



Cognitive Science 43 (2019) e12779

© 2019 The Authors. Cognitive Science published by Wiley Periodicals, Inc. on behalf of Cognitive Science Society. All rights reserved.

ISSN: 1551-6709 online

DOI: 10.1111/cogs.12779

How Polysemy Affects Concreteness Ratings: The Case of Metaphor

W. Gudrun Reijnerse,^a Christian Burgers,^b Marianna Bolognesi,^c
Tina Krennmayr^d

^a*Centre for Language Studies, Radboud University Nijmegen*

^b*Department of Communication Science, Vrije Universiteit Amsterdam*

^c*Faculty of Medieval and Modern Languages, University of Oxford*

^d*Department of Language, Literature, and Communication, Vrije Universiteit Amsterdam*

Received 4 September 2018; received in revised form 24 June 2019; accepted 1 July 2019

Abstract

Concreteness ratings are frequently used in a variety of disciplines to operationalize differences between concrete and abstract words and concepts. However, most ratings studies present items in isolation, thereby overlooking the potential polysemy of words. Consequently, ratings for polysemous words may be conflated, causing a threat to the validity of concreteness-ratings studies. This is particularly relevant to metaphorical words, which typically describe something abstract in terms of something more concrete. To investigate whether perceived concreteness ratings differ for metaphorical versus non-metaphorical word meanings, we obtained concreteness ratings for 96 English nouns from 230 participants. Results show that nouns are perceived as less concrete when a metaphorical (versus non-metaphorical) meaning is triggered. We thus recommend taking metaphoricity into account in future concreteness-ratings studies to further improve the quality and reliability of such studies, as well as the consistency of the empirical studies that rely on these ratings.

Keywords: Metaphor; Concreteness; Familiarity; Ratings; Norming data

Correspondence should be sent to W. Gudrun Reijnerse, Centre for Language Studies, Radboud University Nijmegen, PO Box 9103, 6500 HD Nijmegen, The Netherlands. E-mail: g.reijnerse@let.ru.nl

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

1. Concreteness and metaphoricity

Concreteness ratings (e.g., Brysbaert, Warriner, & Kuperman, 2014; Spreen & Schultz, 1966; Toglia & Battig, 1978) are a popular approach used in a variety of disciplines to operationalize the difference between concrete and abstract words and concepts words and concepts, including in cognitive science (Connell & Lynott, 2012; Siakaluk, Pexman, Sears, Wilson, Locheed, & Owen, 2008), (experimental) psychology (Kaushanskaya & Rehtzigel, 2012; Pexman, Heard, Lloyd, & Yap, 2017), psycholinguistics (Ferreira, Göbel, Hymers, & Ellis, 2015; Van Rensbergen, Storms, & De Deyne, 2015), and cognitive linguistics (Dunn, 2015).

Studies that collect concreteness ratings typically define the distinction between concrete and abstract words in terms of sense experience in that “[a]ny word that refers to objects, materials or persons should receive a high concreteness rating [and] any word that refers to an abstract concept that cannot be experienced by the senses should receive a low concreteness rating” (Spreen & Schultz, 1966, p. 460). Thereby, words like “banana” are rated as relatively concrete while words like “idea” are rated as relatively abstract (Brysbaert et al., 2014). Empirical research has shown that relatively concrete words are more easily processed than abstract words in a variety of tasks, including word recognition (e.g., Strain, Patterson, & Seidenberg, 1995; but see Brysbaert, Mandra, McCormick, & Keuleers, 2018, and Brysbaert, Stevens, Mandra, & Keuleers, 2016, for contrasting findings), memory tasks (e.g., Jefferies, Frankish, & Lambon Ralph, 2006), comprehension tasks (e.g., Kounios & Holcomb, 1994), and production tasks (e.g., Wiemer-Hastings & Xu, 2005).

Validation of concreteness ratings consistently shows high levels of correlation across studies (typically around $r = .90$; e.g., Brysbaert et al., 2014; Friendly, Franklin, Hoffman, & Rubin, 1982; Gilhooly & Logie, 1980a). Yet, what seems to be overlooked in the majority of these studies is that many words are homonymous or polysemous, and hence have multiple meanings (see, e.g., Britton, 1978). Because words are generally presented in isolation (without definition or context), participants may rate different meanings, “possibly leading to spurious average values that do not represent anyone subject’s judgment” (Gilhooly & Logie, 1980b, p. 428).

