for a discussion, see Newman, 2010).

2.3. Visual feature model

We constructed a model based on visual information using ImageNet as the source of the data encoding visual information (De Deyne, Perfors, et al., 2016); it is currently the largest labeled image bank and includes over 14 million images. It consists of images for the concrete nouns represented in WordNet, a language-inspired knowledge base in which synonymous words are grouped in *synsets* and connected through a variety of semantic relations (IS-A, HAS-A, etc.). With 21,841 synsets, ImageNet captures a large portion of the concrete lexicon represented in WordNet. A small part of synsets encoded in the WordNet is shown in Fig. 1.

Visual features were obtained by applying ResNet (He et al., 2016), a pre-trained supervised convolutional neural network, to each concrete synset that had at least 50 images. The 2,048-dimensional activation of the last layer (before the softmax layer) of the network is taken as a visual feature vector as it contains higher level features. Finally, we obtained a single 2,048-dimensional vector to represent each WordNet synset by

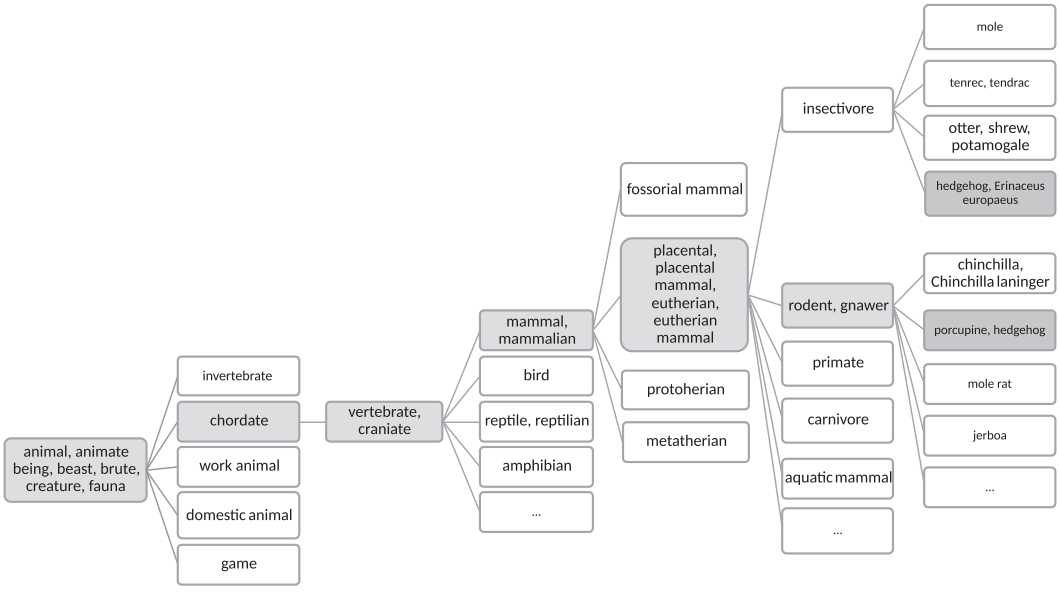

Fig. 1. Part of the WordNet hierarchy for concrete synsets that are covered by ImageNet. Synsets can occur at different hierarchical levels and are labeled by one or multiple words that can overlap with other branches in the tree; this is illustrated for the case "hedgehog."

averaging the feature vectors from its associated individual images. Each image vector was then mapped to a WordNet synset.

The words in this study corresponded to synset labels that could occur both as inner nodes (e.g., *mammal*) or leaf nodes (e.g., *hedgehog*) depending on their level of abstraction. In some cases, a word corresponded to a synset with multiple labels. For example, in Fig. 1 the word *hedgehog* is found in two different synsets (one labeled "*hedgehog*, *Erinaceus europaeus*" and one "*porcupine*, *hedgehog*"). To map these synsets to a single word, the synset labels were split and the corresponding image vectors averaged. Of the 18,851 parsed synset labels in ResNet, 4,449 labels were shared with the SWOW-EN2018 cue words. Of these 4,449 cues, 910 cues mapped onto more than one synset and were averaged. Previous research on high-dimensional distributional language models has shown that point-wise mutual information (PMI), which assigns larger weights to specific features, improves model predictions (De Deyne et al., 2019; Recchia & Jones, 2009). An exploratory analysis showed that this was also the case for the image vectors; we therefore use weighted PMI image vectors throughout this work.

2.4. Affective feature model

Affective factors like valence or arousal capture a significant portion of the structure in the mental representation of both adjectives (e.g., De Deyne, Voorspoels, Verheyen, Navarro, & Storms, 2014) and nouns (e.g., Osgood et al., 1957).The validity of the

subjective judgments of affective factors is supported by recent work that demonstrated that human affective ratings predict the modulation of the potential neural activity signal in areas associated with affective processing (Vigliocco et al., 2013). According to the AEA theory (Vigliocco et al., 2009), these affective factors provide the necessary multimodal grounding for abstract concepts, which lack any physical referents. We constructed the features of the affective model based on human ratings for valence, arousal, and dominance for nearly 14,000 words (Warriner, Kuperman, & Brysbaert, 2013). The ratings by males and females were treated as separate features and supplemented with three additional features for valence, arousal, and dominance from a more recent study based on 20,000 English words (Mohammad, 2018). The norms from the latter study were somewhat different than those from Warriner et al. (2013) in two ways. First, they used best–-worst scaling, which resulted in more reliable ratings. Second, they operationalized arousal differently, resulting in ratings that were less correlated with valence. Each individual word was thus represented by nine features: valence, arousal, and potency judgments for men and women from Warriner et al. (2013) and valence, arousal, and dominance judgments from Mohammad (2018). For example, the representation for the word *rose* would be: [5.9 3.1 6.4 8.1 2.6 6.0 7.8 2.5 3.5]. The first six values correspond to the male ratings for valence, arousal, and potency and the female ratings for valence, arousal, and potency (Warriner et al., 2013). The last three values correspond to rescaled valence, arousal, and dominance for men and women from Mohammad (2018). None of the features were perfectly correlated with each other, which allowed them to each contribute.

2.5. Multimodal models that combine linguistic models with experiential representations

To investigate how adding visual or affective information to our linguistic or word association model affects their performance, we created multimodal models that incorporate both kinds of information. There are multiple efficient ways to combine different information sources; these include auto-encoders (Silberer & Lapata, 2014), Bayesian models (Andrews et al., 2014), and cross-modal mappings (Collell, Zhang, & Moens, 2017). We employ a late fusion approach in which features from different modalities are concatenated to build multimodal representations. This approach performs relatively well (e.g., Bruni et al., 2014; Johns & Jones, 2012; Kiela & Bottou, 2014) and enables us to investigate the relative contribution of the modalities directly. This is achieved by a single dataset-level tuning parameter β, ranging from 0 to 1, which allows us to vary and quantify the relative contribution of the different modalities. The multimodal fusion M of the modalities a and b corresponds to $M = \beta \times v_a \oplus (1-\beta) \times v_b$ where $\oplus$ denotes concatenation, v_a is the vector representation for the linguistic or word association information, and v_b is the vector representation for the affective or visual information. Since the features for different modalities can have different scales, they were normalized using the L_2-norm prior to concatenation, which puts all features in a unitary vector space (Kiela & Bottou, 2014).

3. Study 1: Basic-level triadic comparisons

The first study compares how well each of the models above predicts human similarity judgments. Participants rated similarity in a triadic comparison task in which they were asked to pick the most related pair out of three words. There were two conditions, one in which the words were CONCRETE and one in which they were ABSTRACT. Compared to pairwise judgments using rating scales, the triadic task has several advantages: Humans find relative judgments easier, it avoids scaling issues, and it leads to more consistent results (Li, Malave, Song, & Yu, 2016).

3.1. Method

3.1.1. Participants

Forty native English speakers between 18 and 49 years old (21 females, 19 males, average age 35) were recruited in the CONCRETE condition and forty native English speakers aged 19–46 years (16 females, 24 males, average age = 32) in the ABSTRACT condition. The participant sample size (and number of stimuli) was informed by a previous work on triadic judgments (De Deyne, Navarro, et al., 2016). Data were collected in two online studies through Prolific Academic. We first collected for the abstract study and when enough participants completed the task, the concrete task was administered. All participants signed an informed consent form and were paid £6/h. The procedures were approved by the Ethics Committee of the University of Adelaide.

3.1.2. Stimuli

The CONCRETE stimuli consisted of 100 triads constructed from a subset of 300 nouns present in the lexicons of all of our models and for which valence and concreteness norms were available (Brysbaert et al., 2014; Warriner et al., 2013). All 300 words were selected from a set of superordinate categories identified in previous work (e.g., Battig & Montague, 1969; De Deyne et al., 2008). Approximately half of the triads belonged to natural kind categories (*Fruit*, *Vegetables*, *Mammals*, *Fish*, *Birds*, *Reptiles*, *Insects*, *Trees*) and the other half to human-made categories (*Clothing*, *Dwellings*, *Furniture*, *Kitchen utensils*, *Musical instruments*, *Professions*, *Tools*, *Vehicles*, *Weapons*). Each triad was composed of three basic-level words that belong to the same superordinate category (e.g., *falcon*, *flamingo*, *penguin*) and was constructed by randomly combining category exemplars, subject to the constraint that none of the words occurred more than once across any of the triads. Table A1 of Appendix A contains a list of all the stimuli.

The ABSTRACT stimuli consisted of 100 triads and were also constructed from words that were present in the lexicons of all our models and where norms for concreteness and valence were available (see concrete triads). To ensure that the words were abstract, we removed any words with concreteness ratings of over 3.5 on a 5-point scale (Brysbaert et al., 2014). The average concreteness was 2.6. Moreover, all words were well known, with at least 90% of participants in the word association study indicating that they knew

the word. The resulting set of 29 categories covered a variety of abstract domains, including emotions, attitudes, abilities, social groups, and beliefs. A list of the stimuli together with their category label is presented in Table A2 of Appendix A.

To identify superordinate and basic-level words for abstract categories, we used WordNet (Fellbaum, 1998), selecting stimuli corresponding to categories in the WordNet hierarchy at a depth of 4–8 in the taxonomy. This ensured that the exemplars were neither overly specific nor overly general. For example, the triad *bliss–madness–paranoia* consists of three words defined at depths 8, 9, and 9 in the hierarchical taxonomy, with the most specific shared category label or hypernym at depth 6.

3.1.3. Procedure

For each of the 100 triads, participants were instructed to select the pair of words (out of the three) that was most related in meaning, or to indicate if any of the words is unknown. They were asked to only consider the word meaning, ignoring superficial properties like letters, sound, or rhyme. The method and procedure for ABSTRACT stimuli were identical to that for the concrete except that the example given to participants now contained abstract words. The concrete task took 8 min, and the ABSTRACT one, 10 min.

3.1.4. Behavioral results

All three words in over 99% of concrete triads and 96% of abstract triads were considered known by participants. Similarity judgments were calculated by counting how many participants chose each of the three potential pairs. Because the number of judgments varied depending on whether participants judged them to be unknown, they were converted to proportions. The Spearman split-half reliability was .92 for concrete triads and .90 for abstract triads. These reliabilities provide an upper bound of the correlations that can be achieved with our models.

3.2. Model evaluation

Human triad preferences were predicted by calculating the Pearson correlation with the model preferences for either the distributional linguistic, word association, visual, and affective vectors for each of the three word pairs in each triad. The model preferences were calculated using the cosine similarities for all three pairs and rescaling them to sum to one.[5] Note that in all analyses that follow, no results are available for the visual feature model of abstract words for the obvious reason that abstract words are not found in ImageNet. The correspondences between the human preferences and the model predictions measured through Pearson correlations are shown in Table 1.

