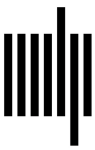
This is a section of doi:10.7551/mitpress/12200.001.0001

The Open Handbook of Linguistic Data Management

Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister

Citation:

The Open Handbook of Linguistic Data Management
Edited by: Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, Lauren B. Collister
DOI: 10.7551/mitpress/12200.001.0001
ISBN (electronic): 9780262366076
Publisher: The MIT Press
Published: 2022



The MIT Press

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

42 Managing Semantic Norms for Cognitive Linguistics, Corpus Linguistics, and Lexicon Studies

Bodo Winter

1 Introduction

Bloomfield (1933:140) famously noted that "the statement of meanings is the weak point in language study." One issue is that meaning is difficult to measure objectively. It is relatively straightforward to quantify linguistic phenomena that have surface manifestations, such as words or particular grammatical constructions. Meaning, however, is much more elusive because it ultimately resides within the language user's head.

This chapter will discuss data management in the domain of semantics. It will be argued that meaning can be fruitfully studied using *norms*. This term is used by psycholinguists when they collect ratings for linguistic items, generally words. As an example of a norm data set, consider the emotional valence norms collected by Warriner, Kuperman, and Brysbaert (2013), who asked hundreds of native English speakers to rate words for how good or bad they are. The word *vacation* received a rating of 8.53 on the nine-point rating scale for these emotional valence norms, which is 3.47 valence points above the mean emotional valence of 5.06. This indicates that the word *vacation* overall appeared to be very positive to the raters in this study. The same word received a rating of 3.14 on Brysbaert and colleagues' five-point concreteness scale (Brysbaert, Warriner, & Kuperman 2014), which is quite close to the average concreteness ($M = 3.03$), indicating that native speakers felt that the word *vacation* was neither particularly concrete nor particularly abstract.

The "norming" of stimuli is standard procedure within psycholinguistics. However, the resultant norms are also increasingly being investigated as an object of study in their own right, or they are used in conjunction with corpora as means of quantifying particular semantic dimensions. This chapter will present two linguistic examples where researchers commonly do *not* use norms to elucidate some of the problems that may arise for norm-less semantics with respect to the reproducibility of these studies (section 2); followed by an overview of some common norm data sets (section 3); and, finally, a discussion of methodological challenges of norm-based research and how norm-based linguistics fits within contemporary efforts to facilitate reproducible research (section 4).

2 Norm-less semantics: Two examples

2.1 Corpus linguistics: Semantic prosody

It is widely known that word meaning depends on context. Take, for example, the verb *to cause*, which, when seen in isolation appears to be a rather neutral term. However, when using a corpus to look at the contexts this term tends to occur in (e.g., via a concordancer), it becomes apparent that by and large, only bad things get caused (Stubbs 2001), which is exemplified by the concordance lines from the Corpus of Contemporary American English (COCA; Davies 2009) shown in table 42.1.

The idea that words consistently occur in certain types of attitudinal or emotional contexts has been dubbed *semantic prosody* in the British tradition of corpus linguistics (Hunston 2007; Louw 1993; Stewart 2010; Whitsitt 2005), and the fact that language users can so greatly mischaracterize the "connotation" of words when they introspect on them in isolation is a major argument for using corpus methods when looking at word meaning. In this field, other headwords that have been studied with respect to semantic prosody include *to set in* (Sinclair 1991) and *utterly* (Louw 1993), both of which tend to occur in negative contexts. As another example, consider the plural form *days*, for which Louw (1993) claims

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

Table 42.1
Ten concordances of *cause* (verb) + noun from the Corpus of Contemporary American English (Davies 2009)

. . . that the flapping of a butterfly's wings can	cause	tornadoes.
. . . because he lacked intent to	cause	injury or property damage.
. . . were not welcome were ones that were only meant to	cause	offense without furthering a dialogue or conversation.
Space is full of all sorts of garbage that can	cause	problems, including some of the stuff we send up . . .
. . . relief against a use of a mark that might	cause	dilution by blurring or dilution by tarnishment . . .
. . . that the manufacturer's use of the mark is likely to	cause	dilution by blurring or tarnishment.
. . . and especially intramedullary instrumentation	cause	bone marrow extravasation and release of fat emboli . . .
. . . for a free Samsung-sponsored concert—but also	cause	crowd concerns.
Interfascicular dissection may additionally	cause	a disruption in the segmental vascular supply . . .
. . . to educate themselves about the factors that	cause	dogs—all dogs—to bite.

that it tends to occur in contexts invoking a sense of nostalgia.

