Semantic distance predicts metaphoricity and creativity judgments in synesthetic metaphors

Winter, Bodo; Strik Lievers, Francesca

DOI:
10.1075/msw.00029.win

Citation for published version (Harvard):
Winter, B & Strik Lievers, F 2023, 'Semantic distance predicts metaphoricity and creativity judgments in synesthetic metaphors', *Metaphor and the Social World*, vol. 13, no. 1, pp. 59-80.
https://doi.org/10.1075/msw.00029.win

Semantic distance predicts metaphoricity and creativity judgments in synesthetic metaphors

Bodo Winter & Francesca Strik Lievers

Submission for special issue "Methods in metaphor research"

Abstract

This paper discusses a way of operationalizing metaphoricity quantitatively using a numerical measure of the semantic distance between two domains. We demonstrate the construct validity of this measure with respect to metaphoricity and creativity judgments in the domain of synesthetic metaphors—expressions such as *sweet melody* and *loud color* that involve combinations of terms from conceptually distinct sensory modalities. In a pre-registered study, we find that a continuous measure of sensory modality difference predicts metaphoricity and creativity judgments. While our results use synesthetic metaphors as a test case, it is possible to extend the application of our measure of semantic distance to other metaphorical expressions. In addition to demonstrating the utility of this measure, this work also demonstrates the utility of rating data in the domain of metaphor research.

Keywords: metaphor identification; norms; ratings; synesthetic metaphor; perception; perceptual metaphor; cosine similarity

1 Introduction

Most theoretical proposals in metaphor research characterize metaphors as involving two terms, conceptual domains, or frames that are conceptually distant from each other. For example, Holyoak and Stamenković (2018, p. 641) characterize metaphor as "the use of language to describe one thing in terms of something else that is conceptually very different." Researchers have used semantic distance between two domains as a proxy to aid in the identification of metaphors (Wan, Ahrens, et al., 2020). Conceptual distance by itself is not sufficient to identify metaphors (as it also characterizes other figures, see Prandi, 2017), but given that it is a necessary component of metaphor, conceptual distance should help in predicting naïve language users' intuitions about what constitutes metaphor. This basic idea pertains to all metaphors, but we explore it here in the context of synesthetic metaphors in adjective-noun phrases such as *sweet fragrance* and *smooth music* (Shen, 1997; Strik Lievers, 2015; Ullmann, 1959; Williams, 1976). Focusing on metaphors involving perceptual terms allows us to make use of existing databases where words have been rated for how strongly they relate to particular senses (Lynott et al., 2019; Lynott & Connell, 2009).

In their most typical expression, synesthetic metaphors involve a pairing of an adjective and a noun, each of which refers to a distinct sensory modality. In contrast, non-metaphoric adjective-noun sensory pairs involve two terms associated with the same sensory modality, such as *loud music* and *sweet taste*. In the past, researchers have used the semantic distance between the adjective and the noun term as a proxy for metaphoricity (Winter, 2019a) or metaphorical creativity (Chersoni et al., 2019). However, both of these studies *assumed* (but did not test) that semantic distance is a proxy of metaphoricity/metaphorical creativity. Here, we provide a crucial test of the assumption that domain distance actually taps into these constructs, assessing whether semantic distance predicts naïve language users' metaphoricity ratings (Study 1) and creativity ratings (Study 2). At the same time, these studies also demonstrate the usefulness of semantic ratings for metaphor identification (Wan, Ahrens, et al., 2020; Wan, Xing, et al., 2020).

Synesthetic metaphors have traditionally not received as much attention within the field of metaphor research, especially not within conceptual metaphor theory (Gibbs, 1994; Lakoff & Johnson, 1980). However, they provide an ideal test case for looking at the relation between semantic distance and metaphor. This is because for synesthetic metaphors, the semantic distance between the two terms can be operationalized as how much the two terms differ in their sensory modalities, which is easily quantifiable thanks to the availability of a large number of datasets containing ratings for how much particular sensory words correspond to the five senses (e.g., the word *blue* is rated to be high in visual strength but low in olfactory strength). Such modality ratings (Lynott et al., 2019; Lynott & Connell, 2009) make it possible to quantify directly how similar or dissimilar two terms are with respect to the sensory modalities they evoke. However, the principles that we outline here are not specific to synesthetic metaphors and can be extended to any measure of a word's meaning, including word vectors generated via distributional semantics (see Lenci, 2018, for review). No matter how one chooses to represent meaning, it is important to test the idea that semantic distance is involved in metaphor. In the discussion section, we give pointers to how the method discussed here can be applied to metaphors that are not synesthetic metaphors.

2 Background

2.1 Modality ratings

Linguists and cognitive scientists have been able to make a number of generalizations about sensory language. For example, it has been found that taste and smell words are overall less neutral and more evaluative than words for the other senses (Dubois, 2000; Krifka, 2010; Winter, 2016), or that within perception verbs, vision is often the source of semantic extension, with visual verbs used to metaphorically describe perception in other modalities (Evans & Wilkins, 2000; Viberg, 1983). Much of this work rests on having a clear understanding of what a "taste word" or a "sight word" is (Ronga, 2016; Winter, 2019b). Unfortunately, researchers very rarely discuss this fundamental classificatory issue, taking it for granted that an analyst can easily determine whether a word belongs to sight, sound, touch, taste, or smell.[1] This may be quite uncontroversial in cases such as the verb *to see* and the adjective *blue*, both of which are clearly visual. For many other cases however, this classification is not as easy, and resultingly, different analysts may classify particular sensory terms differently. As a case in point, consider the highly multisensory word *harsh*. Is it auditory, tactile, or perhaps even gustatory? After all, *harsh* is found in such common expressions as *harsh sound*, *harsh feeling*, and *harsh taste*.

Ronga (2016) discussed the problem of sensory modality classification in the context of synesthetic metaphors, giving examples how some of the same sensory terms have been classified as belonging to different senses by different researchers. Winter (2019a) argues that the lack of clear criteria for associating words with particular sensory modalities diminishes the reproducibility of research on sensory language, in line with general concerns about the questionable practice of relying exclusively on the introspective judgments of individual analysts who may be theoretically biased (Dąbrowska, 2016b, 2016a; Gibbs, 2007).

Bloomfield (1933, p. 140) famously noted that "the statement of meanings is the weak point in language study". Luckily, these days, word meaning can be captured in more reliable ways, thereby also creating new avenues for metaphor research. Specifically, there has been much work demonstrating the efficacy of distributional semantics, which refers to a class of approaches that computationally infer meaning via context (Günther et al., 2019; Landauer & Dumais, 1997; Lenci, 2008; Lund & Burgess, 1996), following Firth's (1957, p. 179) widely cited credo that "you shall know a word by the company it keeps". These approaches can quantify how similar two words are in meaning by looking at whether the words occur in similar contexts.

Another approach of quantifying meaning is to ask native speakers to rate words for particular semantic dimensions of interest. For example, words can be rated for concreteness (Brysbaert et al., 2014), emotional valence (Warriner et al., 2013), size (Scott et al., 2019), roughness (Stadtlander & Murdoch, 2000), motion (Medler et al., 2005), or how much they correspond to different body parts (Lynott et al., 2019). The ease with which data can be crowdsourced these days has made a large number of norming "megastudies" (Keuleers & Balota, 2015) available, many of which make new linguistic analyses possible (Winter, 2021), such as performing cluster analyses on semantic spaces defined by sets of ratings (Winter, 2019a).

