Visual and affective grounding in language and mind

Simon De Deyne[1], Danielle J. Navarro[2], Guillem Collell[3], and Amy Perfors[1]

[1]University of Melbourne, School of Psychological Sciences, 3010 VIC, Australia
[2]University of New South Wales, School of Psychology, 2052 NSW, Australia
[3]KU Leuven, Department of Computer Science, 3001 Heverlee, Belgium

Abstract

One of the main limitations in natural language-based approaches to meaning is that they are not grounded. In this study, we evaluate how well different kinds of models account for people's representations of both concrete and abstract concepts. The models are both unimodal (language-based only) models and multimodal distributional semantic models (which additionally incorporate perceptual and/or affective information). The language-based models include both external (based on text corpora) and internal (derived from word associations) language. We present two new studies and a re-analysis of a series of previous studies demonstrating that the unimodal performance is substantially higher for internal models, especially when comparisons at the basic level are considered. For multimodal models, our findings suggest that additional visual and affective features lead to only slightly more accurate mental representations of word meaning than what is already encoded in internal language models; however, for abstract concepts, visual and affective features improve the predictions of external text-based models. Our work presents new evidence that the grounding problem includes abstract words as well and is therefore more widespread than previously suggested. Implications for both embodied and distributional views are discussed.

Keywords: multimodal representations; semantic networks; grounded cognition; distributional semantics; affect

When you look up the word *rose* in a dictionary, one definition is "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 how precisely are they represented at all? Different theories give different answers to this question depending on how much language or non-linguistic sensory representations contribute to meaning. In embodied theories, meaning derives from somatosensation, vision, olfaction, and perhaps even internal affective states. By contrast, lexico-semantic views stress the contribution of language, suggesting that the meaning of *rose* can be derived in a recursive fashion by considering the meaning of words in its linguistic context like bush, red, and flower (Firth, 1968). Both views are extremes, and current theories of semantics tend to take an intermediate position in which both language and non-linguistic representations contribute to meaning, depending on the situation. These theories propose another solution to the symbol grounding problem: to understand the meaning of a rose and avoid circularity, symbols need to grounded in their world referents (Harnad, 1990).

One idea is that language is a symbolic system that represents meaning via the relationships between (amodal) symbols, but is also capable of capturing embodied representations since these symbols make reference to perceptual representations. This *symbol interdependency hypothesis* proposed by Louwerse (2011) has been supported by a host of empirical findings and has contributed to a growing consensus that meaning is represented in both symbolic and sensory grounded representations (Andrews, Frank, & Vigliocco, 2014; Riordan & Jones, 2011; Vigliocco, Meteyard, Andrews, & Kousta, 2009). Although language has the capacity to capture sensory properties, this capacity is imperfect; 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). Many theories propose that this not need be true, at least in the case of abstract words which lack physical referents (Paivio, 2013). However, in recent years it has been suggested that sensorimotor experience, emotional experience, and sociality could provide grounding even for abstract words (Borghi et al., 2017).

Just as the grounding problem suggests that models incorporating only linguistic information will fare less well than those based on linguistic and sensory input, other considerations suggests that the opposite may also be true – that models incorporating only sensory input will fare less well than those based on both. For instance, it is not clear how sensory-based models might capture the meaning of concepts like “romance” or “purity”, which are connotations of the word *rose* that are not directly derived from sensorial impressions.

In this paper, we revisit the question about how language encodes sensory properties, making use of the distinction between *external* language and *internal* language (Chomsky, 1986; Taylor, 2012). External models treat language as an entity that exists in the world, consisting of a set of utterances made by a speech community, and (partially) measured by large text corpora. To date, most attempts

---

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/

to investigate symbol grounding have relied on "external" language models. By contrast, internal models view language in terms of a stored mental representation, a body of knowledge possessed by the speakers, and (partially) measured using a variety of methods including semantic differentials (Osgood, Suci, & Tannenbaum, 1957), feature elicitation (De Deyne et al., 2008; McRae, Cree, Seidenberg, & McNorgan, 2005), or word association responses (De Deyne, Navarro, & Storms, 2013; Kiss, Armstrong, Milroy, & Piper, 1973; Nelson, McEvoy, & Schreiber, 2004). In psychology, this is referred to as the experiential tradition (Andrews, Vigliocco, & Vinson, 2009).[1]

Drawing the distinction between models of external language and internal language (henceforth denoted *E-language* and *I-language* respectively) allows us to assess the severity of the symbol grounding problem, and – in particular – to determine whether the issue is more problematic for external language models. Does a text corpus encode the same "implicit" perceptual knowledge of the world as a word association data set? What can we learn about the nature of representations by contrasting the behavior of external and internal language models when they are made into multimodal models by augmenting them with grounded sensory information?

Grounding for concrete concepts

Previous work on multimodal grounding has tended to rely on I-language representations to capture sensory (usually visual) properties. For example, some studies have used conceptual features derived from feature elicitation tasks, in which participants are asked to list meaningful properties of the concept in question. These tasks typically elicit entity features (e.g., an *apple* <is red>), functional features (e.g., an *axe* is <used for chopping>) or taxonomic features (e.g., *judo* is a <martial art>), and the task itself can be viewed as a kind of "censored" word association task in which a subset of relations are considered relevant to the task and others (particularly thematic, situational features) are not. Using this approach, previous research has integrated E-language models (e.g., topic models) with additional information derived from elicited features for nouns or verbs (Andrews et al., 2014; Johns & Jones, 2012; Steyvers, 2010).

More recently, researchers in machine learning and computational linguistics have used models that derive (mostly visual) sensory representations from large image databases instead of I-language feature-listing tasks. These models range from hybrid approaches, in which visual features are obtained by human annotators (Silberer, Ferrari, & Lapata, 2013), to more detailed approaches using a bag-of-visual words derived from annotated low-level visual features such as using scale-invariant feature transformations (SIFT) (e.g. Bruni et al., 2014). More recently, deep convolutional networks are often used for this purpose, as they typically outperform the low-level representations captured by simpler feature-extraction methods (e.g. Lazaridou, Pham, & Baroni, 2015). These studies usually focus on grounding in the visual modality, and this is reflected in the type of concepts that are modeled (typically concrete nouns). In this study we restrict ourselves to visually grounded models

[1]The E-language approach is sometimes referred to as the distributional tradition, but this might be somewhat ambiguous, as a distributional approach can also be applied to I-language or experiential approaches

to model concrete concepts, though we agree with authors who have noted that auditory input is relevant to sensorimotor grounding (Kiela & Clark, 2015).

Grounding for abstract concepts

Abstract concepts have attracted attention in recent years because distributional and embodied approaches make different predictions about how they are represented. They pose a particular challenge to embodied theories, as abstract concepts like "opinion" do not have clearly identifiable referent (Borghi et al., 2017). In contrast, distributional theories can easily explain how abstract concepts are represented in terms of their distributional properties in language (Andrews et al., 2014). Beyond the theoretical issues, however, there are good reasons to study abstract concepts, simply because they are so prevalent in everyday language. For example, 72% of the noun or verb tokens in the British National Corpus are rated by human judges as more abstract than the noun *war*, which many would already consider to be quite abstract (Hill & Korhonen, 2014).

Even though most work on multimodal models has focused on concrete words, some researchers have started to question whether a distributional account based on language is sufficient to explain how abstract concepts are represented. According to *strong* embodiment theories, concrete and abstract concepts do not substantially differ because they are both grounded in the same systems that are engaged during perception, action and emotion (Borghi et al., 2017). Other theories like the conceptual metaphor theory explain differences between abstract and concrete theories in terms of metaphors derived from concrete domains to provide grounding to abstract meaning (Lakoff & Johnson, 2008).

Here we will mainly focus on a recent view that highlights the role of affect in grounding abstract concepts. According to the *Affective Embodiment Account* (AEA), the acquisition and representation of abstract concepts are grounded in internal affective states (Vigliocco et al., 2009). As claimed by the AEA, affect should be considered another type of information that is grounded in experience. Moreover, this affective grounding is not limited to emotion words, but extends to most words that have affective associations (Kousta, Vinson, & Vigliocco, 2009). Evidence for this proposal comes from data on lexical decision showing a processing advantage for abstract words that could be accounted for by emotional valence even when confounding factors like imageability or context availability are taken into account (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011). The theory is also supported by neuro-imaging data showing that human ratings for affective associations for abstract words predict modulation of the BOLD signal in areas associated with affective processing (Vigliocco et al., 2013).

Further evidence suggesting that distributional approaches do not fully capture the representation of abstract concepts comes from a study in which external language-based measures were compared with internal subjective measures to predict the voxel activation in a representational similarity analysis (Wang et al., 2017). In this study, external distributional representations were derived from word co-occurrences modelled with `word2vec` and LSA (see further), whereas I-language

representations of abstract words were assessed using a property rating task including affective features like valence, arousal, emotion, and social interaction. E-language and experiential properties were found to encode distinct types of information and were only weakly correlated with each other (r = .32). When Wang et al. (2017) related the fMRI BOLD signal in a word familiarity judgment task with these E-language and I-language representations, they found dissociable neural correlates for both of them. Notably, I-language features were found to be sensitive to areas involved on emotion processing. A subsequent principal component analysis of the whole-brain activity pattern showed that the first neural principal component, capturing most of the variance in abstract concepts, was associated with valence information. This effect was stronger than other factors such as emotion or social interaction. This first component was correlated significantly with the I-language feature but was not correlated with E-language.

These findings also align with behavioral findings showing that valence is the most important dimension for representing adjectives (De Deyne, Voorspoels, Verheyen, Navarro, & Storms, 2014) and other words (Recchia & Louwerse, 2015). Taken together, they suggest that E-language and I-language encode different aspects of meaning, with the latter involved in affective processing.