3.2.1. Distributional linguistic versus word association model comparisons

We first compared the performance of the distributional linguistic model against the baseline word association model. The word association model showed a high correlation with the triad preferences in both the CONCRETE, $r(298) = .76$, CI_{95} [0.71, 0.80], and ABSTRACT tasks, $r(298) = .82$, CI_{95} [0.78, 0.86]. The correlations for the distributional

Table 1
Pearson correlations and confidence intervals for unimodal and multimodal models. The top panel shows the performance when v_a corresponds to the distributional linguistic model, while the middle panel v_a corresponds to the word association baseline. The bottom panel corresponds to purely experiential model where v_a corresponds to the affective model and is added for completeness. In each panel, the unimodal columns show the performance of that model (v_a) as well as the two experiential models (v_b) on either the concrete or the abstract words. The best-fitting multimodal models combining v_a and v_b were found by optimizing the correlation for mixing parameter β and are shown in column $r_{v_{ab}}$. The improvement due to adding experiential information ($r_{v_{ab}} - r_{v_a}$) is shown in column Δr

		v_a = Distributional linguistic model									
		Unimodal					Multimodal				
Dataset	n	r_{v_a}	CI_{95}	v_b	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
Concrete	300	.64	[0.57, 0.70]	Visual	.67	[0.60, 0.73]	.48	.75	[0.70, 0.80]	.12	[0.07, 0.17]
Concrete	300	.64	[0.57, 0.70]	Affect	.21	[0.10, 0.32]	.50	.68	[0.62, 0.74]	.04	[0.02, 0.08]
Abstract	300	.62	[0.54, 0.68]	Affect	.51	[0.43, 0.59]	.58	.74	[0.69, 0.79]	.13	[0.08, 0.19]
		v_a = Word association model									
		Unimodal					Multimodal				
Dataset	n	r_{v_a}	CI_{95}	v_b	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
Concrete	300	.76	[0.71, 0.80]	Visual	.67	[0.60, 0.73]	.35	.81	[0.77, 0.85]	.05	[0.03, 0.08]
Concrete	300	.76	[0.71, 0.80]	Affect	.21	[0.10, 0.32]	.38	.78	[0.74, 0.82]	.02	[0.00, 0.05]
Abstract	300	.82	[0.78, 0.86]	Affect	.51	[0.43, 0.59]	.05	.82	[0.78, 0.86]	.00	[−0.01, 0.01]
		v_a = Affective model									
		Unimodal					Multimodal				
Dataset	n	r_{v_a}	CI_{95}	v_b	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
Concrete	300	.21	[0.10, 0.32]	Visual	.67	[0.60, 0.73]	.45	.73	[0.67, 0.77]	.52	[0.41, 0.63]

Note. Note that the confidence intervals for Δr are based on testing significant differences for dependent overlapping correlations based on Zou (2007). This approach increases the power to detect an effect compared to Fisher's r to z procedure which assumes independence.

linguistic model were considerably lower: $r(298) = .64$, CI_{95} [0.56, 0.70] for CONCRETE triads and $r(298) = .62$, CI_{95} [0.54, 0.68] for ABSTRACT triads. To test whether the difference in correlation was significant, confidence intervals for correlation differences were calculated for dependent overlapping variables (Zou, 2007) using the `cocor` package in R (Diedenhofen & Musch, 2015). The 95% confidence interval for the difference in correlation was CI_{95} [0.06, 0.19] for the CONCRETE triads and CI_{95} [0.14, 0.27] for the ABSTRACT triads, suggesting that the word association model better predicted human similarity judgments for both kinds of words.

3.2.2. Visual and affective multimodal model comparisons

For the multimodal models, we were primarily interested in understanding how visual information contributed to the representation of CONCRETE words and how affective

information contributed to the representation of ABSTRACT words. In contrast to abstract words, it is also feasible to investigate a combination of both visual and affective information for concrete words. For completeness we include this scenario for CONCRETE word with v_b = Affect as part of the results reported in Table 1.

The confidence intervals in the last column of Table 1 indicate that for CONCRETE words, adding visual information helped both the distributional and the word association models. The $r_{v_{ab}}$ values corresponding to the visual multimodal model are higher than the r_{v_a} values, and in both cases this difference was significant as the confidence interval of the difference did not include zero. However, visual information improved performance of the distributional linguistic model more (0.12 vs. 0.05, respectively). The bootstrapped confidence interval of the difference between Δrs, CI_{95} [0.01, 0.12], did not include zero, suggesting this difference was significant.[6] This suggests that the word association model may incorporate some visual information that the distributional linguistic model does not. For completeness, Table 1 also shows the results for concrete triads using a multimodal model based on affective information. Affective information by itself only weakly predicts the preferences in concrete triads ($r(298) = .21$), but it offers a small multimodal gain when combined with the distributional linguistic model ($\Delta r = .04$, CI_{95} [0.02, 0.08]) and the word association model ($\Delta r = .02$, CI_{95} [0.00, 0.05]).

For ABSTRACT words, we found that affective information improved the performance of the distributional linguistic model substantially (from $r_{v_a}(298) = .62$ to $r_{v_{ab}}(298) = .74$). This improved performance is consistent with the AEA hypothesis by Vigliocco et al. (2009) and suggests that the decisions made by people in the triad task were based in part on affective information. Indeed, the affective model predictions alone did capture a significant portion of the variability in the abstract triads, $r(298) = .51$, CI_{95} [0.43, 0.59], which is remarkable considering that the model consists of only nine features. Interestingly, affective information did not improve the performance of word association model, suggesting that word association data already incorporates affective information. Moreover, the multimodal gain for the distributional linguistic model (0.13, compared to 0.00 for word associations) was significantly larger, with a bootstrapped confidence interval of the difference between Δrs CI_{95} [0.07, 0.19].

To further explore the effect of adding visual or affective information, Fig. 2 plots the performance of each model as a function of the β-weighted proportion of experiential information added when the words are either CONCRETE (left panel) or ABSTRACT (right panel). The word association model performs best, with small improvements from visual features ($\Delta r = .05$, $\beta = 0.35$, see Table 1). For the CONCRETE triads, we examine the effect of adding visual information. The distributional linguistic model performs worse but improves slightly more when visual features are added (r increase of .07). For completeness, we also included a model where v_a is affective and v_b is visual for the concrete triads. For the concrete triads, the affective model, shown in orange and included in Table 1, is easily the worst of the three. For the ABSTRACT triads, we examine the effect of adding affective information. When this is done, the word association model does not improve, but the distributional linguistic model improves considerably.

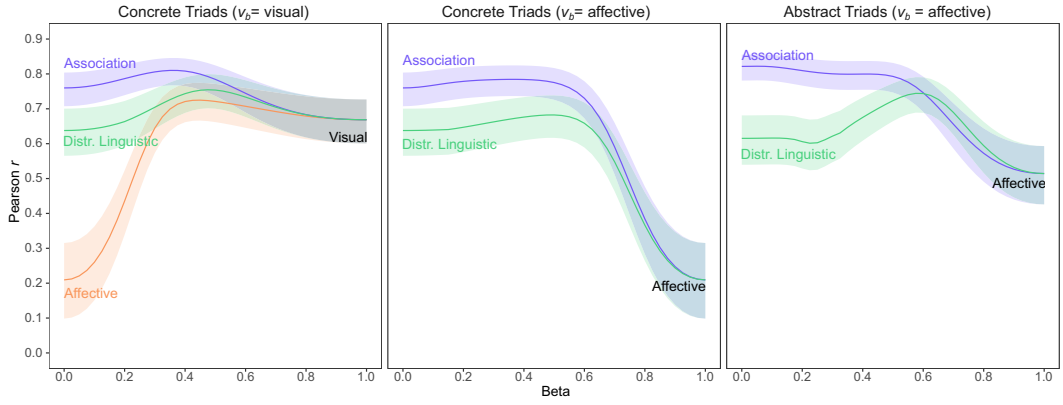

Fig. 2. The effect of adding visual or affective experiential information to predict triadic preferences for CONCRETE (first and second panels) and ABSTRACT (third panel) word pairs. Each panel shows the unimodal distributional linguistic and word association r_{v_a} correlations on the left side of the x-axis and the unimodal experiential (affective or visual features) r_{v_b} correlations on the right side. Intermediate values on the x-axis indicate multimodal models. In the first panel, visual information is added: Larger β values correspond to models that weight visual feature information more. In the second and third panels, affective information is added: Larger β values correspond to models that weight affective information more. Peak performance for all models usually occurs when about half of the information is experiential.

Overall, word association models outperformed distributional linguistic models at the optimal β value, even when these models included visual (respectively 0.81 vs. 0.75, CI_{95} [0.02, 0.09]) or affective (respectively 0.82 vs. 0.74, CI_{95} [0.03, 0.13]) information.

3.3. Robustness

Thus far, our results support the hypothesis that the model of meaning based on word associations captures visual information in concrete words and affective information in abstract words. The performance of the distributional linguistic model was worse than the word association baseline for both CONCRETE and ABSTRACT words. The distributional linguistic models were most improved by adding affective information, consistent with the AEA hypothesis that the meaning of abstract words, like concrete words, relies on experiential information. To what degree do these findings reflect the specific choices we made in setting up our models? To address this question, we tested how robust our results were when compared against alternative models based on different corpora and different embedding techniques.

3.3.1. Distributional linguistic models

First, we investigated whether the linguistic model performance was due to the specific embeddings used. To test this, we chose `GloVe` embeddings as an alternative distributional linguistic model (Pennington, Socher, & Manning, 2014). In contrast to `word2-vec` embeddings, `GloVe` does not incrementally learn embeddings but is based on a factorization of a global word co-occurrence matrix, which can lead to improved

predictions in certain tasks. We used the published word vectors for a model of comparable size to our corpus consisting of 6 billion (B) tokens derived from the GigaWord 5 and Wikipedia 2014 corpus. We also included an extremely large corpus consisting of 840B words from the Common Crawl project.[7] As before, the distributional linguistic vectors (v_a) were combined with visual or affective information (v_b) to create a multimodal model that was optimized by fitting values of β. Fig. 3 shows the unimodal distributional linguistic model correlations and the optimal multimodal correlations.

For unimodal predictions of concrete items, the `Glove-840B` was better than `word2vec` (CI_{95} [−0.15, −0.04]) and `Glove-6B` (CI_{95} [−0.17, −0.07]), but the smaller `Glove-6B` was not significantly different from `word2vec`. The multimodal predictions of CONCRETE items followed the same pattern, with only `GloVe-840B` better than `word2vec` (CI_{95} [−0.07, −0.01]), and `Glove-6B` (CI_{95} [−0.07, −0.02]). For the unimodal prediction of ABSTRACT items, `word2vec` outperformed both `GloVe-6B` (CI_{95} [0.02, 0.14]) and `GloVe-840B` (CI_{95} [0.01, 0.13]). However, once affect was added in the multimodal model, the distinct models did not obtain significantly different correlations.

Does adding experiential information improve performance for the `GloVe` models, as it did for `word2vec`? For CONCRETE items for `Glove-6B`, it appeared to: Predictions using `Glove-6B` were significant (the $\Delta r = .13$ had a confidence interval that did not overlap with zero, CI_{95} [0.09, 0.19]). The same was true for `Glove-840B` ($\Delta r = 0.06$, CI_{95} [0.03, 0.10]). The pattern was the same for ABSTRACT words: Adding experiential information resulted in a significant improvement for both `GloVe-6B` ($\Delta r = .18$, CI_{95} [0.12, 0.24]) and `GloVe-840B` ($\Delta r = .17$, CI_{95} [0.12, 0.24]).

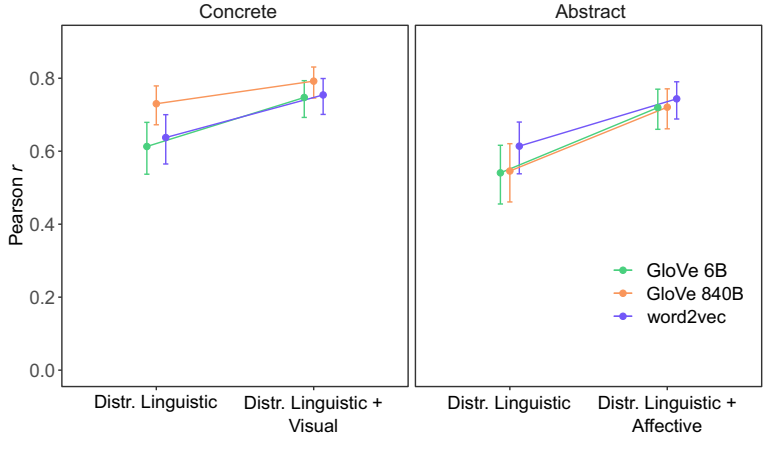

Fig. 3. Evaluation of alternative distributional linguistic models on CONCRETE and ABSTRACT words in the triad task. The figure shows the correlations and 95% confidence intervals for unimodal and multimodal (visual left, affective right panel) models using the standard `word2vec` based on 2B token corpus and two embeddings based on `GloVe`, trained on a corpus of either of 6B and 840B tokens.

