Behav Res (2017) 49:471–483
DOI 10.3758/s13428-016-0717-1

Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor

Eileen R. Cardillo[1,3] · **Christine Watson**[2] · **Anjan Chatterjee**[1,3]

Published online: 8 March 2016
© Psychonomic Society, Inc. 2016

Abstract As the cognitive neuroscience of metaphor has evolved, so too have the theoretical questions of greatest interest. To keep pace with these developments, in the present study we generated a large set of metaphoric and literal sentence pairs ideally suited to addressing the current methodological and conceptual needs of metaphor researchers. In particular, the need has emerged to distinguish metaphors along three dimensions: the grammatical class of their base terms, the sensorimotor features of their base terms, and the syntactic form in which the base terms appear. To meet this need, we generated nominal metaphors (and matched literal sentences) using entity nouns as the base terms, with the intention that they be used in concert with already published sets of predicate metaphors or nominal metaphors using event nouns. Using the results of three norming experiments, we provide 120 pairs of closely matched metaphoric and literal sentences that are characterized along 14 dimensions: 11 at the sentence level (length, frequency, concreteness, familiarity, naturalness, imageability, figurativeness, interpretability, ease of interpretation, valence, and valence judgment reaction time), and three related to the base term (visual, motion, and auditory imagery). These items extend previously published stimuli, filling an extant gap in metaphor research and allowing for tests of new behavioral and neural hypotheses about metaphor.

Keywords Figurative language comprehension · Nominal metaphors · Sentence norms

Despite its poetic associations, metaphoric language plays a vital and frequent role in everyday language. Most obviously, a metaphor may be preferred over a literal expression in order to foster fresh insight or capture attention by virtue of its novelty. By expressing familiar concepts in new ways, subtleties of meaning may emerge. Metaphors also make powerful learning tools. As the oft-cited likening of the orbit of electrons around an atom to the orbit of planets around the sun illustrates, we more readily understand new concepts through familiar ones. Perhaps most importantly, metaphors may allow us to conceptualize and communicate about abstract concepts (Jamrozik, McQuire, Cardillo, & Chatterjee, in press; Lakoff & Johnson, 1980, 1999). As a vehicle to abstraction, metaphor transcends the domain of sonnets and novels, revealing itself to be a unique and essential feature of human cognition (Gentner, 2003; Gibbs, 1994).

This consensus about the importance of metaphoric thought and language contrasts with the disagreement concerning how we understand it. Several cognitive models of metaphor comprehension have attempted to explain the process. According to one view, comprehension initially relies on a comparison in which the conceptual similarities between two superficially different domains (base and target) are mapped and aligned (Gentner et al., 2001; Gentner & Wolff, 1997). In a different account, comprehension is essentially an act of categorization, whereby a base term is taken as a prototypical member of a newly constructed category, of which the target is also a member (Glucksberg, 2003; Glucksberg &

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0717-1) contains supplementary material, which is available to authorized users.

✉ Eileen R. Cardillo
eica@mail.med.upenn.edu

[1] Center for Cognitive Neuroscience, University of Pennsylvania, 3720 Walnut Street, B51, Philadelphia, PA 19104-6241, USA

[2] Moss Rehabilitation Hospital, Philadelphia, PA, USA

[3] Department of Neurology, University of Pennsylvania, Philadelphia, PA 19104, USA

Keysar, 1990). Alternatively, a process of sensorimotor abstraction may drive comprehension. In this view, the metaphoric sense of the base term is the conceptual core remaining after irrelevant, literal sensorimotor features have been attenuated (Chatterjee, 2008; Chen, Widick, & Chatterjee, 2008). These proposals may not be mutually exclusive. Different processes may predominate at different stages as the metaphoric senses of base terms become conventionalized in their use (Bowdle & Gentner, 2005), and/or different metaphor types may rely on different mechanisms (Cardillo, Schmidt, Kranjec, & Chatterjee, 2010; Schmidt, Kranjec, Cardillo, & Chatterjee, 2010).

These uncertainties regarding cognitive mechanisms are paralleled by uncertainties regarding neural mechanisms. Early patient research suggested a right hemisphere (RH) specialization for figurative language, but it lacked the methodological rigor common to current neuropsychological and neuroimaging investigations (Schmidt et al., 2010). This traditional view is also not well supported by neuroimaging studies, which outnumber patient studies at this point. Functional magnetic resonance imaging (fMRI) and positron emission tomography studies suggest a motley set of critical areas, with poor consensus regarding hemispheric lateralization, not to mention regional differences within hemispheres. Recent meta-analyses, however, have indicated that metaphor comprehension is largely a left-hemisphere-mediated process, with the RH playing a weaker and less consistent role (Rapp, Mutschler, & Erb, 2012). Rather than resolve the question of hemispheric specialization, the current challenge to the field is determining the functional roles of the many regions involved and the conditions that recruit them.

Clarifying these uncertainties about neural substrates entails stimulus optimization at three levels. The first level is psycholinguistic: Lexical and sentential properties that impact processing difficulty must be controlled by careful stimulus design, selection, and/or statistical analyses. Comparisons between metaphors and literal sentences, or familiar and novel metaphors, are common in neuroimaging studies, but standards for matching them vary widely. The second is methodological: Task-related differences are especially important to better characterize in neuroimaging studies and to distinguish from figurativeness-related processes. Passive reading, imagery generation, and decisions about valence, semantic relatedness, or plausibility—all of which are tasks used in fMRI studies of metaphor—entail additional and different cognitive demands beyond metaphor comprehension. Task differences likely contribute significantly to the current poor consensus regarding the neural basis of metaphor comprehension and the necessity of RH engagement (e.g., Yang, Edens, Simpson, & Krawczyk, 2009). The third level is conceptual: Not only is figurativeness sometimes confounded with difficulty (Schmidt & Seger, 2009; Yang et al., 2009), but it may interact with other stimulus properties in theoretically and neurally significant ways.

Previously, we highlighted the three conceptual questions we believe are the most likely to contribute to differences between neural studies and the most useful to investigate in future metaphor research in cognitive neuroscience (Cardillo et al., 2010): (1) Might previous findings about hemispheric lateralization reflect differences in novelty, rather than figurativeness? (2) Are metaphoric words understood by reference to their literal sensory and motor properties, thereby entailing reactivation of sensorimotor cortices for their comprehension? (3) Do metaphors of different syntactic forms (e.g., predicate vs. nominal) rely on different cognitive and neural mechanisms for comprehension? To assist addressing these questions, we previously generated 560 matched literal and metaphorical sentences of different syntactic forms (nominal, predicate) and sensory modalities (auditory, motion) and normed them on ten sentence-level characteristics (frequency, concreteness, length, familiarity, naturalness, imageability, figurativeness, interpretability, valence, and time required to make a valence judgment; Cardillo et al., 2010).

