



Evaluating the predication model of metaphor comprehension: Using word2vec to model best/worst quality judgments of 622 novel metaphors

Parastoo Harati¹ · Chris Westbury¹ · Milad Kiaee²

Accepted: 4 February 2021 / Published online: 1 April 2021
© The Psychonomic Society, Inc. 2021

Abstract

In this paper our goal is to undertake a systematic assessment of the first, most widely known, and simplest computational model of metaphor comprehension, the predication model developed by Kintsch (*Cognitive Science*, 25(2), 173–202, 2000). 622 metaphors of the form “x is a y” were selected from a much larger set generated randomly. The metaphors were judged for quality using best/worst judgments, which asks judges to pick the best and worst metaphor from among four presented metaphors. The metaphors and their judgments have been publicly released. We modeled the judgments by extending Kintsch’s predication model (2000) by systematically walking through the parameter space of that model. Our model successfully differentiated metaphors rated as good ($> 1.5z$) from metaphors rated as bad ($< -1.5z$; Cohen’s $d = 0.72$) and was able to successfully classify good metaphors with an accuracy of 82.9%. However, it achieved a true negative rate below chance at 36.3% and had a resultantly low kappa of 0.037. The model could not distinguish unselected random metaphors from those selected by humans as having metaphorical potential. In a follow-up study we showed that the model’s quality estimates reliably predict metaphor decision times, with better metaphors being judged more quickly than worse metaphors.

Keywords Computational modeling · Vector space model · Comprehension · Metaphor quality · Word2vec · Figurative language processing

In spite of the ubiquity of metaphors, these linguistic units are challenging to study due to the combination of their intricacies, their unconstrained semantic characteristics, and their multitude of forms (Kövecses, 2017). The need for a more systematic approach in the study of metaphors will be fulfilled if the stimuli used in metaphor studies are more standardized.

Recent research into metaphor processing suggests that metaphors are processed in the same manner as literal linguistic units. Sperber and Wilson (2008) believe that metaphors are not part of a discrete and distinct category of language use, but rather are utterances that differ in the strength of their implicature. In other words, metaphors are considered as cases

that take place at one end of a continuum that ranges from literally true to non-literal analogues.

Creating computational models to explain and simulate the human interpretation of metaphorical statements has long been a matter of interest (see reviews in Veale et al., 2016; Reid & Katz, 2018). By modeling metaphor judgments computationally, we are forced to identify abstract similarities between metaphors that may superficially appear very different. The challenge in modeling metaphor comprehension arises from the fact that words in the target and source positions may not share many relevant features (Bowdle & Gentner, 2005). For example, in the metaphor *life is a highway*, the limited number of features that are relevant to both the target *life* and the source domain *highway* makes it hard to find a simple solution to associate the target and source easily.

In this paper our goal is to undertake a systematic assessment of the first, most widely known, and simplest computational model of metaphor comprehension the predication model developed by Kintsch (2000). That model is based on an algorithm that accounts for one of the major challenges of metaphor modeling: the lack of an adequate number of relevant features of the source that are shared by the target.

✉ Parastoo Harati
p.harati@ualberta.ca

¹ Department of Psychology, University of Alberta,
Edmonton, Alberta, Canada

² Department of Mechanical Engineering, University of Alberta,
Edmonton, Alberta, Canada

Kintsch's algorithm models metaphor as a subordinate of argument-predicate type sentences and is applicable to other types of figurative language (Kintsch, 2001). We systematically vary the parameters of the model, to optimize and quantify its performance in distinguishing metaphors that are rated good and bad by humans.

Computational modeling

Introduction

The predication algorithm envisions metaphor meaning as flowing through a spreading activation network of semantically related words (Kintsch, 2000). The stimulation of spreading activation of meaning is done using a vector model of language in the following steps. First, the m nearest cosine neighbors of the source word are identified. Kintsch (2000) used LSA (Landauer & Dumais, 1997) to make vectors from a corpus of 11 million word tokens. He proposed a number between 500 and 1500 as the best value for m , the semantic neighborhood size to compute, though without offering any empirically grounded justification. Second, the k neighbors most similar to the target word among the m neighbors of the source word are identified. Kintsch suggested $k = 5$. In the third step, those k vectors are averaged together with the target vector to make a new vector. The cosine distance between this new vector and the vector of the source word was proposed as a measure of metaphor comprehensibility.

Kintsch's predication algorithm explains metaphor comprehension in a manner that is in line with the theory of categorization within psychology (Utsumi, 2011). The categorization theory posits that metaphor comprehension is explained through class inclusion. The source is considered to be a member of an ad hoc category of which the target is a prototypical member (Glucksberg & Keysar, 1990). For instance, when comprehending the metaphor *Life is a highway*, the target of *life* is taken to belong to a "long and challenging" category, a category to which the source word *highway* also belongs.

Kintsch (2000) introduced the model, but that introduction was purely conceptual and included no data except some limited simulation data from just a single metaphor, "My lawyer is a shark." Kintsch and Bowles (2002) assessed how the model represents human judgments of metaphors more rigorously. They asked 24 participants to rate the difficulty of comprehending each of 30 metaphors on a 5-point scale. The model was tested on a small set of 26 metaphors: 13 metaphors that were rated as "easy to understand" (a rating of 2 or lower; average = 1.75) and 13 that were rated as "difficult to understand" (a rating of 3 or higher; average = 3.68). Four intermediate metaphors were discarded. Using the cosine

distance between the vector representing the meaning of the metaphor model and the vectors for the component words as an estimate of difficulty, the model did not succeed in distinguishing easy from difficult metaphors. The average cosine distance between the vectors of the target and source words also did not differentiate between easy and difficult metaphors. The model did succeed in distinguishing easy from difficult metaphors based on the average distance of the vectors of the closest neighbors of the source from the vector of the target, though Kintsch and Bowles noted that the correlation between rated difficulty and this measure ($r = -0.46$) was "statistically significant but not very high" (p. 258). These results hint at the potential of the predication algorithm, but the results are not very strong. Other analyses suggest that the predication algorithm appropriately selects the relevant features of the metaphor's meaning (Al-Azary & Buchanan, 2017). However, research so far has neglected to test other relevant features used by the predication algorithm in order to build a better model of metaphor comprehension.

In this study, we replicate Kintsch's predication algorithm (2000) and assess the model using human judgments of a large number of metaphors. We also extend the model by considering the effect of using different parameter settings for the metaphor predication algorithm and by expanding the range of the parameters used by Kintsch (2000) and Kintsch and Bowles (2002).