The issue of multiple meanings and concreteness ratings was raised in the early 1980s (e.g., Gilhooly & Logie, 1980a; Toglia & Battig, 1978; see also Theijssen, van Halteren, Boves, & Oostijk, 2011), when research showed that at least 30% of words in English texts have multiple meanings (e.g., Britton, 1978). However, this idea never gained ground in concreteness-ratings studies (exceptions are Gilhooly & Logie, 1980b; Scott, Keitel, Becirspahic, Yao, & Serono, 2019; Wollen, Cox, Coahran, Shea, & Kirby, 1980). When different meanings of the same word refer to equally concrete (or abstract) entities, this may not affect the word’s concreteness rating (Theijssen et al., 2011). A case in point is the noun “organ,” which can refer to a body part ($M = 5.39$, $SD = 1.59$; Gilhooly & Logie, 1980b) or to a musical instrument ($M = 5.64$, $SD = 1.65$; *ibid*). In both cases, “organ” refers to an object that can be experienced by the senses.

However, the issue of multiple meanings becomes a more serious threat to the validity of concreteness ratings when one meaning is more concrete than another. In this report, we specifically connect this with metaphoricity, because metaphors typically align an abstract meaning to a more concrete one within the same word (Lakoff & Johnson, 1980). A case in point is the noun “device”, which has a mean concreteness score of 4.49 ($SD = 1.65$; Gilhooly & Logie, 1980a) when presented in isolation. However, “device”, can refer to a piece of equipment (non-metaphorical, relatively concrete: $M = 5.08$, $SD = 1.91$; Gilhooly & Logie, 1980b) or to a method (metaphorical, relatively abstract: $M = 3.43$, $SD = 1.71$; Gilhooly & Logie, 1980b).

Furthermore, large-scale corpus research has shown that metaphors are ubiquitous in language, given that around 1 in 8 words (i.e., 12,6% of all words) across different genres is used metaphorically (Steen et al., 2010). Therefore, concreteness ratings for words with a concrete and a more abstract meaning may be conflated (Theijssen et al., 2011; see Gilhooly & Logie, 1980b; Toglia & Battig, 1978).

This may (in part) be resolved by explicitly alerting participants to the intended meaning of items to be rated. To the best of our knowledge, however, metaphoricity has not been taken into account in concreteness-ratings studies. Therefore, this report investigates how triggering metaphorical versus non-metaphorical meanings of words affects perceived concreteness scores.

2. Method

2.1. Materials

We selected 96 nouns from the VU Amsterdam Metaphor Corpus (Steen et al., 2010).¹ These were part of a metaphorical domain construction (Sullivan, 2013), consisting of a metaphorical noun modified by a non-metaphorical classifying adjective (e.g., “economic growth”, “financial crash”; Reijniere, Burgers, Krennmayr, & Steen, 2018). Thereby, these nouns are polysemous.

Based on the available sense descriptions for British English in the *Macmillan English Dictionary online*, two definitions were formulated for each noun: one for a metaphorical meaning and one for a non-metaphorical meaning (see online Appendix A at: osf.io/xmjzf). For the noun “growth”, for instance, the metaphorical definition was “an increase in the success of a business or a country’s economy, or in the amount of money invested in them” (*Macmillan* sense description 2. Hereafter: MM2, etc.). The non-metaphorical definition was “an increase in the size, number, or development of a living thing” (MM3).²

2.2. Design and instrumentation

We randomly split the nouns into two sets of 48 nouns (Set A, Set B). Participants rated all 96 nouns in the dataset once, either with metaphorical definitions for the nouns

in Set A, and non-metaphorical definitions for the nouns in Set B, or vice versa. Nouns were presented in random order to prevent order effects. To ensure that metaphorical and non-metaphorical definitions were equally familiar to participants, we also measured perceived familiarity (Gilhooly & Logie, 1980b; Theijssen et al., 2011).

Ratings were made on a 7-point Likert scale ranging from “*very unfamiliar*” to “*very familiar*” for familiarity and from “*very abstract*” to “*very concrete*” for concreteness. Unfamiliar words were defined as “words that you have not seen, heard, or used very often” and familiar words as “words that you have seen, heard, or used very often” (based on Noble, 1953). We defined abstract words as “refer[ring] to concepts that you cannot experience with your senses” and concrete words as “refer[ring] to concepts that you can see, hear, feel, smell, or taste” (based on Spreen & Schultz, 1966). For counterbalancing purposes, the order of ratings (concreteness or familiarity first) was reversed for half of the participants.