It is important to stress that rating data sets are still based on introspective judgments and therefore, essentially a subjective form of data. But in contrast to the judgments of individual linguists, rating datasets have several advantages (Winter, 2019a, 2021). First, they are created by judgments from naïve language users without

[1] The five senses are merely used as a shorthand here, given that this is a folk model of the senses that is common in the West. However, physiological research cannot tell us how many senses there are, and empirical researchers as well as philosophers agree that there is no set number of senses (see discussion in Winter, 2019a).

the theoretical biases of linguists, and they are created independently of specific investigations (e.g., the modality ratings have not been collected with the application to synesthetic metaphors in mind). Second, by collecting judgments from a large number of participants, one can tap into the "wisdom of the crowd", getting a more reliable measure by averaging over idiosyncratic judgments. Third, rating data can make results more comparable across studies if different researchers rely on the same established rating datasets. Fourth and finally, average ratings are scalar, which opens up new avenues for analyzing semantics, such as ways of quantifying domain distance in a continuous fashion, as we do below.

Distributional semantics is arguably more general than semantic ratings in that the approach is not focused on one specific dimension of meaning. Instead, the whole meaning of the word is taken into account, inferred via its usage across many different contexts. Distributional semantics is corpus-based and thus does not rely on introspective judgments of isolated words, which could be seen as an advantage. However, without any extra data sources, the primary outcome of distributional semantics is semantic similarity data, e.g., whether terms A and B occur in similar contexts or not. One advantage of rating data over distributional semantics is that it allows pinpointing particular semantic dimensions of interest, such as whether a word belongs to a given sensory modality or not. Ultimately, both distributional semantics and introspective judgments approximate the latent construct of "meaning" in different ways. As meaning is something that cannot be directly measured objectively, researchers have to approximate the latent construct of meaning via introspective judgment on words (rating data) or via distributional patterns in text (distributional semantics). Both of these approaches operationalize the concept of "meaning" differently and are thus, ultimately, complementary to each other.

Asking native speakers to rate words for specific semantic dimensions of interest has a long history (Osgood et al., 1957; Paivio et al., 1968). Lynott and Connell (2009) were the first to use this methodology to collect sensory modality ratings for adjectives. Specifically, participants were asked to rate English words for how much they correspond to each one of the five senses. Participants were given a scale from 0 to 5 with a separate slider for each sensory modality. This allows for a word to be associated with multiple different sensory modalities. For example, the word *abrasive* was rated to be highest in touch (3.7), but also relatively high in sight (2.9), and somewhat lower in sound (1.7). It was rated to be close to zero in taste (0.6) or smell (0.6). Using modality ratings, we can circumvent the classificatory problems raised by Ronga (2016). Moreover, it becomes possible to treat modality association as a continuous quality, rather than straitjacketing multisensory words into particular sensory modalities (Lynott & Connell, 2009; Winter, 2019a).

2.2 Synesthetic metaphors and cosine similarity

Strik Lievers (2017) defined synesthetic metaphors in terms of conceptual conflict (Prandi, 2017). For example, when the clearly auditory descriptor *loud* is used to modify the clearly visual noun *color*, there is a crossmodal conceptual conflict, with no consistent relation between loudness and color in their purely auditory and visual interpretation. In contrast, when two words from the same sensory modality are combined, there is no crossmodal conceptual conflict, such as with the expressions *loud noise* and *dark color*.

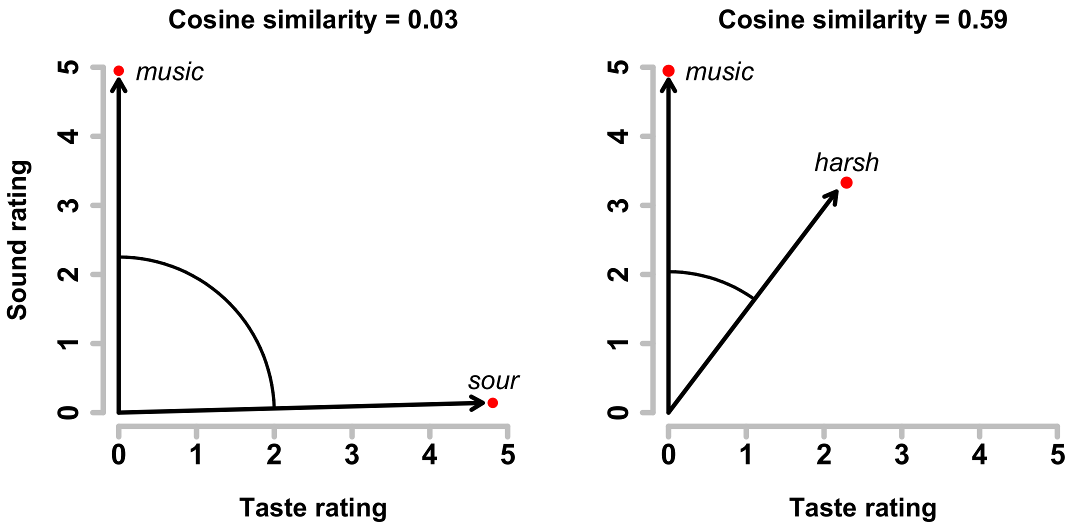

Figure 1. Two-dimensional subset of the sensory modality space (taste and sound dimensions only) with three words (*sour, harsh, music*) represented as directional vectors pointing to locations in this space; because *sour* and *music* point into quite different directions, the angle between the two vectors is larger than for *harsh music*, which translates into a lower cosine similarity

The semantic distance between the adjective and noun can be quantified using sensory modality ratings, where each word can be represented by a five-valued vector that points to a particular location in a five-dimensional "sensory modality space". This is visualized for two dimensions, taste and sound, in Figure 1. The cosine similarity is a quantitative measure of the similarity between two vectors $\vec{a}$ and $\vec{b}$, defined as follows:

$$(1)\ similarity = \cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}||\vec{b}|}$$

What the "content" of each vector is differs between applications. Here, $\vec{a}$ and $\vec{b}$ are the modality vectors of the adjective and noun defined by the ratings, with each vector quantifying the degree to which a word is associated with the five senses. In a distributional semantic application, $\vec{a}$ and $\vec{b}$ would be word vectors derived from corpora.

The cosine similarity measure itself corresponds to the angle between two vectors, such as those shown in Figure 1. Cosine similarity ranges from 0 (vectors are orthogonal, pointing in completely different directions) to 1 (vectors are on top of each other, pointing to exactly the same location in space). Following Strik Lievers' (2017) definition of synesthetic metaphor in terms of conflicting sensory domains, synesthetic metaphors should have lower cosine similarities (e.g., *sweet music*) than non-metaphoric combinations of two words from the same modality (e.g., *blue color, loud noise, abrasive feeling*). Winter (2019a) uses cosine similarity of modality rating data to test aspects of the distribution of synesthetic metaphors in corpora. Chersoni et al. (2019) similarly used cosine similarity to find synesthetic metaphors in corpora, and Wan et al. (2020) successfully combined it with other conceptual features within a metaphor detection task, not restricted to synesthetic metaphors (see also Wan, Xing, et al., 2020). These approaches took cosine similarity as a means of tapping into the construct of metaphor for granted. Here, we provide the crucial behavioral test that this measure actually corresponds to people's intuitions about metaphor.