Although progress has been made, these questions about neural organization largely remain outstanding. At this point, the most headway has been made regarding the question of how hemisphericity, figurativeness, and novelty interact. Both the graded salience hypothesis (GSH; Giora, Zaidel, Soroker, Batori, & Kasher, 2000) and the coarse coding hypothesis (CCH; Jung-Beeman, 2005) predict a special role for the RH in metaphor processing. The CCH attributes this lateralization to the RH's specialization for processing semantically distant relationships, which are common in metaphoric expressions. In contrast, the GSH suggests that this lateralization reflects RH dominance for processing low-salience meanings, a property often confounded with figurativeness. Salient meanings, by contrast, are lexicalized, context-independent, and prominent senses of a word. According to the GSH, salience is a composite construct but is "determined primarily by frequency of exposure and experiential familiarity" with a particular word sense (Giora, 2002, p. 491). Given the importance of familiarity to salience, the GSH predicts that the RH processes novel metaphors but the LH is sufficient for understanding familiar ones (Giora et al., 2000). In partial support of these hypotheses, the emerging picture seems to be that *both* hemispheres, and especially bilateral prefrontal cortex, are recruited when deriving a new figurative sense. In a novel familiarization task, Cardillo et al. (2012) parametrically varied participant experience with novel nominal and predicate metaphors to test for a right–left shift in lateralization with increased familiarity. Instead, their results indicated that the more familiar a metaphor was, the less it engaged a subset of the same regions necessary for comprehending the metaphor for the first time [inferior frontal gyrus (IFG) bilaterally, left posterior middle temporal gyrus (pMTG), and right postero-

lateral occipital cortex]. Studies directly comparing metaphors at either end of the familiarity continuum have similarly implicated bilateral IFG, and more extensive RH activation, in the processing of novel metaphors (Desai, Conant, Binder, Park, & Seidenberg, 2013; Forgács, Lukács, & Pléh, 2014; Mashal, Vishne, & Laor, 2014; Subramaniam, Beeman, Faust, & Mashal, 2013; Subramaniam, Faust, Beeman, & Mashal, 2012). Meta-analyses have confirmed that RH involvement is more likely when metaphoric expressions are novel (Rapp et al., 2012), but two major uncertainties remain: What are the functional roles of these RH areas, and are they *necessary* for comprehension, or do they merely play a facilitatory role when LH processing demands are high? Studies of novel metaphor comprehension in patients with brain injury affecting novelty-sensitive areas have yet to be done, but this line of research promises insight into both issues.

The answers to the other two conceptual questions—the importance of metaphor type and sensorimotor grounding for metaphor comprehension—remain unclear. On the one hand, a number of studies now support the hypothesis that words, even when used metaphorically, are processed in part by the same sensorimotor regions relevant for perceiving or executing the sensory and motor features relevant to their literal senses. Texture metaphors (*She had a rough day*) engage somatosensory cortex (Lacey, Stilla, & Sathian, 2012), taste metaphors (*She received a sweet compliment*) engage primary and secondary gustatory cortex (Citron & Goldberg, 2014), arm action metaphors (*The Congress is grasping the new state of affairs*) engage secondary motor areas for action planning and coordination (Desai et al., 2013), and motion verb metaphors (*The man fell under the spell*; Chen et al., 2008) and fictive motion sentences (*The pipe goes into the house*; Saygin, McCullough, Alac, & Emmorey, 2010; Wallentin, Lund, Ostergaard, Ostergaard, & Roepstorff, 2005; Wallentin, Østergaard, Lund, Østergaard, & Roepstorff, 2005) engage primary motion perception area MT+/V5 or the secondary motion association cortex, pMTG. On the other hand, embodied cognition accounts (e.g., Gallese & Lakoff, 2005; Gibbs, 2006) posit that figurative extensions of action verbs should also engage primary and supplementary motor cortex, a prediction that has not generally been supported (Chen et al., 2008; Wallentin, Lund, et al., 2005; Wallentin, Østergaard, et al., 2005). Rather, motor cortex engagement appears to be modulated by familiarity: Highly conventional figurative uses of verbs, as are found in idioms, do not engage M1/SMA (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Raposo, Moss, Stamatakis, & Tyler, 2009), but less familiar metaphoric senses may (Desai et al., 2011; Obert et al., 2014). Similarly, Cardillo et al. (2012) observed that pMTG activation was graded by familiarity with action verb metaphors. These studies provide preliminary support for previous hypotheses that the degree of sensorimotor engagement is determined by metaphor novelty, such that sensorimotor areas are more strongly recruited for understanding novel figurative senses of words than for conventional ones (Aziz-Zadeh & Damasio, 2008; Cardillo et al., 2010). The initial abstraction of the metaphoric sense of a base term may require activating its literal sense, via sensorimotor simulation, before selecting only those more abstract, conceptual features relevant to the figurative extension. With repeated exposure to the metaphoric sense, a simulation becomes less necessary, and the newly learned metaphorical sense can be accessed directly without reliance on sensorimotor grounding. Observations of similar modality-specific activation graded by metaphor familiarity in domains other than action/motion are needed to establish confidence in this conclusion.

Clarifying the current neuroimaging literature also requires model building and testing that incorporate both noun- and verb-based metaphors orthogonal to object- and action-based semantics. Currently, action semantics and verb figurativeness are generally conflated, as are object semantics and noun figurativeness. For instance, Chen et al. (2008) concluded that the greater activity in pMTG for metaphorical and literal motion verb sentences (*The man fell under her spell*; *The child fell under the slide*) than for abstract sentences (*The merchant was greedy and gluttonous*) reflected the motion features of the action verbs—but the conflation of base term grammatical class and action semantics leaves open the possibility that pMTG is recruited for verb or event processing more generally, rather than by motion features per se (Bedny, Caramazza, Grossman, Pascual-Leone, & Saxe, 2008). To test this possibility, three kinds of contrasts are needed: (1) a comparison of noun-based metaphors and verb-based metaphors, both generated from base terms with salient motion; (2) a comparison of verb-based and/or noun-based metaphors with and without motion qualities (i.e., coming from different sensory modalities); and (3) a comparison of noun-based metaphors with motion qualities that do and do not refer to events. The latter two contrasts have not, to our knowledge, been tested. A direct comparison of predicate and nominal metaphors using nominalized verbs as the base terms in the nominal condition (Cardillo et al., 2012) is the closest approximation to the first proposed contrast. These results indicated no difference in how strongly the two metaphor types recruited pMTG, suggesting that base term semantics are a more important determinant of neural processing than syntactic structure or base term grammatical class. However, the base terms in this study were primarily, but not exclusively, motion-related, again leaving open the possibility that pMTG is responsive to event semantics more generally rather than to motion features in particular. To resolve the current ambiguity regarding the relative importances of syntax, grammatical class, and base term semantics in determining

the neural basis of metaphor, existing stimuli can be used to test the first two contrasts, but new items are required in order to test the third critical contrast.

The purpose of this study was to extend existing metaphor stimuli to meet this need. We generated and normed nominal metaphors with entity semantics so that, in conjunction with our previously published stimulus set (Cardillo et al., 2010), we now provide a superset of nominal and predicate metaphors that will allow us to fully teasing apart the influences of semantics, grammatical class, and syntactic structure. The stimuli of Cardillo et al. (2010) consist of 240 nominal literal–metaphor sentence pairs and 240 predicate literal–metaphor sentence pairs, using base terms with either salient auditory or motion features (for examples, see Table 1). Critically, the base term in these nominal sentences is always a nominalized verb, to maximize the semantic similarity of the base terms in the two metaphor types. By closely matching the sensorimotor features of nominal and predicate base terms, these items are optimized for detecting processing differences related to grammatical class (noun vs. verb) and syntactic construction (nominal vs. predicate), but not event semantics. In the present study, we generated nominal literal–metaphor pairs using entity nouns with salient auditory or motion features as the base terms (for examples, see Table 1). *Entity* nouns refer to static and concrete persons, places, or things. Thus, a comparison of nominal-event and nominal-entity items is optimized to detect processing differences related to sensorimotor features (dynamic, action events vs. static, concrete entities), while holding grammatical class and syntactic construction constant. The stimuli were generated using an extensive norming procedure identical to the one established in Cardillo et al. (2010), in order to allow easy mixing of items from the two sets, if desired. In Norming Study 1, we established values for nine sentence characteristics that impact comprehension on the basis of the individual words within each sentence. In Norming Study 2, we acquired six offline ratings related to the overall meaning of each sentence. In Norming Study 3, we acquired two online measures related to the overall meaning of each sentence.