2.3. Procedure

Data were collected online through Qualtrics. Participants first provided informed consent, after which they received instructions about the rating task. Participants were informed that the entire range of options could be used, that there were no correct or incorrect answers, and that they should give their true judgment, work fairly quickly—but not carelessly—and carefully read the definitions of the words before providing their judgments. Finally, each participant received the same practice question in which they were asked to rate the noun “house”, defined as “a building for living in, usually where only one family lives” (MM1) for both concreteness and familiarity (or vice versa). After this practice question, the survey started.

All 96 nouns were displayed in random order, with one noun per page. As a reminder, the definitions of concreteness/abstractness and (un)familiarity were displayed at the top of each page. After rating all nouns on familiarity and concreteness, participants were asked about general demographics (age, gender, education level, nationality, native language), thanked, paid USD2 for completing the survey, and debriefed. No further items were measured. On average, completing the survey took 28 min and 17 s.

2.4. Participants

Participants were sampled through Amazon’s Mechanical Turk on July 15–16, 2015. The MTurk HIT Approval Rate was set to 95% to ensure high-quality work, and only MTurk workers located in the United States could participate. In earlier work on concreteness ratings, scholars strived to obtain at least 30 participants rating each individual word (Brysbaert et al., 2014). To make sure that an extension of previous findings has sufficient statistical power, it has been recommended to include at least 2.5 times the number of original observations (Simonsohn, 2015), which means that we needed at least 75 observations per word per group (non-metaphor, metaphor), leading to a total of 150 participants. However, when participants are randomly assigned to conditions, some may

drop out, leading to an unequal number of participants per groups. We therefore strove for a minimum sample size of 200 completes.

In total, 236 participants completed the survey. We excluded participants without the US nationality or English as their first language. A total of 230 participants met the inclusion criteria ($M_{\text{age}} = 35.48$, $SD_{\text{age}} = 11.34$, range: 18–69; 47% female; 70.4% completed an (under)graduate degree).³

2.5. Data analysis

Data were analyzed using R (R Core Team, 2017). The R packages *lme4* (Bates, Maechler, Bolker, & Walker, 2015) and *sjPlot* (Lüdtke, 2017) were used to fit a linear mixed effects model to the perceived concreteness ratings, with definition (non-metaphorical versus metaphorical) as a fixed independent variable, and order (concreteness or familiarity first) as fixed control variable. As random effects, intercepts for participants and words were first included (Model 1), after which the by-participant and by-word random slopes for the effect of metaphor were added (Model 2; Winter, 2013). Here, p -values were obtained by means of likelihood ratio tests of the full model (with the effect) against the reduced model (without the effect). Data and data-analytical procedures of the analyses reported in this paper are publicly accessible at osf.io/xmjzf.

3. Results

3.1. Main analyses

Table 1 displays the model for familiarity. The fitted model indicated a significant effect of word meaning (non-metaphorical versus metaphorical) on perceived familiarity of words. The mean perceived familiarity score was on average 0.09 points \pm 0.05 (SE) lower for metaphorical (vs non-metaphorical) word meanings. We also found small effects for order and the interaction of metaphor and order. Nevertheless, it should also be noted that the actual effect is small, given that we talk about a shift of 0.09 points on a 7-point scale. It could be the case that these are ceiling effects as both the metaphorical ($M = 6.16$, $SD = 0.70$) and non-metaphorical meanings ($M = 6.29$, $SD = 0.65$) were relatively familiar.

Table 2 displays the model for concreteness. The fitted model indicated a significant effect of word meaning (non-metaphorical vs. metaphorical) on perceived concreteness: The mean perceived concreteness score was on average 1.62 points \pm 0.12 (SE) lower for metaphorical (vs. non-metaphorical) word meanings. The effect of metaphor on concreteness is much larger than on familiarity, as reflected in average concreteness scores for metaphorical ($M = 3.32$, $SD = 0.89$) and non-metaphorical word meanings ($M = 4.96$, $SD = 0.66$). We also found a trend for order. More important, we found a positive relation between familiarity and concreteness, with an increase in perceived familiarity with 1 point being associated with an average increase of concreteness by 0.15 points \pm 0.01

Table 1
Fixed effects estimates and variance-covariance estimates for models for the predictors of familiarity