Method

Procedure

We used a set of 623 novel metaphors. This large set was used to try to minimize human bias in their selection by including a wide range of metaphors. We developed our novel metaphors using random computational generation. Nouns were drawn from lists distributed with the freeware program "JanusNode" (JanusNode.com), a scriptable text generation toy that includes many noun lists. We used this program to generate many tens of thousands of metaphors using an "x is a y" template. From these, we initially hand-selected 1507 unique metaphors that were judged by the selector to have a possible metaphorical interpretation, irrespective of the quality of that interpretation. We hand-corrected these where necessary by adding appropriate articles before the nouns. Examples of accepted metaphors include *Alcohol is a leash*, *Togetherness is a devil*, *A neurosis is a riddle*, and *A parent is a parachute*.

Many of the selected metaphors used the same target or source word, sometimes many times. For example, 23 metaphors were accepted that used the target word *love*, 21 that used the target word *God*, 11 that used the source domain *friend*, and seven that used the source domain *magician*. We selected a subset of metaphors by hand so

that no target or source was used more than three times. A second rater then evaluated the selected metaphors, being asked to eliminate those that met any one of the following criteria:

- It seems stupid.
- It seems literally true.
- It is too obvious.
- It is trite.
- It makes no sense.
- You don't know what one or more of the words means.
- It seems bizarre.
- It is offensive.
- It is not even a metaphor.
- What it implies is not true.
- You have heard it before.

Following this round of elimination, we ended up with 623 novel metaphors that satisfied the first and second authors as being novel and admitting of a possible interpretation. Of these, four used the same target three times (i.e., 4 sets \times 3 = 12 metaphors), 139 used the same target two times, three used the same source three times, and 157 used the same source two times. Exactly 400 metaphors (64.2%) used unique words in both positions. One low-rated metaphor (*A genital is a circus*) was removed from our set after we had collected ratings, because the obscure singular word *genital* is not generally acceptable in English.

For collecting human-judgment ratings of goodness, we used the best/worst scaling method (Kiritchenko & Mohammad, 2016). On each trial, participants were presented with four metaphors that appeared vertically on the screen and were instructed to choose the best (first column) and the worst (second column) metaphor from the tuple. They indicated their choice by clicking on a radio button, which would not allow users to choose the same metaphor for both categories. Only after choosing both a best and worst metaphor were participants able to click on a button to move to the next trial. The four metaphors were random, but computationally optimized (using software released by Hollis, 2018) so that all items were compared to all other items. The stimuli were displayed in 26-point Times New Roman font.

The use of best/worst scaling has two main advantages over rating scales. One is that it asks raters to make two decisions that are often easier than using a rating scale: to pick the best and the worst metaphor from among a set of four. The other is that it generates implicit rank information in a very efficient way. Although each participant makes only two decisions per trial, we obtain rank information for three of the four pairwise items in each trial, with only the two unselected metaphors being unranked against each other.

Stimuli were presented to the participants using custom-designed software running under Apple's OS

10.15 connected to 17.1" LCD monitors with 1280 \times 1024 pixel resolution. The data were gathered in testing rooms constructed to reduce outside noise. Participants were shortly briefed about the task before signing the consent forms. They were shown written instructions on the screen, which was followed by verbal instructions as they finished reading the instructions.

Results from the best/worst scaling choices were scored using value learning, as outlined in Hollis (2018). Value learning is an algorithm that learns the expected value of a given metaphor V_A (\sim its probability of being chosen as best), by updating that metaphor's estimated probability of being chosen as better or worse than another metaphor towards its given value in each pairwise comparison (i.e., towards 1 if the metaphor is chosen as best; towards 0 if it is chosen as worst). In our study, each metaphor that was chosen as best or worst participated in three such pairwise comparisons per best/worst trial, since there were four metaphors in each trial. The rule for updating the value score V_A of the metaphor selected as best or worst is:

$$V_{A-UPDATED} = V_A + \alpha \beta (\gamma - V_A)$$

Note that this is an error-correction formula, similar to many other learning models: a previous prediction (V_A) is corrected for prediction error ($\alpha \beta (\gamma - V_A)$). β is a self-adjusting (consistently decreasing) learning rate parameter intended to minimize artifacts that may be introduced from the order of the input data. γ is the observed value in the current trial (i.e., 1 if the metaphor was chosen as best or 0 if it was chosen as worst).

The final parameter α is a salience weight that takes into account the relationship of the metaphor that was selected as either best or worst to each of the other three metaphors in a trial. It is defined in terms of the a priori odds that the current metaphor is expected to win the matchup, expressed as an odds ratio. The odds ratio O_A for a metaphor with current value V_A is:

$$O_A = V_A / (1.0 - V_A)$$

The salience parameter α puts this odds ratio in the context of the odds ratio of the three competitor metaphors on each trial, each of which has its own odds ratio O_B . The parameter is bounded between 0 and 1, by being defined as 1 minus O_A (the odds ratio of the current metaphor being chosen as best) divided by the sum of the odds ratios of both metaphors in the pairwise matchup:

$$\alpha = 1.0 - (O_A / (O_A + O_B))$$

Unexpected wins (which are more informative than expected wins, by definition of Shannon information) are thus weighted more strongly than expected wins. For example, a

metaphor with an odds ratio O_A of 1 (which is as likely to be chosen as best as to be chosen as worst) that is chosen as best against a metaphor with an odds ratio O_B of 9 (9:1 odds of being chosen as best) will be given a salience weight of $1 - 1/9 = 0.89$, a high weight because such a win was improbable according to the a priori odds for each metaphor. If the two metaphors were reversed, the metaphor chosen as best would be given a salience weight of $1 - 9/10 = 0.10$, a low weight because such a win was already likely according to the a priori odds for each metaphor.

Value learning returns scores bounded between 0 and 1, where 1 would indicate that an item is unbeatable (always chosen as best), 0 would indicate that it is always beatable (always chosen as worst) and 0.5 would indicate that it is equally likely to be chosen as best or worst.

Participants

Sixty-two students from the University of Alberta participation pool participated in this study, which was enough to have every metaphor rated 32 times, a number suggested as optimal (due to decreasing returns on accuracy after that) in Hollis and Westbury (2018).