To summarize, the unimodal results indicate that a very large corpus based on 840 billion (B) tokens improves performance for concrete items but results in lower correlations for abstract ones. This suggests that visual language about concrete entities might be relatively underrepresented in all but the largest corpora. The current word2vec model based on 2 billion (2B) performs favorably compared to the 6B GloVe embeddings trained on a corpus more similar in size. For the multimodal comparisons, regardless of the nature of the distributional linguistic model, adding experiential information improved performance. Overall, the findings are robust regardless of the corpus or embedding method used. In other words, the results for these distinct distributional linguistic models are not very different, despite both architectural (word2vec vs. GloVe) and corpus differences (2B words for the current corpus, 6B or 840B words used to train GloVe).

3.3.2. Word association model

There were fewer parameters and degrees of freedom in the word association model than the distributional linguistic model, since its behavior is determined by a single activation decay parameter α, which we set at 0.75 in line with previous work (De Deyne, Perfors, et al., 2016). However, this might have had some effect on model performance: Especially for basic-level comparisons, a high value of α might introduce longer paths, which might add more thematic information at the expense of shorter category-specific paths. For this reason, we also evaluated performance for other values of α.

Fig. 4 shows that performance was reasonably robust over all values of α. Smaller α values did occasionally improve the results, but only for the unimodal results with CONCRETE words: $r_{\alpha=0.05}(298) = .80$ compared to the default $r_{\alpha=0.75}(298) = .76$, $\Delta r = -.04$, CI_{95} [−0.07, −0.02]. For ABSTRACT words, the optimal performance was obtained when $\alpha = 0.85$ and was not significantly different from the default setting ($\alpha = 0.75$). One interpretation of this is that the representation of concrete words does not reflect indirect paths as much as the representation of abstract concepts. Regardless, this pattern suggests that even the modest improvement found when visual information was added to the word association model was probably somewhat overestimated relative to what would have been obtained using the optimal value for α.

4. Study 2: Pairwise similarity experiments

Study 1 evaluated the performance of distributional linguistic and word association models when participants perform relative similarity judgments between basic-level concrete or abstract words. To see if these findings generalize to a broader set of concepts and a different paradigm, we evaluated the same models on multiple datasets containing pairwise semantic similarity ratings, including some that were explicitly collected to compare language-based and (visual) multimodal models.

Unlike the CONCRETE triad task in Study 1, most of the existing datasets include a wide range of concrete semantic relations rather than just taxonomic basic-level ones. According to Rosch et al. (1976), "basic-level categories are shown to be the most inclusive

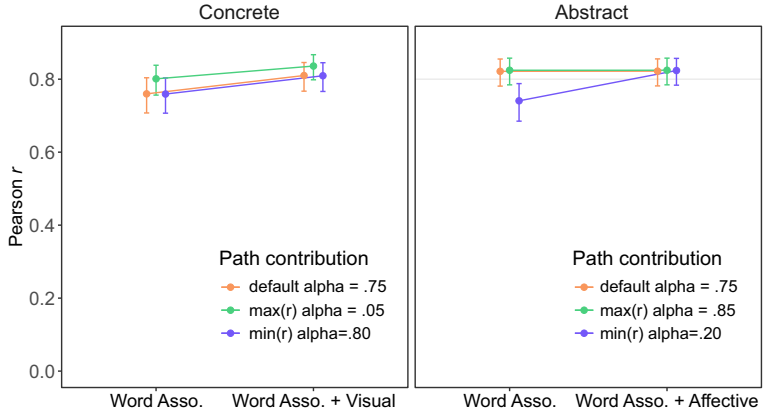

Fig. 4. Evaluation of alternative word association models on CONCRETE and ABSTRACT words in the triad task. The left panel multimodal model includes visual information; the right multimodal panel includes affective information. The length of the random walk was varied by setting α, and the maximal and minimal values of *r* were overall similar regardless of α.

categories for which a concrete image of the category as a whole can be formed and to be the first categorizations made during the perception of the environment" (p. 382). With this in mind, investigating the similarity of items from different categories (e.g., *butter—croissant*; *passion—justice*) might be a relatively insensitive way of gauging the effect of additional visual or affective information.

Fortunately, the large size of previously published datasets allows us to impose restrictions and compare items where both words belong to the same basic-level category, as well as to evaluate only abstract concepts. By comparing concepts of different types, we will be able to investigate whether the results from Study 1 only apply to concrete concepts on the basic level or whether they generalize further. In addition, while most of the new datasets contain primarily concrete nouns, some of them include a sufficient number of abstract words as well. Given the finding in Study 1 that affective information is important to the representation of those concepts, it is essential to determine whether this finding replicates and generalizes to different tasks.

4.1. Datasets

We consider five different datasets of pairwise similarity ratings.[8] Three of these datasets, the MEN data (Bruni, Uijlings, Baroni, & Sebe, 2012), the MTURK-771 data (Halawi, Dror, Gabrilovich, & Koren, 2012), and the SimLex-999 data (Hill, Reichart, & Korhonen, 2016), are commonly used as a general benchmark for semantic and distributional models. Two more recent datasets were additionally included because they allow us to more directly address the role of visual and affective information. The first one was the Silberer2014 dataset (Silberer & Lapata, 2014), which was collected with the specific purpose of evaluating visual and semantic similarity in multimodal models. The second dataset was SimVerb-3500 (Gerz, Vulić, Hill, Reichart, & Korhonen, 2016), which

contains a substantial number of abstract verbs. It thus allows us to extend our findings beyond concrete nouns to verbs and investigate whether the important role of affective information in abstract concepts found in Study 1 replicates here. Each of the datasets is slightly different in terms of procedures, stimuli, and semantic relations of the word pairs being judged. The next section briefly explains these differences and reports on their internal reliability, which sets a bound on the maximal correlation of the models we want to evaluate.[9]

4.1.1. MEN

The MEN dataset (Bruni, Uijlings, et al., 2012) consists of 3,000 word pairs constructed specifically for testing multimodal models, and thus most words were concrete. The words were selected randomly from a subset of words occurring at least 700 times in the ukWaC and WaCkypedia corpus. Next, semantic vectors derived from these corpora models were used to derive cosine values from the first 1,000 most similar items, 1,000 pairs were sampled from the 1,001–3,000 most similar items, and the last 1,000 items from the remaining items. As a result, the MEN consists of concrete words that cover a wide range of semantic relations. The estimated reliability that serves as an upper bound was $\rho = 0.84$ (Bruni, Uijlings, et al., 2012).

4.1.2. MTURK-771

The MTURK-771 dataset (Halawi et al., 2012) consists of 771 word pairs and was constructed to include various types of relatedness. It consists of frequent nouns taken from WordNet that are synonyms, have a meronymy relation (e.g., *leg—table*), or a holonymy relation (e.g., *table—furniture*). The authors converted WordNet to an undirected graph and only included words with graph distances between 1 and 4. The variability of distance and type of relation suggests that this dataset is quite varied. The reliability, calculated as the correlation between randomly split subsets, was 0.90.

4.1.3. SimLex-999

The SimLex-999 dataset (Hill et al., 2016) consists of 999 word pairs and is different from all other datasets in that participants were explicitly instructed to ignore (associative) relatedness and only judge "strict similarity." It also differs from previous approaches by using a more principled selection of items consisting of adjective, verb, and noun concept pairs covering the entire concreteness spectrum. A total of 900 word pairs were selected from all associated pairs in the USF association norms (Nelson et al., 2004) and supplemented with 99 unassociated pairs. None of the pairs consisted of mixed part-of-speech. In this task, associated non-similar pairs in this list would receive low ratings, whereas highly similar (but potentially weakly associated) items would be rated highly. The reported average pairwise agreement was higher for abstract than concrete concepts ($\rho = 0.70$ vs. $\rho = 0.61$). This is relevant for current purposes as a separate analysis for abstract concepts is reported below. The overall inter-rater reliability calculated over split-half sets was 0.78 (Gerz et al., 2016).

4.1.4. Silberer

The Silberer dataset (Silberer & Lapata, 2014) consists of all possible pairings of the nouns present in the McRae et al. (2005) concept feature norms. For each of the words, 30 randomly selected pairs were chosen to cover the full variation of semantic similarity. The resulting set consisted of 7,569 word pairs. In contrast to previous studies, the participants performed two rating tasks, consisting of both visual and semantic similarity judgments. The inter-rater reliability, calculated as the average pairwise ρ between the raters, was 0.76 for the semantic judgments and 0.63 for the visual judgments.

4.1.5. SimVerb-3500

The SimVerb-3500 dataset (Gerz et al., 2016) consists of 3,500 word pairs and was constructed to remedy the bias in the field toward studying nouns, and thus consists of an extensive set of verb ratings. Like the SimLex-999 dataset, it was designed to be representative in terms of concreteness and constrained the judgments explicitly by asking participants to judge similarity rather than relatedness. The items were selected from the University of South Florida association norms and the VerbNet verb lexicon (Kipper, Snyder, & Palmer, 2004), which was used to sample a large variety of classes represented in VerbNet. Inter-rater reliability obtained by correlating individuals with the mean ratings was high ($\rho = 0.86$). Note that since the visual feature model trained on ImageNet only contains nouns, SimVerb-3500 cannot be used to study the impact of visual information.

4.2. Evaluation of multimodal visual models

The similarity judgments are predicted by the word2vec distributional linguistic model and word association model, each respectively combined with the visual information to create the multimodal model. As in Study 1, the relative contribution of either the distributional linguistic word associations representation versus visual information will be determined by the best fit of the mixing parameter β.

4.2.1. Datasets involving diverse semantic comparisons

We first consider the three datasets corresponding to a mixed list of word pairs covering a variety of taxonomic and relatedness relations. Of these, the MEN and MTURK-771 datasets are most similar to each other, since they consist of pairs that include both similar and related pairs across various levels of the taxonomic hierarchy. The first set of comparisons involves multimodal models with visual information, and consequently, the analyses are restricted to those pairs that are present in the linguistic, word association, and visual models. The results reveal that adding visual information only improved the word association model for the MEN data but not for MTURK-771 (Δr respectively .02 CI_{95} [0.01, 0.02] and .00, CI_{95} [−0.01, 0.02]). However, visual information significantly improved the distributional linguistic model on both datasets (Δr respectively .03 CI_{95} [0.02, 0.04] and .04 CI_{95} [0.01, 0.08]).

The SimLex-999 dataset is slightly different than all others, in that participants were explicitly instructed to *only* evaluate strict similarity, ignoring any kind of associative

relatedness between the two items. The unimodal word association model performed far better than the distributional linguistic model on this dataset ($r(298) = .72$ vs. .43, $\Delta r = .29$, CI_{95} [0.22, 0.48]). Adding visual information did not improve the performance of the association model but resulted in considerable improvement in the distributional linguistic model ($r(298) = .14$, CI_{95} [0.07, 0.21]). A potential explanation for the difference between datasets is that SimLex-999 focused on strict similarity, whereas MEN and MTURK-771 covered a broader range of semantic relations, including thematic ones, for which visual similarity is of limited use. Note the relative small number of cases ($n = 300$) in SimLex-999. Table 2 reflects the fact that abstract word pairs are not encoded in the visual model, whereas most items are encoded in the affective model in Table 4.