A major methodological issue in the study of semantic prosody is that concordances are usually hand-classified for their emotional connotation, such as whether they are overall positive or negative. As noted by Bednarek (2008:122), it is "difficult to establish objectively" whether something is negative or positive. Similarly, Stewart (2010:91) notes in a critique of semantic prosody research that "one analyst's meat is another analyst's poison."

If the analysis of a semantic prosody relies on subjective introspection without any clear criteria for what makes a positive or negative context, it is strictly speaking not *reproducible*. That is, if the same data set is given to a different researcher (who has not conducted the original study), they may draw different conclusions. Without inter-reliability checks (which are not often conducted in at least some areas of corpus linguistics and cognitive linguistics), linguistic intuitions are by definition not reproducible because they are contingent on one person's beliefs.

Moreover, the cumbersome method of hand-classifying concordances for whether they are positive or negative has also been a constraining factor with respect to the scope of research on semantic prosody, which has been criticized for focusing on the meaning of only a few isolated headwords (Stewart 2010; Whitsitt 2005). This focus on a small set of linguistic items is a natural outgrowth of the fact that classifying concordances by hand is a very time-consuming process.

Norms provide a solution to both of these issues; they allow reproducible research on semantic prosody that is furthermore *scalable*, which affords more generalizable claims that expand beyond isolated examples. While it is a time-consuming and expensive process to collect a norm data set, once it has been established, it can be used for a whole range of applications by different researchers. For example, the above-mentioned emotional valence norms can be used in conjunction with corpora to see whether the words surrounding a given headword in a concordance line are overall positive or negative, which was an approach taken by some researchers (Dilts & Newman 2006; Snefjella & Kuperman 2016; Winter 2016). In these studies, the subjective judgment has been "outsourced" to a norming study. This means that given the same concordances and the same norming data set, different researchers will reach the same conclusions, thus ensuring reproducibility. The scalability of this type of research is demonstrated, for example, by Winter (2016), who showed that taste words have overall more positive semantic prosodies than smell words. Such claims—characterizing hundreds of words rather than a few isolated headwords—are practically impossible when hand-classifying concordances.

Thus, norms provide an opportunity to free the phenomenon of semantic prosody from the shackles of a time-consuming classification process that is difficult to reproduce and that is difficult to extend to larger chunks of the English lexicon.

2.2 Cognitive linguistics: Sensory language and perceptual metaphor

As another example of norm-less linguistics, consider perceptual language, including the study of perception verbs such as *to see* and *to hear* (Evans & Wilkins 2000; Matlock 1989; Sweetser 1990; Viberg 1983), or the study of perceptual adjectives as they occur in metaphorical expressions such as *smooth melody* (Ronga et al. 2012;

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

Strik Lievers 2015; Williams 1976; Winter 2019). As one particular finding in this field, consider the hypothesis that there is a "hierarchy of the senses," following the order touch > taste > smell > sight/sound (Ullmann 1959). This hypothesis was formed based on the observation that sensory adjectives are more likely to extend from "lower" senses to the "higher" senses, as evidenced by expressions such as *smooth melody* (touch to sound) and *rough smell* (touch to smell), as opposed to the ill-formed *squealing feeling* (sound to touch) or *barking taste* (sound to taste).