Results

The range of ratings is shown in Fig. 1. The value learning scores cover a range between 0.25 (for the metaphor *An accordion is an electrocuting*, which perhaps only makes sense if you hate accordions) and 0.76 (for the metaphor *A door is a beginning*). Although our metaphors thereby spanned a wide range of judged quality, as they were intended to, note that the bounding limits of 0.25 and 0.76 also suggest that there is imperfect agreement about their quality, since no metaphor was universally recognized as best or worst.

Without knowing the variance of judgments for individual items, it is difficult to evaluate how useful a set of judgment norms are. Since individual items may have overlapping confidence intervals, knowing only their means is not very useful. Although value-scored best/worst judgments do not allow for the calculation of variance, we can estimate the standard error by treating the value scores as if they were the results of a poll on whether or not the item was best. This is not perfectly accurate because, as outlined above, value scoring does not compute a simple average value of binary votes, since it weights each vote by how surprising it is given the previous history of observations. However, an item's value score is similar to the probability that the item would be voted best against any randomly selected item. This allows us to estimate the standard error of measurement on each item using the standard error of proportions $(pq)^{0.5} / N^{0.5}$, where p is the probability an item would be voted best (the value score), q is the probability it would not be voted best ($1 - p$), and N is the number of times an item participated in a vote (i.e., how many times it was seen). Each item appears 32 times per experiment. On scores that range between 0.25 and 0.76, the estimated standard error of measurement ranges between 0.032 (for items at extremes) and 0.044 (for items in the middle). The 95% confidence interval per item is then bounded between 0.062 and 0.086. Rounding up a little, we can conclude that items with a difference of at least 0.09 in their value scores will not have overlapping 95% confidence intervals.

The ten best and worst metaphors are shown in Table 1. The full set of metaphors and their ratings is available from <https://osf.io/ye2nm/>.

Modeling human judgments

To replicate and extend Kintsch's model, we used the word2vec skip-gram model (Mikolov et al., 2013) run over

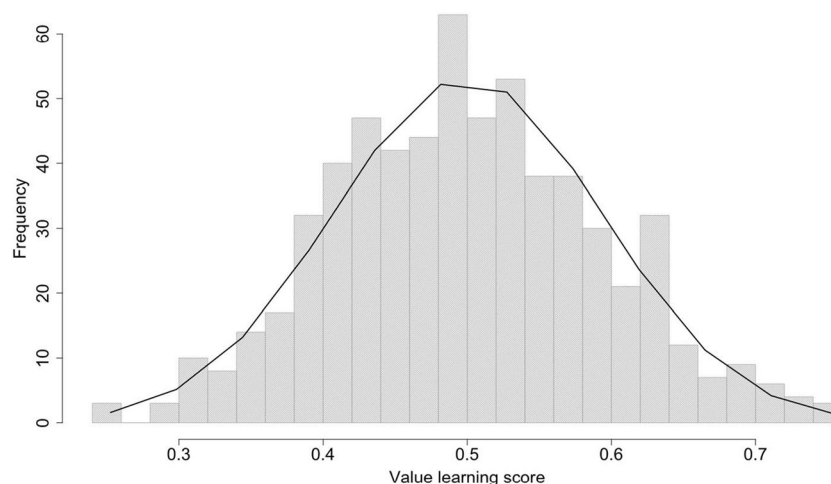


Fig. 1 Distribution of value learning scores for ratings of metaphor goodness

Table 1 The ten metaphors rated best (left) and worst (right) by participants. Higher value learning scores indicate metaphors estimated to be better

Metaphor	Value learning score	Metaphor	Value learning score
A door is a beginning.	0.76	An accordion is an electrocuting.	0.25
Art is a reflection.	0.75	A mommy is a devastation.	0.25
A parent is a parachute.	0.74	A penis is a chameleon.	0.26
A past is a guide.	0.73	A schoolgirl is an ocean.	0.28
Hope is a battle.	0.73	A cuddle is a telescope.	0.30
Cruelty is a virus.	0.73	A chicken is a theory.	0.30
Greed is an illness.	0.72	Gravity is a monkey.	0.30
Language is a tool.	0.71	Womanliness is a policeman.	0.30
An election is a zoo.	0.71	A television is a yoke.	0.31
Poetry is a garden.	0.71	A job is a urinal.	0.31

the Google News corpus, which contains over three billion words, consisting of three million word types. The matrix is available from: <https://drive.google.com/file/d/0B7XkCwpI5KDYNINUTTISS21pQmM/edit?usp=sharing>.

Word2vec has two main advantages over the LSA model used by Kintsch (2000). One is that it is more psychologically plausible, since it is based on a simple learning algorithm that attempts to predict the local context, and that incrementally builds the matrix. LSA is based on the psychologically implausible ideas that we keep a running tally of how often words co-occur and can perform some sort of statistical dimension reduction of the co-occurrence matrix (see Levy & Goldberg, 2014, for how word2vec deals with this problem by its very nature). The second is that predictive models such as word2vec achieve “a thorough and resounding victory against their count-based counterparts” (Baroni et al., 2014, p. 238) when compared on a variety of relevant performance measures such as synonym detection, semantic decision, and conceptual classification (see also Mandera et al., 2017, but also, for a dissenting view in the focal domain, Altszyler et al., 2016).

For each word type, the word2vec skip-gram model generates a vector of length 300 (by convention), which encodes the predicted context (two words on either side) of that word, using a neural network with one hidden layer of 300 nodes. We used a minimum occurrence count of 10, a window size of 2, and negative sampling, $k = 10$.

Method

We slightly modified Kintsch’s (2000) algorithm for metaphor comprehension, using word2vec vectors instead of his LSA vectors, and focusing only on neighbors that were shared between m_1 and m_2 . The model works as follows:

1. Compute the cosine distances between the vectors for the target and source word and every other word in the word2vec matrix.
2. Extract the common neighbors among the m closest neighbors of each word. We will discuss the value of m below.
3. Extract the k words among the m common neighbors that are closest to the source word (or fewer, if there are less than k). Kintsch set $k = 5$, but we will consider other values below.
4. Average together the k vectors with the vector of the target word. We will also consider averaging them with the vector of the source word, though Kintsch did not (but see Utsumi, 2011).
5. Compute the cosine between the vector for the source word, and the new averaged vector computed in Step 4 (or between the two averaged vectors, if the common neighbors were averaged with the vectors of both the source and target word).

The basic idea behind this model is that we are able to get an idea of the strength of the relationship between the vectors of the source and target words used in a metaphor, by “nudging” them in the direction of their common meaning in that metaphor, as estimated by the vectors of their overlapping neighbors. The nudging occurs when we average the common neighbors with the vectors of one or both words from the metaphors. It is similar to explaining to someone that when you said that a third person was funny, you meant “funny haha,” not “funny peculiar.” The predication of the ambiguous word *funny* by the word *haha* pushes the interpretation of that word towards the desired interpretation of humor.