An increasing number of metaphor researchers think of "metaphoricity" as something that is at least in principle gradable (see discussions in Dunn, 2015; Hanks, 2006; Müller, 2009). This measure of cosine similarity could be interpreted as a continuous measure of "metaphoricity", corresponding to the theoretical view that metaphors are not categorical. According to this view, the further two semantic domains are away from each other in a metaphoric expression, the more metaphorical is such expression in a scalar manner. However, using cosine similarities does not commit oneself to thinking about metaphoricity in a continuous fashion and is compatible with a categorical interpretation of metaphor. Instead, cosine similarity could be interpreted as a measure of metaphorical creativity (Chersoni et al., 2019), where expressions with high cosine similarity (similar modalities) are less creative metaphors than expressions with low cosine similarity. Creativity can fruitfully be seen as a scalar construct, where something can be "more" or "less" creative. Whether a scalar value of metaphoricity has to be seen as involving a scalar/continuous analysis of metaphor is more controversial. For example, whereas Strik Lievers (2017) sees synesthetic metaphors as a category, Winter (2019a, 2019b) thinks of it in continuous terms. However, both approaches are compatible with the notion that synesthetic metaphors have to be associated with the quantitative measure of cosine (dis)similarity, given that this is an explicit measure of domain distance between modalities.

Here, we performed two experiments. Study 1 investigates metaphoricity judgments; Study 2 investigates creativity judgments. Besides providing a test case for the idea that semantic distance as operationalized via cosine similarity predicts these two rating scales, this also allows us to assess whether metaphoricity judgments are correlated with creativity judgments, which is theoretically expected given that metaphor and creativity are often thought to be connected (Indurkhya, 1992; Gentner et al., 1997; Leung et al., 2012; Hidalgo-Downing, 2015).

3 Methods

3.1 Pre-registration and data availability

All hypotheses, materials, and statistical analyses were pre-registered prior to data collection on the Open Science Framework. All materials, data, and analysis code are available in the publicly accessible OSF repository: https://osf.io/wcuqd/?view_only=0b9012f275654129abcb3cd78c967df5

3.2 Stimulus creation

We selected adjectives from the Lynott and Connell (2009) modality ratings for 423 property descriptors.[2] We selected the top 5 most exclusive adjectives per sensory modality, thereby yielding a dataset of 25 adjectives. Some of these adjectives however were deemed unsuitable, which led us to the replacements detailed in the online OSF repository. The final set of adjectives used to construct the stimuli is shown in Table 1.

Sight	Sound	Touch	Taste	Smell
blue	*beeping*	*hard*	*malty*	*fragrant*
bright	*loud*	*scaly*	*minty*	*odorous*
crimson	*noisy*	*smooth*	*savory*	*perfumed*
dark	*shrill*	*warm*	*sour*	*scented*
silver	*squealing*	*woolly*	*sweet*	*smelly*

Table 1. Sensory adjectives used in this study

For nouns, we used the Lancaster norms for 40,000 English words (Lynott et al., 2019) rather than the noun norms collected by Lynott and Connell (2013). This is for the following reason: Because adjectives are overall more exclusively tied to particular sensory modalities than nouns (Winter, 2019a), it is hard to find nouns that denote sensory impressions while also relating to specific modalities very strongly. Therefore, it is necessary to use a larger set to select from, thereby maximizing the potential of finding suitable nouns. Detailed selection procedure for the nouns is explained in the online OSF repository. The nouns shown in Table 2 were chosen as a basis for generating adjective-noun pairs. As has been noted in the previous literature, there are not many nouns for touch concepts (Popova, 2005), and there are generally few taste and smell words in English (Levinson & Majid, 2014; Winter et al., 2018), which explains the unequal number of nouns across the five sensory modalities.

Sight	Sound	Touch	Taste	Smell

[2] The original the Lynott and Connell (2009) dataset was preferred over the much more extensive new Lancaster norms (Lynott et al., 2019) because it includes only adjectives that are clearly sensory descriptors. Many of potentially relevant words from the Lynott and Connell (2009) norms do not occur in the Lancaster norms, such as the onomatopoeias *banging*, *beeping*, and *squealing*.

brightness	*noise*	*touch*	*flavor*	*aroma*
color	*music*	*feeling*	*taste*	*odor*
gleam	*melody*	*contact*		*smell*
lighting	*harmonics*			*scent*
neon	*chatter*			
darkness	*echo*			
glimmer				

Table 2. Sensory nouns used in this study

We then exhaustively paired all selected adjectives with all selected nouns, computing the cosine similarity between the two as described in Section 2.2. From this superset of all adjective-noun pairs, we selected five random pairs for each decile of the cosine similarity scale, e.g., five pairs with similarities ranging from 0 to 0.1, five pairs ranging from 0.1 to 0.2, and so on. The total set of final adjective-noun pairs with their respective cosine similarity values is shown in Table 3.

	Pair	**Cosine**		**Pair**	**Cosine**
1	*malty melody*	0.05	26	***bright touch***	0.51
2	*crimson scent*	0.05	27	***warm gleam***	0.51
3	*perfumed touch*	0.07	28	*blue chatter*	0.51
4	*perfumed color*	0.09	29	***warm flavor***	0.55
5	*loud flavor*	0.09	30	***sour scent***	0.60
6	***fragrant touch***	0.12	31	***smooth color***	0.63
7	*noisy taste*	0.14	32	***warm taste***	0.65
8	*perfumed brightness*	0.14	33	***hard color***	0.65
9	*smooth scent*	0.16	34	*smelly taste*	0.65
10	***warm music***	0.19	35	***sour odor***	0.69
11	*crimson taste*	0.21	36	*scented flavor*	0.71
12	*smelly feeling*	0.22	37	*scaly glimmer*	0.74
13	*bright noise*	0.24	38	*woolly sight*	0.75
14	***warm smell***	0.25	39	*malty odor*	0.76
15	*fragrant gleam*	0.27	40	***fragrant taste***	0.79
16	*squealing touch*	0.31	41	*woolly feeling*	0.85
17	***warm scent***	0.31	42	*scaly feeling*	0.85
18	***smooth flavor***	0.35	43	***smooth feeling***	0.89
19	***loud color***	0.35	44	***warm feeling***	0.89
20	*beeping neon*	0.37	45	***sweet flavor***	0.90
21	*silver feeling*	0.40	46	***perfumed aroma***	0.94
22	***warm odor***	0.43	47	***hard touch***	0.96
23	*warm glimmer*	0.44	48	***dark glimmer***	0.98
24	*hard chatter*	0.44	49	***beeping noise***	0.99
25	***smooth taste***	0.47	50	***sour taste***	0.99

Table 3. Adjective-noun pair stimulus pairs used in this experiment and their cosine similarities; sorted by cosine similarity (least similar to most similar); bold pairs are attested in the Corpus of Contemporary American English