Current study

The goal in this work is to investigate how – and to what extent – E-language and I-language models are able to capture the meaning of concrete and abstract words; and moreover, to measure the extent to which their ability to do so can be improved by the addition of experiential input. Our work is motivated by several observations. First, recent performance improvements by distributional lexico-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). Recent improvements in extracting visual features have been similarly striking (Chatfield, Simonyan, Vedaldi, & Zisserman, 2014; He, Zhang, Ren, & Sun, 2016; Lazaridou et al., 2015). Together, this would suggest that "multimodal" models that integrate these external sources of information might provide insight about how humans use environmental cues (language or visual) to represent semantic concepts.

We focus on two kinds of concepts. For the first, we propose to focus on basic level concepts (*apple*, *guitar*,...) belonging to common categories (*fruit*, *music instruments*,...). Not only is this taxonomic level the most informative and learned early on in life, it also encodes most perceptual properties (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Previous work has also shown that judgments for concepts at this level provide a considerable challenge for text co-occurrence models, suggesting that this is where they lack the crucial perceptual grounding (e.g., De Deyne, Peirsman, & Storms, 2009).

Our second aim is to look beyond concrete noun concepts, and computationally test the AEA hypothesis that multimodal (especially affective) grounding is required for abstract concepts. For I-language models, we might expect that much of this knowledge is implicitly encoded in the training

data (e.g., word associations), and if so adding visual or affective features to the training data would have little impact on the performance of I-language models. In contrast, we might suspect that this information may be missing from text corpora; if so the performance of E-language models should be improved by adding visual or affective grounding to the training data.

Alternatively, if it is indeed the case that E-language does encode perceptual and affective features in a sufficiently rich way, the representation of abstract concepts might primarily rely on verbal information (Paivio, 2014), and there would be little impact on the performance of an E-language model. Given recent advancements in E-language in accounting for a host of lexical and semantic behavior (see Mandera et al., 2017, for example), this possibility is not unreasonable. This would be in line with the symbolic interdependency hypothesis, according to which language captures sensory properties as well (Louwerse, 2011) and might rely on the quality of the language input such as better language corpora and algorithms to learn from it (such neural embedding models).

In the remainder of the article we will provide a brief overview of how our linguistic and experiential models are constructed. We then evaluate how the experiential information (visual or affective) augments performance of the I-language and E-language models on a similarity judgment task. In the first study we present two experiments that focus on basic level comparisons among concrete and abstract words and explain how they are constructed to tap into visual and emotional properties of meaning. In the second study, we will investigate if these findings generalize to other datasets that have been used as benchmarks for language models.

Models

Language

External. External language models capture the distributional properties in the language environment as is, rather than providing a mental model of how this information is encoded.

Corpus. The language corpus used for training our external language model was constructed to provide us with a reasonably balanced set of texts that is representative of the type and amount of language a person experiences during a lifetime — including both formal and informal language as well as spoken and written language. This corpus is fully described in De Deyne, Perfors, and Navarro (2016). To summarize, it consists of four parts: (1) a corpus of English movies subtitles, (2) written fiction and non-fiction taken from the Corpus of Contemporary American (COCA, Davies, 2009), (3) informal language derived from online blogs and websites available through COCA, and (4) SimpleWiki, a concise accessible version of Wikipedia. The resulting corpus consisted of 2.16 billion tokens and 4.17 million types. Together, this aims to encompass knowledge that is likely available to the average person but remains sufficiently generous in terms of the quality and quantity of data so that models incorporating it would perform similarly to the existing state-of-the-art.

Word2Vec embeddings. Word embedding models have recently been proposed as psychologically motivated alternatives 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 techniques is `word2vec` (Mikolov et al., 2013). In contrast to count-based models word embedding models like `word2vec` try to predict the word from context (CBOW) or the context from the word (skip-gram), which leads to improved results accounting for lexical and semantic processing (Baroni, Dinu, & Kruszewski, 2014; De Deyne, Perfors, & Navarro, 2016; Mandera et al., 2017). To train `word2vec` on this corpus we used a continuous bag of words (CBOW) architecture in which the model is given the surrounding context for a word (i.e., the other words in a sliding window) and is trained to predict that word. The other parameters were taken from previous work 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, & Navarro, 2016). This corresponds to the best performing model on a wide range of similarity judgment studies and show performance comparable with other published embeddings (De Deyne, Perfors, & Navarro, 2016). A number of these studies will be re-analysed in Part 2 of this paper.

Internal. Internal language models use language (in the form of word associations) to derive a proxy of the internal representations used by people. It is thus partially a linguistic model, but probably incorporates experiential or affective information as well. The free association procedure is useful because it does not censor the type of responses. This makes it suitable for capturing the representations of all kind of concepts (including abstract ones) and all kind of semantic relations (including thematic ones). It also avoids dissociating the lexicon in two different types of entities (concepts and features) which allows us to represent these data in a unimodal graph.

Word association data. The current data were collected as part of the *Small World of Words* project,[2] an ongoing crowd sourced project to map the mental lexicon in various languages. The data are those reported by De Deyne, Perfors, and Navarro (2016) and consist of 88,722 fluent English speakers who were asked to give three different responses to each cue included in a short list of cues (between 14 and 18 words). The current dataset consisted of a total of 12,292 cues for which at least 300 responses were collected for every cue. In line with previous work, we constructed a semantic weighted graph from these data where each edge corresponds to the association frequency between a cue and target word. This 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, that is 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 (2018), 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. These random walks implement the idea of spreading activation over a semantic network. To limit the contribution of long paths, a decay parameter (α = .75) is defined in line with all our previous work. This

[2] `https://smallworldofwords.org/project/`

algorithm is similar to other approaches (Austerweil, Abbott, & Griffiths, 2012), but differs by taking an additional PPMI weighting of the graph with indirect paths G_{rw} to avoid a frequency or degree bias and to reduce spurious links (see Newman, 2010, for a discussion).

Experiential models

Visual information. In this study, we used ImageNet (Russakovsky et al., 2015) as a source of visual information; it is currently the largest labeled image bank including over 14 million images. It is centered on 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 included, ImageNet covers a large portion of the concrete lexicon. Visual features providing the perceptual grounding to our models were obtained from Collell and Moens (2016). They employed a pre-trained supervised convolutional neural network (CNN) model trained to classify the 1000 different types of objects of the ImageNet Challenge (ILSVRC2012). Visual features from each image are extracted using the forward pass of a pre-trained ResNet CNN model (He et al., 2016). The 2048-dimensional activation of the last layer (before the softmax) is taken as a visual feature vector of for image, as it contains higher level features. Finally, a single 2048-dimensional vector to represent each concept is obtained by averaging the feature vectors from its associated individual images.

The image vectors in ImageNet and the trained models refer both to leaf nodes and inner nodes of the ImageNet tree. Each of the images in ImageNet maps onto a specific WordNet synset. At this point, the synsets do not necessarily map onto a single word. Moreover, some of the synset labels do not map onto the basic-level concept most people would refer to. For example the word *hedgehog* is found in both synsets n01893825 (*hedgehog, Erinaceus europaeus, Erinaceus europeaeus*) and n02346627 (*porcupine, hedgehog*). Because we are interested in a comparison with I-language models, we only kept labels that were part of the set of 12,000 cues from the Small World of Words word association data (SWOW-EN2018, De Deyne et al., 2018). When words occurred in multiple synset labels their vectors were averaged. Returning to the example, this means that n01893825 would be labeled as hedgehog, and n02346627 would refer to both porcupine and hedgehog. The vector for hedgehog would then be composed of the average vectors of the n01893825 and n02346627 synsets. This reduced the number of ResNet vectors for 18,851 synsets to a subset of 4449 shared with the SWOW-EN2018 cue words. Of these 4449 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., 2018; Recchia & Jones, 2009). An exploratory analysis showed that this is also the case for the image-vectors and in the remainder of the text we will make use of these weighted PMI image vectors.

Affective information. Previous studies following on the seminal work by Osgood have shown that affective factors like valence or arousal capture a significant portion of the structure represented

in the mental lexicon (e.g., De Deyne et al., 2014; Osgood et al., 1957). In agreement with the AEA theory by Vigliocco et al. (2009), we expect that affective factors provide necessary grounding to abstract concepts, which lack any physical referents.[3] To provide grounding for abstract words in multimodal models, we derived six affective features from Warriner, Kuperman, and Brysbaert (2013) who collected ratings of valence, dominance and potency judgments for nearly 14,000 words from a balanced group of males and females. These features were supplemented with three features for valence, arousal and dominance lexicon for 20,000 English words (Mohammad, 2018). These recent norms were somewhat different than those from Warriner et al. (2013) in two ways: they used best-worst scaling, which resulted in more reliable ratings, and operationalized dominance differently, resulting in ratings that were less correlated with valence. The final affective feature model consisted of words each represented by a one-dimensional vector consisting of nine elements: valence, arousal, and potency judgments for men and women from (Warriner et al., 2013) and valence and arousal and dominance from Mohammad (2018). None of the features were perfectly correlated with each other, which allowed them to each contribute.

Creating multimodal models

In order to investigate how adding visual or affective information to our language models affects their performance, we create multimodal models that incorporate both kinds of information. Recent studies have proposed efficient ways to combine different information sources including auto-encoders (Silberer & Lapata, 2014), Bayesian models (Andrews et al., 2014) or cross-modal mappings (Collell, Zhang, & Moens, 2017). In this work we used 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. We included a tuning parameter β to vary the relative contribution of the different modalities. That is, the multimodal fusion M of the modalities a and b corresponds to $M = \beta * v_a \bigoplus (1-\beta) * v_b$ where $\bigoplus$ denotes concatenation and v_a and v_b are the respective vector representations. Because the features for different modalities can have different scales, they were normalized using the L_2-norm, which puts all features in a unitary vector space (Kiela & Bottou, 2014).