Method

Construction of stimuli

An initial pool of 312 sentences of nominal syntactic form was generated). Each sentence consisted of two noun phrases linked by a copula (i.e., "An X is a Y"), with zero, one, or two modifying adjectives. The adjectives were not designed as part of the experimental manipulation but, rather, to make the items comparable in terms of overall length, frequency, or concreteness, and/or to clarify the metaphorical extension of the base term, since most of the metaphors were unfamiliar. For all four conditions, most of the items had one adjective ($n = 47–51$). Half of the sentences expressed a literal meaning, and half expressed a metaphorical meaning.

To generate these items, 78 concrete nouns with salient auditory properties and 78 concrete nouns with salient motion properties were first selected as base terms (i.e., the word to be extended metaphorically). Next, for each noun, both a literal sentence and a metaphorical sentence were created, resulting in 156 literal–metaphor sentence pairs. In this way, in each pair an identical noun implied a literal or a figurative interpretation, depending on its context. Critically, all of the metaphors involving auditory base terms were designed such that no sound was implied by the figurative interpretation of the sentence. Likewise, all metaphors involving motion base

Table 1 Examples of each sentence type

Modality	Sentence Type	Literal	Metaphorical
Auditory	Predicate*	The lecturer droned for many hours.	The contract droned for many pages.
		The anxious author shrieked at the mouse.	Her pale skin shrieked in the sun.
	Nominal-Event*	Her immediate remark was a snigger.	The book was a sexist snigger.
		The conversation was a hushed whisper.	His glance was a furtive whisper.
	Nominal-Entity	The prize was an upright hoover.	His mind was a hungry hoover.
		The solemn song was an anthem.	The sitcom was a national anthem.
Motion	Predicate*	The girl tangoed with the instructor.	The garlic tangoed with the ginger.
		The model tottered in high heels.	The cake shop tottered on bankruptcy.
	Nominal-Event*	The snake's move was a slither.	The deal was a greedy slither.
		His gait was a confident swagger.	His yacht was a rich swagger.
	Nominal-Entity	The police evidence was a bullet.	The coffee was a caffeine bullet.
		The tourist attraction was a geyser.	His temper was a faithful geyser.

* Example stimuli from Cardillo et al. (2010), included for comparison

terms were designed such that no physical or fictive motion was implied by the figurative interpretation of the sentence.

Overview of the norming studies

For all norming tasks, we replicated exactly the procedures (tasks, sample sizes, selection criteria, and analyses) of Cardillo et al. (2010), to facilitate combining the items into a single superset of stimuli to be sampled, if desired. The initial pool of sentences was normed both offline and online, at both the word and sentence levels, in order to characterize its psycholinguistic and semantic properties and to highlight any problematic items. Before norming, three measures of length (number of characters, number of words, and number of content words) were calculated for each sentence, as well as average frequency and concreteness scores based on the values for the content words of the sentence (i.e., nouns, verbs, and adjectives). Frequency values were calculated using the popular measure established by Kučera and Francis (1967), as well as using values from the more recent and larger corpus, SUBTLEXus (Brysbaert & New, 2009). Concreteness values were taken from the MRC Psycholinguistic Database (Coltheart, 1981) and the University of South Florida Norms (Nelson, McEvoy, & Schreiber 1998). For those words for which concreteness ratings were not found in either of these databases, we collected our own (Norming Study 1).[1] These participants also judged the strength of the auditory and visual imagery associated with all of the base terms, to ensure a valid manipulation of sensory modality. A different set of participants rated the base terms from this candidate stimulus set and from Cardillo et al. (2010) for strength of motion imagery. A further set of individuals normed the stimuli at the sentence level, interpreting them as well as rating them in terms of familiarity,[2] naturalness, imageability, figurativeness, and ease of interpretation (Norming Study 2). Additionally, a valence judgment task was administered to a final group of individuals to generate an online measure of comprehension difficulty for each item (Norming Study 3). Given the sensitivity of fMRI to reaction time (RT) differences between conditions, coupled with the fact that valence judgment is currently the most frequently used task in fMRI studies of metaphor (see Table 1 in Bohrn, Altmann, & Jacobs 2012), valence RT provides an important dimension for neuroimagers. Note that we added ease of interpretation to this norming study to provide researchers a subjective measure of interpretation difficulty to complement our objective measure (interpretability score) and our online measure (valence RT).[3] We anticipate that which measure(s) will be most important to control will vary with study design and question.

[1] We collected concreteness values on the assumption that concreteness judgments for common words would not differ much, despite a span of several decades between MRC, South Florida, and our norming study. We did not, however, include redundant items to confirm this consistency across time, so we cannot be certain of their stability.

[2] We chose to norm items in terms of familiarity rather than salience because (1) familiarity is a strong determinant of salience, making it a reasonable proxy; (2) familiarity is a simpler construct than salience, making its effects easier to interpret; and (3) familiarity is frequently investigated in individual cognitive and neural studies, making it more useful for comparisons across studies. For individuals specifically interested in salience, we direct them to Roncero and de Almeida (2014).

Norming study 1: words

Participants Sixty native English speakers were recruited from the University of Pennsylvania community in compliance with the procedures established by the university's Institutional Review Board and were compensated $15 or given course credit for their participation. Forty of the participants (mean age = 19.0 years, *SD* = 1.7; 30 females, ten males; mean education = 13.0 years, *SD* = 1.6) made concreteness ratings and judged the base terms for auditory and visual imagery (20 participants rated one half of the items, the other 20 participants rated the other half), and 20 participants (mean age = 20.9 years, *SD* = 2.3; 16 females, four males; mean education = 14.7 years, *SD* = 1.4) rated the base words for motion imagery.

Stimuli The initial pool of sentences contained 953 content words, 204 of which lacked published concreteness values and thus required norming (Appendix A). All base terms were rated for their associated auditory, visual, and motion imagery (*n* = 156; Appendix B). For participants who received the motion imagery list, all base terms from Cardillo et al. (2010) were also included (*n* = 226; Appendix C), since these had not been collected previously and doing so allows for future studies that combine sentences from the norming studies.

Task An Excel workbook was generated with separate worksheets corresponding to the four rating tasks (concreteness, auditory imagery, visual imagery, and motion imagery) and one line per worksheet corresponding to each item. For concreteness, participants were instructed to rate the words in terms of their accessibility to one or more of the senses, using a scale from 1 (*very abstract*) to 7 (*very concrete*). For auditory imagery, participants were instructed to rate the words in terms of the speed and "ease or difficulty with which they arouse a particular sound," using a scale from 1 (*no sound*) to 5 (*clear sound*). For visual imagery, participants were instructed to rate the words in terms of the speed and "ease or difficulty with which they arouse a mental picture or visual image," using a scale from 1 (*no image*) to 5 (*clear image*). For motion imagery, participants were instructed to rate the

[3] We thank an anonymous reviewer for this suggestion.

words in terms of the speed and "ease or difficulty with which they arouse visual motion," using a scale from 1 (*no motion*) to 5 (*clear motion*). In all cases, the instructions were coupled with several examples and explanations (see the supplemental materials for more detail).[4]

Data analysis For all words, ratings were averaged over the 20 participants for each of the judgments. The 204 new concreteness values supplemented the previously published values for the other 749 content words in the stimuli set. These individual concreteness ratings were then used to determine an average concreteness value for each of the 312 candidate sentences (i.e., the sum of the concreteness values associated with each content word in any particular sentence, divided by the number of content words in that sentence). The imagery ratings of the base terms indicated three problematic base terms: One base term used in the auditory conditions (*cab*) elicited stronger motion than auditory imagery, and two used in the motion conditions (*sprinkler* and *zipper*) elicited stronger auditory than motion imagery. Overall, visual imagery was consistently strong in the motion base terms. By contrast, its strength varied widely in the auditory base terms, since some referred to palpable objects (e.g., instrument names) whereas others referred to intangible musical concepts (e.g., *rhythm*, *beat*, *melody*).