3.3 Word frequencies

Metaphoricity could be affected by metaphor familiarity. To control for this, we extracted adjective frequencies, noun frequencies, and adjective-noun pair frequencies from the Corpus of Contemporary American English (COCA). There were 25 pairs that were attested in the corpus, and 25 pairs that were not. It is important to note that stimuli with higher cosines were also more likely to be attested in COCA (for discussion, see Winter, 2019a). The pairs that were attested had an average cosine of $M = 0.62$ ($SD = 0.27$). In contrast, the pairs that were not attested had an average cosine of only $M = 0.38$ ($SD = 0.27$). A mixed Bayesian logistic regression model with corpus attestation as the response and cosine as the predictor (brms: Bernoulli, prior on cosine coefficient = normal(0, 2); random intercepts for adjectives and nouns) indicated that there was strong support for an effect of cosine similarity on attestation (logit slope: -3.67, 95% credible interval: [-6.40, -1.10], $p(\beta > 0) = 0.002$).[3] This finding replicates the previously made observation that adjective-noun pairs with higher modality similarity are overall more frequent in a corpus than genuinely crossmodal combinations, such as synesthetic metaphors (Winter, 2019a).

3.4 Participants

Participants were recruited on Amazon Mechanical Turk via CloudResearch and received 1.40 USD as reimbursement. Participants had to have completed at least 100 HITs (Human Intelligence Tasks) with an approval rate of at least 90%. We additionally used the "CloudResearch approved participants" feature of CloudResearch to get high-quality Turkers who have successfully completed HITs with CloudResearch prior to the experiment. Location was restricted to the United States. For Experiment 1, we collected data from 50 participants (23 female, 27 male; average age 39, range 19-67). For Experiment 2, we collected data from 48 participants (17 female, 31 male; average age 36, range 22-53).

3.5 Procedure

All participants saw all 50 items in randomized order. Participants were asked to rate "how metaphoric a particular expression is". They were not given a definition of metaphor (as we wanted their judgements to reflect the common understanding of metaphoricity), but they received an example of a metaphoric/creative and a

[3] Since half of the stimulus set has an attestation of zero and overall frequencies were fairly low, a model with the binary distinction "attested" versus "not attested" performs just as well as a model that uses the raw frequencies. We therefore use the simpler measure of attestation for adjective-noun pairs.

literal/uncreative expression, which was *cold anger* as opposed to *cold ice*. We chose these examples because it involved a sensory word similar to the ones used in our study, but not in a synesthetic metaphor expression. The reasoning behind this choice of instructions was that we did not want to bias participants strongly in favor of our hypothesis by providing actual examples of synesthetic metaphors.

In our instructions, we also emphasized to participants that the word *feeling* ought to be interpreted with respect to tactile sensations felt through the skin, not as an emotional term. The questions displayed beneath each adjective-noun pair were: *To what extent is this expression literal or metaphoric?* (Study 1) And: *How creative is this expression?* (Study 2). Response options were labelled "very literal", "literal", "moderately literal", "neutral", "moderately metaphoric", "metaphoric", "very metaphoric" (Study 1), and "very uncreative", "uncreative", "moderately uncreative", "neutral", "moderately creative", "creative", "very creative" (Study 2).

3.6 Data exclusion

Only complete records were considered (i.e., we did not analyze drop-outs). All participants correctly answered a catch trial with a clear objective solution ("2 + 3 = ____"), which meant that no exclusions had to be made on the basis of this criterion. In addition, we pre-registered that we wanted to exclude "straightliners" (Kim et al., 2019; Zhang & Conrad, 2014) that gave the same response too many times, for which our criterion was that participants could not have more than 80% of all responses to be the same response. Here too, no participants had to be excluded on this criterion either. Taken together, the fact that neither exclusion criteria led to any exclusions suggests high data quality.

3.7 Statistical analysis

All data was analyzed with R version 4.0.2 (R Core Team, 2019). The tidyverse package version 1.3.0 (Wickham et al., 2019) was used for data processing and visualization. Bayesian models were computed with the brms package version 2.14.4 (Bürkner, 2017). The main statistical model is a mixed cumulative ordinal regression model (Bürkner & Vuorre, 2019) with rating as the response variable (ordinal-valued: 1 to 7). In the case of Study 1, this ordinal response variable was rated metaphoricity; in the case of Study 2 it was rated creativity. Other than this change in the response variable, both studies used the same model specification. The primary predictor of interest was cosine similarity, continuously-valued ranging from 0 to 1. Each model also included three additional predictors: whether the adjective-noun pair was attested in COCA or not (binary categorical variable: attested versus not attested), the log10 frequency of the adjective, and the log10 frequency of the noun. Random effects included random intercepts for participant, adjective, noun, and adjective-noun pair. In addition, we included by-participant-varying random slopes in the cosine predictor, thereby allowing participants to respond differently to the cosine manipulation. The model was estimated with MCMC and 5,000 iterations (3,000 warmup samples excluded). All models converged well (Rhat = 1.0) and posterior predictive checks revealed that the model could have predicted the data distribution well.

4 Results

4.1 Experiment 1: Metaphoricity

Figure 2 shows what is known as a “joy plot”, which is a set of density curves visualizing the distribution of cosine similarity values for each of the 7 different response categories, from “very literal” to “very metaphoric”. The picture clearly shows that literal responses cluster around adjective-noun pairs with high cosine values (little domain distance), and metaphoric responses cluster tend to have lower cosine values (high domain distance). The distributions do not only differ in their means, however, but also in their spread. Figure 2 clearly shows that “very literal” responses were disproportionately given to high cosine similarity adjective-noun pairs. On the other hand, all intermediate response options (from “metaphoric” to “literal”) have very wide distributions. It is noteworthy that the metaphoric endpoint (“very metaphoric”) is also more spread out than the corresponding literal endpoint. This means that a wider range of cosine similarity values were accepted as “very metaphoric,” indicating overall higher variability across adjective-noun pairs for this response category.