At the basis of any generalization of sensory language is a categorization of sensory words according to perceptual modality, such as classifying *smooth* as a touch word, or classifying *squealing* as a sound word. Such classifications are often assumed to be self-evident, which may in fact be the case for some isolated examples. However, there are also numerous examples that are hard to classify. For example, how is one to categorize dimension words such as *long* and *thick* according to sensory modality? And how does one deal with highly multisensory words, such as *harsh*, which can be used to describe a *harsh sound*, a *harsh feeling*, or even a *harsh smell* or *harsh taste*? The literature in this field is ripe with examples where different researchers have classified the same words differently. For example, Ronga et al. (2012) mention how the same dimension words are either classified as sight-related (Williams 1976) or touch-related (Popova 2005). These examples show that it is generally not straightforward to classify words according to senses, and criteria for doing so need to be made explicit in order for research on perceptual language to be reproducible.

Luckily, there are by now multiple data sets of sensory norms for English (Lynott & Connell 2009, 2013; Speed & Majid 2017; Winter 2016). These norms have been used to make research on perceptual words more reproducible. For example, Winter (2019) re-analyzed data on perceptual metaphor using norms, showing that the hierarchy of the senses needs to be reinterpreted when using more reproducible research methods. Strik Lievers and Winter (2018) used sensory norms to show that sound concepts are more lexically differentiated in the verbal as opposed to adjectival domain (there is a disproportionate number of sound verbs in English). Winter, Perlman, and Majid (2018) used sensory norms to argue that the English language is overall visually dominant.

3 Norm data sets and examples of norm-based linguistics

There is by now a wealth of linguistic norms available. Many of these norms have so far only been applied to psycholinguistic experiments, which leaves a lot of room for future corpus linguistic research or analyses of the lexicon using norms.

Collecting norms has a long history in the language sciences. The first large norming studies were conducted in the 1950s and 1960s, such as those discussed in *The Measurement of Meaning* (Osgood, Suci, & Tannenbaum 1957). One of the oldest norm data sets that was used for a long time in psycholinguistics was Paivio's, which included concreteness, imageability, and meaningfulness ratings for 925 English words (Paivio, Yuille, & Madigan 1968).

These days, norms are often collected via crowdsourcing platforms in large "megastudies" (Keuleers & Balota 2015), such as the above-mentioned concreteness norms, which have been collected for 40,000 English words using data from over 4,000 participants. Another fruitful semantic dimension for norming has been emotional valence, for which a large norming data set exists for about 14,000 English words (Warriner, Kuperman, & Brysbaert 2013). As mentioned, researchers have begun to apply these norms to the study of semantic prosody, but an even more wide-spread application is the domain of *opinion mining* or *sentiment analysis* in computer science and in industry, where rating data are used to classify texts such as online reviews with respect to whether they are overall positive or negative.

As a result of the collective efforts of hundreds of researchers, there is by now a wealth of freely available data sets available. Some perhaps unexpected dimensions of meanings that have been "normed" include roughness, hardness, and size of touch adjectives (Stadtlander & Murdoch 2000), the color and motion-relatedness of nominal concepts (Medler et al. 2005), or the graspability and pain-relatedness of object terms (Amsel, Urbach, & Kutas 2012). There are also norms for dimensions that are not strictly speaking exclusively semantic, such as the Bochum English Countability Lexicon that includes expert annotator's ratings for whether nouns are mass or count (Kiss et al. 2016).

A lot of norming data sets are traditionally published in the journal *Behavior Research Methods*, although a bewildering amount of norming data is hidden in

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

published studies that are otherwise focused on answering substantive rather than methodological questions. These data sets may be difficult to find. An issue with the massive amount of data that is already available is that many particular data sets don't have a lot of visibility, which also means that researchers within certain subareas of linguistics may not know that norms are an available methodological option for answering their research questions. For example, corpus linguists studying semantic prosody may continue to hand-classify concordances simply because they are unaware that emotional valence norms exist.

Luckily, there are by now a few websites that allow easy access to norm data sets. One of them is the LAB, the Linguistic Annotated Bibliography (Buchanan, Valentine, & Maxwell 2018). Another one is the language goldmine (languagegoldmine.com). These websites facilitate searching for specific norm data sets, and they encourage exploration of the wealth of available data sets that exist.