	Model 1			Model 2			Model 3			Model 4		
	B	95% CI	p	B	95% CI	p	B	95% CI	p	B	95% CI	p
Fixed parts												
Intercept	6.30	6.19–6.41	<.001	6.30	6.19–6.41	<.001	6.41	6.27–6.54	<.001	6.42	6.28–6.55	<.001
Metaphor	-0.14	-0.16 to -0.11	<.001	-0.14	-0.22 to -0.05	.003	-0.14	-0.22 to -0.05	.003	-0.09	-0.19 to -0.00	.044
Order							-0.24	-0.41 to -0.08	.004	-0.26	-0.43 to -0.10	.002
Metaphor* Order										-0.09	-0.14 to -0.04	<.001
Random parts												
σ^2	0.736			0.687			0.687			0.687		
$\tau_{00,participants}$	0.441			0.409			0.392			0.392		
$\tau_{00,words}$	0.119			0.137			0.137			0.137		
ICC _{responseld}	0.340			0.331			0.322			0.322		
ICC _{words}	0.092			0.111			0.113			0.113		
Observations	22080			22080			22080			22080		
R^2/Ω^2	.442/.442			.483/.482			.483/.482			.482/.482		
Evaluation												
Log likelihood	-28,584			-28,016			-28,012			-28,007		
χ^2 deviance	137.07***			1,136.67***			7.93**			10.81**		
df deviance	1			4			1			1		

Note. Evaluation of model 1 compares the model with metaphor to the null model with only random effects for participants and words.
* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 2
Fixed effects estimates and variance-covariance estimates for models for the predictors of concreteness

	Model 1			Model 2			Model 3			Model 4		
	B	95% CI	p	B	95% CI	p	B	95% CI	p	B	95% CI	p
Fixed parts												
Intercept	4.97	4.75–5.18	<.001	4.97	4.71–5.22	<.001	5.05	4.79–5.32	<.001	4.09	3.79–4.39	<.001
Metaphor	-1.64	-1.68 to -1.60	<.001	-1.64	-1.88 to -1.41	<.001	-1.64	-1.88 to -1.41	<.001	-1.62	-1.85 to -1.39	<.001
Order							-0.19	-0.35 to -0.02	.029	-0.14	-0.31–0.02	.089
Familiarity										0.15	0.13–0.17	<.001
Random parts												
σ^2	2.439			2.037			2.037			2.023		
$\tau_{00,participants}$	0.452			0.400			0.393			0.367		
$\tau_{00,words}$	0.940			1.452			1.452			1.413		
ICC-ResponseId	0.118			0.103			0.101			0.096		
ICC-words	0.245			0.373			0.374			0.372		
Observations	22,080			22,080			22,080			22,080		
R^2/Ω_0^2	.466/.466			.560/.560			.560/.560			.563/.563		
Evaluation												
Log likelihood	-41,725			-40,135			-40,132			-40,049		
χ^2 deviance	5,384.40 ^{***}			3,180.00 ^{***}			4.70 [*]			166.95 ^{***}		
df deviance	1			4			1			1		

Note. Evaluation of model 1 compares the model with metaphor to the null model with only random effects for participants and words. Adding the two-way interaction of metaphor*order did not improve model fit; this interaction is thus not included in the reported models. * $p < .05$; ** $p < .01$; *** $p < .001$.

(*SE*). Nevertheless, we found the effects of metaphor both without (Model 3) and with controlling for familiarity (Model 4).

Mean perceived familiarity and concreteness ratings for individual nouns are given in online Appendices B and C, respectively, at osf.io/xmjzf.

3.2. *Additional analyses*

Our main results raise questions about the potential dominance of metaphorical or non-metaphorical meanings in existing concreteness-ratings databases in which words are presented in isolation. To address this issue, we ran additional analyses to compare our findings to those of Brysbaert et al. (2014). In Brysbaert et al.'s database, data are available for 94 out of the 96 nouns from our sample.⁴

Results from these additional analyses showed that the concreteness scores for the metaphorical meanings of the nouns in our dataset are statistically significantly lower (i.e., more abstract) than those of Brysbaert et al. (2014) for 62 out of the 94 nouns (66%). By contrast, the scores for the non-metaphorical meanings in our dataset are statistically significantly higher (i.e., more concrete) than those of Brysbaert et al. (2014) for 28 out of 94 nouns (29.8%). In addition, the non-metaphorical meanings in our dataset are significantly lower (i.e., more abstract) than those of Brysbaert et al. (2014) for 8 out of 94 nouns (8.5%). For the remaining nouns, no significant difference between the scores in our dataset and those in Brysbaert et al.'s was found (i.e., the 95% CI for these nouns included zero).