Of the remaining datasets, the Silberer one is most directly relevant to evaluate the role of visual information. The words in this study consisted of concrete nouns taken from the McRae feature generation norms (McRae et al., 2005). However, since words across different categories were compared, the ratings occur between entities specified at different taxonomic levels. This might provide less of a challenge for distribution-based language models to predict due to large perceptual differences. In contrast to the other studies, two types of ratings were collected—semantic and visual. Semantic ratings are similar to

Table 2
Pearson correlation and confidence intervals for correlation differences Δr between unimodal and multimodal **visual** models. The top part of the table shows the results for the distributional linguistic model, whereas the bottom part shows the results for the word association model

		v_a = Distributional linguistic, v_b = Visual								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MEN	942	.79	[0.77, 0.82]	.66	[0.62, .70]	0.38	.82	[0.80, 0.84]	.03	[0.02, 0.04]
MTURK-771	260	.67	[0.59, 0.73]	.49	[0.39, 0.58]	0.38	.71	[0.64, 0.76]	.04	[0.01, 0.08]
SimLex-999	300	.43	[0.33, 0.52]	.54	[0.45, 0.61]	0.55	.56	[0.48, 0.64]	.14	[0.07, 0.21]
Silberer (Sem.)	5,799	.73	[0.71, 0.74]	.78	[0.77, 0.79]	0.53	.82	[0.82, 0.83]	.10	[0.09, 0.10]
Silberer (Vis.)	5,777	.59	[0.57, 0.61]	.74	[0.73, 0.75]	0.65	.75	[0.74, 0.76]	.16	[0.15, 0.17]
Average		.55		.55			.64		.08	
		v_a = Word association, v_b = Visual								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MEN	942	.79	[0.76, 0.81]	.66	[0.62, 0.70]	0.45	.81	[0.78, 0.83]	.02	[0.01, 0.02]
MTURK-771	260	.74	[0.68, 0.79]	.49	[0.39, 0.58]	0.33	.75	[0.69, 0.79]	.00	[−0.01, 0.02]
SimLex-999	300	.72	[0.66, 0.77]	.54	[0.45, 0.61]	0.43	.73	[0.68, 0.79]	.01	[−0.01, 0.03]
Silberer (Sem.)	5,799	.84	[0.83, 0.85]	.78	[0.77, 0.79]	0.58	.87	[0.86, 0.88]	.03	[0.03, 0.04]
Silberer (Vis.)	5,777	.73	[0.72, 0.74]	.74	[0.73, 0.75]	0.65	.79	[0.78, 0.80]	.06	[0.05, 0.07]
Average		.71		.55			.73		.02	

other studies and involve participants judging similarity. Visual ratings, however, only involve judging the similarity of the appearance of the concepts. One might thus expect that models based on visual features will predict visual similarities better than semantic ratings.

The results indicate that for the word association model, adding visual information resulted in significant improvement; however, the improvements were small (relative to the distributional model) even for the visual ratings ($\Delta r = .03$, CI_{95} [0.03, 0.04] for the semantic judgments and $\Delta r = .06$, CI_{95} [0.05, 0.07] for the visual judgments). For the distributional linguistic model, adding visual information resulted in an improved prediction for both judgments, especially the visual ones ($\Delta r = .10$, CI_{95} [0.09, 0.10], for the semantic judgments and $\Delta r = .16$, CI_{95} [0.15, 0.17] for the visual judgments). Consistent with both of these sets of findings, the word association model better predicts people's similarity judgments than the distributional linguistic judgments (the difference was .11 CI_{95} [0.10, 0.12] for the semantic and .14, CI_{95} [0.13, 0.16] for the visual judgments).

To summarize, the contribution of visual information in the multimodal word association model was relatively small compared to the distributional linguistic model. The average gain across datasets was .02 for the former compared to .08 for the latter (see Table 2).

4.2.2. Comparisons at the basic level

The complete Silberer dataset that we evaluated above contains both basic-level within-category comparisons like *dove—pigeon* as well as broader comparisons across categories like *dove—butterfly*. Since Study 1 involved only basic-level comparisons within superordinates, in order to provide a better comparison between it and the Silberer data, we did an additional analysis only on the basic-level terms in the Silberer data. To achieve this, we annotated the Silberer2014 dataset with superordinate labels for common categories taken from Battig and Montague (1969) and De Deyne et al. (2008) such as *bird*, *mammal*, *musical instrument*, *vehicle*, *furniture*, and so on. We then restricted the analysis to only the basic-level terms within those superordinate categories. Words for which no clear common superordinate label (see above) could be assigned were not included. This reduced the number of pairwise comparisons from 5,799 to 1,086, which is still sufficiently large for the current purposes.

As the two right panels of Fig. 5 demonstrate, the overall correlations were lower for basic-level terms than those for the complete set of items. For the word association model, the difference was respectively .19, CI_{95} [0.16, 0.23] and .21, CI_{95} [0.17, 0.26]. For the distributional linguistic model, the results were respectively a difference of .27, CI_{95} [0.22, 0.34] and .24, CI_{95} [0.19, 0.30]. These results support the idea that compared to mixed lists of items, comparisons at the basic level present a considerable challenge. Moreover, adding visual information results in a larger improvement (relative to the unimodal model) for basic-level items. For the semantic comparisons using distributional linguistic representations, the multimodal difference Δr was .10 for the full datasets versus .16 for the basic-level data (see Tables 2 and 3). A bootstrapped test confirmed that the

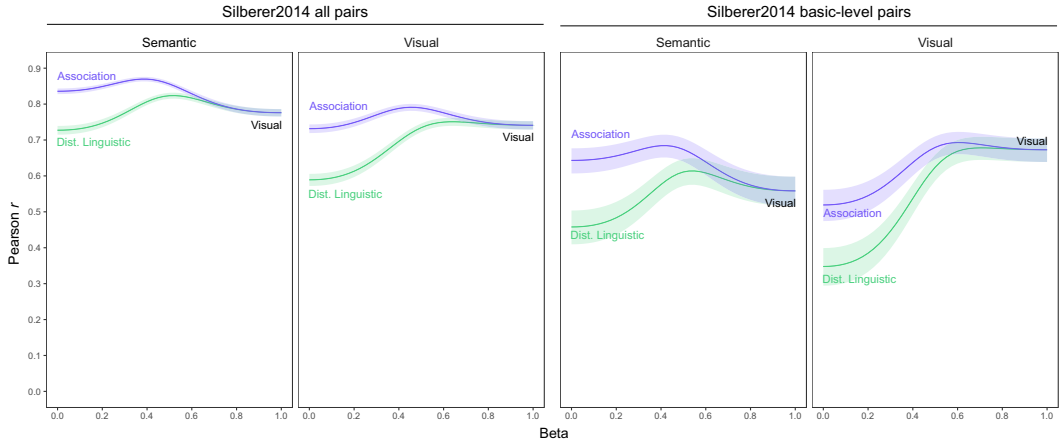

Fig. 5. Results of multimodal models (created by combining either distributional linguistic or word association models with visual features) based on pairwise similarity ratings from the Silberer dataset. The plots show correlations between human judgments and models together with 95% confidence intervals (shaded). The two left panels show the semantic and visual judgments for all items, whereas the two right panels show performance on the subset of basic-level items in which the word pairs belong to the same superordinate category.

difference between the full and basic-level correlation was significant, CI_{95} [−0.03, −0.09].

Finally, in line with Study 1, we found that the visual information improved performance more for the distributional linguistic model than the word association model: The difference between $\Delta r = .16$ and $\Delta r = .04$ was significant, CI_{95} [0.10, 0.16]. For completeness, Table 3 and Fig. 5 also show the results for the visual judgments. The main finding here is that when comparisons are restricted to visual judgments at the basic level, the unimodal visual model (when $\beta = 1.0$) is clearly superior to both the word association ($r_{v_b} - r_{v_a} = .15$, CI_{95} [0.11, 0.20]) and distributional linguistic ($r_{v_b} - r_{v_a} = .33$, CI_{95} [0.27, 0.37]) model, in line with earlier argument that basic-level comparisons rely strongly on visual information.

4.3. Evaluation of multimodal affective models

To investigate the effect of affective information, we supplemented the datasets in the previous section with SimVerb-3500 (Gerz et al., 2016), which contained pairwise similarity ratings for verbs as described before. Most words in SimVerb-3500 (2,926 out of 3,500) were included in the affective norms of Warriner et al. (2013). The results are shown in Table 4. In contrast to the findings for visual information and the ABSTRACT triads in Study 1, most of our datasets show no improvement when affective information is added. The only exceptions are the SimLex-999 and SimVerb-3500 datasets, where adding affective information improved the distributional linguistic model, $\Delta r = .07$, CI_{95} [0.04, 0.10] and $\Delta r = .11$, CI_{95} [0.008, 0.13], respectively.

Table 3
Replication of Table 2 restricted to basic-level word pairs in the Silberer dataset

		v_a = Distributional linguistic, v_b = Visual								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
Basic-Sem.	1,086	.46	[0.41, 0.50]	.56	[0.52, 0.60]	0.55	.61	[0.58, 0.65]	.16	[0.12, 0.19]
Basic-Vis.	1,086	.35	[0.29, 0.40]	.67	[0.64, 0.70]	0.70	.68	[0.64, 0.71]	.33	[0.28, 0.38]
		v_a = Word association, v_a = Visual								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
Basic-Sem.	1,086	.64	[0.61, 0.68]	.56	[0.52, 0.60]	0.43	.68	[0.65, 0.71]	.04	[0.03, 0.06]
Basic-Vis.	1,086	.52	[0.47, 0.56]	.67	[0.64, 0.70]	0.55	.69	[0.66, 0.72]	.17	[0.14, 0.21]

Again we find that the word association model provides better estimates of pairwise similarity than the distributional linguistic model, except for the MEN dataset, where they are on par (CI_{95} [−0.01, 0.02]). The unimodal word association correlation for the additional SimVerb-3500 dataset, $r(2924) = .64$, was similar in magnitude to the SimLex-999 task ($r(911) = .68$) and was somewhat lower than the other datasets. For the distributional linguistic model, the SimVerb-3500 correlations was also moderate, $r(2924) = .33$, which might reflect the difficulty of this model in handling strict similarity and accurately representing verb meaning.

To conclude, on average the multimodal affective gain was limited in both the distributional linguistic and word association model (respectively .03 and .02, see Table 4). These results are different from Study 1, but there are at least two reasons why the affective information may have provided less benefit in this study. First, not all datasets here involved comparisons between basic-level items within a superordinate category. Second, none of the datasets were constructed to investigate abstract words, and it is for these that we might expect affective information to be most important.

4.3.1. The role of affect in abstract words

Unlike the abstract condition in Study 1, some of the pairwise datasets contain both concrete and abstract words. To test whether the presence of concrete words masked the effect of the multimodal affective model, we screened how abstract the stimuli were in each dataset using the concreteness norms from Brysbaert et al. (2014). As in Study 1, we only included similarity judgments for which the average concreteness rating of both words in the pair was smaller than 3.5 on a 5-point scale. MEN only had 41 pairs (2.1% of words) and no pairs matched in the Silberer dataset. The MTURK-777, SimLex-999, and SimVerb-3500 datasets had a reasonable number of abstract words (respectively 121 or 18.6%, 42.8% or 391 words, and 67.4% or 1,973 words) and for these datasets we tested whether adding affective information results in a larger improvement (relative to

Table 4
Pearson correlation and confidence intervals for correlation differences Δr between unimodal and multimodal **affective** models. The top part of the table shows the results for the distributional linguistic model, whereas the bottom part shows the results for the word association model

		v_a = Distributional linguistic, v_b = Affect								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MEN	1,981	.80	[0.78, 0.81]	.31	[0.27, 0.35]	0.45	.80	[0.78, 0.81]	.00	[0.00, 0.00]
MTURK-771	653	.70	[0.66, 0.74]	.26	[0.19, 0.33]	0.53	.71	[0.67, 0.75]	.01	[0.00, 0.02]
SimLex-999	913	.45	[0.39, 0.50]	.33	[0.27, 0.39]	0.65	.63	[0.47, 0.56]	.07	[0.04, 0.10]
SimVerb-3500	2,926	.33	[0.30, 0.36]	.33	[0.30, 0.36]	0.68	.44	[0.41, 0.47]	.11	[0.08, 0.13]
Silberer (Sem.)	5,428	.74	[0.73, 0.75]	.21	[0.19, 0.24]	0.33	.74	[0.73, 0.76]	.00	[0.00, 0.00]
Silberer (Vis.)	5,405	.60	[0.58, 0.61]	.16	[0.14, 0.16]	0.30	.60	[0.58, 0.61]	.00	[0.00, 0.00]
Average		.60		.27			.63		.03	
		v_a = Word association, v_b = Affect								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MEN	1,981	.80	[0.78, 0.82]	.31	[0.27, 0.35]	0.45	.81	[0.79, 0.82]	.01	[0.00, 0.01]
MTURK-771	653	.77	[0.73, 0.80]	.26	[0.19, 0.33]	0.45	.77	[0.74, 0.80]	.00	[0.00, 0.01]
SimLex-999	913	.68	[0.65, 0.71]	.33	[0.27, 0.39]	0.50	.69	[0.65, 0.72]	.01	[0.00, 0.01]
SimVerb-3500	2,926	.64	[0.62, 0.66]	.33	[0.30, 0.30]	0.50	.65	[0.63, 0.67]	.01	[0.01, 0.02]
Silberer (Sem.)	5,428	.84	[0.84, 0.85]	.21	[0.19, 0.24]	0.23	.84	[0.84, 0.85]	.00	[0.00, 0.00]
Silberer (Vis.)	5,405	.73	[0.72, 0.74]	.16	[0.14, 0.19]	0.00	.73	[0.72, 0.74]	.00	[0.00, 0.00]
Average		.71		.27			.73		.02	

the unimodal model) for abstract pairs compared to all pairs. The results are shown in Table 5 and Fig. 6.