Besides the above-mentioned norm-based studies on sensory linguistics and semantic prosody, it is worth pointing out a few more studies that have used norms to showcase the utility of this approach. Warriner and Kuperman (2015) used emotional valence norms to test the *Pollyanna hypothesis*, which is the hypothesis that overall, speakers have a prosocial need to talk about positive things more often than about negative things—this hypothesis was supported by looking at both the type and token frequencies of positive words. The English language has more positive words than negative words overall, and positive words are also used more frequently than negative words in various corpora. In another norm-based study, Lupyan and Winter (2018) used concreteness norms to argue that language is much more abstract than is commonly assumed by "embodied" approaches to cognition.

There are also a number of norming studies for figurative language, for example, Katz et al. (1988) normed more than four hundred literary and non-literary metaphors for various semantic dimensions, such as metaphor goodness, comprehensibility, or metaphoric imagery (for a replication, see Campbell & Raney 2016). Littlemore et al. (2018) normed metaphors for goodness and then correlated these norms with ratings from other norm data sets to show that metaphors judged to be "good" exhibit asymmetries in word frequency and emotional valence, but not concreteness.

A by-now quite extensive line of research has used *iconicity norms*. These norms quantify the degree to which participants feel that a word's form resembles its meaning. For example, the English words *hissing*, *click*, *humming*, *gurgle*, *beep*, and *screech* tend to receive high ratings in these norms, whereas for words such as *moss*, *beginning*, *faucet*, *onion*, and *atom*, participants do not feel that the word sounds like what it means (Perry, Perlman, & Lupyan 2015; Perry et al. 2017). Iconicity norms were first collected for signed languages (Caselli et al. 2017; Grote 2013; Vinson et al. 2008), and this idea was subsequently extended to spoken languages, including English and Spanish (Perry, Perlman, & Lupyan 2015). These norms have led to a number of interesting findings. For example, the norms were useful in showing that iconicity disproportionately resides in the perceptual part of the English vocabulary (Sidhu & Pexman 2018; Winter et al. 2017) or that words with more semantic neighbors are less likely to be iconic (Sidhu & Pexman 2018). Iconicity norms have also been used to show that children's language is relatively more iconic compared to adult's language (Perry, Perlman, & Lupyan 2015) and that adults increase the frequency of iconic words when talking to their children (Perry et al. 2017). The iconicity norms have also been used to compare signed and spoken languages (Perlman et al. 2018).

4 Methodological issues and reproducibility in the context of norms

Norms facilitate reproducibility because given the same norm data set, a different researcher can reproduce an analysis, not having to rely on another researcher's subjective evaluation of particular linguistic items. That said, norm data are still subjective, as they rely on native language user judgments. The word *banker*, for example, may be judged to be neutral by some but negative by others. The idea of using a norm data set rather than relying on the single linguist's intuition is that one can benefit from the "wisdom of the crowd" effect, where individual differences in subjective judgment are less influential due to averaging over many people's intuitive responses.

That said, the subjectivity of norms needs to be kept in mind when doing analyses with norm data sets. For example, in the case of iconicity norms, it is not always clear what native language users base their iconicity

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

judgments on. For example, Perry, Perlman, and Lupyan (2015) collected norms on a scale from –5 to +5, with –5 indicating that a word sounds like "the opposite of what it means." Many of the words with negative iconicity scores are questionable, such as *dandelion* (iconicity: –2.8), *silent* (–2.17), and *would* (–2.1). It is not clear why participants felt that these words sounded like the opposite of their respective meanings. As a result of this, when Sidhu and Pexman (2018) used these norms, they made the analytical decision to exclude words at the lower end of the scale.