Study 1: Basic level triadic comparisons

In our first study we compare these different models to human performance in a triadic comparison task, in which people were are asked to pick the most related pair out of three words. 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). In previous work we have shown that this task can be used for a wide range

[3] Often affective and emotive factors are confounded in the literature. Emotions are not considered to be psychological primitives in the same sense valence and arousal are, and therefor we focus on affect.

of semantic comparisons between Dutch words (De Deyne, Navarro, et al., 2016): people make consistent judgments regardless of whether all of the words involved belong to basic categories (*lemon – lime – apple*), share a domain (*drill – violin – bicycle*), or are weakly related (*Sunday – vitamin – idiot*).

In contrast to concrete concepts, the taxonomic structure of abstract concepts has received less attention in the literature. To identify a set of stimuli that could be used to compare I- and E-language models we identified a subset of words that were present in both datasets and were also part of the extensive concreteness norm set for English words in Brysbaert, Warriner, and Kuperman (2014). To derive triads that are relatively neutral towards the different models we used WordNet (Fellbaum, 1998) to identify common categories for basic level concepts. 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 hypernym at depth 6 (*entity* > *abstraction* > *attribute* > *state* > *condition* > *psychological state*).

Method

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. All the procedures were approved by the Ethics Committee of the University of Adelaide. The participants were recruited online through Prolific Academic ©, signed an informed consent form, and were paid £6 /hour.

Stimuli. The CONCRETE stimuli consisted of 100 triads constructed from the subset of nouns present in the lexicons of all three models for which valence and concreteness norms were available in Warriner et al. (2013) and Brysbaert et al. (2014). Furthermore, all 300 words belonged to a varied set of common 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 man-made categories (*clothing*, *dwellings*, *furniture*, *kitchen utensils*, *musical instruments*, *professions*, *tools*, *vehicles*, *weapons*). The triads were constructed by randomly combining category exemplars so none of the words occurred more than once across any of the triads. A list of the stimuli together with their category label is presented in Appendix 2.

The ABSTRACT stimuli were also constructed from the subset of words that were present in the lexicons of all three models as well as included in available concreteness and valence norms. Next, words were screened based on concreteness, with a cutoff set at 3.5 on a 5-point scale with 5 indicating the most concrete instances (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 stimuli corresponded to categories in the WordNet hierarchy at a depth of 4 to 8 in the hierarchic taxonomy. This way exemplars were included from categories

that were not overly specific or general. Nineteen different categories are included: *ability*, *action*, *activity*, *attitude*, *belief*, *cognition*, *feeling*, *idea*, *knowledge domain*, *location*, *magnitude*, *person*, *physical condition*, *possession*, *psychological state*, *social group*, *sound*, *statement*, and *time period*.

As before, within each category, a triad was sampled from all possible triad combination without repeating a word across all 300 triads. The method and procedure were identical to that for the CONCRETE except for the example given to participants which now contained abstract words. A list of the stimuli together with their category label is presented in Appendix 3.

Procedure. Participants were instructed to select the pair of words (out of the three) were 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.

Behavioral results. Judging the 100 CONCRETE triads took on average 8 minutes. All words were known by over 99% of participants. The ABSTRACT triad task was slightly more difficult, taking on average 10 minutes to complete, with all words known by over 96% percent of people. The judgments were pooled by counting how many participants choose each of the three potential pairs. Because the number of judgments varied depending on whether people judged them to be unknown, the 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.

Evaluation

The triad preferences were predicted by calculating the cosine similarity between the distributions for each of the three triad pairs. The preference scores was calculated by rescaling the similarities for all three pairs to sum to one.[4] The correspondences between the human preferences and the model predictions are shown in Table 1. The I-language model showed a strong correlation with the triad preferences in both the CONCRETE (r = .76) and ABSTRACT task (r = .82). The results for the E-language model were considerably lower: r = .64 for CONCRETE triads and r = .62 for ABSTRACT triads.

For the grounded 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. However, for completeness Table 1 also evaluates performance of the models on CONCRETE words when affective information is added.

It is evident that for CONCRETE words, adding visual information helped both the I-language and E-language models (r_{max} values are higher than r values). However, visual information improved performance of the E-language model more, suggesting that I-language models may already incorporate some of that information. Overall, I-language models still outperformed E-language models, even when visual or affective information was included.

[4] For the E-model, we additionally rescaled the preferences between the 0 - 1 range as `word2vec` produces a small proportion of negative similarities.

t

Table 1

Pearson correlations and confidence intervals for unimodal I- and E-language (v_a) and visual and affective modalities (v_b). Best-fitting multimodal models combining linguistic and experiential information were found by optimizing the correlation for mixing parameter β and are shown in column r_{max}. Confidence intervals for the difference ($\Delta r = r_{max} - r_{v_a}$) were obtained using the confidence method for overlapping dependent correlations from Zou (2007).

Dataset	v_a	r_{v_a}	CI	v_b	r_{v_b}	CI	β	r_{max}	CI	Δr	CI
Concrete	I	.76	.71, .80	Vis.	.67	.60, .73	.35	.81	.77, .85	.05	.03, .08
Concrete	I	.76	.71, .80	Aff.	.21	.10, .32	.38	.78	.74, .82	.02	.00, .05
Concrete	E	.64	.57, .70	Vis.	.67	.60, .73	.48	.75	.70, .80	.12	.07, .17
Concrete	E	.64	.57, .70	Aff.	.21	.10, .32	.50	.68	.62, .74	.04	.02, .08
Abstract	I	.82	.78, .86	Aff.	.51	.43, .59	.05	.82	.78, .86	.00	-.01, .01
Abstract	E	.62	.54, .68	Aff.	.51	.43, .59	.58	.74	.69, .79	.13	.08, .19

For ABSTRACT words, we found that affective information improved the performance of the E-language model substantially (from $r = .61$ to $r_{max} = .74$). This improved performance is consistent with the proposal by Vigliocco et al. (2009) and also suggests that the decisions made by people in the triad task were based in part on affective information. Consistent with this, the affective model predictions did capture a significant portion of the variability in the abstract triads, $r = .51, CI = [.43, .59]$, which is remarkable considering that the model consists of only nine features. Interestingly, affective information did not improve the performance of I-language models, suggesting that word association data already incorporates affective information.

In order to further explore the effect of adding visual or affective information, Figure 1 plots the performance of each model as a function of how much additional information is added when the words are either CONCRETE (left panel) or ABSTRACT (right panel). Within each panel, the left side depicts models with no visual information ($\beta = 0$) and the right corresponds to models with only visual information ($\beta = 1$). As before, we see that adding visual information improves the I-language association model only slightly whereas there is a more considerable improvement for the E-language model. Peak performance for all multimodal models occurred when about 40% to 50% of the information was visual. For ABSTRACT words, the left side of the panel depicts models with no affective information ($\beta = 0$) and the right side reflects only affective information ($\beta = 1$). There is no improvement for the association model but a considerable improvement for the E-language model when affective information is added. Peak performance for the multimodal E-language occurred when around 60% of the information was affective.

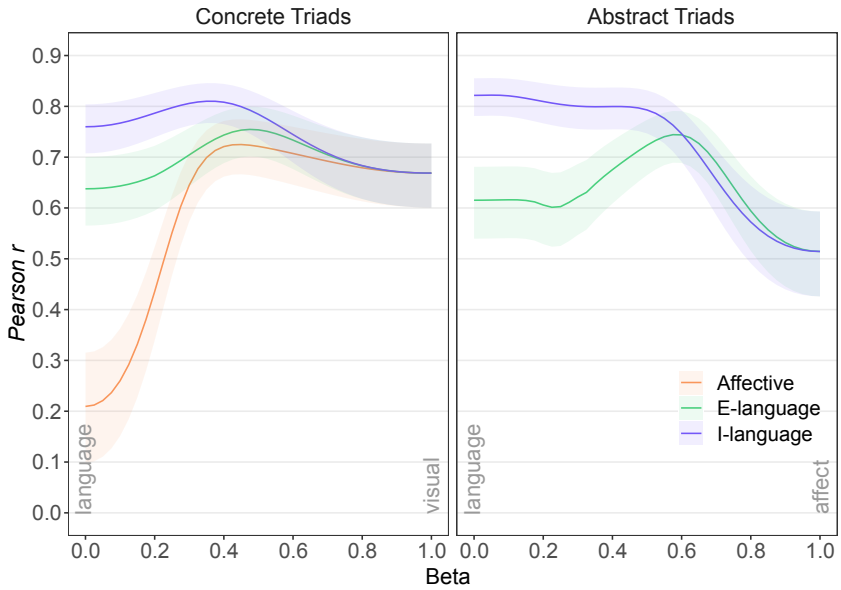

Figure 1. The effect of adding experiential information to predict performance on the triad tasks involving CONCRETE (left panel) and ABSTRACT (right trial) words. In the left panel, visual information is added: larger β values correspond to models that weight visual feature information more. The I-language model performs best, with small improvements from visual features (Δr = .03). The text-base E-language model performs worse, but improves substantially when visual features are added (r increase of .07). The emotion model is easily the worst of the three. In the right panel, affective information is added: larger β values correspond to models that weight affective information more. The I-language model is easily superior to the text-based E-language model, and it gains little to nothing from the added affective features. The E-language model, however, gains considerably when affective information is added.

Robustness

So far our results support the hypothesis that the grounding problem is less pronounced in I-language models. Performance of E-language models was worse on ABSTRACT words than CONCRETE words, but I-language models performed similarly for both kinds of words. Moreover, E-language models were most improved by adding affective information, consistent with the AEA hypothesis that abstract words, like concrete words, need affective grounding.

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 tested against alternative models based on different corpora and different embedding techniques.