Norming study 2a: sentences

Participants Forty participants were recruited from the University of Pennsylvania community in compliance with the procedures established by the university's Institutional Review Board, and were compensated $25 or given course credit for their participation. All participants were native English speakers, and none had participated in Norming Task 1. Because of the large number of items to be evaluated and concerns about fatigue, 20 participants made judgments on half of the sentences (mean age = 20.5 years, *SD* = 2.9; 11 females, nine males; mean education = 15.0 years, *SD* = 1.8) and 20 participants made judgments on the other half of the sentences (mean age = 22.1 years, *SD* = 3.2; ten females, ten males; mean education = 15.5 years, *SD* = 2.0).

Stimuli All 312 candidate sentences were assessed.

Task The items were randomly divided in half, and for each of these subsets an Excel workbook was generated with separate worksheets corresponding to the five norming tasks (familiarity, naturalness, imageability, figurativeness, and interpretation), with one line per worksheet corresponding to each item. In this way, participants saw both literal and metaphorical items, with 20 responses collected for each item.

For the familiarity task, participants were instructed to rate their frequency of experience with the sentence and its meaning, using a scale from 1 (*very unfamiliar*) to 7 (*very familiar*). For the naturalness task, participants were instructed to rate each sentence for how "natural and normal" it seemed, using a scale from 1 (*very unnatural*) to 7 (*very natural*). For the imageability task, participants were instructed to rate "how quickly and easily each sentence brings a visual image to mind," using a scale from 1 (*no image*) to 7 (*clear*, *immediate image*). For the figurativeness task, participants were instructed to rate how literal an interpretation each sentence suggested, using a scale from 1 (*very literal*) to 7 (*very figurative*). For the interpretation task, participants were instructed to write the meaning of each sentence using their own words (full instructions can be found in the supplemental materials). The familiarity, naturalness, imageability, and figurativeness ratings were collected for both literal and metaphorical sentences. Given the difficulty of restating a concrete, literal sentence in novel words and the absence of any theoretical relevance for such descriptions, interpretations were only collected for the metaphors.

Data analysis To generate familiarity, naturalness, imageability, and figurativeness ratings for each item, averages were calculated. Several steps were necessary to determine the interpretability of each item. First, for each metaphor, two of the authors (E.C., C.W.) and a third researcher independently judged the number of interpretations that reflected a plausible figurative construal of the sentence.[5] In contrast, blank, nonsensical, literal, or uninformative (e.g., "Just what it says") interpretations were not taken to indicate metaphoric comprehension. To encourage consistent evaluations, the judges were first trained on 100 interpretations from Cardillo et al. (2010).

The three judges evaluated 3,120 interpretations, resulting in 9,360 plausibility judgments. The average pairwise percent agreement between judges was 83.9 %, with the two judges being in greater agreement with each other (88.2 %) than with the third researcher (82.2 % and 81.1 %, respectively). An interrater reliability analysis using the Kappa statistic was used to further determine consistency among the judges. A Fleiss Kappa for multiple raters of .53 (*SE* = .01, 95 % CI = .51–.55, $p < .00001$) indicated moderate agreement between

[4] The instructions were slightly modified directions from Paivio and colleagues (1968), and also were very similar to those used in the two other major sources of concreteness and imageability norms in the MRC Psycholinguistic Database (i.e., Gilhooly & Logie, 1980; Toglia & Battig, 1978). The exact instructions for all tasks can be found in the supplemental materials.

[5] For some sentences, all interpretations reflected a single meaning; for many others, the responses indicated multiple or overlapping meanings. Given the plausibility of more than one interpretation in the absence of context and the subjectivity of determining where one meaning ends and another begins, rather than tally the incidences of the most common interpretation, *any* plausible figurative interpretation was taken to indicate that the metaphor had been understood.

the judges. The strong level of agreement between two of the three judges (Cohen's Kappa = .62) reveals the sensitivity of the plausibility judgment to characteristics of the judges (e.g., differences in research backgrounds). The stronger three-way agreement in our previous study (Cardillo et al., 2010), with no difference in task training but more similar research experience, suggests that the richness of information provided by the interpretation task may be best leveraged when the judges are given additional training.

As in our previous study, an interpretability score for each participant was calculated by dividing the number of their interpretations that were deemed plausible by at least two of the judges by the total number of items assessed by that participant (# plausible interpretations/all possible interpretations). This assessment revealed poor overall comprehension by four participants in one list and by two participants in the other list (>30 % of their interpretations were not considered plausible and/or were left blank), so their data were excluded from subsequent analyses. To generate an interpretability score for each item, the number of interpretations deemed plausible by at least two of the judges was divided by the total number of interpretations for that item (# plausible interpretations/all possible interpretations). These results indicated that 28 metaphors failed to reach the minimum desired comprehensibility criteria (70 % plausible interpretations) established in our previous study (Cardillo et al., 2010).

Norming study 2b: sentences

Participants Twenty participants (mean age = 22.5 years, *SD* = 3.7; 16 females, four males; mean education = 16.0 years, *SD* = 2.5) were recruited from the University of Pennsylvania community in compliance with the procedures established by the university's Institutional Review Board and were compensated $10 for their participation. All participants were native English speakers, and none had participated in the other norming tasks.

Stimuli All 312 candidate sentences were assessed.

Task Items were presented in an Excel workbook, with one line corresponding to each item. Participants were instructed to rate each item for its ease of interpretation, using a scale from 1 (*easy to interpret*) to 7 (*difficult to interpret*). The full instructions are reported in the supplemental materials.

Data analysis To generate an ease-of-interpretation score for each item, averages were calculated.

Norming study 3: online comprehension

Participants Twenty participants (ages 18–22 years, 13 females, seven males) were recruited from the University of Pennsylvania undergraduate community in compliance with the procedures established by the university's Institutional Review Board and were compensated $15 or given course credit for their participation. All of the participants were native English speakers, and none had participated in Norming Task 1 or 2.

Stimuli All 312 candidate sentences were assessed.

Task Sentences were presented centrally in black 18-point font on a white background, using E-Prime 1.1 software on a Dell Inspiron laptop. Sentences were displayed for 3,000 ms and separated by a 1,000-ms intertrial interval. Participants were instructed to read each sentence and then to judge its emotional valence, using the 'f' key to indicate a positive valence and the 'j' key to indicate a valence that was *not* positive, defined as either neutral or negative. They were informed that there was no right or wrong answer and were encouraged to respond as quickly as possible. Twelve practice trials preceded four blocks of experimental trials. Each participant received a different random order of items and saw each sentence only once.

Data analysis For every sentence, RTs were averaged across participants and the proportion of positive valence judgments was calculated.

Results

To determine the reliability of the norms, we calculated intraclass correlation coefficients for all dimensions requiring a subjective rating. All measures indicated high agreement across raters (Table 2), whether all raters rated all items (i.e., two-way random average measures for ease of interpretation, concreteness, and base term imagery) or different raters rated different items (i.e., one-way random average measures for familiarity, naturalness, imageability, and figurativeness).

Table 2 Intraclass correlation coefficients for the rating tasks

Dimension	Intraclass Correlation Coefficient
Base term auditory imagery	.969
Base term visual imagery	.975
Base term motion imagery	.962
Concreteness	.969
Familiarity	.857
Naturalness	.886
Imageability	.872
Figurativeness	.975
Ease of interpretation	.957

The results of the norming studies were used to eliminate problematic stimuli from the initial pool. The imagery ratings from Norming Study 1 and the interpretability scores from Norming Study 2 indicated that 31 metaphor–literal sentence pairs to be discarded, resulting in 60 nominal-auditory and 66 nominal-motion items remaining possible sentence pairs. To generate a final stimulus set with equal numbers of items in each condition, the six items in the nominal-motion condition with the lowest interpretability values were also discarded. The lexical and sentential characteristics of the final set of 240 sentences are summarized in Table 3 (see the Electronic supplementary material for the full set of items and their norming values).