This example shows that it is important to consider *construct validity* for norming studies, which refers to the question of whether a study measures the construct that one intends it to measure. While it may be theoretically well defined that a word sounds like what it means (such as the onomatopoetic word *beep* mimicking the corresponding sound), it is not so clear what it means for a word to sound the opposite of what it means. As another example of construct validity, consider a norming study conducted by Engelthaler and Hills (2018), who collected ratings for whether words were humorous or not, with words such as *nitwit*, *tinkle*, and *egghead* receiving high ratings, compared to words such as *trauma*, *oxide*, and *cleaver*. Their rating study was word-based, but the literature on humor generally considers humor to be something that arises over sequences of words. Thus, it is not clear whether the humor ratings actually measure that what humor researchers call humor, and the correspondence between the norms and specific theories in humor research is not straightforward. As another example of potentially limited construct validity, consider the fact that sensory norms have also been collected for highly abstract words. What does it *mean* to ask a participant to rate words such as *freedom* or *vulnerability* as relating to sight, touch, sound, taste, or smell? And, can we trust the resulting ratings?

Another issue that stems from the fact that norms are subjective has to do with infrequent and little-known words. Winter (2019) lists several examples in the sensory norm data sets commonly used in psycholinguistics for which participants clearly misunderstood a word, presumably because they did not know its meaning very well. For example, the word *brackish* was classified by participants as predominantly touch-related (Lynott & Connell 2009) even though dictionary definitions list its meaning as "slightly salty." Similarly, the word *clamorous* has an auditory meaning listed in dictionaries, but was rated to be higher in tactile than in auditory strength (Lynott & Connell 2009). Given that both of these words are rather infrequent, it seems likely that word knowledge (and its lack thereof) needs to be taken into account. Analyses should not be performed on words that are not sufficiently known by participants. After all, what did participants rate if they did not know the meaning of these words? To alleviate these concerns, the analyst may want to take data from large-scale word knowledge studies into account (Brysbaert et al. 2016; Keuleers et al. 2015).

Norming studies should also compute *inter-rater reliability statistics* (e.g., Engelthaler & Hills 2018), such as the intra-class coefficient for continuous data. *High inter-rater reliability* means that participants of a norming study agree with each other, and *low inter-rater reliability* means that participants differ in their judgments. Low inter-rater reliability could suggest that participants used different criteria for performing judgments or that the overarching construct is not well defined, with different participants interpreting the instructions differently. Here, it is important to keep in mind that the agreement of raters may not be uniform across all the words from a norm data set. Pollock (2018) brings up the important issue that the standard deviation of ratings is higher for words in the middle ranges of norming studies. For example, it may be clear that *murder* is negative and that *happiness* is positive, which means that different participants rate these words very consistently and the standard deviations across participants are low for these words. But a seemingly more neutral word such as *banker* may be rated positively by some participants and negatively by others. The word would end up with a neutral score, even though some participants may have felt very strongly about this word. Thus, researchers doing norm-based linguistics should keep the standard deviation of norms in mind, with Pollock (2018) recommending that researchers may want to exclude words with very high standard deviations.

The standard deviation of a particular word's ratings may also be higher when words are highly polysemous, which is another factor that needs to be taken into account. Given that norming studies generally present words in isolation, the context may not be enough to disambiguate certain meanings. This invites the possibility that different participants rate different meanings of the

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

same word form. As an example of this, consider the fact that the noun *firm* was rated to be high in tactile strength in Lynott and Connell (2013). This was presumably the case because participants rated the (much more frequent) adjective *firm* (to the touch), even though the noun sense was clearly implied by the study's focus on nouns.

All of these constraints of norm data need to be recognized, but they should not cause researchers to shy away from using norms for their research, so as long as norms are used cautiously. Any method is characterized by advantages and disadvantages, and the advantages of norms are that they allow researchers to avoid hand-classification, which facilitates reproducibility (if using the same norm data set) and generalizability (via the opportunity to analyze many more words automatically). In general, given that judgments for any individual word may be off, norm-based linguistics is best done on large sets of words, so that the noise inherent in the judgment for particular words is less influential.