First, we investigated whether the text-model performance was due to the specific embeddings used. To test this, we chose `GloVe` embeddings as an alternative E-language model (Pennington, Socher, & Manning, 2014). In contrast to `word2vec` embeddings, `GloVe` takes into account the global

structure in which words occur, which can lead to improved predictions. We used the published word vectors for a model of comparable size to our corpus based on 6 billion tokens derived from the GigaWord 5 and Wikipedia 2014 corpus. We also included an extremely large corpus consisting of 840 billion words from the Common Crawl project.[5] As before, the language vectors were combined in a multimodal visual or affective model and the correlations were optimized by fitting values of β. Figure 2 shows the unimodal E-language correlations and the optimal multimodal correlations.

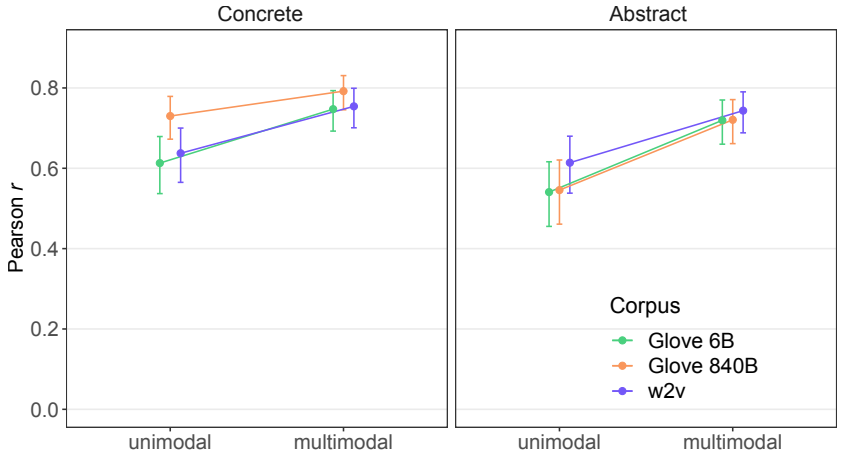

Figure 2. Evaluation of alternative E-language models on CONCRETE and ABSTRACT words in the triad task. A very large corpus based on 840B tokens improves performance for concrete items, but results in lower correlations for abstract triads. The current `word2vec` model based on 2B words performs favorably compared to the 6B `GloVe` embeddings. Overall findings are robust regardless of the corpus or embedding method used.

As Figure 2 illustrates, the results for these different E-language models are comparable, despite both architectural (`word2vec` vs `GloVe`) and corpus differences (2B words for the current corpus, 6B or 840B words used to train `GloVe`). Regardless of the E-language model, adding experiential information improved performance, more markedly so for ABSTRACT words. Only the model derived from an extremely large corpus 840 billion words improved overall, but this improvement was restricted to the CONCRETE words. This suggests that language about concrete entities might be relatively under-represented in all but the largest corpora. However, it is difficult to draw any strong conclusions from this as even the best model showed only a moderate performance on ABSTRACT concepts – even though it is commonly assumed that abstract words are primarily acquired through language. If anything, this pattern of results may demonstrate some of the risks inherent in using models that are so large they no are no longer representative of the kind of linguistic inputs people might be exposed to.

For the I-language model, there were fewer parameters and experimental degrees of freedom than the E-language model: the I-language model is determined by single activation decay parameter α, which we set at .75 in line with previous work (De Deyne, Navarro, et al., 2016). However, this might

[5]Both pre-trained vectors can be obtained from `https://nlp.stanford.edu/projects/glove/`

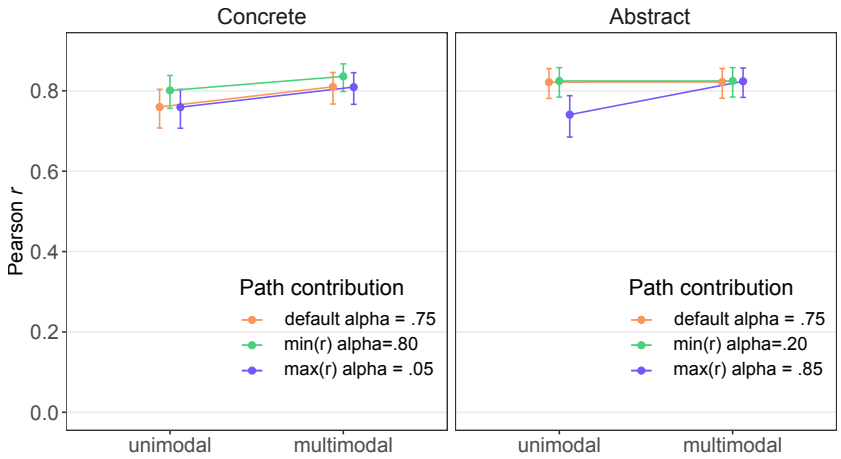

Figure 3. Evaluation of alternative I-language models on CONCRETE and ABSTRACT words in the triad task. The length of the random walk α was varied, and maximal and minimal values of r were overall similar regardless of α. Optimal performance was obtained when $\alpha = 0.1$ for CONCRETE words and $\alpha = 0.85$ for ABSTRACT words, suggesting that default settings ($\alpha = 0.75$) were close to optimal for abstract words but not for concrete words. One implication is that the representation of concrete words does not reflect indirect paths as much as the representation of abstract concepts.

have had some effect on model performance: especially for basic level comparisons, a high value of alpha might introduce longer paths which might add more thematic information at the expense of shorter category-specific paths. For this reason, we also calculated the results for other values of alpha. Figure 3 shows that this was indeed the case, and smaller α values further improved the results: $r = .80$ for $\alpha = .05$. This suggests that even the modest improvement found when visual information was added to the I-language model was somewhat overestimated when using the default value for α.

Study 2: Pairwise Similarity Experiments

In the previous study we evaluated visual and affective grounding when people perform relative similarity judgments between basic-level concrete or abstract words. To see if these findings generalize to a larger set of concepts and a different paradigm, we evaluated the same models on multiple datasets containing pairwise semantic similarity ratings, including some that were collected specifically to compare language-based and (visual) multimodal models.

Unlike the CONCRETE triad task in Study 1, most of the existing datasets include a cover wide range of concrete semantic relations rather than just taxonomic basic level ones. As we argued earlier, this might be a relatively insensitive way of gauging the effect of grounded visual or affective representations because perceptual (and potentially affective) properties are especially important at the basic level (Rosch et al., 1976). However, the extensive size of these datasets allows us to impose restrictions on the semantic relations under consideration. This has several advantages. First, it will

allow us to investigate whether the results from Study 1 only apply to concrete concepts on the basic level. Second, while most of the new datasets contain mostly 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 important to determine whether this finding replicates and generalizes to different tasks.

Datasets

We consider here five different datasets of pairwise similarity ratings.[6]. Three are more typical, and include 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). Two more recent datasets were additionally included because they permit is to better directly address the role of visual and affective grounding. 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 words. It thus allows us to extend our findings beyond concrete nouns to verbs and investigate whether the important role of affective grounding found in Study 1 replicates here.

Each of the datasets is slightly different in terms of procedure, stimuli and semantic relations of the word pairs being judged. The next section explains these differences and also reports on their internal reliability, which sets a bound on the prediction of the models we want to evaluate.

MEN. The MEN dataset (Bruni, Uijlings, et al., 2012) was constructed specifically for the purpose of 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, a text corpus model was used to derive cosine values from the first 1,000 most similar items, 1,000 pairs were sampled from the 1001-3000 most similar items and the last 1000 items from the remaining items. As a result, the MEN consists of concrete words that cover a wide range of semantic relations. The reported reliability of $\rho = .84$ was obtained by correlating two subsets of ratings.

MTURK-771. This MTURK-771 dataset (Halawi et al., 2012) was constructed to include various types of relatedness. It consists of nouns taken from WordNet that are synonyms, have a meronymy relation (e.g., *leg – table*) or a holonymy relation (e.g., *table – furniture*). Graph distance varied between 1 and 4. Both words in each pair had to be frequent, achieved by only keeping pairs that occurred with at least 40 million search results on the Yahoo! Web search engine. The variability

[6]Most of these datasets were also used in a previous study in which evaluate count-based and prediction based E-language 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).

of depth and type of relation suggests that this dataset is quite varied. The reliability, calculated as the correlation between random split subsets, was .90.

SimLex-999. The SimLex-999 dataset (Hill et al., 2016) 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 a low ratings, whereas highly similar (but potentially weakly associated) items would be rated highly. Inter-rater agreement between all pairwise individual respondents was $\rho = .67$. Of interest for our work is that the inter-annotator agreement was higher for abstract than concrete concepts $\rho = .70$ vs $\rho = .61$ and lower for nouns compared to adjectives or verbs $\rho = .61$ vs .79 and .72. An inter-rater reliability of .78 was calculated over split-half sets (comparable to the other datasets) in a subsequent study (Gerz et al., 2016).

Silberer2014. The Silberer dataset (Silberer & Lapata, 2014) consisted 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. Inter-rater reliability, calculated as the average pairwise ρ between the raters, was .76 for the semantic judgments and .63 for the visual judgments.

SimVerb-3500. The SimVerb-3500 dataset (Gerz et al., 2016) was constructed to remedy the bias in the field towards 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 associative relatedness. Items were selected from the USF 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, $\rho = .86$.

Visual grounding

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. As the first two panels of Figure 4 demonstrate, adding visual information did not improve the I-language models and only slightly improved the E-language models.[7]

[7] Correlations for all datasets are reported fully in Appendices 4 and 5.

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. Interestingly, the I-language model performed far better than the E-language model on this dataset. Visual grounding did not improve the performance of the I-language model but resulted in considerable improvement in the E-language model (Δr = .14). One (speculative) possibility 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.