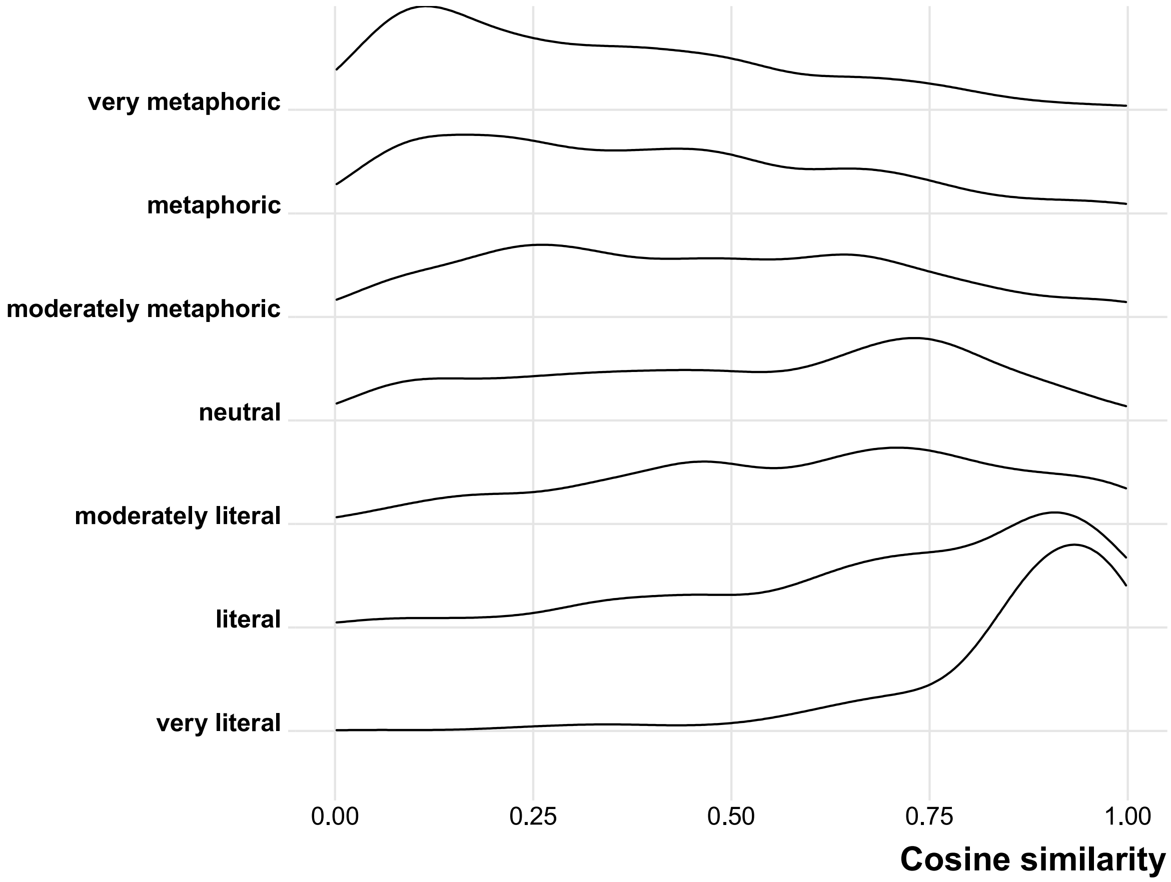

Figure 2. Joy plot showing the distribution of cosine similarities from “very literal” responses (bottom) to “very metaphoric” responses (top)

This conclusion is also corroborated by the Bayesian mixed ordinal regression model. The coefficients of this model are shown in Figure 3. These coefficients express the degree to which a particular variable in the model (cosine similarity, adjective frequency, noun frequency, corpus attestation) impacts metaphoricity ratings. Most importantly, the coefficients shown in Figure 3 clearly reveal a strong effect of cosine similarity on the response, where adjective-noun pairs with higher cosine similarity values (= less domain distance) received lower metaphoricity ratings (coefficient: -5.0, *SE* = 0.76). The 95% credible interval of this coefficient was far away from zero [-6.45, -3.49], with the posterior probability of this effect being negative being exactly $p(\beta_1 < 0) = 1.0$. This indicates that not a single posterior sample showed a cosine similarity effect with the opposite sign, suggesting that we can be very certain in the effect of cosine similarity, given this model and data.

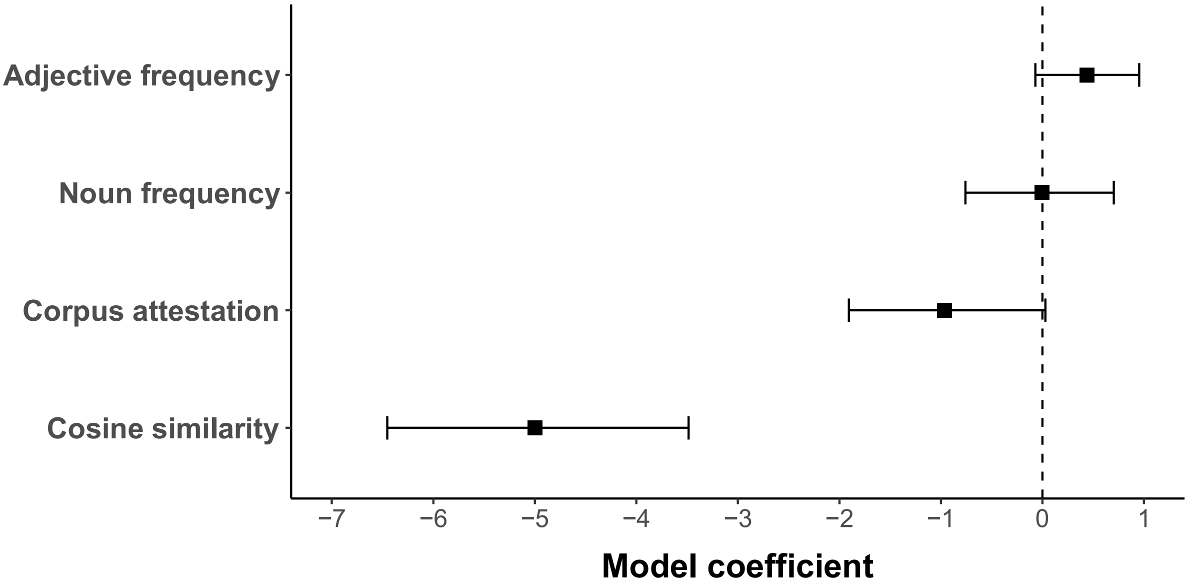

Figure 3. Coefficients of the Bayesian ordinal mixed regression model; values further to the right mean that the coefficient is associated with more metaphoric responses; values further to the left are associated with more literal responses; black squares show posterior means of the coefficients; intervals show 95% Bayesian credible intervals

In addition, there was a clear tendency for attested as opposed to unattested pairs to be judged as less metaphoric / more literal (coefficient: -0.96, *SE* = 0.50). The 95% credible interval of this coefficient slightly overlapped with zero [-1.91, +0.03], and was associated with a relatively high probability of being negative, $p(\beta_1 < 0) = 0.97$. Moreover, there was a tendency for more frequent adjectives to be judged as *more* metaphoric (coefficient: +0.44, *SE* = 0.26). The 95% credible interval of this coefficient slightly overlapped with zero [-0.07, +0.95], and was associated with a relatively high probability of being positive, $p(\beta_1 > 0) = 0.95$. This finding may reflect the fact that more frequent words are more often used as source domains for metaphor, as demonstrated for synesthetic metaphors in Winter (2019a) and for metaphors in general in Winter and Srinivasan (2021).

Finally, there was no indication that noun frequency mattered at all (coefficient: -0.01, *SE* = 0.37), with this coefficient firmly covering zero [-0.76, +0.70], with a low posterior probability of this effect being below zero, $p(\beta_1 < 0) = 0.50$. When all factors are taken together, the model described 61% of the variance in metaphoricity ratings (95% credible interval: [59%, 62%]).

4.2 Experiment 2: Creativity

Moving on to Experiment 2, Figure 4 shows a joy plot for the cosine similarity distributions as a function of response, from "very uncreative" (bottom) to "very creative" (top). A very similar tendency to the metaphoricity ratings is observed, with more creative metaphors having lower cosines (higher domain distance), and less creative metaphors having higher cosines (lower domain distance). However, the density curves are overall less crisp and more strongly overlapping.