In part due to the methodological concerns outlined herein, it is desirable to replicate specific analyses with different norm data sets. In this context, it is important to distinguish replication from reproducibility (see Gawne & Styles, chapter 2, this volume). Whereas a replication refers to the process of conducting the same experiment or study again with new data, reproducible research is the more basic requirement that even for a given study, another researcher can come to the same conclusion, in other words, all analytical decisions are transparent (see Berez-Kroeker et al., chapter 1, this volume). As an example of how replication can be done within the remit of norm-based linguistics, consider Winter's (2016) analysis of taste and smell language. The claim that taste and smell words are more emotional than other types of sensory words was substantiated by using three different emotional valence data sets. This shows that this result is not contingent on the specific norm set chosen by the analyst, and it shows how one can perform the same analysis using multiple different norm data sets to ensure that a particular claim rests on a firm foundation.

Finally, it is worth pointing out that general standards of reproducible research are even more important in the domain of norms than they may be in other areas of the language sciences. Because norm-based research is easy to conduct if the analyst has the relevant analytical and computational skills, and because there are myriad decisions to make in observational and often exploratory research with norms (e.g., whether to exclude words with high standard deviations or not), it is important to be maximally transparent about one's analytical decisions. At a bare minimum, this requires sharing one's data and code, which should preferably be done via a publicly accessible repository, such as via the Open Science Framework. This way, other researchers can follow all analytical steps, and if they disagree with specific analytical decisions, they have the option of performing a re-analysis.

In addition, it is worth stressing that norm-based research—even though it is entirely observational—can still be *preregistered*. That is, the researcher specifies their analysis plan in advance, so that other researchers can know whether a researcher has deviated from their analysis (for a discussion of preregistration in linguistics research, see Roettger 2019). Preregistration does not mean straitjacketing the analyst, as the registration report can be updated when unforeseen problems arise in a data analysis. However, preregistration allows clearly demarcating the boundary between confirmatory (hypothesis-testing) and exploratory (hypothesis-generating) research.

5 Conclusion and outlook

Traditionally, a large amount of linguistic theorizing has been based on the intuitions of linguists. The overreliance on introspective judgments from single individuals has been criticized by a number of researchers across the language sciences. However, rather than throwing the baby out with the bath water, intuitions should still play an important role in linguistics. Dąbrowska (2016:55) mentions that introspective judgments still "provide the most direct source of information about some aspects of language, notably meaning." The recommendation defended here is that rather than relying on a single linguist, intuitions should be aggregated over hundreds of individuals. Although, as discussed, this approach is not without its caveats (see also Pollock 2018), it is still better than relying *exclusively* on the hand classification of meanings, which lacks reproducibility and scalability.

To make a norm-based linguistics feasible in the long run, more norm data sets need to be collected. The big picture idea for the future of norm-based linguistics is that any important dimension of meaning that is of interest to the analyst can be quantified via collecting suitable ratings. Future research also needs to branch out

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

to other languages. There exist norming data sets of various lexical characteristics for such languages as German (Schmidtke et al. 2014; Vo et al. 2009), Polish (Riegel et al. 2015), Dutch (Speed & Majid 2017), and Chinese (Chen et al. 2019), all of which are from major world languages. Much more work needs to be done so that norm-based linguistics is not restricted to the analysis of English and major world languages.

Though norms may thus prove no panacea, they are an important component of the methodological tool kit of modern linguistics. As emphasized by Dąbrowska (2016:57), "when it comes to understanding something as complex as human language, it will be most productive to use every method that is available."

References

Amsel, B. D., T. P. Urbach, and M. Kutas. 2012. Perceptual and motor attribute ratings for 559 object concepts. *Behavior Research Methods* 44 (4): 1028–1041. https://doi.org/10.3758/s13428-012-0215-z.

Bednarek, M. 2008. Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory* 4 (2): 119–139.

Bloomfield, L. 1933. *Language*. Chicago: Chicago University Press.

Brysbaert, M., M. Stevens, P. Mandera, and E. Keuleers. 2016. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology* 7:1116.

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

Buchanan, E. M., K. D. Valentine, and N. P. Maxwell. 2018. LAB: Linguistic Annotated Bibliography—A searchable portal for normed database information. *Behavior Research Methods* 51 (9): 1878–1888.