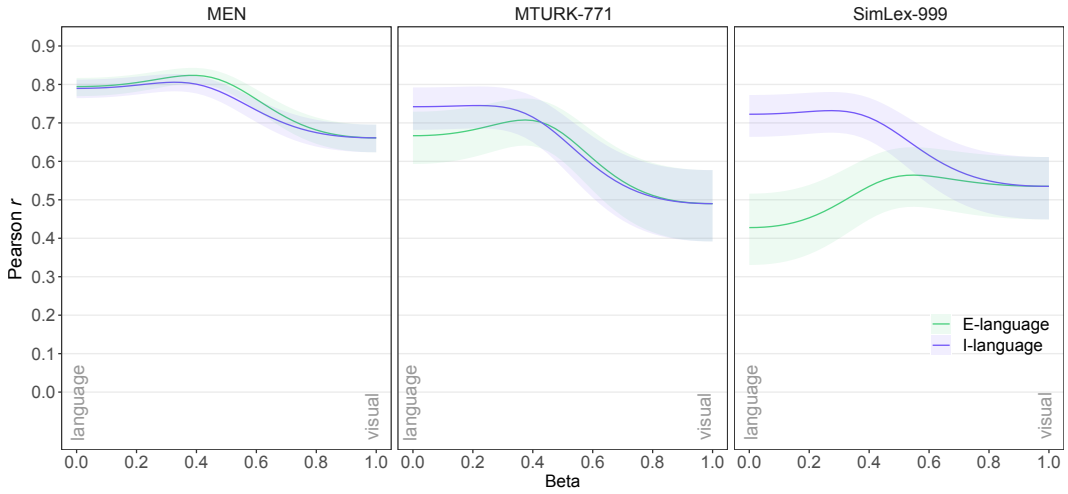

Figure 4. Effects of adding visual information to language models in predicting pairwise similarity ratings for datasets consisting of diverse semantic comparisons. In contrast to the first two panels (MEN, MTURK-771), the SimLex-999 consists of strict similarity ratings (see text). Adding visual information did not improve performance of the I-language models (which performed better overall) but did improve the E-language model, especially on the SimLex-99 dataset.

Comparing visual and semantic judgments. The most relevant study for the current purpose is the Silberer2014 data in which all words consisted of concrete nouns taken from the McRae feature generation norms (McRae et al., 2005). However, since words across different categories were compared, not all ratings reflect basic-level comparisons. In contrast to the other studies, two types of ratings were collected, semantic ratings, in which participants judged the similarity like most other studies, and visual ratings. In the latter task, participants only judged the similarity of the appearance of the concepts. If the visual features capture these properties, we expect it to provide a relatively better fit for these judgments in comparison to models in which visual features are not fully encoded such as the E-language model.

The results for the Silberer2014 dataset are shown in Figure 5 and Appendices 4 and 5. As before, the I-language models better predict people's similarity judgments, this time for both all semantic judgments as well as just visual judgments. Visual information resulted in significant improvement, but affective information did not (see Appendix 5). However, the improvements are

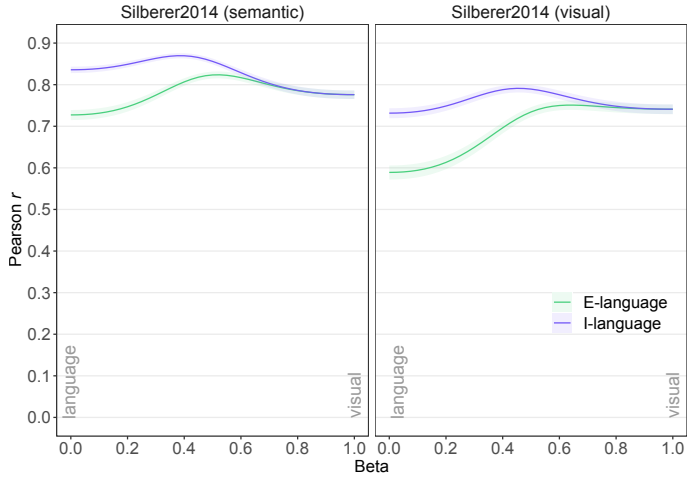

Figure 5. Results of visual grounded language models based on pairwise similarity ratings from the Silberer2014 dataset. The left panel shows semantic judgments, whereas the right panel shows judgments where participants were instructed to consider visual similarity.

quite small, even when participants were explicitly instructed to judge visual similarity ($\Delta r = .03$ for the semantic judgments and $\Delta r = .06$ for the visual judgments).

Overall performance was lower for the E-language model, but relatively better on semantic judgments than visual judgments. Adding visual information results in an improved prediction in both models, which is especially pronounced for visual judgments ($\Delta r = .09$ for the semantic judgments and $\Delta r = .16$ for the visual judgments). As in the SimLex-999 dataset, using a unimodal visual model results in better performance than using a unimodal E-language model, which suggests that visual representations provides a better approximation of the meaning of concrete entities than representations derived from external language corpora.

That said, the complete Silberer2014 dataset contains both basic level within-category comparisons like *dove – pigeon* as well as broader comparisons like *dove – butterfly*. However, Study 1 involved only basic level comparisons. In order to provide a better comparison, 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. Words for which no clear superordinate label could be assigned were not included. This reduced the number of pairwise comparisons 5,799 to 1,086, which is still sufficiently large for our purposes.

As Figure 6 demonstrates, the overall correlations were lower, supporting idea that the basic-level presents a considerable challenge to language-based models. However, the qualitative patterns remain identical. The I-language models performed better across the board and adding visual information improved the I-language models less ($\Delta r = .14$) than it did E-language models ($\Delta r = .32$), although in both cases the improvements were larger than on the full Silberer2014 dataset. As before, we

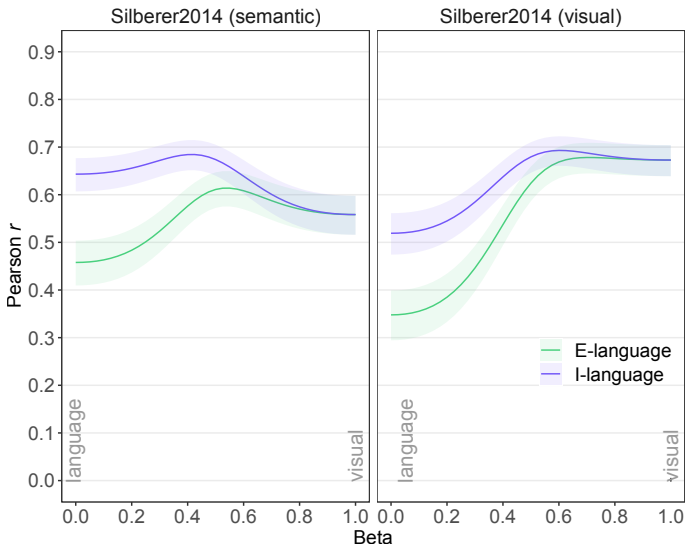

Figure 6. Results of adding visual information to language models for a subset of basic level items from the Silberer2014 dataset. As before, the left panel shows semantic judgments, whereas the right panel shows judgments where participants were instructed to consider visual similarity only. Visual information improves performance of the E-language models for both kinds of judgments, and the I-language models for visual judgments only.

find that the model which includes visual information only (r = .56) outperforms the E-language model (r = .47) for the semantic Silberer2014 judgments. It also outperformed the I-language model, albeit only slightly. This is fairly unsurprising given the fact the basic level encodes more perceptual information than comparisons across different categories.

Affective grounding

To investigate the effect of affective grounding, we took the previous datasets and supplemented them with SimVerb-3500 (Gerz et al., 2016), which contained pairwise similarity ratings for verbs. A sizable subset of these items was included in the affective norms of Warriner et al. (2013). The results are shown in Figure 7. As before, we find that the I-language model provides better estimates of pairwise similarity than the E-language models, except for the MEN dataset, where they are on par. However, in contrast to the findings for visual information and the ABSTRACT triads in Study 1, most of our datasets this time show no improvement when affective information is added. The only exceptions are the SimLex-999 and SimVerb-3500 datasets, where adding affective information slightly improved the E-language model.

There are at least two reasons why the affective information may have provided less benefit here than in the triad task. First, not all datasets involved comparisons between coordinate items at the basic level. 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. The only datasets that contain

a reasonable number of abstract words are SimLex-999 and SimVerb-3500, which are the only ones for which affective information improves performance.

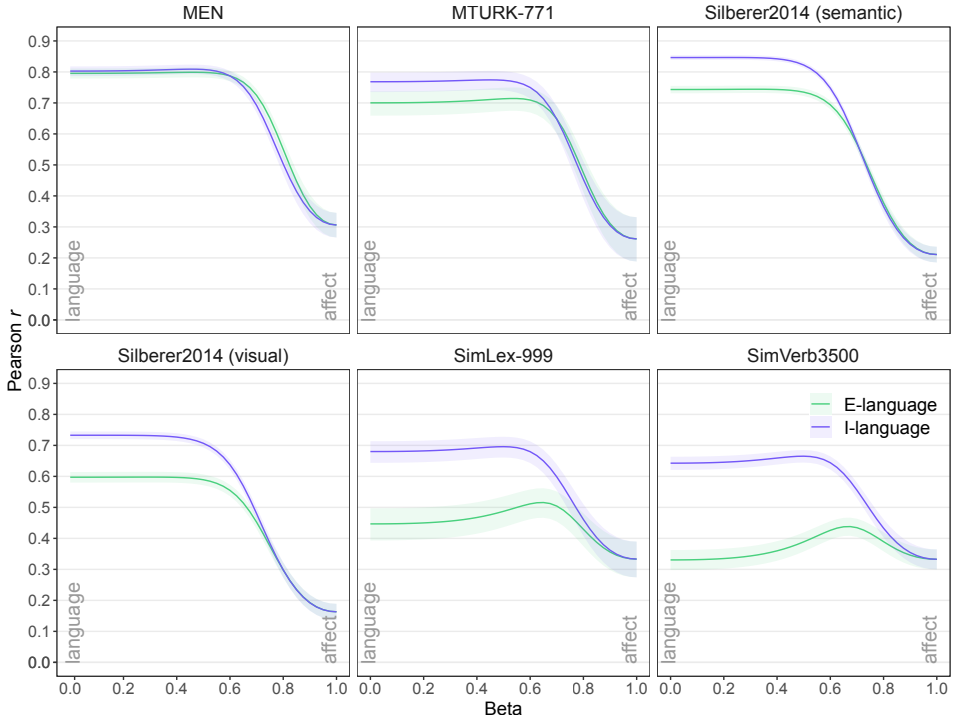

Figure 7. Results of adding affective information to language models for all datasets. Affective information does not improve performance of either kind of language model, except for slight improvements to E-language models for the datasets incorporating verb similarities (bottom row).