We did not calculate statistical differences between the sentence types, because our aim is for the stimulus set to be sampled in ways that control for condition differences or for the norming data to be used to covary out the influences of nuisance variables, rather than for the set to be used in its entirety. Nonetheless, the overall means confirmed desired differences between the conditions, as well as indicating some areas in which control is likely to be necessary. As intended, the base terms consisting of nouns with motion properties were rated as having more salient motion imagery than auditory imagery, and the base terms consisting of nouns with auditory properties were rated as having more salient auditory imagery than motion imagery. Unsurprisingly, given the visual nature of motion perception, the visual imagery provoked by the motion base terms was also greater than the auditory imagery provoked by those terms, and was weaker than the auditory imagery provoked by the auditory base terms.

At the sentence level, the literal and metaphorical sentences were similar in length and frequency. However, the literal sentences of both modalities were judged not only to be less figurative (as they should be, by definition), but also more concrete, familiar, natural, easy to interpret, and imageable than their matched metaphorical sentences. This same pattern (with the exception of the concreteness difference) was observed in our previously normed stimuli (Cardillo et al., 2010). The greater familiarity and naturalness of the literals highlights the difficulty of constructing novel literal sentences without inadvertently evoking metaphoric interpretations. For example, consider the novel literal sentence "He's sitting deep in the bubbles" in Van Lancker and Kempler's (1987) proverb study; it is not obvious that a participant wouldn't interpret this sentence figuratively if it was presented in the context of many other metaphors. Rather than generate novel literal sentences that might unintentionally be construed metaphorically, we erred on the side of creating more familiar literal items, knowing that their familiarity could be covaried out if necessary. In doing so, however, our literal items turned out to be easier to understand. We posit that this is the lesser of two

Table 3 Summary of final stimulus characteristics by sentence types

	Literal			Motion			Metaphorical			Motion		
	Auditory						Auditory					
	M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
Base auditory imagery	4.1	0.6	2.2–5.0	2.0	0.6	1.1–3.4	4.1	0.6	2.2–5.0	2.0	0.6	1.1–3.4
Base visual imagery	3.3	1.4	1.3–5.0	4.3	0.7	1.9–5.0	3.3	1.4	1.3–5.0	4.3	0.7	1.9–5.0
Base motion imagery	2.2	0.6	1.3–3.9	3.3	1.0	1.4–4.7	2.2	0.6	1.3–3.9	3.3	1.0	1.4–4.7
Concreteness	465	66	310–596	492	78	201–609	444	89	45–588	461	59	317–583
Frequency1	90	120	4–589	73	101	1–577	66	80	2–438	86	118	0–549
Frequency2	85	121	1–653	54	70	2–334	67	106	0–484	67	92	2–405
# Characters	31.9	4.5	22–41	32.6	4.5	22–42	33.4	5.8	21–42	33.4	4.7	25–43
# Words	5.9	0.5	4–7	5.9	0.5	4–7	4.8	0.6	4–7	6.0	0.5	4–7
# Content words	3.0	0.4	2–4	3.0	0.5	2–4	3.0	0.5	2–4	3.1	0.5	2–4
Interpretability	–	–	–	–	–	–	0.91	0.07	.72–1.0	0.93	0.1	.75–1.0
Ease of interpretation	1.3	0.5	1.0–7.0	1.2	0.3	1.0–7.0	3.5	0.8	1.0–7.0	3.3	0.8	1.0–7.0
Familiarity	5.6	0.6	3.6–6.7	5.6	0.6	4.2–6.8	4.1	1.0	2.4–6.4	4.3	0.9	2.8–6.3
Naturalness	5.8	0.7	3.6–7.0	5.9	0.5	4.8–6.9	4.2	0.9	2.4–5.9	4.4	0.9	2.7–6.4
Imageability	5.2	1.1	2.7–6.9	6.1	0.5	4.7–6.9	4.0	0.8	2.4–6.5	4.0	0.7	2.9–5.7
Figurativeness	1.9	0.7	1.0–3.9	1.8	0.4	1.0–3.0	6.0	0.5	4.9–6.7	6.1	0.5	4.3–6.7
Valence RT (ms)	1,178	162	918–1,801	1,167	160	835–1,634	1,268	172	975–1,763	1,236	167	842–1,680
Valence positive ratio	0.25	0.3	0.0–1.0	0.17	0.2	0.0–.70	0.35	0.3	0.0–.95	.26	0.3	0.0–90

Frequency1 = values from Kučera and Francis (1967); Frequency2 = $SUBTL_{WF}$ values from Brysbaert and New (2009)

evils, when faced with the possibility of literals being unintentionally interpreted metaphorically. We do not believe this to be an inherent difference between metaphoric and literal expressions, so much as a confound that will be important to include as a covariate in any study with unmatched items.

By contrast, the consistent observation of reduced imageability for metaphors relative to literals, across stimulus sets and metaphor types, suggests that this pattern may be inherent to metaphorical language rather than a confound. Abstract concepts, by their nature, are not imageable (e.g., *justice*, *peace*), and it has been argued that we rely on figurative expressions in order to talk about them (Lakoff & Johnson, 1980). However, not all low-imageability concepts are abstract (e.g., auditory concepts like *symphony*, *melody*, *song*, etc.). For these items, a metaphorical sense may not necessarily be any less imageable than its literal sense. Alternatively, this pattern may be an accidental artifact of the particular words chosen or of our creating auditory and motion metaphors that did not imply any literal sound or motion, respectively—a methodological necessity for testing sensorimotor hypotheses that may have inadvertently biased us to generate metaphors encoding more abstract, low-imageability meanings. Until future research clarifies the relationship of imagery to figurativeness, our data suggest that matching metaphors and literals on imageability may require an especially large set of items from which to sample.

As in our previous study, when considering the metaphors separately by modality, interpretability was very high for both auditory and motion metaphors. The only difference between modalities regarded imageability in the literal condition: Auditory literal items were rated as being less imageable than motion literal items, an unsurprising finding considering that the base terms of auditory items were rated as having weaker visual and motion associations.

To further explore the relationships between the sentence-level factors of theoretical interest, the seven values collected for each sentence (familiarity, naturalness, imageability, figurativeness, interpretability, ease of interpretation, and valence RT) were correlated with each other, both collapsed across modalities (Table 4) and separately (Table 5). The results indicated several expected relationships based on our prior study. As we previously observed, familiarity and naturalness were highly correlated, again indicating that these constructs are either conceptually indistinguishable or at least so difficult to disentangle that we suggest researchers not concern themselves with naturalness, in favor of the more theoretically relevant construct of familiarity. Sentences rated higher in familiarity and naturalness also tended to evoke greater visual imagery, were perceived as less figurative, and were more easily understood. Yet these patterns are not sufficiently strong that they could not be orthogonalized with careful item selection (see the Discussion section). Critically, as with our previous stimulus set, it is possible to disentangle comprehension difficulty level from figurativeness. Despite the relative novelty of the metaphors in this stimulus set, the correlations between interpretability and figurativeness, and between ease of interpretation and figurativeness, were both not significant. Also notable, valence RT did not significantly correlate with any of the sentence ratings or the interpretability scores. Balancing conditions in terms of time-on-task is critical in fMRI studies; this pattern suggests that selecting stimuli differing in characteristics of interest but not in processing time should not be a challenge.'

By and large, the auditory and motion metaphor sets showed very similar relationships between the sentence-level factors when correlations were calculated separately for each (Table 5).