For the word association model, the multimodal gain for all pairs was $\Delta r = .00$ for MTURK-771 and $\Delta r = .01$ for both SimLex-999 and SimVerb-3500 (see Table 4). These values were smaller than the corresponding Δrs of .01, .06, and .04 for the subset of abstract items of the respective datasets (see Table 5). A bootstrapped test confirmed that the difference between the correlation for all versus abstract items using the word association model was significant for both SimLex-999, CI_{95} [−0.07, −0.02] and SimVerb-3400, CI_{95} [−.02, −.00] but not for MTURK-771, CI_{95} [−0.02, 0.01].

For the distributional linguistic model, the multimodal gain was $\Delta r = .01$ for MTURK-771, $\Delta r = .07$ for SimLex-999, and $\Delta r = .11$ for SimVerb-3500 for all pairs (see Table 4). These values were smaller than the corresponding Δrs of .02, .21, and .19 for the subset of abstract items of the respective datasets (see Table 5). A bootstrapped test confirmed that the difference between the correlation for all versus abstract pairs was significant for both SimLex-999, CI_{95} [−0.03, −0.09] and SimVerb-3500, CI_{95} [−0.03, −0.09], but not for MTURK-771, CI_{95} [−0.03, 0.02].

Table 5
Replication of the results reported in Table 4 restricted to **abstract** word pairs in the MTURK-771, SimLex-999, and SimVerb-3500 datasets

		v_a = Distributional Linguistic, v_b = Affect								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MTURK-771	121	.67	[0.56, 0.76]	.31	[0.14, 0.46]	0.53	.70	[0.59, 0.78]	.02	[−0.01, 0.06]
SimLex-999	336	.43	[0.34, 0.52]	.51	[0.42, 0.58]	0.68	.64	[0.58, 0.70]	.21	[0.14, 0.28]
SimVerb-3500	1,466	.28	[0.23, 0.32]	.40	[0.36, 0.45]	0.73	.47	[0.43, 0.51]	.19	[0.15, 0.23]
Average		.46		.41			.60		.14	
		v_a = Word association, v_b = Affect								
		Unimodal				Multimodal				
Dataset	n	r_{v_a}	CI_{95}	r_{v_b}	CI_{95}	β	$r_{v_{ab}}$	CI_{95}	Δr	CI_{95}
MTURK-771	121	.77	[0.69, 0.83]	.31	[0.14, 0.46]	0.45	.78	[0.70, 0.84]	.01	[−0.02, 0.04]
SimLex-999	336	.67	[0.61, 0.73]	.51	[0.42, 0.58]	0.58	.74	[0.69, 0.78]	.06	[0.03, 0.10]
SimVerb-3500	1,466	.65	[0.62, 0.68]	.40	[0.36, 0.45]	0.53	.69	[0.66, 0.71]	.04	[0.03, 0.05]
Average		.70		.41			.71		.04	

Finally, in line with Study 1, we found that the affective information improved performance more for the distributional linguistic model than the word association model in all datasets but MTURK-771. For SimLex-999 the difference between $\Delta r = .06$ (associations) and $\Delta r = .21$ (distributional linguistic) was significant, CI_{95} [0.08, 0.21], and so was the difference for SimVerb-3500, $\Delta r = .04$ (associations) and $\Delta r = .19$, CI_{95} [0.11, 0.19] (distributional linguistic).

So far, the role of affective multimodal information in abstract words was consistent with the predicted pattern of a relatively larger gain for the distributional linguistic model. However, the current findings for MTURK-771 indicate only a small multimodal gain. While the number of abstract stimuli was not only smaller in MTURK-771 compared to the other two datasets, further inspection also showed that the stimuli might have been less varied. If the pairs are more neutral in terms of affect, then this might account for the relatively small gain in the multimodal affective model. To verify whether this was the case, we compared the distribution of valence in Warriner et al. (2013) with the three datasets. For all 13,915 normed words in the Warriner norms, the 9-point scale valence ratings quantiles were 1.26 (0%), 4.25 (25%), 5.20 (50%), 5.95 (75%), and 8.53 (100%). For MTURK-771, 91.7% of words fell in the middle range (25%–75%), whereas for Simlex-999 and SimVerb-3500 this was respectively 65.4% and 44.8% for SimVerb-3500. This suggests that the relative multimodal gain in MTURK-771 might be due to a restricted subset of abstract items in terms of emotion.

In line with the averages across datasets in Table 6, we conclude that the role of affective information in multimodal models was consistent with Study 1 and extends its results to a different task and a more varied set of words, including verbs.

4.4. Robustness

As in Study 1, we here perform a series of additional analyses to evaluate the extent to which our results depend on specific modeling choices. We focused on alternative distributional linguistic models and report the results for GloVe trained on 6 billion (GloVe-6B) and 840 billion GloVe-840B tokens. If the results from Study 1 replicate, we would expect performance to be similar to the distributional linguistic model based on word2vec, except when the corpus size is extremely large. In that case, we

Table 6
Comparison between Study 2 results and previously published studies. The correlations reported are between the human ratings and either the unimodal distributional linguistic model in question (v_a), the visual model in that study (v_b), and the multimodal model that combines both v_a and v_a (v_{ab})

Silberer2014 (Semantic judgments)

Study	v_a description	v_b description	r_{v_a}	r_{v_b}	$r_{v_{ab}}$
Silberer et al. (2013)	feature ratings	visual attributes ImageNet	.71	.49	.68
Lazaridou et al. (2015)	word2vec	CNN-features	.62	.55	.72
De Deyne, Navarro, Collell and Perfors (2020)	word2vec	CNN-features	.73	.78	.82
De Deyne et al. (2020)	word associations	CNN-features	.84	.78	.87

Silberer2014 (Visual judgments)

Study	v_a description	v_b description	r_{v_a}	r_{v_b}	$r_{v_{ab}}$
Silberer et al. (2013)	feature ratings	visual attributes ImageNet	.58	.52	.62
Lazaridou et al. (2015)	word2vec	CNN-features	.48	.54	.63
De Deyne et al. (2020)	word2vec	CNN-features	.59	.74	.75
De Deyne et al. (2020)	word associations	CNN-features	.73	.74	.79

MEN

Study	v_a description	v_b description	r_{v_a}	r_{v_b}	$r_{v_{ab}}$
Bruni et al. (2014)	count model	Bag-of-visual features	.73	.43	.78
Kiela and Bottou (2014)	word2vec	SIFT features	.62	.40	.70
Kiela and Bottou (2014)	word2vec	CNN-features	.62	.63	.72
Lazaridou et al. (2015)	word2vec	CNN-features	.68	.62	.76
De Deyne et al. (2020)	word2vec	CNN-features	.79	.66	.82
De Deyne et al. (2020)	word associations	CNN-features	.79	.66	.81

SimLex-999

Study	v_a description	v_b description	r_{v_a}	r_{v_b}	$r_{v_{ab}}$
Lazaridou et al. (2015)	word2vec	CNN-features	.29	.54	.53
De Deyne et al. (2020)	word2vec	CNN-features	.43	.54	.56
De Deyne et al. (2020)	word associations	CNN-features	.72	.54	.73

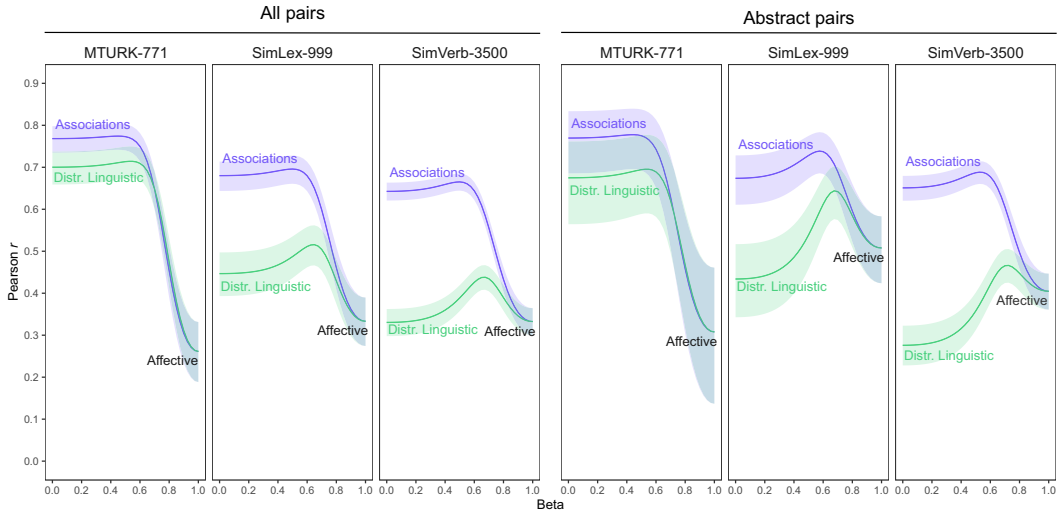

Fig. 6. Investigation of the role of affective information for abstract words comparing the full set (all pairs) with a subset of abstract words taken from the SimLex-999 and SimVerb-3500 datasets. Results qualitatively replicate the finding in Study 1 that adding affective information improves the performance of the distributional linguistic model but not the word association model when the abstractness of the words is considered.

would expect performance to be improved for comparisons that involve more visual categories.

The results, shown in Fig. 7A, are consistent with this: The overall correlations using the 6 billion word GloVe-based linguistic model ($r = .60$ on average) and the one using 840 billion words ($r = .66$ on average) were similar to those obtained using the word2vec-based one ($r = .64$ on average). When visual information was added, correlations were virtually identical for all three distributional linguistic models (between $r = .73$ and $r = .74$).

In Study 1, the GloVe-based models had the best performance for concrete words when based on extremely large corpora (840B words), which suggested that the improved quality of the model primarily reflected taking advantage of the information contained in such large texts rather than the specific embedding technique or parameters. This is consistent with previous research indicating that GloVe improves with size (Pennington et al., 2014). Here as well, we find that comparisons at the basic level, which presumably encode more perceptual properties, were better predicted with a larger corpus (see Fig. 7 C). In line with Study 1, the distributional linguistic model based on the larger corpus did not lead to substantial improvements for abstract concepts (Fig. 7D) compared to the 6B corpus model. Regardless of the model, correlations improved markedly when affective information was included (Δr between .18 and .19 for all three models). This is consistent with our initial findings of Study 1 ($\Delta r = .13$) suggesting that distributional linguistic models specifically lack the affective information that is important for representing abstract concepts.