In his original theoretical paper on the model, Kintsch (2000) proposed a common neighborhood size of $500 < m < 1500$. Kintsch and Bowles (2002) averaged together the closest five of 500 neighbors of the source with the target vector, using the LSA model (Landauer & Dumais, 1997).

In the current study, we refer to this parameter set as the “5-0-500.” This label indicates that Kintsch’s model averaged the target vector with the five closest neighbors of the source ($k1 = 5$), did not average the source vector with any common neighbors ($k2 = 0$), and looked at the 500 closest neighbors of the source word to identify the five neighbors ($m = 500$). Kintsch suggested that this is the best parameter set for his algorithm, as the semantic neighborhood of the target has to be a large number because the target and the source may be quite unrelated, but it should not be too large, as the model may pick up terms that are irrelevant to the intended meaning (Kintsch, 2000).

Kintsch’s claim about the size of the computed common neighborhood of the target is only briefly referred to in his paper. There is no systematic explanation for exclusion of neighborhood sizes out of the initially proposed range (i.e., $500 < m < 1500$). In this study, we systematically varied the $k1$ - $k2$ - m parameter sets for Kintsch’s (2000) algorithm, where $k1$ is the number of common neighbors averaged with the target vector, $k2$ is the number of common neighbors averaged with the source vector, and m is the number of neighbors of each word that were searched to find common neighbors.

We initially varied the $k1$ and $k2$ parameters in steps of 5 from 0 to 20, while varying the m parameter in steps of 500 from 1500 to 3000. We decided on an initial value of m as 1500 since values lower than this often result in there being no shared neighbors between the target and source words.

Results

To adjudicate between the models, we initially focused on maximizing the effect size (Cohen’s d) by contrasting the estimates for metaphors that humans had rated as clearly good (> 1.5 SDs above average, $N = 41$, the top 6.6%) against those rated as clearly bad (< -1.5 SDs below average, $N = 37$, the bottom 5.9%). The results are shown in Table 2.

There are three clear results visible in this table. The first result is shown using bold text in Table 2. For all values of m , the model with the largest effect size was always the model that averaged the common neighbors with the vectors of both the source and target word. The second result is that the effect size increases as m increases. Searching larger neighborhoods for shared neighbors results in better prediction of metaphor quality. The third result is that the models are relatively insensitive to changing values of $k1$ and $k2$. With $m = 3000$ and limiting ourselves to only the models in which $k1 = k2$, the average effect size (Cohen’s d) across all values of $k1$ and $k2$ is 0.546, with a standard deviation of just 0.0047. For the two values of $m > 1500$, the (marginally) largest effect size was seen when both $k1$ and $k2$ were equal to 5.

These three results simplify the modeling problem because they suggest that we can fix the values of both $k1$ and $k2$ to 5 and vary only m . We increased m from 3000 to 6000 in steps

Table 2 Effect sizes in contrasting model estimates for metaphors rated high ($> 1.5z$) or low ($< -1.5z$) by humans

Parameters	m=1500	m=2500	m=3000
0-5-m	0.402	0.420	0.439
5-0-m	0.389	0.476	0.518
5-5-m	0.422	0.518	0.552
0-10-m	0.414	0.383	0.422
10-0-m	0.410	0.451	0.500
10-10-m	0.433	0.496	0.548
0-15-m	0.429	0.391	0.500
15-0-m	0.435	0.472	0.413
15-15-m	0.460	0.509	0.543
0-20-m	0.428	0.390	0.494
20-0-m	0.435	0.475	0.409
20-20-m	0.460	0.508	0.542

The first parameter is the number of common neighbors averaged with the vector for the source word. The second parameter is the number of common neighbors averaged with the vector for the target word. The third parameter, m , is the number of neighbors that were examined in looking for common neighbors between the source and target word. The best result for each value of $k1$ and $k2$ is shown in bold

of 500. The results are shown in Fig. 2. The maximum Cohen’s d was obtained when $m = 4500$, which showed a large effect size of $d = 0.78$. Across all 622 metaphors, the correlation of this model’s estimates with all the human judgments was also the maximum seen among the models, $r = 0.18$ ($p = 3.12e-0.6$, one-tailed, but we note that this is a

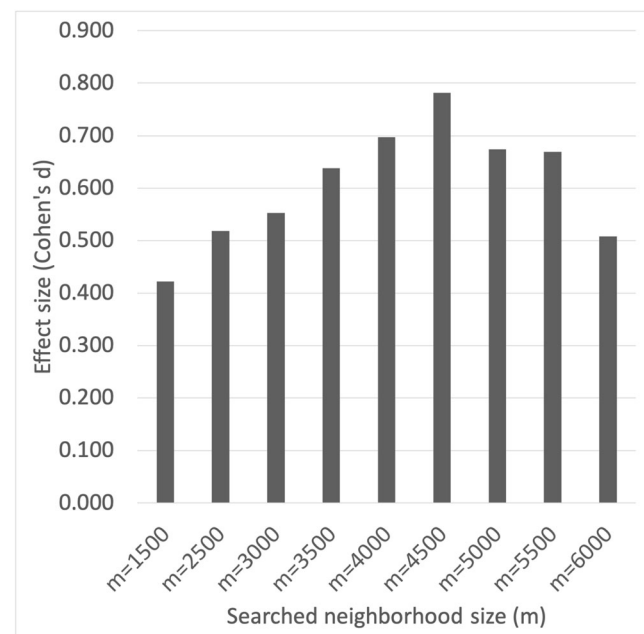


Fig. 2 Effect sizes (Cohen’s d) for model estimates of metaphors judged good ($> 1.5z$) contrasted with metaphors judged bad ($< -1.5z$), by the size to m , the neighborhood searched for common neighbors of the source and target, with $k1$ and $k2$ set to 5

dubious measure since the model's estimates are not normally distributed). This evidence in support of the model is closely analogous to Kintsch and Bowles' (2002) evidence in favor of their model, inasmuch as they also pitted known good metaphors against known bad metaphors and showed that the model could distinguish the two to some extent. In the remainder of the paper, we will present other measures that show why this is weak evidence by which to judge a model of metaphor quality.