To investigate whether these differences from Study 2 resulted from the fact that these datasets did not mainly contain abstract words, we screened each word in these datasets using the concreteness norms from Brysbaert et al. (2014). As before, 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, resulting in 391 words from SimLex-999 and 1973 from SimVerb-3500. The results, shown in Figure 8, show that for abstract words, affective grounding substantially improves the performance of the E-language model but only slightly improves the I-language model. This is consistent with Study 1 and extends its results to a different task, and a more varied set of words including verbs.

Comparison to previous work

Because the datasets in Study 2 have appeared in a number of recent studies, we can perform an empirical comparison between our results and theirs.[8] The Silberer2014 dataset consists of semantic

[8]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 = 5799 for the semantic judgments, n = 5777 for the visual

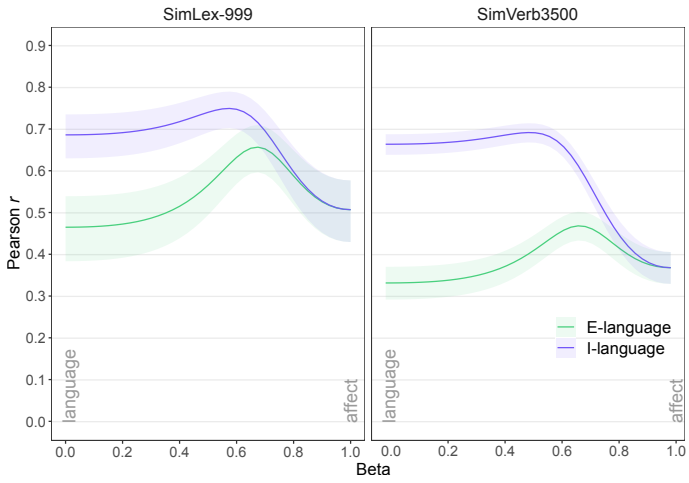

Figure 8. Replication of role of affective information for abstract words from Study 1 using only abstract words taken from the SimLex-999 and SimVerb-3500 datasets.

and visual similarity ratings taken from noun pairs presented in the McRae et al. (2005) feature norms. In the study by Silberer et al. (2013), one of the multimodal representations was derived by combining a distributional model induced from these feature norms with a visual model that was trained to predict visual attributes for the corresponding images in ImageNet. The feature norms are especially of interest because they provide an alternative to the word association-based I-language model. For one thing, they allow more precise propositional statements that capture something about the semantic relation as well (a duck <has a> bill, <is a> bird, etc). Moreover, in contrast to word associations, their instructions appeal to core properties that *define* meaning, rather than allow any kind of verbal response to contribute (De Deyne, Verheyen, Navarro, Perfors, & Storms, 2015).

How do our results compare? For semantic judgments, Silberer et al. (2013) reported a correlation of $r = .71$ with a pure feature-based model, and $r = .49$ for a purely visual model (ours were $r = .84$ for association-based and $r = .78$ for visual). For visual judgments, they reported correlations of $r = .58$ for the feature-based model and $r = .52$ for the visual model (ours were $r = .73$ for association-based and $r = .74$ for visual). When the information from visual and feature-based models were combined in a multimodal representation, Silberer et al. (2013) reported a correlation of $r = .68$ for the semantic judgments, representing a slight decrease from the performance of the purely feature-based model (i.e., $r = .71$). For the visual judgments, the multimodal model produced a correlation of $r = .62$ a slight improvement on the performance of the feature-based model (i.e., $r = .58$). By way of comparison, our multimodal model produced correlations of $r = .87$ for the semantic judgments and $r = .79$ for visual judgments.

Taken together, this comparison suggests that the I-language model derived from word association data performs better than a model based on feature norms even when most words were concrete

judgments) and MEN (n = 942). We will therefore assume some robustness even when some items are missing.

nouns. Similarly, the visual features used in the current study (extracted using convolutional neural networks; CNN) outperformed the image vectors derived in Silberer and Lapata (2014).

There are other studies that are comparable to ours in the sense that they combine visual features with E-language representations. One of the first studies of this kind evaluated performance on the MEN dataset (Bruni et al., 2014). In that paper, the E-language model relied on a sliding window approach to count words in 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). Correlations to the MEN dataset for the E-language model were $r = .73$ (compared to $r = .79$ obtained in our study) and for the image model were $r = .43$ (compared to $r = .66$ here). When the visual information was added to the E-language model, the correlation rose to $r = .78$ (compared to $r = .82$ for our model).

In a second study using the MEN dataset, an E-language model was trained on Wikipedia and used the skip-gram `word2vec` model, and two different methods were considered for the visual features (Kiela & Bottou, 2014). One method used a bag-of-visual words approach with SIFT features, and the other used a CNN to derive features from ImageNet. The E-language model alone achieved correlations of $r = .62$ (compared to our $r = .79$). Interestingly, the method of adding visual information matters: using SIFT-based features as the visual model produced a correlation of $r = .40$ whereas their CNN-based features led to a correlation of $r = .63$ (ours was $r = .66$). The multimodal model results were .70 and .72 for SIFT and CNN-based features, lower than our multimodal r = .82.

More recent work allows an even closer comparison to the current study since it uses a CNN approach for the visual features and skip-gram for the E-language representations (Lazaridou et al., 2015). Although their E-language model is not psychologically motivated – it is based on the entire English Wikipedia – it provides an approximation of what can be encoded through a slightly less natural language register. Their E-language model alone achieved correlations of $r = .68$ for MEN (compared to our $r = .79$), $r = .29$ for SimLex-999 (compared to our $r = .43$), and $r = .62$ and $r = .48$ (compared to our $r = .73$ and $r = .59$) for the semantic and visual judgments in Silberer2014, respectively. Their CNN visual model alone achieved correlations of $r = .62$ for MEN (compared to our $r = .66$), $r = .54$ for SimLex-999 (compared to our $r = .54$), and $r = .55$ and $r = .54$ (compared to our $r = .77$ and $r = .73$) for the semantic and visual judgments in Silberer2014, respectively. Finally, their best-performing multimodal model gave correlations of $r = .76$ (MEN), $r = .53$ (SimLex-999), $r = .72$ (Silberer2014 semantic), and $r = .63$ (Silberer2014 visual) — an indication of consistent improvements, although smaller on the Silberer2014 datasets than the others and smaller than those reported here (see Appendix 5).

Altogether, the comparison with previously work show that our E-language and visual models are at least comparable with state-of-the-art and usually offer notable improvements.

Robustness

As in Study 1, in order to evaluate the extent to which our results depend on specific modelling choices, we consider a variety of alternative E-language and I-language models. In order to do this, we evaluated the models on the Study 2 datasets using the word pairs shared for the GloVe vectors based on 6 billion and 840 billion tokens and our word2vec based model. If the results from Study 1 replicate, we expect performance to be similar to the E-language model based on word2vec, except when the corpus size is extremely large. In that case, we would expect the visual but not the affective grounding problem to be reduced.

The results, shown in Figure 9A, demonstrate that the overall correlations using the 6 billion word GloVe-based E-language model alone ($r = .60$ 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 E-language 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 most perceptual properties were better predicted with a larger corpus (see Figure 9C). In line with Study 1, the E-language model based on the larger corpus did not lead to substantial improvements for abstract concepts (Figure 9D) compared to the 6B corpus model. Regardless of the model, correlations remained moderate and 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 E-language specifically lacks affective-based information for abstract concepts.

General Discussion

In two studies we investigated how the grounding problem affects meaning derived from language in abstract and concrete concepts. By contrasting internal and external language, we were able to identify what information the linguistic environment provides, and how this information might be co-determined by imagery and affect in internal language. In both studies, we found that grounded affective and visual information is needed to supplement models based on external language in order to account for meaning. Our findings replicated and extended previous work addressing visual grounding. We additionally identified a novel and substantial effect of affective information for abstract concepts.