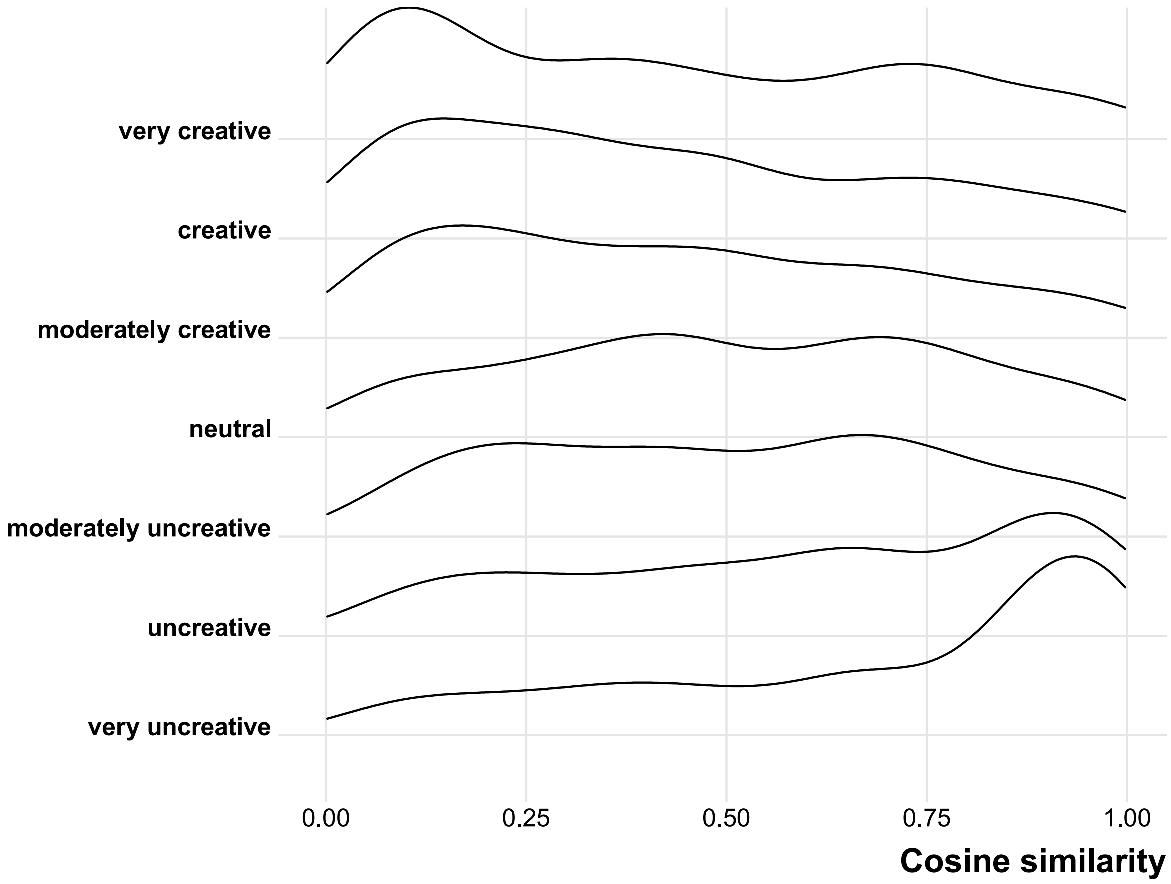

Figure 4. Joy plot showing the distribution of cosine similarities from "very uncreative" responses (bottom) to "very creative" responses (top)

The coefficients of the corresponding Bayesian mixed ordinal regression model are shown in Figure 5. As before, these quantify the extent to which a particular predictor influences ratings, in this case creativity ratings. The scale of

Figure 5 is the same as that for metaphoricity ratings in Figure 3, which visually highlights the fact that all coefficients in the creativity dataset are closer to zero than in the metaphoricity one. This indicates that cosine similarity more strongly corresponds to metaphoricity than metaphorical creativity.

There was a strong effect of the cosine similarity of the adjective-noun pair, with higher cosine pairs judged to be less creative (coefficient: -1.88, *SE* = 0.50). The 95% credible interval of this coefficient also did not overlap with zero [-2.85, -0.91] and had a very high probability of being negative, $p(\beta_1 < 0) = 0.99$. Again, the effect of cosine similarity on creativity ratings was overall much less strong than the effect of cosine similarity on metaphoricity ratings, with the cosine similarity coefficient for the creativity judgments (-1.88) being less than half of the cosine similarity coefficient for the metaphoricity judgments (-5.0).

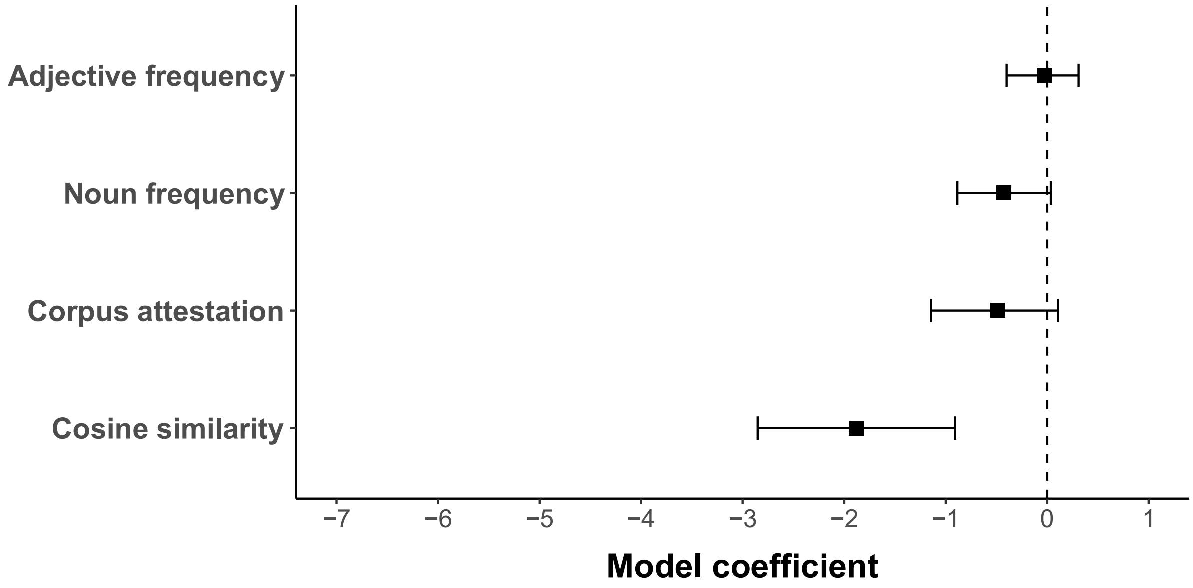

Figure 5. Coefficients of the Bayesian ordinal mixed regression model; values further to the right mean the coefficient is associated with more creative responses; values further to the left are associated with more uncreative responses; black squares show posterior means of the coefficients; intervals show 95% Bayesian credible intervals

In contrast to Experiment 1, there was no reliable effect of adjective frequency (+0.07, *SE* = 0.18), with the 95% credible interval of this coefficient firmly spanning zero, [-0.30, +0.43], and a low posterior probability of this effect being positive, $p(\beta_1 > 0) = 0.44$. There was a weak tendency for adjective-noun pairs with more frequent nouns to be judged as less creative (-0.34, *SE* = 0.24), but the coefficient of this effect included zero, [-0.84, +0.12]. The posterior probability of this coefficient being below zero was, $p(\beta_1 < 0) = 0.97$. Finally, adjective-noun pairs that were attested in a corpus were judged to be less creative (coefficient: -0.49, *SE* = 0.32). The 95% credible interval of this coefficient spanned zero [-1.14, +0.10], but was associated with a relatively high probability of being negative, $p(\beta_1 < 0) = 0.94$.