The distribution of 5.5.4500 values with respect to the best/worst human judgment data is shown in Fig. 3. Using cutoff values of 1.5z (= a value score of 0.64) on the 5.5.4500 estimates and a cutoff of $> 0.5z$ (more likely good than bad) on the human judgments, the model achieves a true positive rate (sensitivity) of 82.9% (exact binomial $p = 2.4e-06$) but a true negative rate (specificity) below chance at 36.3% (exact binomial $p = 8.5e-07$, in the wrong direction). The model's precision (ratio of true positives to all positives) is only 8.42%, reflecting the high false positive rate. It is more successful at correctly rejecting metaphors judged not-good by humans, with a 96.7% negative predictive value. The high negative predictive value displays the model estimates are good indicators of metaphors low quality. In other words, when a metaphor is bad it will have a low model estimate. Cohen's kappa (κ , a measure of inter-judge agreement) is just 0.037, indicating only slight overall agreement between the human judges and the computational model. The F-score (a global measure of model accuracy bounded between 0 and 1) was low, at 0.15. All in all, the model performs poorly, especially at the most important task of identifying good metaphors.

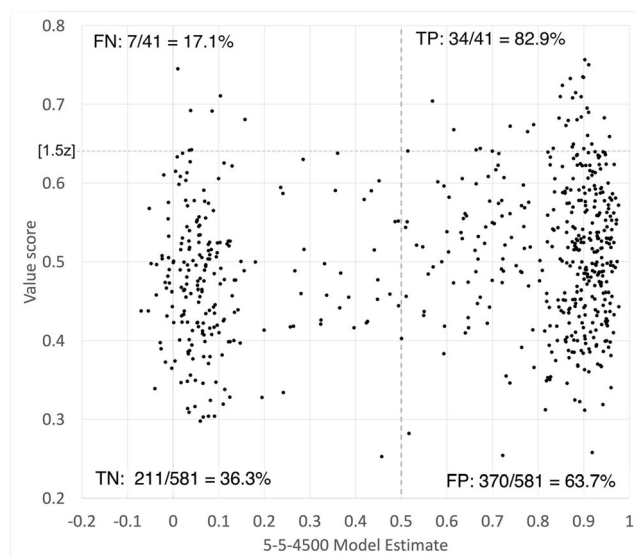


Fig. 3 Distribution of 5-5-4500 model estimates (x-axis) graphed against human best/worst value scores (y-axis). Vertical is the mean for human judgments; horizontal is the 1.5z cutoff for model estimates. TP/FP = true/false positive rate; TN/FN = true/false negative rate

The distribution of 5.5.4500 values by good, middle, and bad metaphors is shown in Fig. 4 (which also includes the distribution of 622 randomly generated “x is a y” phrases that are discussed below). Figure 4 clearly shows that the difficult problem of differentiating middle-range metaphors from good or bad metaphors using the 5.5.4500 estimates would not be possible, because there is a total overlap of middle-range metaphors at both extremes.

Metaphors that were judged good by humans but estimated bad by the model include *A parent is a parachute*, *Pessimism is a spear*, *Imagination is a tornado*, and *An election is a zoo*. Metaphors judged bad by humans but estimated good by the model include *A bed is a bus*, *A bird is a flower*, *A penis is a chameleon*, and *A mother is a devastation*.

Discussion

This study extended the model of metaphor difficulty developed by Kintsch (2000). Both our study and Kintsch and Bowles (2002) examined the “is-a” form of metaphors (e.g., *Fascination is a lantern*). Kintsch and Bowles (2002) tested their model superficially, without examining their parameter settings, using a weak five-point rating scale for just 26 metaphors, applying vectors made from a small corpus of 11 million words, and without considering classification accuracy. In our study we considered over 620 metaphors ordered by judged quality, used vectors from a corpus of over three billion tokens, and have confronted (though certainly not solved) the problem of classification accuracy.

In this study we also systematically explored the parameter settings and expanded the range for the parameters to optimize our developed model. The systematic exploration of parameters gives us an objective view of how to optimize parameter settings and what parameters may influence the model development.

This study differs from other computational studies of metaphor quality in terms of method of item development, as we used computer-generated novel metaphors as our experimental metaphoric items, in order to minimize human bias introduced to studies by human-created experimental items.

As we have noted above, the full set of metaphors and their best/worst ratings developed and used in this study are available to the public. This will contribute to having a more standard and homogeneous set of stimuli that can be used in the future by metaphor researchers. Such stimuli allow for more comparative research on metaphors between different studies. The model developed in this study does well at identifying metaphors that are rated highly ($> 1.5z$) by humans (82.9% true positive rate), but does poorly at identifying metaphors rated lower than that (36.3% true negative rate). It has a low precision of 8.41% and a low Cohen's kappa of 0.037. Contrary to previous observation for low model estimates (discussed above), this low precision indicates that the high

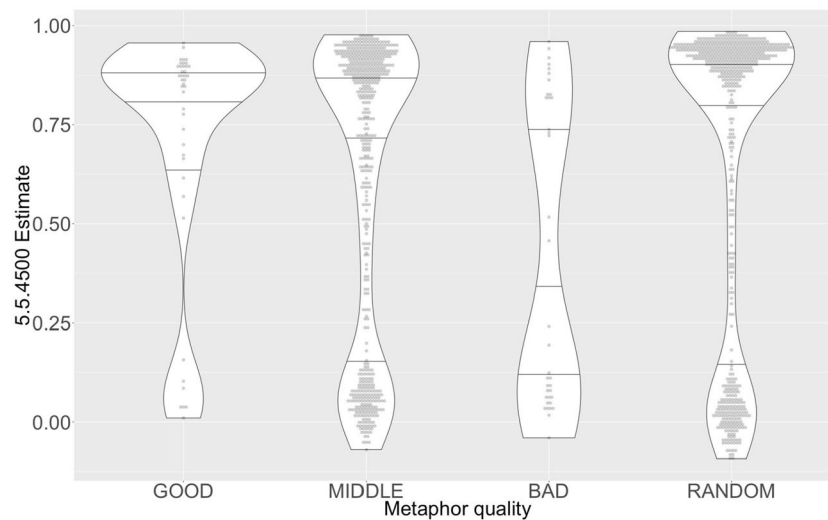


Fig. 4 5.5.4500 estimates for the good ($> 1.5z$), middle, and bad ($< -1.5z$) metaphors judged by humans and for random noun-noun pairs

model estimates are not suggestive of high quality in metaphors. These poor results suggest that the predication algorithm fails to account for some relevant characteristics of metaphors other than just their abstracted semantic relationship. It has been pointed out by several studies (Reid & Katz, 2018; Terai & Nakagawa, 2012; Utsumi, 2011) that the categorization view of metaphor comprehension cannot completely account for human metaphor comprehension.