One of our most consistent findings was that using internal language models derived from word associations resulted in accurate predictions for both concrete and abstract concepts defined at a basic level. This is remarkable because the comparisons at the basic level may especially rely on

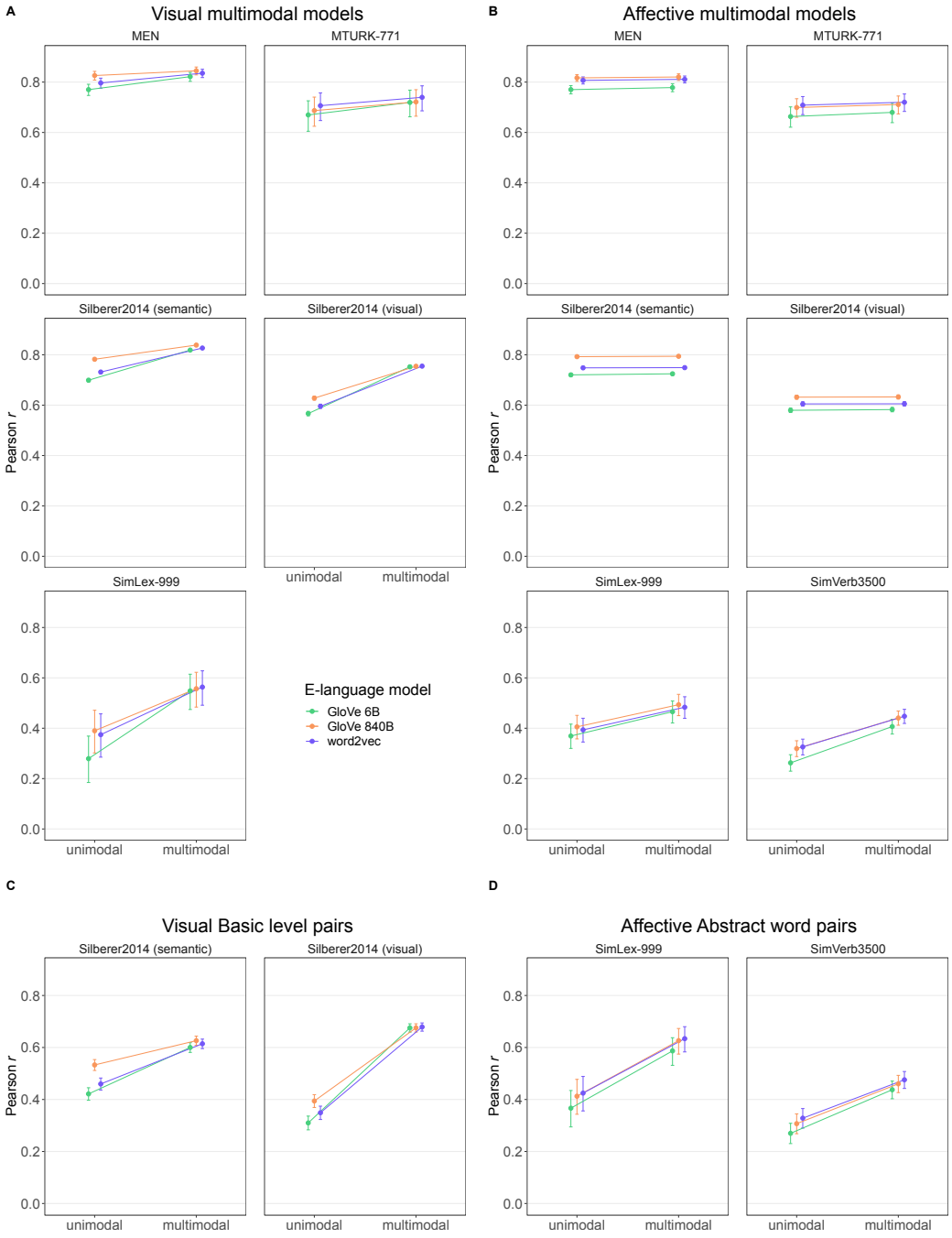

Figure 9. A comparison of the effect of visual and affective grounding in three different E-language models: `word2vec` (W2V) and `GloVe`-6B (6 billion words) and `GloVe`-840B (840 billion words). Top panels show visually grounded (Panel A) and affectively grounded (Panel B) multimodal model predictions for optimal mixing parameters β. The bottom panels show the findings when considering a subset of concrete word pairs at the basic level and words (Panel C) and abstract affective words (Panel D).

grounded visual or affective information: concrete words rely on perceptual features and abstract words incorporate many affective properties. Despite this, providing visual or affective grounding improves the results of the internal model only slightly. Taking into account that most variance is accounted for, this suggest that in common categories like *fruit*, *tools*, *feelings* or *states*, such properties are sufficiently encoded in models built from word associations. This contrasts with our external language models, which benefited far more obviously from additional visual and affective features.

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 pronounced 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 external language models in two of the studies (see Figure 8). This is difficult to explain given that it is often assumed that abstract concepts should be predominantly acquired through language exposure according the distributional hypothesis. Instead, our findings support the Affective Embodiment Account (AEA) by Vigliocco et al. (2009) which suggest that affect is considered to be another type of experiential information (along with vision and other sensorimotor information).

A relatively recent model that learns image features using neural networks accounts for a reasonable amount of variance and this work presents an improvement over previously reported studies using similar datasets. More surprisingly, image-vectors by themselves provided a better account for the basic-level triad preferences than the E-language model (see Figure 1). This finding was also found for two similarity judgment tasks consisting of strict similarities (SimLex-999) and the large-scale dataset on similarity ratings for concrete nouns from the McRae data (Silberer & Lapata, 2014) (see Figures 5 and 5). In both cases visual features provide a better account than text features, which raises some questions about whether the language modality is the primary source through which basic-level representations are acquired. Equally disconcerting is that the only way to improve the performance of E-language on concrete concepts requires extremely large corpora outstretching human information processing capacities.

It is unlikely that our 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, procedures, and corpora. 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.

Finally, the results from Study 2 also support for the idea that the level of comparison matters: the performance of the models is quite comparable when the semantic relation varies widely, and starts to become more differentiated for strict similarity ratings, and comparisons at the basic level. Even part-of-speech can strongly determine how well the models perform, with large differences found between I-language ($r = .70$) and E-language ($r = .30$) predictions for verbs.

Implications for uni- and multi-modal E-language models

In recent years, a new generation of lexico-semantic 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 prediction models only partially capture meaning, especially at levels determined by sensory or affective properties. We suggest that many of the previous evaluation tasks in the literature do not require accurate modal specific representations of meaning. In those cases no advantage for grounding information would be found. In this study, we tried to counteract this in two ways. First, we attempted to improve the psychological plausibility of the models (for example by choosing a representative text corpus). Second, we tried to select appropriate levels of comparison like the basic level where perceptual similarity is an important organizing factor Rosch et al. (1976). Of course, our work relies on the assumption that the current models are reasonable approximations of what meaning can be derived from language. It remains possible that better E-language models will reduce the need for additional grounding. However, in both Study 1 and 2, our results did not depend on the size of the corpus or the way the embeddings were obtained. Furthermore, previous findings for a variety of multimodal models suggested that both the E-language and visual models we used are the current state-of-the-art.

Similarly, our results also hinge on the modal-specific visual and affective representations we used. We derived these according to both theoretical considerations (are the models 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?). Given the recency of the models we used, further improvements are to be expected, which could indicate that perceptual and affective information are even more important than we estimate. Even so, 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.

The current work is an example of how contrasting I- and E-language can be used to evaluate different theories of representation. Do word associations reflect word co-occurrence statistics, as argued by some? For example, in a recent study by Nematzadeh, Meylan, and Griffiths (2017) using topic and word embedding models the correlation with word association strength was .27 for the best performing model, which if anything indicates that a large portion of the variance remains unexplained. Despite some overlap between E-language and I-language, their differences might be more interesting, indicating that they tap into different types of information. Along these lines, text models fail to predict word associations not because of unreliability in either type of language but because word associations tap into different types of representations. Underlying representational differences could explain why text models struggle to predict the color of objects, such as the fact that grass is green (Bruni, Boleda, Baroni, & Tran, 2012) and do not capture the association between

concrete concepts and their typical attributes (Baroni & Lenci, 2008).

While large shared evaluation datasets have proven to be useful tools to benchmark incrementally improving language models, the high correlations for some of them (e.g., MEN) can be somewhat misleading, despite the fact that the sample sizes are quite large. In particular, the lack of abstract concepts and verbs, and the limited number of basic-level comparisons for the mostly concrete concepts might inflate the performance of many of these models to a point where only ceiling effects are obtained. In those cases, the need for deeper grounding might be obscured. The limitations of E-language models are not necessarily restricted to the basic level as well. Recent work has shown that E-language only capture a fraction of the variance in a task that covers the other end of the spectrum of semantic relatedness when humans are asked to judge remote unassociated words, whereas an I-language model provided a good account for these remote triads when spreading activation was considered (De Deyne, Navarro, et al., 2016; De Deyne, Perfors, & Navarro, 2016). If so, the lack of sensory specific grounding is compounded by the fact that inference over semantic relations in text-models only partially captures human semantic reasoning.

Affect and the grounding of abstract words. In line with the AEA account we found substantial improvements when adding affective grounding to external language models' accounts of abstract concepts. Moreover, the improvements compared to visual multimodal accounts were substantial. As far as we know, this is a novel finding showing how affective grounding can improve the representation of these concepts.

Even though the performance for E-language models for abstract words is fairly poor, other factors than affect could contribute to the problem. While affect, and especially valence, has been proposed as a psychological primitive involved not only in abstract words but central to our attitudes, decision making and even the perception of everyday objects (Barrett & Bliss-Moreau, 2009), other features including properties about social interaction (Barsalou & Wiemer-Hastings, 2005), morality (Moll, Zahn, de Oliveira-Souza, Krueger, & Grafman, 2005), or emotions might also be of importance. For one, the Osgood affective model based on three dimensions remains somewhat coarse thanks to its low-dimensional nature. Instead of three factors, richer representations of affect as emotions have also been proposed. 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. In this study, we investigated this possibility in an additional analysis using the NRC Emotion lexicon, which contains judgments for the Ekman emotions for over 14,000 English words (Mohammad & Turney, 2013) to instantiate a broader emotion modality. We found very limited evidence for any contribution of emotions above that of affect: they did not capture the similarities derived from Experiment 1 or the rated similarity datasets in Study 2 as well, despite having more features.

These findings converged with recent neuro-imaging findings in which affect but not emotions were found to determine how abstract concepts were instantiated in the brain (Wang et al., 2017, see introduction).

A second issue is the question of how communication through language is used to learn about the core affective properties of words and to what extend an embodied approach is required to understand the meaning of the affect of words. A priori, language should play an important role because both affect and emotions derived from it are not only expressed as internal states, but play a social role in communication as well. However, affect is also communicated in non-verbal ways, through facial expressions (Cacioppo et al., 2000). In line with the previous point about the role of affect and emotions, research suggests that non-linguistic 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 be partially acquired 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 provide useful cues about valence in the form of the word, and evidence shows that affective congruency between sound and meaning leads to a processing advantage in word recognition (Aryani & Jacobs, 2018). While both factors are likely to contribute to learning affect and the meaning of abstract words, it is unlikely that any factor in itself is sufficient. For one, most of the abstract words a person knows are acquired through reading (Landauer, 2007), and therefore it is not directly clear how acoustic information would provide affective grounding in this case. Moreover, in line with the AEA proposal, the acquisition of affective concepts might rely on embodied processes involving empathy, where people put themselves in someone else's situation and imagine how they would feel.