Overall, the model explained much less variance than the model for Experiment 1: 38%, with a 95% credible interval of [36%, 40%].

4.3 Correlation between Experiment 1 and Experiment 2

To assess whether the metaphoricity ratings from Experiment 1 are correlated with the creativity ratings from Experiment 2, we computed Pearson's *r* for items-based averages. The correlation was high (*r* = 0.84) and the 95% credible interval of this correlation coefficient far away from zero: [0.69, 1.00]. Figure 6 shows the correlation.

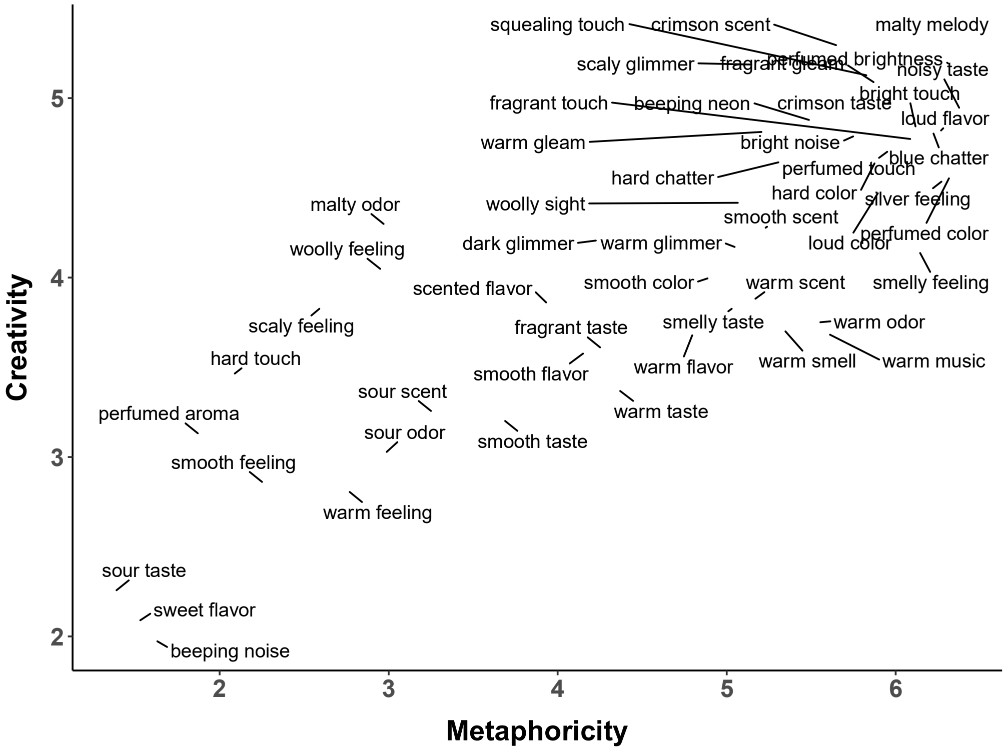

Figure 6. Metaphoricity ratings (Experiment 1) correlate strongly with creativity ratings (Experiment 2)

5 Discussion

Across two experiments, we have shown that a quantitative measure of domain distance — operationalized via cosine similarity of modality ratings — correlated with people's intuitions regarding metaphoricity (Experiment 1) and creativity (Experiment 2). Adjective-noun pairs with higher domain distance (lower cosine similarity) were judged to be more metaphoric, and more creative, although much stronger results were obtained for metaphoricity ratings than for creativity ratings. Consistent with the observation that metaphor plays an important role in creativity and metaphors are often thought to be creative expressions, we found that metaphoricity ratings and creativity ratings were correlated with each other.

Our experiment makes two methodological contributions. First, we show that domain distance is a useful measure to predict naïve language users' metaphoricity ratings. This may also provide support for methods that use it for identifying metaphor, as demonstrated by Wan et al. (Wan et al., 2020). Although domain distance is not a sufficient condition for metaphoricity, it is a necessary condition, and therefore a valuable tool for searching for metaphors in text. Second, we demonstrate the utility of using rating data more generally. Synesthetic metaphors are a great test case to make both of these methodological points because of the availability of modality ratings, as well as because for these metaphors, it is clear what constitutes "domain distance", namely, distance in perceptual modalities. It should be stated however, that the ideas presented here can be extended to other types of metaphor. Domain distance as one of the clues for metaphor is compatible with most theoretical positions in this field and cosine similarity can be computed even if modality rating data are not available. One way is to use other semantic rating data (e.g., concreteness ratings); another is to use distributional semantic vectors (Wan, Ahrens, et al., 2020). Thus, the method we discuss here is in no way constrained to synesthetic metaphors.

The results we present here also provide important insight for the study of synesthetic metaphor. As discussed above, Winter (2019a, 2019b) proposes that it is useful to think about the senses in a continuous manner: rather than straitjacketing words into particular sensory modalities, modality ratings allow a more continuous perspective on whether a word belongs to a sensory modality or not. Modality ratings also allow the same word to be associated with multiple senses. Together with the present results, this perspective on sensory language would mean that there are also gradations of synesthetic metaphors. The literature on synesthetic metaphors discusses cases such as *sweet fragrance* and *sweet melody* on equal terms, classifying both as synesthetic metaphor. But these expressions differ quite radically in their cosine similarity: the domain distance is much higher for *sweet melody* (i.e., lower cosine similarity) than for *sweet fragrance*. This corresponds to the fact that the latter expression involves two highly integrated sensory modalities: taste and smell. The results of Experiment 1 can be seen as consistent with the idea that a continuous treatment of synesthetic metaphors were words are not "forced" to be of any one sensory modality is possible. However, even if one is not theoretically committed to a continuous treatment of metaphoricity in this domain, cosine similarity can be used to find clear cases of metaphor (those that have the biggest domain distance), as demonstrated by Chersoni et al. (2019).

It should be noted that the precise meaning of cosine similarity depends on the specific phenomenon investigated. Vecchi et al. (2017) looked at several semantic deviance measures generated from distributional semantics (including cosines) for adjective-noun pairs more generally. They show that, for example, a cosine measure of semantic deviance between the adjective and the noun predicts human acceptability ratings.

Altogether, we hope that studies like this 1) inspire metaphor researchers to find new ways of quantifying and operationalizing the notion of metaphor, and 2) inspire researchers to make more use of the large amount of rating datasets that are freely available, as these could prove to be an invaluable tool for metaphor research in the future.