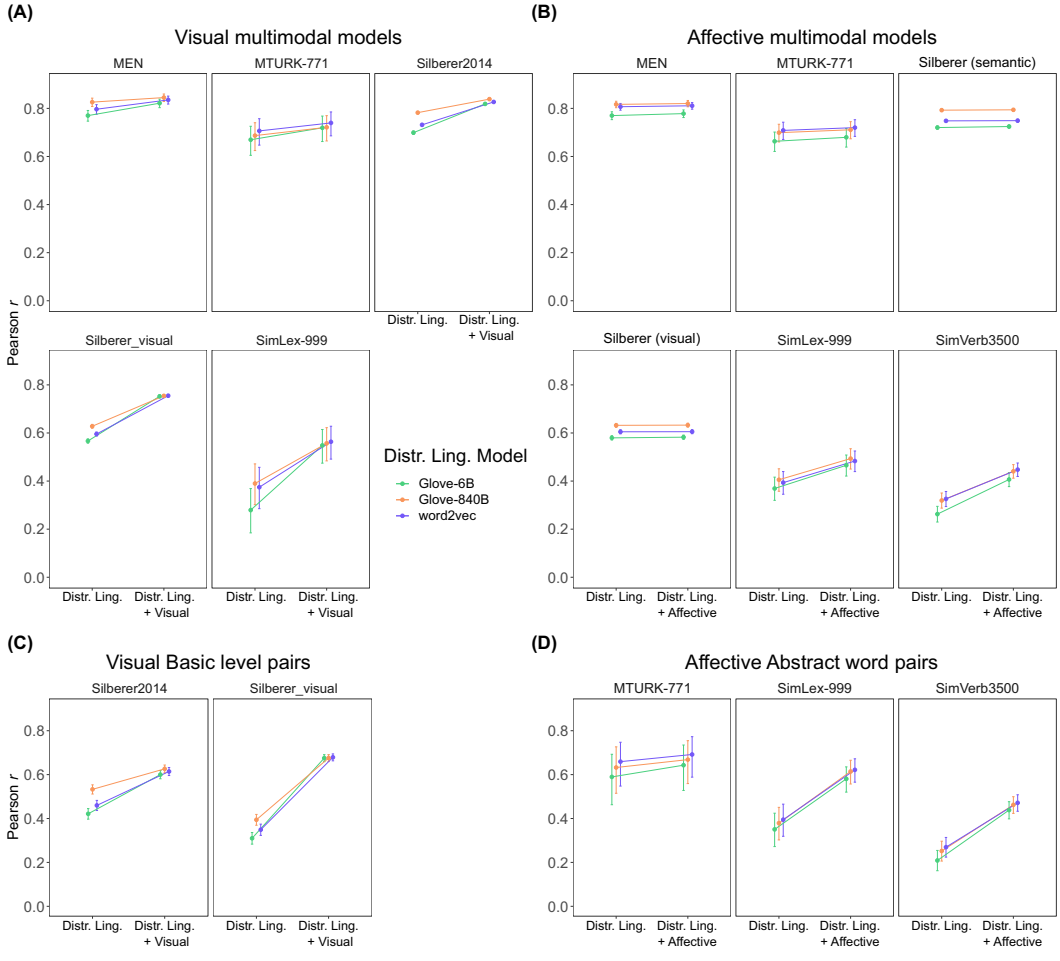

Fig. 7. A comparison of the unimodal and multimodal model performance in three different distributional linguistic models: `word2vec` and `GloVe-6B` (6 billion words) and `GloVe-840B` (840 billion words). Multimodal models reflect the optimized Pearson correlation for mixing parameters β. Top panels show Pearson correlations and 95% confidence intervals for the visual multimodal models (Panel A) and affective multimodal models (Panel B). The bottom panels show the findings when considering a subset of concrete word pairs at the basic level (Panel C) and abstract affective words (Panel D).

4.5. Comparison to previous work

Because the datasets in Study 2 have appeared in several recent studies that investigated multimodal visual models, we can compare between our results and theirs.[10] An overview of the results is shown in Table 6. For the purpose of current comparisons, we focus on the findings reported in four studies: Silberer et al. (2013), Bruni et al. (2014), Lazaridou et al. (2015), and Kiela and Bottou (2014). Each of these studies compared both unimodal distributional linguistic and visual feature models with multimodal models that combined both types of information.

In the study by Silberer et al. (2013), multimodal representations were derived by combining a distributional model induced from human feature generation norms with a visual model that was trained to predict visual attributes for the corresponding images in ImageNet. The feature norms are of direct interest because they provide an alternative experimental measure to contrast with the word association model. Feature norms allow more precise propositional statements that capture something about the semantic relationships (a duck <has a> bill, <is a> bird, etc.). Moreover, in contrast to word associations, the instructions in feature-rating tasks appeal to core objective properties that *define* meaning and exclude affective and attitudinal aspects of meaning (De Deyne, Verheyen, Navarro, Perfors, & Storms, 2015).

The second study by Bruni et al. (2014) consists of a count-based linguistic model using a sliding window and a large text corpus constructed from Wikipedia and other material taken from the internet. For the visual information, a scale-invariant image features transformation (SIFT) was used to extract features, which were then treated as visual "words" in a bag-of-words representation. This *bag-of-visual words* approach was applied to an image corpus of 100K labeled images derived from the ESP-Game data set (Von Ahn, 2006).

The third study by Kiela and Bottou (2014) used a distributional linguistic model that was trained on Wikipedia and relied on the skip-gram `word2vec` model. Two different methods were considered for the visual features: One consisted of a bag-of-visual words approach with SIFT features, and the other used a CNN to extract features from ImageNet.

The last study by Lazaridou et al. (2015) provides an even closer comparison to the current study since it uses a CNN approach for the visual features and skip-gram for the distributional linguistic representations. Although their linguistic model is unlikely to reflect typical language-exposure—it is based on the entire English Wikipedia—it provides an approximation of what can be encoded through a slightly less natural language register.

As can be seen from Table 6, our results are competitive for both the unimodal distributional linguistic and visual feature models as well as the multimodal combination of both models. We obtained correlations as high and often higher than previous work. While we cannot rule out the possibility that a distributional linguistic model with different parameters trained on the same input might provide higher correlations and potentially reduce the multimodal benefit, the overall pattern of results is fairly robust to choices of parameters, distributional linguistic model, and corpus. Finally, in the case of word associations, the correlations are higher than studies that relied on feature ratings used by Silberer et al. (2013), which are considered a gold standard.

5. General discussion

In two studies, we investigated how linguistic and multimodal representations of meaning that include experiential features (visual and affective) can capture human semantic

judgments of abstract and concrete concepts. In both studies, we find that multimodal representations, which combine linguistic with affective or visual information, better predict human similarity judgments. Our findings replicate and extend previous work addressing visual multimodal models (Bruni et al., 2014; Kiela & Bottou, 2014; Lazaridou et al., 2015; Silberer & Lapata, 2014) and identified a novel and substantial effect of affective information for abstract concepts. Our results also identify systematic factors that determine to what degree experiential information improves on the performance of unimodal models.

5.1. What factors determine the performance in multimodal models?

A first factor that determines the degree to which multimodal models improve over unimodal ones is the nature of the experiential modality. In this study, we focused on visual and affective experiential information. In a series of original experiments, we found that adding experiential information to a distributional linguistic model improved its correlation with human ratings by .12 for concrete words (for visual information) and .13 for abstract words (for affective information). In Study 2, these findings were replicated using an extensive set of previously published similarity ratings. These results indicate that the need to supplement distributional linguistic models with affective information is especially important for abstract concepts.

The second factor is the level of comparison: The performance of the models is quite comparable when the semantic relationship extends beyond the basic level, but becomes more differentiated at the basic level. This suggests a potential shortcoming of existing benchmarks: Since they evaluate concepts across taxonomic levels, they might underestimate the relative improvement offered by multimodal models over unimodal ones. Indeed, when we reanalyzed the largest available dataset (Silberer & Lapata, 2014) from Study 2 to focus only on comparisons between basic-level concepts, the visual multimodal model showed much larger improvements than when the full item set was included.

A third factor is the nature or modality of the experiential models. The current results hinge on the visual and affective representations we derived. Given the recency of the models we used, further improvements are to be expected, which could indicate that perceptual and affective information is even more important than that estimated here. For example, in the case of affect, the Osgood model based on three dimensions remains somewhat coarse due to its low-dimensional nature. Instead of three factors, richer representations of affect have also been proposed. This includes properties about social interactions (Barsalou & Wiemer-Hastings, 2005), morality (Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005), and emotions. For example, Ekman (1992) distinguishes six basic emotions (*joy*, *sadness*, *anger*, *fear*, *disgust*, and *surprise*), whereas Plutchik (1994) extends this list with *trust* and *anticipation* as well. Some of these emotional features were included in recent studies to map the meaning of abstract words. One example is the work by Crutch, Troche, Reilly, and Ridgway (2013) in which an abstract conceptual feature rating task was used where participants judged abstract words on nine cognitive dimensions using a Likert-like scale. To explore properties beyond core affect, we ran a

preliminary analysis to investigate this possibility using the NRC Emotion lexicon, which contains judgments for the Ekman emotions for over 14,000 English words (Mohammad & Turney, 2013). We found very limited evidence for any contribution of emotions above that of affect: They did not capture the similarities derived from Study 1 or the rated similarity datasets in Study 2 as well, despite having more features. Finally, if we look at the absolute performance, we see that the correlations are high for some datasets, suggesting that room for improvement is somewhat limited.

A fourth factor that needs to be considered is the nature of the distributional linguistic model. Our work relies on the assumption that the current models are reasonable approximations of what meaning can be derived from text-based linguistic corpora.

We derived these models according to both theoretical considerations (are they appropriate given the words a human knows and is exposed to across the lifespan?) as well as empirical ones (are the models on par with those reported in the literature?). It remains possible that better distributional linguistic models might capture more visual or affective information, which would reduce the difference with a multimodal model. However, in both Study 1 and Study 2, our results for the most part did not depend on the size of the corpus or the way the embeddings were obtained; the only (relatively minor) improvements occurred when using extremely large corpora outstretching human information processing capacities. Furthermore, comparing our results to previous work suggests that both our distributional linguistic models and the visual feature models are representative of the current state of the art.

Finally, there are a host of task and other stimuli factors that may have influenced our results and could be considered in future work. For example, part of speech can strongly determine how well different models perform: We observed large differences between word association and distributional linguistic model performance based on whether the items were verbs or not (see Table 4). However, it is unlikely that our main findings are an artifact of the procedure or the specific stimuli. Study 2 demonstrated that the same qualitative patterns emerge when considering different tasks, stimuli, and procedures. This includes similarity judgments in triadic comparisons for basic-level categories, relatedness judgments for a variety of semantic relations, strict similarity judgments in which participants were to ignore any kind of relatedness, and visual judgments instead of semantic judgments.

5.2. Implications for distributional accounts based on word co-occurrences

In recent years, a new generation of lexico-semantic word co-occurrence models based on word embeddings trained to *predict* words in context has been proposed as an alternative to earlier models that simply *count* word co-occurrences. The improvements from prediction models are considered groundbreaking in how well they account for behavioral measures such as similarity (Baroni et al., 2014; Mandera et al., 2017). The current results might temper such conclusions as more stringent tests show that even these prediction models only partially capture meaning. We suggest that many of the previous evaluation tasks in the literature do not rely on visual or affective information, since they

do not focus on the basic objects belonging to the same superordinate category (where visual information is more relevant) or abstract categories (where affective information is more relevant). In those cases, no advantage for experiential information would be found, but that reflects the test items and not how people represent words in general.

The limitations of distributional linguistic models are evident when considering abstract concepts as well. The fact that affective representations consisting of nine features provide a better account of abstract words in two large datasets (SimLex-999 and SimVerb-3500) raises questions about the extent to which the meaning of abstract words is really fully captured by word co-occurrence models. This conclusion is consistent with other work showing that distributional linguistic models only capture a fraction of the variance in a task in which people are asked to judge unrelated words (De Deyne, Navarro, et al., 2016; De Deyne, Perfors, et al., 2016). Altogether, these results suggest that previously obtained high correlations between predictions derived from distributional linguistic models and similarity ratings may have been to some extent an artifact of the comparisons being evaluated. In particular, the lack of abstract concepts and verbs, as well as the limited number of basic-level comparisons for the mostly concrete concepts, might have inflated the performance of many of these models to a point where only ceiling effects were obtained. In those cases, the efficacy of language-only models may have been exaggerated.

5.3. Implications for models based on word associations

To provide context to the results obtained from the distributional linguistic model, we used a model derived from word associations as a baseline measure. One of our most consistent findings was that models using word associations resulted in correlations close to the maximum allowed by the reliability of the similarity ratings. This is remarkable because the comparisons at the basic level may especially rely on experiential visual or affective information: Concrete words rely on perceptual features and abstract words incorporate many affective properties. Despite this, providing additional visual or affective information improved the results of the word association model only slightly. This contrasts with the distributional linguistic models, which benefited far more clearly from additional visual and affective information.