Details of these analyses are available in Appendices D (tables) and E (forest plots) at osf.io/xmjzf.

4. **Conclusion and discussion**

Previous studies have suggested that polysemy may affect concreteness ratings (e.g., Gilhooly & Logie, 1980b; Theijssen et al., 2011; Toglia & Battig, 1978). We show that this may be particularly true for words that have a metaphorical meaning, because metaphors typically describe something abstract in terms of something more concrete (e.g., Lakoff & Johnson, 1980; see also Veale, Shutova, & Klebanov, 2016). Results of our analyses showed that metaphoricity decreased perceived concreteness of nouns, in that nouns are perceived as less concrete when a metaphorical (vs. non-metaphorical) meaning is triggered. We also demonstrate that this effect of metaphoricity remains when controlling for the (small) difference in familiarity between metaphorical and non-metaphorical meanings. In addition, follow-up analyses in which we examined which meanings (metaphorical or non-metaphorical) are more dominant in existing concreteness-ratings studies suggest that the non-metaphorical meaning of the nouns in our dataset may be slightly more dominant in standard concreteness-ratings studies than their metaphorical counterparts.

More generally, our results are a proof of concept, emphasizing the importance of taking metaphoricity into account in future concreteness-ratings studies. We recommend that such future studies use larger datasets in order to provide more concreteness ratings for metaphorical versus non-metaphorical meanings of nouns. This can be done by providing definitions (as in the present study) or by presenting words in context (Theijssen et al., 2011).

While this study mainly focused on the impact of metaphoricity on concreteness ratings of nouns, future research could consider looking at other word characteristics (e.g., homonymy), other word classes (e.g., verbs), and/or other rating variables (e.g., aptness; Thibodeau, Sikos & Durgin, 2018). In a recent study, Scott et al. (2019) collected ratings for isolated and disambiguated words (e.g., isolated: “shell”, disambiguated: ‘shell (sea)’, ‘shell (military)’) on nine scales, including concreteness, arousal, and valence. Their preliminary analyses suggest that ratings differ when ambiguous versus disambiguated word meanings are presented.

Taken together, such research can further improve the quality and reliability of word-ratings studies, as well as the consistency of the empirical studies relying on this approach to design stimuli and conduct experiments.

Acknowledgments

The contribution of W. Gudrun Reijniere was supported by grant no. 322-89-006, awarded by the Netherlands Organization for Scientific Research (NWO), www.nwo.nl. The authors thank Dr. Stefan Frank for his assistance with setting up the analyses in R.

Notes

1. See <http://ota.ahds.ac.uk/headers/2541.xml> (Oxford Text Archive).
2. For seven nouns in our sample, no metaphorical definition was present in the dictionary (atom, capital, doodling, household, millipede, terrain, and thicket). For these, we used additional sources such as online encyclopedias and sense descriptions for related words to establish a metaphorical definition.
3. One hundred thirteen participants rated the metaphorical definitions of the nouns in Set A and the non-metaphorical definitions of the nouns in Set B; 117 participants rated the non-metaphorical definitions of the nouns in Set A and the metaphorical definitions of the nouns in Set B.
4. No scores are available for the nouns ‘malaise’ and ‘spin-off’ in Brysbaert et al.’s (2014) database, and these two nouns are therefore not part of this additional analysis. To enable comparison between our scores and those of Brysbaert et al. (2014), we standardized the Brysbaert scores by transforming their 5-point scale into a 7-point scale. Then, we calculated Cohen’s d to examine the standardized difference between (1) Brysbaert’s scores and the scores for the metaphorical meaning of the

nouns from our sample, and (2) Brysbaert's scores and the scores for the non-metaphorical meaning of the nouns from our sample.