References

Bloomfield, L. (1933). *Language*. Chicago University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Chersoni, E., Lievers, F. S., & Huang, C.-R. (2019). Semantic distance and creativity in linguistic synaesthesia. *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, 370–378.

Dąbrowska, E. (2016a). Cognitive Linguistics' seven deadly sins. *Cognitive Linguistics*, *27*(4), 479–491. https://doi.org/10.1515/cog-2016-0059

Dąbrowska, E. (2016b). Looking into introspection. In G. Drożdż (Ed.), *Studies in Lexicogrammar: Theory and applications* (pp. 55–74). John Benjamins.

Dubois, D. (2000). Categories as acts of meaning: The case of categories in olfaction and audition. *Cognitive Science Quarterly*, *1*(1), 35–68.

Dunn, J. (2015). Modeling abstractness and metaphoricity. *Metaphor and Symbol*, *30*(4), 259–289.

Evans, N., & Wilkins, D. (2000). In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language*, *76*(3), 546–592.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press.

Gentner, D., Brem, S., Ferguson, R., & Wolff, P. (1997). Analogy and creativity in the works of Johannes Kepler. In T. B. Ward, S. M. Smith, & J. Vaid (Eds.), *Creative Thought: An Investigation of Conceptual Structures and Processes.* (pp. 403–459). American Psychological Association.

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

Gibbs, R. W. (2007). Why cognitive linguists should care more about empirical methods. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, & M. Spivey (Eds.), *Methods in Cognitive Linguistics* (pp. 2–18). John Benjamins.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Hanks, P. (2006). Metaphoricity is gradable. In A. Stefanowitsch & S. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 17–35). Mouton de Gruyter.

Hidalgo-Downing, L. (2015). Metaphor and metonymy. In R. Jones (Ed.), *The Routledge Handbook of Language and Creativity* (pp. 129–150).

Holyoak, K. J., & Stamenković, D. (2018). Metaphor comprehension: A critical review of theories and evidence. *Psychological Bulletin*, *144*(6), 641. https://doi.org/10.1037/bul0000145

Indurkhya, B. (1992). *Metaphor and Cognition. An Interactionist Approach*. Kluwer. https://doi.org/10.1007/978-94-017-2252-0

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468.

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

Krifka, M. (2010). A note on an asymmetry in the hedonic implicatures of olfactory and gustatory terms. In S. Fuchs, P. Hoole, C. Mooshammer, & M. Żygis (Eds.), *Between the regular and the particular in speech and language* (pp. 235–245). Peter Lang.

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

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*(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, *20*(1), 1–31.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics*, *4*, 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Leung, A. K. -y., Kim, S., Polman, E., Ong, L. S., Qiu, L., Goncalo, J. A., & Sanchez-Burks, J. (2012). Embodied metaphors and creative "acts." *Psychological Science*, *23*(5), 502–509. https://doi.org/10.1177/0956797611429801

Levinson, S. C., & Majid, A. (2014). Differential ineffability and the senses. *Mind & Language*, *29*(4), 407–427. https://doi.org/10.1111/mila.12057

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*(2), 203–208. https://doi.org/10.3758/BF03204766

Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558–564.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior Research Methods*, *45*(2), 516–526.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2019). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 1–21. https://doi.org/10.3758/s13428-019-01316-z

Medler, D. A., Arnoldussen, A., Binder, J. R., & Seidenberg, M. S. (2005). *The Wisconsin perceptual attribute ratings database*. Retrieved from http://www.neuro.mcw.edu/ratings/.

Müller, C. (2009). *Metaphors dead and alive, sleeping and waking: A dynamic view*. University of Chicago Press.

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

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1p2), 1.

Popova, Y. (2005). Image schemas and verbal synaesthesia. In B. Hampe (Ed.), *From perception to meaning: Image schemas in cognitive linguistics* (Vol. 29, pp. 395–419). Mouton de Gruyter.

Prandi, M. (2017). *Conceptual Conflicts in Metaphors and Figurative Language*. Routledge.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ronga, I. (2016). Taste synaesthesias: Linguistic features and neurophysiological bases. In E. Gola & F. Ervas (Eds.), *Metaphor and Communication* (pp. 47–60). John Benjamins.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Shen, Y. (1997). Cognitive constraints on poetic figures. *Cognitive Linguistics*, *8*(1), 33–72. https://doi.org/10.1515/cogl.1997.8.1.33

Stadtlander, L. M., & Murdoch, L. D. (2000). Frequency of occurrence and rankings for touch-related adjectives. *Behavior Research Methods, Instruments, & Computers*, *32*(4), 579–587.

Strik Lievers, F. (2015). Synaesthesia: A corpus-based study of cross-modal directionality. *Functions of Language*, *22*(1), 69–95.

Strik Lievers, F. (2017). Figures and the senses. *Review of Cognitive Linguistics*, *15*(1), 83–101.

Ullmann, S. (1959). *The principles of semantics*. Jackson, Son & Co.

Vecchi, E. M., Marelli, M., Zamparelli, R., & Baroni, M. (2017). Spicy adjectives and nominal donkeys: Capturing semantic deviance using compositionality in distributional spaces. *Cognitive Science*, *41*(1), 102–136. https://doi.org/10.1111/cogs.12330

Viberg, Å. (1983). The verbs of perception: A typological study. *Linguistics*, *21*(1), 123–162.

Wan, M., Ahrens, K., Chersoni, E., Jiang, M., Su, Q., Xiang, R., & Huang, C.-R. (2020). Using conceptual norms for metaphor detection. *Proceedings of the Second Workshop on Figurative Language Processing*, 104–109. https://doi.org/10.18653/v1/2020.figlang-1.16

Wan, M., Xing, B., Qi Su, Liu, P., & Huang, C.-R. (2020). Sensorimotor enhanced neural network for metaphor detection. *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 312–317.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, J. M. (1976). Synaesthetic adjectives: A possible law of semantic change. *Language*, *52*(2), 461–478.

Winter, B. (2016). Taste and smell words form an affectively loaded and emotionally flexible part of the English lexicon. *Language, Cognition and Neuroscience*, *31*(8), 975–988.

Winter, B. (2019a). *Sensory linguistics: Language, perception, and metaphor*. John Benjamins.

Winter, B. (2019b). Synaesthetic metaphors are neither synaesthetic nor metaphorical. In L. J. Speed, C. O'Meara, L. San Roque, & A. Majid (Eds.), *Perception metaphor*. John Benjamins.

Winter, B. (2021). Managing semantic norms for cognitive linguistics, corpus linguistics, and lexicon studies. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management*. MIT Press.

Winter, B., Perlman, M., & Majid, A. (2018). Vision dominates in perceptual language: English sensory vocabulary is optimized for usage. *Cognition*, *179*, 213–220.

Winter, B., & Srinivasan, M. (2021). Why is semantic change asymmetric? The role of concreteness and word frequency in metaphor and metonymy. *Metaphor and Symbol*. https://doi.org/10.1080/10926488.2021.1945419

Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. https://doi.org/10.18148/srm/2014.v8i2.5453