Discussion

The purpose of this study was to augment existing published metaphor stimuli with items that can address additional hypotheses or boost sample size and statistical power when combined with previously published items. To this end, we offer matched literal and metaphoric sentences characterized at both the word and sentence levels on 14 variables of methodological and theoretical relevance. Whether they are used in combination with other stimuli or on their own, our hope is that these items will facilitate the methodological and conceptual precision necessary to address existing ambiguities and emerging questions in cognitive and neural studies of metaphor.

At 240 items, the stimulus set joins three others (Cardillo et al., 2010 [$n = 560$]; Katz, Paivio, Marschark, & Clark, 1988 [$n = 464$]; Roncero & de Almeda, 2014 [$n = 168$]) as one of the largest pools of extensively normed stimuli for studies of metaphor. The earliest stimulus set, by Katz et al. (1988), provides 204 literary and 260 nonliterary nominal metaphors, each rated on ten dimensions (comprehensibility, ease of interpretation, metaphoricity, metaphor goodness, metaphor imagery, subject imagery, predicate imagery, familiarity, semantic relatedness, and number of alternative interpretations). More recently, Roncero and de Almeida (2015) provided 84 nominal metaphors and matched similes, normed in terms of property associations, aptness, familiarity, conventionality, saliency, and connotativeness. Informative in their own right, the Katz stimuli have also been a valuable resource in metaphor research since their publication, as well as the only normed set of literary metaphors we are aware of. However, these items have not been normed on some of the characteristics germane to current debates concerning metaphor. By contrast, Roncero and de Almeida's (2015) carefully crafted metaphor–simile pairs are optimized for addressing current, competing cognitive models of metaphor

Table 4 Correlation coefficients between sentence scales, collapsed across modalities

	FAM	NAT	IMG	FIG	EASE	INT	RT
Familiarity (FAM)		.90**	.54**	–.19*	.79**	.42**	–.15
Naturalness (NAT)			.57**	–.11	.83**	.51**	–.13
Imageability (IMG)				–.06	.43**	.30**	–.12
Figurativeness (FIG)					.17	.13	–.07
Ease of interpretation (EASE)						.37**	.06
Interpretability (INT)							–.12
Valence RT (RT)							

** $p < .01$, * $p < .05$'

comprehension (e.g., comparison vs. categorization mechanisms), and for testing the predictions of the GSH. However, without manipulations of the semantics, grammatical category, or sensory associations of base terms, neither stimulus set can address the specific neural hypotheses outlined above regarding sensorimotor grounding or metaphor type. Nor do these sets include matched literal items, a critical comparison condition for both neuroimaging and patient studies. Our aim is to complement the strengths of these other stimulus sets by filling the extant gap in resources for testing neural hypotheses.

The large number of items and normed properties in the current set maximizes its flexibility. When combined with our previous set, we offer an unprecedented level of standardization—800 items that have been normed in identical fashion on all the same parameters. We suggest that careful selection of items should enable avoiding condition differences on parameters of noninterest and maximizing differences of theoretical relevance. We take the success of researchers selecting items from the similarly normed Cardillo et al. (2010) stimuli as a demonstration of the utility of our approach for a variety of study designs (e.g., behavioral: Bowes & Katz, 2015; Jalal & Ramachandran, 2014; fMRI: Cardillo et al., 2012; eyetracking: Columbus et al., 2015; event-related potentials [ERP]: Schmidt-Snoek, Drew, Barlie, & Agauas, 2015). We suggest that computational approaches to stimulus selection, such as SOS ("Stimulus Optimization Software"; Armstrong, Watson, & Plaut, 2012), be used to maximize the efficiency of this selection process for complex designs. For instance, we used SOS to select three sets each of nominal and predicate metaphors from Cardillo et al. (2010) that did not differ in terms of such critical variables as familiarity, naturalness, imageability, figurativeness, and interpretability (Cardillo et al., 2012), a balancing act that would be difficult or impossible with fewer items and trial-and-error selection.

The present items are also modifiable to suit a variety of tasks and populations: Adding context, comprehension questions, primes, semantically related probes, multiple-choice answers, collecting additional ratings, and so forth, are all possible elaborations. How researchers sample and/or augment the stimuli will depend on their questions of interest, subject populations, and methodology. For example, we recently used SOS to select from both the present items and our previously published set to generate matched sets of nominal-event, nominal-entity, and predicate literal–metaphor pairs, combining them with newly crafted multiple-choice questions to generate

Table 5 Correlation coefficients between sentence scales, separated by modality

	FAM	NAT	IMG	FIG	EASE	INT	RT
	Auditory Metaphors						
Familiarity (FAM)		.90**	.54**	–.19*	–.78**	.31*	–.19
Naturalness (NAT)			.64**	–.15	–.83**	.37**	–.13
Imageability (IMG)				–.13	–.44**	.29*	–.03
Figurativeness (FIG)					.17	.08	.09
Ease of interpretation (EASE)						–.28*	.03
Interpretability (INT)							–.20~
Valence RT (RT)							
	Motion Metaphors						
Familiarity (FAM)		.89**	.46**	–.17	–.81**	.54**	–.07
Naturalness (NAT)			.49**	–.09	–.84**	.64**	–.10
Imageability (IMG)				.02	–.43**	.31*	–.25
Figurativeness (FIG)					.21	.15	–.21
Ease of interpretation (EASE)						–.46**	.09
Interpretability (INT)							–.01
Valence RT (RT)							

** $p < .01$, * $p < .05$, ~$p < .10$

a novel assessment of metaphor comprehension in brain-injured patients (Ianni, Cardillo, McQuire, & Chatterjee, 2014). The final items were selected such that the metaphor conditions were matched for interpretability, figurativeness, familiarity, naturalness, imageability, length, frequency, and concreteness, and the metaphors and literals were matched for familiarity, length, frequency, concreteness, and valence. Schmidt-Snoek and colleagues (2015) took a different approach: collecting cloze probability norms and supplementing items with anomalous sentences generated from the same base terms to optimize their utility for an ERP study of modality effects. Critically, Schmidt-Snoek et al. were able to select a final set of 35 auditory and 35 motion metaphors that did not differ significantly on any of the dimensions normed in Cardillo et al. (2010). Columbus and colleagues (2015) modified a subset of our previously published stimuli in other ways: adding context to investigate its impact on comprehension, adding a neutral continuation to optimize the items for eyetracking measures, and re-collecting familiarity ratings to account for these modifications.

The nominal-entity items presented here are uniquely and optimally designed to explore the role of sensorimotor grounding in metaphor comprehension when they are used in combination with the nominal-event items of Cardillo et al. (2010). However, the potential application of the items is not limited to this topic. We have outlined here and previously (Cardillo et al., 2010) a number of other theoretical issues regarding metaphor novelty, semantics, and type that these items are also well-suited to address. As our understanding of metaphor processing has evolved, so too have our questions of interest. Collecting additional dimensions on these items can enable them to evolve along with our questions and maintain their utility. For instance, the relationship between how apt we find a metaphor, our familiarity with it, and the ease and means with which we understand it remains unsettled—is the career of metaphor model sufficient to account for aptness effects (Bowdle & Gentner, 2005; Jones & Estes, 2006)? How can existing models of metaphor comprehension, based on studies of nominal metaphors, be tested or extended to account for predicate metaphors (Cardillo et al., 2010; Schmidt et al., 2010)? How does linguistic context modulate the role of such factors as familiarity and figurativeness (Giora, 1997)? What makes a metaphor aesthetically beautiful, not merely apt or interpretable (Bohrn, Altmann, Lubrich, Menninghaus, & Jacobs, 2013)? Are aesthetic or cognitive dimensions more predictive of conventionalization and permutation? Does reading creative language like novel metaphors enhance reasoning or other forms of creativity (Bowes & Katz, 2015)? Are individuals with reduced cognitive flexibility (e.g., children, the elderly, or patient populations) less able to appreciate metaphors? What might their difficulties reveal about typical comprehension? It is our hope and intention that these stimuli may be extended by researchers to any of these or other emerging questions, as well.