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>.
- Britton, B. K. (1978). Lexical ambiguity of words used in English text. *Behavior Research Methods and Instrumentation*, *10*(1), 1–7. <https://doi.org/10.3758/BF03205079>.
- Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. <https://doi.org/10.3758/s13428-018-1077-9>.
- Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. <https://doi.org/10.1037/xhp0000159>.
- 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. <https://doi.org/10.3758/s13428-013-0403-5>.
- Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition*, *125*(3), 452–465. <https://doi.org/10.1016/j.cognition.2012.07.010>.
- Dunn, J. (2015). Modeling abstractness and metaphoricity. *Metaphor and Symbol*, *30*(4), 259–289. <https://doi.org/10.1080/10926488.2015.1074801>.
- Ferreira, R. A., Göbel, S. M., Hymers, M., & Ellis, A. W. (2015). The neural correlates of semantic richness: Evidence from an fMRI study of word learning. *Brain and Language*, *143*, 69–80. <https://doi.org/10.1016/j.bandl.2015.02.005>.
- Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*(4), 375–399. <https://doi.org/10.3758/BF03203275>.
- Gilhooly, K. J., & Logie, R. H. (1980a). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395–427. <https://doi.org/10.3758/BF03201693>.
- Gilhooly, K. J., & Logie, R. H. (1980b). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, *12*(4), 428–450. <https://doi.org/10.3758/BF03201694>.
- Jefferies, E., Frankish, C., & Lambon Ralph, M. A. (2006). Lexical and semantic binding in verbal short-term memory. *Journal of Memory and Language*, *54*(1), 81–98. <https://doi.org/10.1016/j.jml.2005.08.001>.
- Kaushanskaya, M., & Rechtzigel, K. (2012). Concreteness effects in bilingual and monolingual word learning. *Psychonomic Bulletin Review*, *19*(5), 935–941. <https://doi.org/10.3758/s13423-012-0271-5>.
- Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning Memory and Cognition*, *20*(4), 804–823. <https://doi.org/10.1037/0278-7393.20.4.804>.
- Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
- Lüdtke, D. (2017). sjPlot: Data visualization for statistics in social science (R package version 2.4.0). Available at <https://CRAN.R-project.org/package=sjPlot>. Accessed April 1, 2018.
- Macmillan English Dictionary online. 2009–2018. Available at <http://www.macmillandictionary.com/>. Accessed December 1, 2017.
- Noble, C. E. (1953). The meaning-familiarity relationship. *Psychological Review*, *60*(2), 89–98. <https://doi.org/10.1037/h0054823>.

- Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: Concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, 49(2), 407–417. <https://doi.org/10.3758/s13428-016-0720-6>.
- R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>. Accessed April 1, 2018.
- Reijnerse, W. G., Burgers, C., Krennmayr, T., & Steen, G. J. (2018). On metaphorical views, dynamite, and doodlings: Functions of domain adjectives in metaphorical domain constructions. *Review of Cognitive Linguistics*, 16(2), 432–455. <https://doi.org/10.1075/rcl.00017.rei>.
- Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Serono, 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>.
- Siakaluk, P. D., Pexman, P. M., Sears, C. R., Wilson, K., Locheed, K., & Owen, W. J. (2008). The benefits of sensorimotor knowledge: Body–object interaction facilitates semantic processing. *Cognitive Science*, 32(3), 591–605. <https://doi.org/10.1080/03640210802035399>.
- Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569. <https://doi.org/10.1177/0956797614567341>.
- Spreen, O., & Schulz, R. W. (1966). Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5), 459–468. [https://doi.org/10.1016/S0022-5371\(66\)80061-0](https://doi.org/10.1016/S0022-5371(66)80061-0).
- Steen, G. J., Dorst, A. G., Herrmann, J. B., Kaal, A. A., Krennmayr, T., & Pasma, T. (2010). *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.
- Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning Memory and Cognition*, 21(5), 1140–1154. <https://doi.org/10.1037/0278-7393.21.5.1140>.
- Sullivan, K. (2013). *Frames and constructions in metaphoric language*. Amsterdam: John Benjamins.
- Theijssen, D., van Halteren, H., Boves, L., & Oostijk, N. (2011). On the difficulty of making concreteness concrete. *Computational Linguistics in the Netherlands Journal*, 1, 61–77.
- Thibodeau, P. H., Sikos, L., & Durgin, F. H. (2018). Are subjective ratings of metaphors a red herring? The big two dimensions of metaphoric sentences. *Behavior Research Methods*, 50(2), 759–772. <https://doi.org/10.3758/s13428-017-0903-9>.
- Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Van Rensbergen, B., Storms, G., & De Deyne, S. (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin Review*, 22(6), 1717–1724. <https://doi.org/10.3758/s13423-015-0832-5>.
- 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>.
- Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science*, 29(5), 719–736. https://doi.org/10.1207/s15516709cog0000_33.
- Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. Available at <http://arxiv.org/pdf/1308.5499.pdf>. Accessed April 1, 2018.
- Wollen, K. A., Cox, S. D., Coahran, M. M., Shea, D. S., & Kirby, R. F. (1980). Frequency of occurrence and concreteness ratings of homograph meanings. *Behavior Research Methods & Instrumentation*, 12(1), 8–15. <https://doi.org/10.3758/BF032083>.