Campbell, S. J., and G. E. Raney. 2016. A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior Research Methods* 48 (1): 330–340.

Caselli, N. K., Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey. 2017. ASL-LEX: A lexical database of American Sign Language. *Behavior Research Methods* 49 (2): 784–801.

Chen, I.-H., Q. Zhao, Y. Long, Q. Lu, and C.-R. Huang. 2019. Mandarin Chinese modality exclusivity norms. *PloS One* 14 (2): e0211336. https://doi.org/10.1371/journal.pone.0211336.

Dąbrowska, E. 2016. Looking into introspection. In *Studies in Lexicogrammar: Theory and Applications*, vol. 54, ed. G. Drożdż, 55–74. Amsterdam: John Benjamins.

Davies, M. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14 (2): 159–190.

Dilts, P., and J., Newman. 2006. A note on quantifying "good" and "bad" prosodies. *Corpus Linguistics and Linguistic Theory* 2 (2): 233–242.

Engelthaler, T., and T. T. Hills. 2018. Humor norms for 4,997 English words. *Behavior Research Methods* 50 (3): 1116–1124.

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

Grote, K. 2013. "Modality relativity": The influence of sign language and spoken language on conceptual categorization. PhD thesis, Hochschulbibliothek der Rheinisch-Westfälischen Technischen Hochschule Aachen.

Hunston, S. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12 (2): 249–268. https://doi.org/10.1075/ijcl.12.2.09hun.

Katz, A. N., A. Paivio, M. Marschark, and J. M. Clark. 1988. Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol* 3 (4): 191–214.

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

Keuleers, E., M. Stevens, P. Mandera, and M. Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology* 68 (8): 1665–1692.

Kiss, T., F. J. Pelletier, H. Husic, R. N. Simunic, and J. M. Poppek. 2016. A sense-based lexicon of count and mass expressions: The Bochum English Countability Lexicon. Paper presented at the 10th Language Resources and Evaluation Conference, Portorož, Slovenia, May 23–28.

Littlemore, J., P. P. Sobrino, D. Houghton, J. Shi, and B. Winter. 2018. What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation. *Metaphor and Symbol* 33 (2): 101–122.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In *Text and Technology: In Honour of John Sinclair*, ed. M. Baker, G. Francis, and E. Tognini-Bonelli, 157–176. Amsterdam: John Benjamins.

Lupyan, G., and B. Winter. 2018. Language is more abstract than you think, or, why aren't languages more iconic? *Philosophical Transactions of the Royal Society B: Biological Sciences* 373 (1752): 20170137.

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

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

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

Matlock, T. 1989. Metaphor and the grammaticalization of evidentials. *Annual Meeting of the Berkeley Linguistics Society* 15:215–225.

Medler, D. A., A. Arnoldussen, J. R. Binder, and M. S. Seidenberg. 2005. *The Wisconsin Perceptual Attribute Ratings Database*. http://www.neuro.mcw.edu/ratings/.

Osgood, C. E., G. J. Suci, and P. H. Tannenbaum. 1957. *The Measurement of Meaning*. Champaign: University of Illinois Press.

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

Perlman, M., H. Little, B. Thompson, and R. L. Thompson. 2018. Iconicity in signed and spoken vocabulary: A comparison between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology* 9:1433. https://doi.org/10.3389/fpsyg.2018.01433.

Perry, L. K., M. Perlman, and G. Lupyan. 2015. Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PloS One* 10 (9): e0137147. https://doi.org/10.1371/journal.pone.0137147.

Perry, L. K., M. Perlman, B. Winter, D. W. Massaro, and G. Lupyan. 2017. Iconicity in the speech of children and adults. *Developmental Science* 21 (3): e12572. https://doi.org/10.1111/desc.12572.

Pollock, L. 2018. Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study. *Behavior Research Methods* 50 (3): 1198–1216.