Our results showing that visual and affective properties for categories like *fruit*, *tools*, *feelings*, or *states* are sufficiently encoded in models built from word associations provides further evidence that word associations are not restricted to the language modality but multimodal in the sense that the "first word that comes to mind" relies on imagery and the recollection of affective states as well.

Instead of treating word associations merely as a dependent variable, the current work is an example of how a dialectic approach (see Taylor, 2012) that contrasts text-based distributional linguistic models with word-association-based models can be used to get a deeper insight into the nature of the underlying representations. The finding that experiential information is encoded in word associations challenges the tacit assumption that word associations can be accurately predicted from language. As a consequence, if natural

language and word associations capture different aspects of meaning, it would be expected that the correlations between them would be moderate at best. The most recent evidence along these lines comes from a study by Nematzadeh, Meylan, and Griffiths (2017) that compared topic and word embedding models with word associations. The best performing model correlated .27 with associative strength. This weak correlation may reflect the fact that word associations tap directly into experiential types of representations. This lack of experiential grounding could also explain why text models struggle to predict the color of objects (Bruni, Boleda, Baroni, & Tran, 2012) and do not capture the association between concrete concepts and their typical attributes (Baroni & Lenci, 2008), both of which are often among the strongest associates to a cue word.

5.4. The contribution of affect to the meaning of abstract words

Consistent with our results, previous research suggests that multimodal models consistently improve performance relative to unimodal ones (e.g., Bruni et al., 2014). However, the gain was especially large for the affective multimodal model account of abstract concepts, even though the underlying representation consisted of a handful of features. Indeed, this simple affective model—consisting of only nine features by itself—provided a better prediction for abstract concepts than linguistic models in both studies (see Tables 1 and 5).

This contribution of affective information is difficult to explain given that it is often assumed that abstract concepts rely only on a verbal code to represent their meaning, whereas concrete words rely on an imagery code as well (Paivio, 1971). If abstract words are predominantly acquired through language exposure, distributional linguistic models should correlate strongly with human judgments of abstract words, and potentially not as strongly with judgments of concrete words. This advantage for abstract words was not supported in Study 1 or in Study2.[11]

An outstanding issue is what sort of learning is required to acquire the core affective properties of words. To what extent is nonverbal or embodied information necessary? One might expect that language should a priori play an important role because affect and emotions play a social role in communication. However, affect can be inferred from a range of signals. Nonverbally, affect is communicated through facial expressions (Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000), and these expressions correspond to an internal (embodied) state. Indeed, research suggests that nonlinguistic cues from facial expressions provide information about core affect (valence and arousal), and might also capture emotions like anger or fear when the context supports this (Barrett & Bliss-Moreau, 2009). Besides facial expressions, affect might also be communicated through auditory aspects of spoken language. The tone of voice and other acoustic cues contribute to the affective state of the listener (Nygaard & Lunders, 2002) and lead to altered lexical processing (Schirmer & Kotz, 2003). Language also provides useful cues about valence in the form of the word; affective congruence between sound and meaning leads to a processing advantage in word recognition (Aryani & Jacobs, 2018).

While all of these factors are likely to contribute to learning affect and the meaning of abstract words, it is unlikely that any factor in itself is sufficient. Consider, for example,

factors such as facial expressions or tone. For many words, it is not directly clear how acoustic or nonverbal information would provide sufficient affective grounding, as many abstract words are acquired through written language only. Estimates by Brysbaert, Stevens, Mandera, and Keuleers (2016) underline this point: A 20-year-old exclusively exposed to social interaction would have encountered about 81,000 word types, whereas a 20-year-old exclusively exposed to text will have encountered 292,000 types. Of course, when visual or acoustic cues are absent, the acquisition of affective concepts can still rely on embodied processes involving empathy, where people put themselves in someone else's situation and imagine how they would feel. This view would be mostly consistent with the AEA (Kousta et al., 2009) as well.

Thus far, we have remained vague about whether affect extends beyond abstract concepts and plays a role in certain concrete concepts as well. Partly, this reflects the fact that investigations about affective grounding are relatively new. The current findings suggest that the correlations between human similarity judgments and affective multimodal models are much smaller for concrete than abstract words. This could be due to the fact that we used a subset of mostly concrete nouns, whereas large effects for valence and arousal have been found for adjective representations (De Deyne et al., 2014). Potentially, many adjectives have more extreme valence or arousal values than most concrete nouns. A second explanation might be that there is a trade-off in accessing the meaning of concrete nouns. According to this scenario, affective information would be an integral part of the meaning of these words, but information from other modalities dominates. Affective information is logically more salient in abstract concepts due to the limited contribution of visual and other perceptual modalities. All this suggests a few interesting avenues in which emotional-laden concrete words and abstract words might be compared in future research.

Acknowledgments

We acknowledge the valuable contributions of Andrew Olney and two anonymous reviewers who reviewed a previous version of this manuscript and Meredith McKague for valuable input and discussion. Salary support for this research was provided to Simon De Deyne from ARC grants DE140101749 and DP150103280. Guillem Collell acknowledges the CHIST-ERA EU project MUSTER (http://www.chistera.eu/projects/muster). Data for the triads, together with the language and image embeddings, are available at https://simondedeyne.me/data/

Notes

1. Often affective and emotive factors are confounded in the literature. Emotions are not considered to be psychological primitives in the same sense that core affect dimensions such as hedonistic valence and arousal are. For instance, in studies

involving facial behaviors, reports of experience, and peripheral nervous system activity, affect tends to be a stronger predictor than emotions categories (Barrett, 2006).

2. There are potentially other parallels between concrete and abstract basic-level concepts; for instance, like concrete basic objects, some basic abstract concepts (e.g., *good*, *bad*) might be learned earlier and more distinct in terms of affect.
3. Some studies have questioned the reported improvements of word embedding models over count models (De Deyne, Perfors, & Navarro, 2016; Levy, Goldberg, & Dagan, 2015). In this study as well, the advantage over count models was rather small. Still, we decided to included word embedding models because this makes it easier to compare them with the most recent studies and benchmarks.
4. https://smallworldofwords.org/project/
5. For the distributional linguistic model, we additionally rescaled the cosine similarity between the 0–1 range as `word2vec` produces a small proportion of negative similarities.
6. Non-parametric bootstrap confidence intervals were calculated using the R `boot` package (Canty & Ripley, 2019) using default settings and a total of 10,000 replications.
7. Both pre-trained vectors can be obtained from https://nlp.stanford.edu/projects/glove/. The Common Crawl project can be accessed from https://commoncrawl.org/
8. Most of these datasets were also used in a previous study which evaluated count-based and prediction based linguistic models (De Deyne, Perfors, & Navarro, 2016). In this work we did not include some smaller datasets reported previously such as the Radinsky dataset (Radinsky, Agichtein, Gabrilovich, & Markovitch, 2011), the Rubenstein and Goodenough dataset (Rubenstein & Goodenough, 1965), and the WordSim dataset (Agirre & Soroa, 2009) because these had fewer comparisons for which both language and visual or affective information was available (for visual, 18, 28, and 78 comparisons respectively; for affective, 18, 22, and 74 comparisons).
9. The different studies sometimes report different types of reliability and use different rating methods of similarity. Across all studies the reported reliability should be taken as a lower bound and procedural differences could explain why certain correlations are somewhat lower: some studies report split half correlations while others averaged correlations over individuals.
10. The comparison of the findings is complicated by the occasional use of different metrics (Pearson r vs. rank correlations) and missing observations across the different studies. However, the number of observations in the studies we consider here tend to be large: SimLex-999 ($n = 300$), Silberer2014 ($n = 5{,}799$ for the semantic judgments, $n = 5{,}777$ for the visual judgments), and MEN ($n = 942$). We will therefore assume some robustness in the measures even when some items are missing.
11. The only case where the correlation was higher for abstract words in comparison to concrete ones was in the comparison between the subset of abstract words in

SimLex-999, $r(391) = .47$ (see Table 5) against all words in SimLex-999 $r(913) = .45$ (see Table 4). However, the confidence interval for $\Delta r = .02$, CI_{95} [−0.11, 0.07] indicates that the null hypothesis of no difference is retained.

References

Agirre, E., & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. In A. Lascarides, C. Gardent, & J. Nivre (Eds.), *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 33–41). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/1609067.1609070

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, *6*, 359–370. https://doi.org/10.1111/tops.12096

Aryani, A., & Jacobs, A. M. (2018). Affective congruence between sound and meaning of words facilitates semantic decision. *Behavioral Sciences*, *8*(6). https://doi.org/10.3390/bs8060056

Austerweil, J. L., Abbott, J. T., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3041–3049). Red Hook, NY: Curran Associates.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In A. Fraser & Y. Liu (Eds.), *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. *1*, pp. 238–247). Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.3115/v1/p14-1023

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, *20*, 55–88.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*, 28–58. https://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. In *Advances in experimental social psychology* (pp. 167–218). Cambridge, MA: Elsevier Academic Press. https://doi.org/10.1016/s0065-2601(08)00404-8

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition* (pp. 129–163). Cambridge, England: Cambridge University Press. https://doi.org/10.1017/cbo9780511499968.007

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monographs*, *80*, 1–45. https://doi.org/10.1037/h0027577

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*, 263–292. https://doi.org/10.1037/bul0000089

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 136–145).

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on multimedia* (pp. 1219–1228). New York: Association for Computing Machinery. https://doi.org/10.1145/2393347.2396422

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116. https://doi.org/10.3389/fpsyg.2016.01116

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911. https://doi.org/10.3758/s13428-013-0403-5

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, discourse. *Discourse Processes*, *25*, 211–257.

Cacioppo, J. T., Berntson, G. G., Larsen, J. T., Poehlmann, K. M., & Ito, T. A. (2000). The psychophysiology of emotion. *Handbook of Emotions*, *2*, 173–191.

Canty, A., & Ripley, B. D. (2019). boot: Bootstrap r (s-plus) functions [Computer software manual]. R package version 1.3-24.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *arXiv Preprint* arXiv:1405.3531. https://doi.org/10.5244/c.28.6

Clark, A. (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences*, *10*, 370–374. https://doi.org/10.1016/j.tics.2006.06.012

Collell, G., Zhang, T., & Moens, M.-F. (2017). Imagined visual representations as multimodal embeddings. In S. Markovitch & S. Singh (Eds.), *Proceedings of the thirty-first AAAI conference on artificial intelligence (AAAI-17)* (pp. 4378–4384). AAAI Press.

Crutch, S. J., Troche, J., Reilly, J., & Ridgway, G. R. (2013). Abstract conceptual feature ratings: The role of emotion, magnitude, and other cognitive domains in the organization of abstract conceptual knowledge. *Frontiers in Human Neuroscience*, *7*, 1–14. https://doi.org/10.3389/fnhum.2013.00186

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190. https://doi.org/10.1075/ijcl.14.2.02dav

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2020). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987–1006. https://doi.org/10.3758/s13428-018-1115-7

De Deyne, S., Navarro, D. J., Perfors, A., & Storms, G. (2016). Structure at every scale: A semantic network account of the similarities between unrelated concepts. *Journal of Experimental Psychology: General*, *145*, 1228–1254. https://doi.org/10.31234/osf.io/s3k79

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations. *Behavior Research Methods*, *45*, 480–498. https://doi.org/10.3758/s13428-012-0260-7

De Deyne, S., Peirsman, Y., & Storms, G. (2009). Sources of semantic similarity. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the Cognitive Science Society* (pp. 1834–1839). Austin, TX: Cognitive Science Society.

De Deyne, S., Perfors, A., & Navarro, D. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In H. Watanabe (Eds.), *Proceedings of the 26th international conference on computational linguistics* (pp. 1861–1870). Osaka, Japan: The COLING 2016 Organizing Committee. https://doi.org/10.3758/s13428-012-0260-7

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*, 1030–1048.

De Deyne, S., Verheyen, S., Navarro, D. J., Perfors, A., Storms, G. (2015). Evidence for widespread thematic structure in the mental lexicon. In R. Dale et al (Ed.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 518–523). Boston, MA: Cognitive Science Society.