So far we have remained vague about whether affect extends beyond abstract concepts. Partly this reflects the fact that investigations about affective grounding are relatively new. Our data presented here suggest that the effect sizes for affective grounding are much smaller for concrete words. This could be due to the fact that we used a subset of mostly abstract nouns, whereas previous results for adjectives have shown large effects for valence and arousal (De Deyne et al., 2014). In any case, any conclusions in this domain are likely to be preliminary.

Overall, this work suggests that while external language models have the general capacity to approximate human semantic cognition, their prediction accuracy is substantially improved by combining them with grounded information in the form of perceptual and affective features. This is especially true for abstract concepts and when the representations are derived from psychologically plausible corpora.

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, USA: Association for Computational Linguistics.

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science*, *6*, 359–370.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463–498.

Aryani, A., & Jacobs, A. M. (2018). Affective Congruence between Sound and Meaning of Words Facilitates Semantic Decision. *Behavioral Sciences*, *8*(6). Retrieved from http://www.mdpi.com/2076-328X/8/6/56

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). Curran Associates, Inc.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 238–247).

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, *20*(1), 55–88.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, *41*, 167–218.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Grounding cognition: The role of perception and action in memory, language, and thought. In D. Pecher & R. A. Zwaan (Eds.), (pp. 129–163). Cambridge University Press.

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.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin*, *143*(3), 263-292.

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.

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). ACM.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

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., et al. (2000). The psychophysiology of emotion. *Handbook of emotions*, *2*, 173–191.

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*.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Collell, G., & Moens, M.-F. (2016). Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 2807–2817).

Collell, G., Zhang, T., & Moens, M.-F. (2017). Imagined visual representations as multimodal embeddings. In *Proceedings of the The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 4378–4384).

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.

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*(2), 159–190.

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. (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 518–523). Cognitive Science Society.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*. (in press)

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.

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.

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 *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 1861–1870). Osaka, Japan.

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., 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.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

Fellbaum, C. (1998). *WordNet: An electronic lexical Database*. MIT Press. (Available at http://www.cogsci.princeton.edu/ wn)

Firth, J. R. (1968). *Selected papers of JR firth, 1952-59*. Indiana University Press.

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*.

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 KnowledgeDdiscovery*

and Data Mining (pp. 1406–1414).

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335–346.

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).

Hill, F., & Korhonen, A. (2014). Learning abstract concept embeddings from multi-modal data: Since you probably can't see what I mean. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 255–265).

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

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*, 103–120.

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).

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).

Kipper, K., Snyder, B., & Palmer, M. (2004). Extending a verb-lexicon using a semantically annotated corpus. In *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). The computer and literacy studies. In A. J. Aitken, R. W. Bailey, & N. Hamilton-Smith (Eds.), (pp. 153–165). 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*(1), 14.

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, *112*(3), 473–481.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 3–35). Lawrence Erlbaum Associates.

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*.

Li, L., Malave, V., Song, A., & Yu, A. (2016). Extracting human face similarity judgments: Pairs or triplets. *Journal of Vision*, *16*, 719–719.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

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.

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.

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.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. , *29*(3), 436–465.

Moll, J., Zahn, R., de Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience*, *6*(10), 799.

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.

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 *Proceedings of the 39th annual meeting of the Cognitive Science Society* (pp. 859–864).

Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press, Inc.

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, *30*(4), 583–593.

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

Paivio, A. (2013). Dual coding theory, word abstractness, and emotion: A critical review of Kousta et al.(2011). *Journal of Experimental Psychology: General*, *142*, 282-287.

Paivio, A. (2014). *Mind and its evolution: A dual coding theoretical approach*. Psychology Press.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). ACL.

Plutchik, R. (1994). *The psychology and biology of emotion.* HarperCollins College Publishers.

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

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.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *Quarterly Journal of Experimental Psychology*, *68*(8), 1584-1598.

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.

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.

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

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... others (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252.

Schirmer, A., & Kotz, S. A. (2003). Erp evidence for a sex-specific stroop effect in emotional speech. *Journal*

of cognitive neuroscience, *15*(8), 1135–1148.

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).

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). ACL.

Steyvers, M. (2010). Combining feature norms and text data with topic models. *Acta Psychologica*, *133*(3), 234–243.

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Press.

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*(7), 1767–1777.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*, 219–247.

Von Ahn, L. (2006). Games with a purpose. *Computer*, *39*(6), 92–94.

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., ... Bi, Y. (2017). Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 1–14.

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.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413.

Appendices

Stimuli Study 1

Table 2

Concrete triad stimuli in Experiment 1.

Natural Kinds

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

Artifacts

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

Table 3

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

Detailed statistics Study 2

Table 4

*Pearson correlation and confidence intervals for correlation differences Δr between unimodal **I-language** r_{v_a} (column 3), and the optimal multimodal model correlation r_{max} (column 7) with sensory modality v_b (Vis. = visual, Aff. = affect).*

Dataset	n	r_{v_a}	CI	v_b	β	r_{max}	CI	Δr	CI
MEN	942	.79	.76, .81	Vis.	0.45	.81	.78, .83	.02	.01, .02
MTURK-771	260	.74	.68, .79	Vis.	0.33	.75	.69, .79	.00	-.01, .02
SimLex-999	300	.72	.66, .77	Vis.	0.43	.73	.68, .79	.01	-.01, .03
SimVerb-3500	133	.46	.31, .58	Vis.	0.23	.46	.33, .60	.00	-.02, .02
Silberer2014 (Sem.)	5799	.84	.83, .85	Vis.	0.58	.87	.86, .88	.03	.03, .04
Silberer2014 (Vis.)	5777	.73	.72, .74	Vis.	0.65	.79	.78, .80	.06	.05, .07
		.71				.73		.02	
Silberer2014 (Basic-Sem.)	1086	.64	.61, .68	Vis.	0.43	.68	.65, .71	.04	.03, .06
Silberer2014 (Basic-Vis.)	1086	.52	.47, .56	Vis.	0.55	.69	.66, .72	.17	.14, .21
MEN	1981	.80	.78, .82	Aff.	0.45	.81	.79, .82	.01	.00, .01
MTURK-771	653	.77	.73, .80	Aff.	0.45	.77	.74, .80	.00	.00, .01
SimLex-999	913	.68	.65, .71	Aff.	0.50	.69	.65, .72	.01	.00, .01
SimVerb-3500	2926	.64	.62, .66	Aff.	0.50	.65	.63, .67	.01	.01, .02
Silberer2014 (Sem.)	5428	.84	.84, .85	Aff.	0.23	.84	.84, .85	.00	.00, .00
Silberer2014 (Vis.)	5405	.73	.72, .74	Aff.	0.00	.73	.72, .74	.00	.00, .00
		.75				.75		.01	
SimLex-999 (Abstract)	391	.69	.63, .74	Aff.	0.58	.75	.70, .79	.06	.04, .10
SimVerb3500 (Abstract)	1973	.66	.64, .69	Aff.	0.50	.69	.67, .71	.03	.02, .04
		.68				.72		.05	

Table 5

*Pearson correlation and confidence intervals for correlation differences Δr between unimodal **E-language** r_{v_a} (column 3), and the optimal multimodal model correlation r_{max} (column 7) with sensory modality v_b (Vis. = visual, Aff. = affect).*

Dataset	n	r_{v_a}	CI	v_b	β	r_{max}	CI	Δr	CI
MEN	942	.79	.77, .82	Vis.	0.38	.82	.80, .84	.03	.02, .04
MTURK-771	260	.67	.59, .73	Vis.	0.38	.71	.64, .76	.04	.01, .08
SimLex-999	300	.43	.33, .52	Vis.	0.55	.56	.48, .64	.14	.07, .21
SimVerb-3500	133	.11	-.06, .28	Vis.	0.43	.14	-.03, .30	.03	-.08, .14
Silberer2014 (Sem.)	5799	.73	.71, .74	Vis.	0.53	.82	.82, .83	.10	.09, .10
Silberer2014 (Vis.)	5777	.59	.57, .61	Vis.	0.65	.75	.74, .76	.16	.15, .17
		.55				.64		.08	
Silberer2014 (Basic-Sem.)	1086	.46	.41, .50	Vis.	0.55	.61	.58, .65	.16	.12, .19
Silberer2014 (Basic-Vis.)	1086	.35	.29, .40	Vis.	0.70	.68	.64, .71	.33	.28, .38
MEN	1981	.80	.78, .81	Aff.	0.45	.80	.78, .81	.00	.00, .00
MTURK-771	653	.70	.66, .74	Aff.	0.53	.71	.67, .75	.01	.00, .02
SimLex-999	913	.45	.39, .50	Aff.	0.65	.63	.47, .56	.07	.04, .10
SimVerb-3500	2926	.33	.30, .36	Aff.	0.68	.44	.41, .47	.11	.08, .13
Silberer2014 (Sem.)	5428	.74	.73, .75	Aff.	0.33	.74	.73, .76	.00	.00, .00
Silberer2014 (Vis.)	5405	.60	.58, .61	Aff.	0.30	.60	.58, .61	.00	.00, .00
		.60				.63		.03	
SimLex-999 (Abstract)	391	.47	.38, .54	Aff.	0.68	.66	.60, .71	.19	.13, .26
SimVerb3500 (Abstract)	1973	.33	.29, .37	Aff.	0.68	.47	.43, .50	.14	.11, .17
		.40				.56		.16	