One problem with such models is that they cannot distinguish true semantic relatedness from metaphorical relatedness. Although we did not include them in our metaphor set, we assume that phrases like *A cat is a dog* or *A robin is a bird* would necessarily be judged by humans as very poor metaphors, since one is a literally untrue statement and the other is a literally true statement. However, these relationships are both given an estimate above 0.98 by the 5-5-4500 model, which is a higher score than the highest-rated metaphor (*A blessing is a smile*, rated 0.96). The model's classification performance is overestimated because our stimuli set did not include any literal and/or metaphorically uninterpretable statements (i.e., statements for which it would be impossible to think of a metaphorical interpretation).

We assessed whether we could improve our model by giving it information about such “shallow” semantic relationships, by computing the 5-5- m scores of all of the metaphors in our set, with m stepped from 50 to 250 in steps of 50. We used linear regression across those five estimates and the 5-5-4500 estimates to predict the human goodness ratings. Only the 5-5-4500 estimates entered into the model. This may be because very few metaphors have high estimates with such small values of m . For example, only five metaphors (0.8%) had a 5-5-50 estimate above 0.5.

It is possible that there may be a “sweet spot” of semantic relatedness for metaphors. Strings of the “ x is a y ” form that have little or no apparent metaphoric potential (e.g., [according to our own intuitions] *Devotion is nutmeg*, *Seriousness is a sapphire*, *A critique is a weaver*) may show $k1$ - $k2$ - m values that distinguish them from strings that do have some metaphoric potential. To explore this possibility, we randomly generated 622 new “Noun1 is a Noun2” strings without any human selection. Although it is possible that some of those strings might be good metaphors, we know from the human effort of selecting our original set that the vast majority of randomly generated strings do not admit of any apparent metaphoric interpretation. The 5.5.4500 estimates of these new phrases are graphed against the estimates for the good, middle, and bad metaphors in Fig. 4. There is a near complete overlap of the distribution of 5.5.4500 estimates among these random metaphors with the distribution among the metaphors, demonstrating that it would not be possible to use these estimates to distinguish acceptable metaphors from non-metaphors. This in turn suggests that the literature on computational models of metaphor has oversimplified the problem (as we did above) by focusing on strings that have first been preselected as acceptable metaphors by humans. The harder problem is going to be to distinguish strings that form acceptable metaphors from strings that do not form acceptable metaphors.

One implication of the model developed here is that it suggests that humans are sensitive to very subtle semantic relationships. Models for predicting human judgments that search fewer than 4500 neighbors of the source and target words perform worse than the model that searches that large neighborhood. In Kintsch's original terms, activation must spread widely for metaphors.

Behavioral study

The 5.5.4500 model we developed predicts human best/worst judgments to a significant extent, at least when good and bad metaphors are directly contrasted (our original criterion for assessing the parameter sets). We conducted an experiment that was intended to cross-validate the model by examining whether its estimates were predictive of human behaviors other than judgment: in particular, whether it could predict decision choices and times in deciding which of two metaphors was better.

Method

Stimuli

The stimuli were the 622 metaphors introduced above.

Procedure

The experiments were conducted in three small testing rooms using custom-written software running on three identical Apple G4 Macintosh Minis running Apple OS 10.15, attached to identical 17.1-in. monitors. The screen resolutions were set to 1280 × 1024 pixels.

After explaining what a metaphor is, participants were shown these instructions, which were also explained by an RA:

We are interested in your intuitions about what makes a metaphor good.

We will show you two metaphors, one above and one below. We ask you to let us know which one seems better to you. If you think the one on the top is better, please press the “k” key. If you think the one below is better, press the “m” key. Note that the “k” key is above the “m” key on the keyboard, so you will press the top key when the phrase on top is better, and the bottom key when the phrase on the bottom is better.

Sometimes you may feel that both metaphors are equally acceptable. In that case, you can press either key randomly.

Please use the first and second fingers of your right hand to press the keys and make your decision as quickly as you can, without sacrificing accuracy.

Following this, an example trial was shown, with the RA still in the testing room. If the participant had no further questions, they were left alone to complete the experiment.

Each participant saw 200 trials, with a self-timed break after half of them had been viewed. On each trial, two randomly selected metaphors were shown in black 24-point Times font, centered on the screen 30% and 50% down from

the top, against a white background. No metaphor was shown more than once to any participant.

Participants

Participants were 50 self-reported native English speakers (35 [70%] female; 43 [86%] self-reported right-handers) who participated after giving written informed consent in return for partial course credit. They had an average [SD] age of 19.7 [2.7] years and an average [SD] of 14.0 [1.2] years of education. Although all reported to have learned English before the age of 5 years, 19 [38%] also reported speaking at least one other language.

Results: RT

The data were analyzed using R (R Core Team, 2020). We removed two participants whose responses were at chance with respect to the best/worst ratings ($p > 0.06$ by exact binomial probability), on the presumption that they may have been responding randomly. To identify outliers, we normalized the RTs to the length of the metaphors by dividing by the sum of the lengths of the source and target words in both metaphors. No participants had an average RT more than three standard deviations from the participant average. We removed 363 trials (3.6% of all data) with RTs that were less than 50 ms per char and 171 trials (1.7%) with RTs that were over 500 ms per char. We also removed 1090 trials (10.9%) that contained a word whose frequency did not appear in our frequency dictionary (Shaoul & Westbury, 2006).

We analyzed the remaining data using linear mixed-effects models.

The model development is summarized in Table 3. The final model is shown in Table 4. Adding the 5.5.4500 estimates to a model that included length and frequency of the target and source words reduced the Akaike information criterion (AIC) values by 24, suggesting that the model with the estimates is much more likely to minimize information loss than the same model without those estimates. The model estimates for each word did not interact. We used the MuMIn R library (Bartón, 2019) to estimate the total variance accounted for by the model. The conditional R^2 (which includes fixed and random effects) was 0.31. The marginal R^2 (attributable to fixed effects only) was 0.031. As shown in Fig. 5, participants are faster to make a goodness decision when the 5.5.4500 estimates for either metaphor were higher (i.e., when those estimates suggested that a metaphor was estimated to be a better metaphor) than when those estimates were lower.