Popova, Y. 2005. Image schemas and verbal synaesthesia. In *From Perception to Meaning: Image Schemas in Cognitive Linguistics*, vol. 29, ed. B. Hampe, 395–419. Berlin: Mouton de Gruyter.

Riegel, M., M. Wierzba, M. Wypych, Ł. Żurawski, K. Jednoróg, A. Grabowska, and A. Marchewka. 2015. Nencki affective word list (NAWL): The cultural adaptation of the Berlin affective word list–reloaded (BAWL-R) for Polish. *Behavior Research Methods* 47 (4): 1222–1236.

Roettger, T. 2019. Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10 (1): article 1.

Ronga, I., C. Bazzanella, F. Rossi, and G. Iannetti. 2012. Linguistic synaesthesia, perceptual synaesthesia, and the interaction between multiple sensory modalities. *Pragmatics and Cognition* 20 (1): 135–167.

Schmidtke, D. S., T. Schröder, A. M. Jacobs, and M. Conrad. 2014. ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods* 46 (4): 1108–1118.

Sidhu, D. M., and P. M. Pexman. 2018. Lonely sensational icons: Semantic neighbourhood density, sensory experience and iconicity. *Language, Cognition and Neuroscience* 33 (1): 25–31.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Snefjella, B., and V. Kuperman. 2016. It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition* 156:135–146.

Speed, L. J., and A. Majid. 2017. Dutch modality exclusivity norms: Simulating perceptual modality in space. *Behavior Research Methods* 49 (6): 2204–2218.

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

Stewart, D. 2010. *Semantic Prosody: A Critical Evaluation*. Abingdon, UK: Routledge.

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

Strik Lievers, F., and B. Winter. 2018. Sensory language across lexical categories. *Lingua* 204:45–61.

Stubbs, M. 2001. *Words and Phrases*. Oxford, UK: Blackwell.

Sweetser, E. 1990. *From Etymology to Pragmatics*. Cambridge: Cambridge University Press.

Ullmann, S. 1959. *The Principles of Semantics*. Glasgow: Jackson, Son & Co.

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

Vinson, D. P., K. Cormier, T. Denmark, A. Schembri, and G. Vigliocco. 2008. The British Sign Language (BSL) norms for age of acquisition, familiarity, and iconicity. *Behavior Research Methods* 40 (4): 1079–1087.

Vo, M. L., M. Conrad, L. Kuchinke, K. Urton, M. J. Hofmann, and A. M. Jacobs. 2009. The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods* 41 (2): 534–538.

Warriner, A. B., and V. Kuperman. 2015. Affective biases in English are bi-dimensional. *Cognition and Emotion* 29 (7): 1147–1167.

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

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

Whitsitt, S. 2005. A critique of the concept of semantic prosody. *International Journal of Corpus Linguistics* 10 (3): 283–305.

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. 2019. *Sensory Linguistics: Language, Perception, and Metaphor*. Amsterdam: John Benjamins.

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

Winter, B., M. Perlman, L. K. Perry, and G. Lupyan. 2017. Which words are most iconic? Iconicity in English sensory words. *Interaction Studies* 18 (3): 433–454. https://doi.org/10.1075/is.18.3.07win.

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023

MIT Press Direct

© 2021 The Massachusetts Institute of Technology

This work is subject to a Creative Commons CC-BY-NC license. Subject to such license, all rights are reserved.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Berez-Kroeker, Andrea L., editor. | McDonnell, Bradley James, editor. | Koller, Eve, editor. | Collister, Lauren B., editor.
Title: The open handbook of linguistic data management / edited by Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Series: Open handbooks in linguistics series | Includes bibliographical references and index.
Identifiers: LCCN 2020044363 | ISBN 9780262045261 (hardcover)
Subjects: LCSH: Computational linguistics. | Natural language processing (Computer science) | Data mining.
Classification: LCC P98 .O64 2021 | DDC 410.285—dc23
LC record available at https://lccn.loc.gov/2020044363

Downloaded from http://direct.mit.edu/books/oa-edited-volume/chapter-pdf/2122694/c038500_9780262366076.pdf by guest on 27 September 2023