Author note This research was supported by a National Institutes of Health grant (No. R01-DC-012511) and by a National Science Foundation subcontract (No. 330161-18110-7341) awarded to A.C. We thank Casey Gorman and Sam Cason for their help collecting and organizing the data; Geena Ianni for assistance determining the plausibility of metaphor interpretations; and Alex Kranjec, Gwenda Schmidt, and Marguerite McQuire for helpful discussions about the topic.

Appendix A: words normed for concreteness

accessory, accompaniment, actress, ad, added, aged, aggressive, alcoholic, anniversary, another, anthem, appetite, approaching, archeological, aria, artifact, artifact, astronomer, attraction, backyard, blizzard, boiling, boomerang, breakthrough, Broadway, bugle, busy, cabbie, cacophony, caffeine, call, campaign, careful, catapult, chainsaw, chariot, cheetah, childish, chord, clapping, classical, collective, comet, coordinator, corruption, counseling, coworker, crate, creative, critique, daughter, defense, digestive, drawn, drinking, dripping, drumbeat, dutiful, editing, editor, electric, emotional, endless, erosion, escalator, explosion, fading, faithful, familiar, familiar, fateful, ferris-wheel, fickle, financial, footwork, footwork, foreboding, forward, foster, friendship, gadget, gazelle, glamour, gong, gossipy, grand, gunfire, gyroscope, hanging, hoover, horn, horse-drawn, inchworm, incriminating, ingredient, inquiry, internship, iPod, jazz, jackhammer, kettle, lance, landslide, laptop, layoffs, leaky, letter, liberal, local, logistical, loudspeaker, lullaby, marathon, medieval, meds, megaphone, memo, mental, merry-go-round, meteor, metronome, musical, national, nearby, nickname, nuisance, pager, parachute, Pavlovian, paycheck, piccolo, planner, plumber, plunger, poem, position, practice, predator, pregnant, pricey, program, pulley, racehorse, rainstorm, recommendation, refund, reliable, resounding, reunion, rhetorical, ringing, rockstar, rollercoaster, sale, satyr, schizophrenic, screenplay, secret, seesaw, self-pitying, signpost, sitcom, skewer, sled, slingshot, slug, snake-charmer, soothing, specialty, spitball, sprinkler, spritely, stallion, startling, steering, stepmother, stolen, stopwatch, stunning, surprise, sweethearts, targeted, technology, teeter-totter, testosterone, thunderstorm, ticking, tidal, tone, torpedo, torrent, treadmill, tumbleweed, tumor, unpredictable, vagrant, vortex, war, weedwhacker, whirlpool, whirlwind, windshield, winged, wiper

Appendix B: words normed for auditory, motion, and visual imagery

ambulance, amplifier, anthem, applause, arrow, ball, bass, beat, blizzard, boomerang, boulder, bullet, cacophony, canary, carousel, catapult, chainsaw, cheetah, chimes, choir, chorus, clock, comet, curtain, dart, drill, drum, echo, elevator, engine, erosion, escalator, explosion, faucet, flood, flute, footsteps, gavel, gazelle, geyser, gong, gun, gunfire, hammer, heartbeat,

hoover, hurricane, hymn, inchworm, iPod, jackhammer, kettle, lance, landing, landslide, laughter, lever, lightning, loudspeaker, lullaby, marathon, melody, merry-go-round, meteor, metronome, molasses, mouse, music, musical, noise, note, pager, parachute, pendulum, piano, piccolo, race, racehorse, rattle, rattlesnake, rhythm, river, rollercoaster, rooster, seesaw, signal, siren, skewer, sled, slingshot, sloth, slug, snail, snowball, song, star, stopwatch, stream, symphony, teeter-totter, thud, thunder, thunderstorm, tides, tone, tornado, torpedo, traffic, trumpet, tsunami, tumbleweed, tune, turtle, vortex, wave, wheel, steering wheel, whirlpool, whirlwind, wrench

Appendix C: words normed for motion imagery only[C1]

argue, babble, balloon, bang, bark, belch, blast, bleat, blubber, blurt, bounce, buzz, cackle, call, canter, cartwheel, chant, charge, chat, cheer, chime, chirp, chop, chuckle, clamber, clamor, clash, clatter, click, climb, clomp, cluck, coast, collapse, coo, cough, crackle, crawl, creep, cry, dance, dart, dash, dig, dive, dodge, drift, drive, drone, drop, drum, fall, fart, fizzle, flip, flit, flop, flounder, flow, flush, fly, gasp, gesture, giggle, glide, groan, growl, grumble, grunt, guffaw, gurgle, hiss, hobble, holler, hoot, hop, howl, huff, hug, hum, hush, inch, jingle, jog, jump, knock, laugh, launch, leap, lift, limp, lope, lumber, lurch, march, moan, mosey, move, mumble, murmur, oink, plod, plow, plummet, polka, pop, pounce, prance, press, puff, pull, punch, purr, push, quack, race, rain, rant, rattle, reel, retreat, ride, roar, roll, run, sail, sashay, scamper, scream, screech, scurry, serenade, shatter, shout, shriek, shuffle, sidle, sigh, sing, sizzle, skulk, skydive, slam-dunk, slap, sleepwalk, slide, slink, slither, slouch, slurp, smash, snake, snap, snarl, sneak, sneeze, snicker, sniff, snigger, snore, snort, sob, spill, spin, splash, spring, sprint, sputter, squawk, squeal, stab, stammer, stampede, stand, stir, stomp, stream, stretch, stroll, strut, stumble, stutter, surf, surge, swagger, swarm, sweep, swim, swing, tailspin, take-off, tango, thunder, tiptoe, toss, totter, traipse, trudge, tug, twitter, voice, wade, wail, walk, waltz, wander, wave, weep, whimper, whine, whinny, whir, whirl, whisper, whistle, whoop, wiggle, wind, worm, wrestle, yawn, yell, yelp, yip, yodel, yowl, zigzag

[C1] Base terms from Cardillo et al. (2010)

References

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). SOS! An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods, 44,* 675–705. doi:10.3758/s13428-011-0182-9

Aziz-Zadeh, L., & Damasio, A. (2008). Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology, 102,* 35–39.

Aziz-Zadeh, L., Wilson, S. M., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology, 16,* 1818–1823. doi:10.1016/j.cub.2006.07.060

Bedny, M., Caramazza, A., Grossman, E., Pascual-Leone, A., & Saxe, R. (2008). Concepts are more than percepts: The case of action verbs. *Journal of Neuroscience, 28,* 11347–11353. doi:10.1523/JNEUROSCI.3039-08.2008

Bohrn, I. C., Altmann, U., & Jacobs, A. M. (2012). Looking at the brains behind figurative language—A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia, 50*(11), 2669–2683. doi:10.1016/j.neuropsychologia.2012.07.021

Bohrn, I. C., Altmann, U., Lubrich, O., Menninghaus, W., & Jacobs, A. M. (2013). When we like what we know—A parametric fMRI analysis of beauty and familiarity. *Brain and Language, 124,* 1–8.

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review, 112,* 193–216. doi:10.1037/0033-295X.112.1.193

Bowes, A., & Katz, A. (2015). Metaphor creates intimacy and temporarily enhances theory of mind. *Memory & Cognition, 43,* 953–963. doi:10.3758/s13421-015-0508-4

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977–990. doi:10.3758/BRM.41.4.977

Cardillo, E. R., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2010). Stimulus design is an obstacle course: 560 matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior Research Methods, 42,* 651–664. doi:10.3758/BRM.42.3.651

Cardillo, E. R., Watson, C. E., Schmidt, G. L., Kranjec, A., & Chatterjee, A. (2012). From novel to familiar: Tuning the brain for metaphors. *NeuroImage, 59,* 3212–3221. doi:10.1016/j.neuroimage.2011.11.079

Chatterjee, A. (2008). The neural organization of spatial thought and language. *Seminars in Speech and Language, 29,* 226–252.