To put this model into context against the “gold standard,” we also predicted RT using the human best/worst judgments (Table 4). This model is not intended to add to our understanding of metaphor processing (since it merely correlates two unknowns; see discussion in Westbury, 2016) but rather to

Table 3 LME model development for predicting pairwise metaphor goodness judgment RTs. All frequencies were logged and all predictors were normalized. Numbers following a dash at the end of predictors refer to metaphor 1 or metaphor 2. The main effects of interest (for human goodness ratings and their 5.5.4500 model estimates) are below the black line

Name	Model specification	AIC	AIC difference	Note
M1	(1 SubjectID)	146515	N/A	Base model
M2	M1 + (1 ORDER)	146381	−134	Accepted
M3	M2 + FREQ-SOURCE-1 + FREQ-TARGET-1 + FREQ-SOURCE-2 + FREQ-TARGET-2	146124	−257	Accepted
M4	M3 + LENGTH-SOURCE-1	146089	−35	Accepted; FREQ-SOURCE-1 dropped out
M5	M4 + LENGTH-TARGET-1	146036	−53	Accepted
M6	M4 + LENGTH-SOURCE-2	146026	−10	Rejected; no convergence.
M7	M6 + LENGTH-TARGET-2	146011	−15	Accepted; FREQ-TARGET-2 dropped out
M8	M7 + GOODNESS-1 * GOODNESS-2	145941	−70	Accepted
M9	M7 + 5.5.4500-1 + 5.5.4500-2	145987	−24	Accepted; Interaction did not enter.
M10	M8 + 5.5.4500-1 + 5.5.4500-2	N/A	N/A	Model reduced to M8.

help assess the quality of the 5.5.4500 model. The 5.5.4500 estimates did not perform as well as the human goodness judgments themselves. The human judgments of both metaphors entered the model in interaction, reducing the AIC value by 70, making the model with the human goodness judgments a much better model than the model with the estimates. The conditional R^2 was 0.32. The marginal R^2 (attributable to fixed effects only) was 0.035.

To see whether the estimates and the human judgments accounted for different portions of the variance, we added the human judgments to the final model that included the 5.5.4500 model estimates. The model estimates dropped out of the model, suggesting that those estimates account for no additional variance beyond that accounted for by the judgments.

Results: Agreement with human judgments

We used binomial linear mixed-effects models to assess agreement with the human judgments (i.e., by defining “correct” as choosing the metaphor that was rated more highly by

Table 4 Final model for predicting pairwise metaphor judgment RTs

Predictor	Estimate	SE	t
(Intercept)	4464.68	179.65	24.85
M1.5.5.4500	−56.52	22.65	−2.5
M2.5.5.4500	−51.96	22.75	−2.28
M1.TARGET-LENGTH	194.23	28.81	6.74
M2.TARGET-LENGTH	178.17	22.72	7.84
M1.SOURCE-LENGTH	173.93	22.75	7.65
M2.SOURCE-FREQ	−160.82	22.85	−7.04
M1.TARGET-FREQ	−108.06	28.82	−3.75

the best/worst judgments). Neither the 5.5.4500 estimates nor the human best/worst judgments entered the model. We also modeled the effect of both the difference and the absolute difference between the estimates for the first and second metaphor, with the same null results.

These null results reflect the fairly poor agreement between the best/worst judgments and the pairwise decisions. The pairwise decisions agreed with the best/worst judgments just 65.3% of the time. This reflects the fact that most metaphors have a middle-range rating, so many metaphors are randomly paired with metaphors that are very close in their judged goodness. We modeled only the subset of 1049 pairs with an absolute difference in best/worst rating of 0.3. In this subset, the agreement between pairwise decision and best/worst judgment was 83.7%. The 5.5.4500 estimates did not enter the model for this subset of data either.

Discussion

The most important finding from this experiment is the finding that faster decisions are made for metaphors that have a higher 5.5.4500 estimate. A high 5.5.4500 estimate suggests that source and target in metaphors have a strong semantic relationship, and therefore need only shallow processing to determine relatedness.

General discussion

Metaphor comprehension has generally been modeled with two classes of model, comparison and categorization models (Gentner & Bowdle, 2001). The comparison model suggests that metaphor comprehension is achieved by finding common properties between conceptual representations of the source and target. Under the categorization model (Glucksberg,

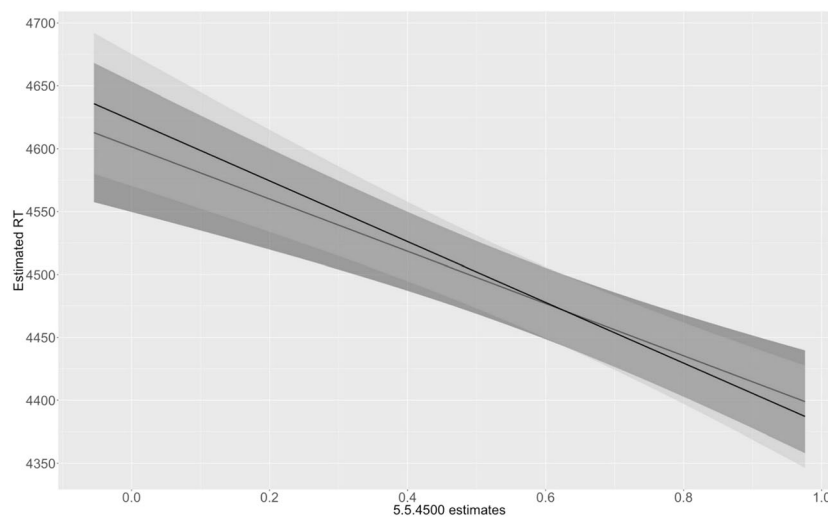


Fig. 5 Human decision time about pairwise metaphor goodness, as a function of the 5.5.4500 model estimates for the first metaphor (darker 95% confidence intervals) or the second metaphor (lighter 95%

confidence intervals). Decisions are faster when either metaphor is estimated by the model to be better

2001, 2003) the target is treated as subordinate to the source, in the sense that the target belongs to an abstract category that is defined by the source. Kintsch's (2000) predication algorithm was noted by Glucksberg (2003) to be consistent with the categorization model. In enhancing Kintsch's model, our best-parameterized model emphasizes the target as well as the source. This makes it a hybrid model of metaphor comprehension, consistent with both the categorization and the comparison models. Our model finds common semantic features through the shared neighbors of the target and source but uses those features to measure the goodness of fit between an abstract category that is defined by adjusting the vectors of both the target and source word vectors.