De Deyne, S., Voorspoels, W., Verheyen, S., Navarro, D. J., & Storms, G. (2014). Accounting for graded structure in adjective categories with valence-based opposition relationships. *Language, Cognition and Neuroscience*, *29*, 568–583. https://doi.org/10.1080/01690965.2013.794294

Deese, J. (1965). *The structure of associations in language and thought*. Baltimore, MD: Johns Hopkins Press.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*, 0121945. https://doi.org/10.1371/journal.pone.0121945

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*, 169–200. https://doi.org/10.1080/02699939208411068

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press. https://doi.org/10.1080/02699939208411068. Available at: http://www.cogsci.princeton.edu/wn

Firth, J. R. (1968). Selected papers of JR Firth, 1952–59. *Indiana University Press*. Bloomington: Indiana University Press. https://doi.org/10.2307/412194

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv Preprint* arXiv:1608.00869. https://doi.org/10.18653/v1/d16-1235

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1406–1414). New York: Association for Computig Machinery. https://doi.org/10.1145/2339530.2339751

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). Piscataway, NJ: IEEE Service Center. https://doi.org/10.1109/cvpr.2016.90

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665–695. https://doi.org/10.1162/coli_a_00237

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*, 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 36–45). Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1005

Kiela, D., & 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* (pp. 2461–2470). Association for Computational Linguistics. https://doi.org/10.18653/v1/d15-1293

Kipper, K., Snyder, B., & Palmer, M. (2004). Extending a verb-lexicon using a semantically annotated corpus. In M. T. Lino, M. F. Xavier, F. Ferreira & R. Costa (Eds.), *Proceedings of the 4th international conference on language resources and evaluation (LREC-2004)* (pp. 1557–1560).

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), *The computer and literacy studies* (pp. 153–165). Edinburgh, UK: Edinburgh University Press.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14. https://doi.org/10.1037/a0021446

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*, 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*, Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226470993.001.0001

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv Preprint* arXiv:1501.02598. https://doi.org/10.3115/v1/n15-1016

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225. https://doi.org/10.1162/tacl_a_00134

Li, L., Song, A., Malave, V., Cottrell, G., & Yu, A. (2016). Extracting human face similarity judgments: Pairs or triplets. *Journal of Vision*, *16*, 719. https://doi.org/10.1167/16.12.719

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1111/j.1756-8765.2010.01106.x

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559. https://doi.org/10.3758/bf03192726

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv Preprint* arXiv:1301.3781.

Mohammad, S. M. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the annual conference of the Association for Computational Linguistics (ACL)* (pp. 174–184). Melbourne, Australia: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*, 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*, 799. https://doi.org/10.1038/nrn1768

Mollin, S. (2009). Combining corpus linguistics and psychological data on word co-occurrence: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5*, 175–200. https://doi.org/10.1515/cllt.2009.008

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, *36*, 402–407. https://doi.org/10.3758/bf03195588

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In G. Gunzelmann, A. Howes, T. Tenbrink & E.J. Davelaar (Eds.), *Proceedings of the 39th annual meeting of the Cognitive Science Society* (pp. 859–864). Austin, TX: Cognitive Science Society.

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, *30*, 583–593. https://doi.org/10.3758/bf03194959

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Champaign, IL: University of Illinois Press.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston. https://doi.org/10.4324/9781315798868

Paivio, A. (2013). Dual coding theory word abstractness and emotion: A critical review of Kousta et al (2011). *Journal of Experimental Psychology: General*, *142*(1), 282–287. https://doi.org/10.1037/a0027004

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1162

Plutchik, R. (1994). *The psychology and biology of emotion*. New York: HarperCollins.

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: Computing word relatedness using temporal semantic analysis. In S. Sadagopan (Ed.), *Proceedings of the 20th international conference on world wide web* (pp. 337–346). ACM Digital Library.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*, 647–656. https://doi.org/10.3758/brm.41.3.647

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*, 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439. https://doi.org/10.1016/0010-0285(76)90013-x

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*, 627–633. https://doi.org/10.1145/365628.365657

Schirmer, A., & Kotz, S. A. (2003). ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience*, *15*, 1135–1148. https://doi.org/10.1162/089892903322598102

Silberer, C., Ferrari, V., & Lapata, M. (2013). Models of semantic representation with visual attributes. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 572–582). Association for Computational Linguistics.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics* (pp. 721–732). Association for Computational Linguistics. https://doi.org/10.3115/v1/p14-1068

Simmons, W. K., Hamann, S. B., Harenski, C. N., Hu, X. P., & Barsalou, L. W. (2008). fMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology – Paris*, *102*, 106–119. https://doi.org/10.1016/j.jphysparis.2008.03.014

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*, 234–243. https://doi.org/10.1016/j.actpsy.2009.10.010

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199290802.001.0001

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology*, *113*, 169–193. https://doi.org/10.1037/0096-3445.113.2.169

Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2013). The neural representation of abstract words: The role of emotion. *Cerebral Cortex*, *24*, 1767–1777. https://doi.org/10.1093/cercor/bht025

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*, 219–247. https://doi.org/10.1515/langcog.2009.011

Von Ahn, L. (2006). Games with a purpose. *Computer*, *39*, 92–94. https://doi.org/10.1109/mc.2006.196

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J. R., Men, W., Gao, J.-H., & Bi, Y. (2017). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, *28*(2), 4305–4318. https://doi.org/10.1093/cercor/bhx283

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413. https://doi.org/10.1037/1082-989x.12.4.399

Appendix A: Stimuli Study 1

Table A1
Concrete triad stimuli in Experiment 1

Natural Kinds

Birds: blackbird–eagle–raven, canary–duck–goose, chicken–parakeet–pigeon, crow–ostrich–owl, falcon–flamingo–penguin, parrot–pelican–turkey. **Body parts:** ear–leg–thumb, elbow–nipple–skin, face–foot–heel, finger–heart–toe, hand–lip–tongue. **Colors:** crimson–pink–yellow, green–khaki–purple. **Crustaceans:** oyster–prawn–shrimp. **Fish:** carp–eel–herring, cod–octopus–tuna, dolphin–shark–whale, goldfish–jellyfish–trout, salmon–squid–swordfish. **Fruit:** apricot–pear–raisin, banana–cherry–pineapple, blueberry–fig–mango, coconut–melon–raspberry grape–lemon–lime, kiwi–peach–plum. **Geological formation:** beach–cave–ravine, crater–glacier–volcano, grass–gully–mountain. **Insects:** ant–cockroach–leech, beetle–flea–termite, butterfly–ladybug–worm, mosquito–moth–wasp, slug–snail–spider. **Mammals:** bear–rat–tiger, beaver–goat–horse, camel–cow–otter, cat–dog–gorilla, deer–hamster–lion, elephant–hyena–panther, giraffe–leopard–sheep, kangaroo–mouse–pony, rabbit–walrus–zebra. **Reptiles:** alligator–cobra–lizard, crocodile–frog–tortoise. **Trees:** cedar–fir–willow. **Vegetables:** artichoke–lettuce–tomato, avocado–cabbage–mushroom, broccoli–eggplant–onion, carrot–radish–turnip, cucumber–spinach–zucchini

Artifacts

Breakfast: bread–muffin–oatmeal, jam–sandwich–toast. **Buildings:** apartment–hotel–temple, cabin–castle–church, office–tent–trailer. **Clothing:** bikini–jacket–sweater, blouse–gown–suit, coat–parka–swimsuit. **Drinks:** beer–milk–tea, champagne–lemonade–wine, coffee–vodka–whiskey. **Electronic devices:** camera–projector–radio, computer–monitor–telephone. **Fabrics**: cotton–fleece–lace, denim–silk–velvet, linen–satin–wool. **Fashion accessories**: bracelet–buckle–purse, button–lipstick–watch, necklace–shawl–umbrella. **Food**: doughnut–fudge–lollipop, hamburger–lasagna–stew, omelet–roll–spaghetti. **Furniture**: bath–bed–dresser, chair–couch–desk, cupboard–stool–table. **Kitchen utensils**: blender–mixer–scissors, bottle–bowl–spoon, fork–spatula–toaster, kettle–oven–plate. **Music instruments**: accordion–banjo–harp, clarinet–drum–piano, flute–harmonica–trombone, guitar–triangle–violin. **Professions**: farmer–lawyer–secretary, gardener–nurse–scientist, pilot–surgeon–teacher. **Sports**: archery–boxing–frisbee, baseball–golf–polo, cricket–squash–tennis. **Tools**: anvil–chisel–hatchet, clamp–crowbar–hoe, rake–spade–wrench. **Vehicles**: airplane–cab–tractor, boat–limousine–scooter, buggy–ferry–yacht, bus–jeep–sled. **Weapons**: bomb–grenade–spear, bow–rope–shotgun, cannon–revolver–shield, dagger–harpoon–stick

Table A2
Abstract triad stimuli in Experiment 1. **Category labels** refer to the most specific common hypernym in WordNet found at depth [*d*]

Ability [5]: aptitude–breadth–invention, daydream–focus–method, fantasy–intellect–talent. **Act** [5]: capture–expansion–pursuit. **Action** [6]: admission–courtesy–removal, debut–progress–violence, flutter–rampage–selection, journey–rush–trick. **Activity** [6]: adoption–work–worship, adventure–training–treatment, arrogance–endeavor–support, betrayal–espionage–hassle, bribery–hoax–struggle, care–monopoly–treason, craft–crusade–raid, crime–research–scramble, custom–education–rehearsal, gaming–restraint–role, mayhem–stealth–theft, mischief–nightlife–violation. **Attitude** [5]: ideology–socialism–taboo. **Basic cognitive process** [6]: attention–memory–vogue. **Belief** [5]: creed–phantom–religion, faith–magic–opinion. **Bias** [8]: bias–prejudice–racism. **Change** [7]: breakup–rotation–voyage, gesture–reform–repair. **Cognition** [4]: estimate–sensation–wisdom, folklore–intuition–regard, ghost–sight–theory, illusion–layout–respect. **Cognitive state** [7]: certainty–disbelief–mystery. **Content** [5]: access–essence–idea, agenda–ignorance–rule. **Cost** [7]: bounty–perk–ransom. **Discipline** [7]: economics–logic–sociology. **Emotion** [6]: happiness–panic–tantrum. **Feeling** [5]: affection–ambition–heartache, amazement–outrage–rapture, anger–ecstasy–relief anguish–dread–joy, awe–ego–love, boredom–devotion–empathy, contempt–grudge–remorse, delight–horror–wonder, desire–relish–vanity, disgust–fondness–grief, dismay–distress–suspense, emotion–enjoyment–envy, fear–jealousy–mood, fetish–lust–wrath, fury–hope–thrill, pity–pride–surprise. **Idea** [6]: fallacy–notion–plan, feature–scheme–tactic. **Location** [4]: boundary–empire–zone. **Magnitude** [5]: depth–majority–size, dimension–limit–number. **Person** [4]: addict–delegate–slob, ancestor–believer–sir, brute–wanderer–weirdo, celebrity–foreigner–maniac, communist–fool–outsider, corporal–expert–youth, counsel–foe–snob, darling–hero–thinker, disciple–fanatic–killer, dreamer–hick–novice, follower–optimist–sinner, graduate–savior–scoundrel, guardian–liar–moron heir–rebel–supporter, patriot–sweetie–whiz. **Physical condition** [6]: addiction–insomnia–plague, complaint–harm–phobia, disease–sickness–thirst, frenzy–handicap–hunger. **Process** [5]: hindsight–insight–sweetness. **Psychological state** [6]: annoyance–insanity–tension, assurance–interest–sanity, bliss–madness–paranoia. **Region** [5]: district–homeland–region, frontier–heaven–paradise, hell–premises–territory. **Science** [8]: algebra–geology–science, astronomy–math–physics. **Social group** [4]: ally–meeting–monarchy, business–charity–reunion, clan–dynasty–industry, enemy–seminar–sorority, minority–regime–utility. **Statement** [5]: bargain–comment–summary, covenant–excuse–remark, evasion–notice–reply. **Time period** [5]: birthday–evening–semester, century–holiday–maturity, childhood–era–year, morning–period–vacation. **Transferred property** [5]: benefit–legacy–welfare, donation–rent–royalty