Chen, E., Widick, P., & Chatterjee, A. (2008). Functional–anatomical organization of predicate metaphor processing. *Brain and Language, 107,* 194–202.

Citron, F. M. M., & Goldberg, A. E. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience, 26,* 2585–2595. doi:10.1162/jocn_a_00654

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505. doi:10.1080/14640748108400805

Columbus, G., Sheikh, N. A., Côté-Lecaldare, M., Häuser, K., Baum, S. R., & Titone, D. (2015). Individual differences in executive control relate to metaphor processing: an eye movement study of sentence reading. *Frontiers in human neuroscience, 8*. doi:10.3389/fnhum.2014.01057

Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., & Seidenberg, M. S. (2011). The neural career of sensory–motor metaphors. *Journal of Cognitive Neuroscience, 23,* 2376–2386. doi:10.1162/jocn.2010.21596

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., & Seidenberg, M. S. (2013). A piece of the action: Modulation of sensory–motor regions by action idioms and metaphors. *NeuroImage, 83,* 862–869. doi:10.1016/j.neuroimage.2013.07.044

Forgács, B., Lukács, Á., & Pléh, C. (2014). Lateralized processing of novel metaphors: Disentangling figurativeness and novelty. *Neuropsychologia, 56,* 101–109. doi:10.1016/j.neuropsychologia.2014.01.003

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory–motor system in conceptual knowledge. *Cognitive Neuropsychology, 22,* 455–479. doi:10.1080/02643290442000310

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge: MIT Press.

Gentner, D., Bowdle, B. F., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 199–253). Cambridge: MIT Press.

Gentner, D., & Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language, 37,* 331–355.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge: Cambridge University Press.

Gibbs, R. W., Jr. (2006). Metaphor interpretation as embodied simulation. *Mind & Language, 21,* 434–458. doi:10.1111/j.1468-0017.2006.00285.x

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12,* 395–427. doi:10.3758/BF03201693

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics, 8,* 183–206.

Giora, R. (2002). Literal vs. figurative language: Different or equal? *Journal of Pragmatics, 34,* 487–506. doi:10.1016/S0378-2166(01)00045-5

Giora, R., Zaidel, E., Soroker, N., Batori, G., & Kasher, A. (2000). Differential effects of right and left hemispheric damage on understanding sarcasm and metaphor. *Metaphor and Symbol, 15,* 63–83.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences, 7,* 92–96.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review, 97,* 3–18. doi:10.1037/0033-295X.97.1.3

Ianni, G. R., Cardillo, E. R., McQuire, M., & Chatterjee, A. (2014). Flying under the radar: Figurative language impairments in focal lesion patients. *Frontiers in Human Neuroscience, 8,* 871. doi:10.3389/fnhum.2014.00871

Jalal, B., & Ramachandran, V. S. (2014). A pilot investigation of "metaphor blindness" in a college student population. *Medical Hypotheses, 82,* 648–651.

Jamrozik, A., McQuire, M., Cardillo, E. R., & Chatterjee, A. (in press). Metaphor: Bridging embodiment to abstraction. *Psychonomic Bulletin & Review.*

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language, 55,* 18–32. doi:10.1016/j.jml.2006.02.004

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences, 9,* 512–518.

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol, 3,* 191–214. doi:10.1207/s15327868ms0304_1

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lacey, S., Stilla, R., & Sathian, K. (2012). Metaphorically feeling: Comprehending textural metaphors activates somatosensory cortex. *Brain and Language, 120,* 416–421.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.

Mashal, N., Vishne, T., & Laor, N. (2014). The role of the precuneus in metaphor comprehension: Evidence from an fMRI study in people with schizophrenia and healthy participants. *Frontiers in Human Neuroscience, 8,* 818. doi:10.3389/fnhum.2014.00818

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from http://w3.usf.edu/FreeAssociation/

Obert, A., Gierski, F., Calmus, A., Portefaix, C., Declercq, C., Pierot, L., & Caillies, S. (2014). Differential bilateral involvement of the parietal gyrus during predicative metaphor processing: An auditory fMRI study. *Brain and Language, 137,* 112–119. doi:10.1016/j.bandl.2014.08.002

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25. doi:10.1037/h0025327

Raposo, A., Moss, H. E., Stamatakis, E. A., & Tyler, L. K. (2009). Modulation of motor and premotor cortices by actions, action words and action sentences. *Neuropsychologia, 47,* 388–396.

Rapp, A. M., Mutschler, D. E., & Erb, M. (2012). Where in the brain is nonliteral language? A coordinate-based meta-analysis of functional magnetic resonance imaging studies. *NeuroImage, 63,* 600–610. doi:10.1016/j.neuroimage.2012.06.022

Roncero, C., & de Almeida, R. G. (2014). The importance of being apt: Metaphor comprehension in Alzheimer's disease. *Frontiers in Human Neuroscience, 8,* 973. doi:10.3389/fnhum.2014.00973

Roncero, C., & de Almeida, R. G. (2015). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods, 47,* 800–812. doi:10.3758/s13428-014-0502-y

Saygin, A., McCullough, S., Alac, M., & Emmorey, K. (2010). Modulation of BOLD response in motion-sensitive lateral temporal cortex by real and fictive motion sentences. *Journal of Cognitive Neuroscience, 22,* 2480–2890.

Schmidt, G. L., Kranjec, A., Cardillo, E. R., & Chatterjee, A. (2010). Beyond laterality: A critical assessment of research on the neural basis of metaphor. *International Journal of Neuropsychology, 16,* 1–5. doi:10.1017/S1355617709990543

Schmidt, G. L., & Seger, C. A. (2009). Neural correlates of metaphor processing: The roles of figurativeness, familiarity and difficulty. *Brain and Cognition, 71,* 375–386.

Schmidt-Snoek, G. L., Drew, A. R., Barile, E. C., & Agauas, S. J. (2015). Auditory and motion metaphors have different scalp distributions: An ERP study. *Frontiers in Human Neuroscience, 9,* 126. doi:10.3389/fnhum.2015.00126

Subramaniam, K., Beeman, M., Faust, M., & Mashal, N. (2013). Positively valenced stimuli facilitate creative novel metaphoric processes by enhancing medial prefrontal cortical activation. *Frontiers in Psychology, 4,* 211. doi:10.3389/fpsyg.2013.00211

Subramaniam, K., Faust, M., Beeman, M., & Mashal, N. (2012). The repetition paradigm: Enhancement of novel metaphors and suppression of conventional metaphors in the left inferior parietal lobe. *Neuropsychologia, 50,* 2705–2719. doi:10.1016/j.neuropsychologia.2012.07.020

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale: Erlbaum.

Van Lancker, D. R., & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain Language, 32,* 265–277.

Wallentin, M., Lund, T. E., Ostergaard, S., Ostergaard, L., & Roepstorff, A. (2005). Motion verb sentences activate left posterior middle temporal cortex despite static context. *NeuroReport, 16,* 649–652.

Wallentin, M., Østergaard, S., Lund, T. E., Østergaard, L., & Roepstorff, A. (2005). Concrete spatial language: See what I mean? *Brain and Language, 92,* 221–233. doi:10.1016/j.bandl.2004.06.106

Yang, F. G., Edens, J., Simpson, C., & Krawczyk, D. C. (2009). Differences in task demands influence the hemispheric lateralization and neural correlates of metaphor. *Brain and Language, 111,* 114–124.