We have made two main contributions.

One is the development and public release of a set of 622 novel "x is a y" metaphors that have been ranked ordered for metaphoric quality using human best/worst judgments. The rank ordering that comes out of best/worst judgments is superior to human scale ratings that have been released for smaller sets of metaphors (e.g., Campbell & Raney, 2016; Katz et al., 1988; Roncero & de Almeida, 2015) because they allow quantified differentiation of the judged quality any two metaphors, rather than putting the metaphors into a small number of bins of goodness, often with large variance, as rating scales do. Because we had many more items that were not preselected to be good metaphors, because best/worst rating naturally forces the human judges to directly contrast individual items as they make each decision, and because the procedure contrasts every item against every other item, we were able to get a clearer measure of which metaphors are judged by humans to be better than which.

Our second contribution has been to undertake a systematic analysis of the co-occurrence model implementation of the predication algorithm originally proposed by Kintsch (2000

as a measure of metaphor comprehension difficulty. Such an analysis has not been undertaken before. The original model was randomly parameterized, and there was little discussion of how the parameter values were chosen, making it hard for researchers to assess its utility. One finding from the current study was that the original developer of the model was correct both in guessing that $k2 = 5$, and in not worrying too much about that guess. We have demonstrated that the model is insensitive to the values of $k1$ or $k2$ (although it performs better at estimating human judgments of metaphor quality when they are both > 0). Although our use of a different co-occurrence model and a much larger corpus than Kintsch means that direct comparisons are not very meaningful, we did find that searching a large set of $m = 4500$ neighbors for overlapping neighbors was superior to his suggested values for m between 500 and 1500 for estimating metaphor goodness. However, even the best-parameterized model performed poorly at distinguishing metaphors judged good by humans from other metaphors.

We have also considered whether the best-parameterized model was able to perform well at a more difficult problem than the problem of ranking human-selected metaphors, the problem of distinguishing a metaphor from a non-metaphor. Unfortunately, it was not. Research on modeling metaphor comprehension has oversimplified the problem by focusing only on preselected metaphors, ignoring the difficult problem of how to pre-select them. Further work will be required to solve this problem.

The models we considered here are the simplest vector models (see Reid & Katz, 2018), focusing on a simple abstracted representation of the semantics of the target and source words by considering how the number of overlapping neighbors moves the relevant vector representations close to each other. They do not consider the semantics of those words

themselves. Future work may find it fruitful to consider: What is it about the semantics of source and target words that makes them suitable for use in a metaphorical comparison?

Research involving semantic space can be varied in different ways that adds to the metaphor research. The results of the models developed in this study are only restricted to a single specific semantic space. Future research may consider incorporating this detail into modeling metaphor quality to make it more applicable and account for more complexity by considering other semantic spaces.

Acknowledgements This work was made possible by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

- Al-Azary, H., & Buchanan, L. (2017). Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory and Cognition*, 45(2), 296–307. <https://doi.org/10.3758/s13421-016-0650-7>.
- Altszyler, E., Sigman, M., Ribeiro, S., & Slezak, D. F. (2016). Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database, 1–14. <https://doi.org/10.1016/j.concog.2017.09.004>.
- Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, 1, 238–247. <https://doi.org/10.3115/v1/p14-1023>.
- Bartón, K. (2019). MuMIn. R package version 1.43.14 <https://CRAN.R-project.org/package=MuMIn>
- Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112(1), 193–216. <https://doi.org/10.1037/0033-295X.112.1.193>.
- Campbell, S. J., & Raney, G. E. (2016). A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior Research Methods*, 48(1), 330–340.
- Gentner, D., & Bowdle, B. F. (2001). Convention, form, and figurative language processing. *Metaphor and Symbol*, 16(3–4), 223–247.
- Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. New York: Oxford University Press.
- 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(1), 3.
- Hollis, G. (2018). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior Research Methods*, 50(2), 711–729.
- Hollis, G., & Westbury, C. (2018). When is best-worst best? A comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior Research Methods*, 50(1), 115–133.
- 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(4), 191–214.
- Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 7(2), 257–266.
- Kintsch, W. (2001). Predication. *Cognitive Science*, 25(2), 173–202. [https://doi.org/10.1016/S0364-0213\(01\)00034-9](https://doi.org/10.1016/S0364-0213(01)00034-9).
- Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol*, 17(4), 249–262.
- Kiritchenko, S., & Mohammad, S. M. (2016). Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, 811–817. <https://doi.org/10.18653/v1/n16-1095>.
- Kövecses, Z. (2017). Levels of metaphor. *Cognitive Linguistics*, 28(2). <https://doi.org/10.1515/cog-2016-0052>.
- 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*, 1(2), 211. <https://doi.org/10.1037/0033-295X.104.2.211>.
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems (NIPS)*, 2177–2185.
- Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, 92, 57–78. <https://doi.org/10.1016/j.jml.2016.04.001>.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Reid, N. J., & Katz, A. N. (2018). Vector Space Applications in Metaphor Comprehension. *Metaphor and Symbol*, 33(4), 280–294. <https://doi.org/10.1080/10926488.2018.1549840>.
- 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(3), 800–812. <https://doi.org/10.3758/s13428-014-0502-y>.
- Shaoul, C., & Westbury C. (2006) USNET Orthographic Frequencies for 1,618,598 types. (2005–2006) Edmonton, AB: University of Alberta (downloaded from <http://www.psych.ualberta.ca/~westburylab/downloads/wlallfreq.download.html>)
- Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge Handbook of Metaphor and Thought*, 84, 105. <https://doi.org/10.1017/CBO9780511816802.007>.
- Terai, A., & Nakagawa, M. (2012). A corpus-based computational model of metaphor understanding consisting of two processes. *Cognitive Systems Research*, 19–20, 30–38. <https://doi.org/10.1016/j.cogsys.2012.03.001>.
- Utsumi, A. (2011). Computational Exploration of Metaphor Comprehension Processes Using a Semantic Space Model. *Cognitive Science*, 35(2), 251–296. <https://doi.org/10.1111/j.1551-6709.2010.01144.x>.
- Veale, T., Shutova, E., & Klebanov, B. B. (2016). Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1), 1–160. <https://doi.org/10.2200/S00694ED1V01Y201601HLT031>.
- Westbury, C. (2016). Pay no attention to that man behind the curtain: Explaining semantics without semantics. *The Mental Lexicon*, 11(3), 350–374. <https://doi.org/10.1075/ml.11.3.02wes>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.