Behavior Research Methods (2023) 55:2522–2531
https://doi.org/10.3758/s13428-022-01912-6

Concreteness ratings for 62,000 English multiword expressions

Emiko J. Muraki[1] · **Summer Abdalla**[2] · **Marc Brysbaert**[3] · **Penny M. Pexman**[1]

Accepted: 16 June 2022 / Published online: 22 July 2022

Abstract
Concreteness describes the degree to which a word's meaning is understood through perception and action. Many studies use the Brysbaert et al. (2014) concreteness ratings to investigate language processing and text analysis. However, these ratings are limited to English single words and a few two-word expressions. Increasingly, attention is focused on the importance of multiword expressions, given their centrality in everyday language use and language acquisition. We present concreteness ratings for 62,889 multiword expressions and examine their relationship to the existing concreteness ratings for single words and two-word expressions. These new ratings represent the first big dataset of multiword expressions, and will be useful for researchers interested in language acquisition and language processing, as well as natural language processing and text analysis.

Keywords Concreteness · Word recognition · Multiword expressions · Idioms

Growing interest in multiword expressions

Language researchers are increasingly interested in multiword expressions. These are sequences of words representing a unitary meaning that could be replaced by a single word (Constant et al., 2017; Hubers et al., 2019), challenging traditional accounts of individual words as the building blocks of language (Contreras Kallens & Christiansen, 2022; Senaldi et al., 2022). There are four main types of such expressions. The first is compound nouns, which in English are often expressed as sequences of words (e.g., washing machine, clothes dryer, ironing board). The second type is particle verbs (e.g., give in, give over, give up). The third type consists of idiomatic expressions where a series of words is used to express a meaning that may not have much to do with the meaning of the individual words (e.g., spill the beans, be on the same page). Finally, there are fixed expressions that consist of familiar strings of words forming a single idea (e.g., thank you, go to bed, bride and groom). Fixed expressions differ from idioms because their meaning is transparent, although the border between both is fuzzy. Therefore, in the remainder of the paper we will consider them as a single category.

There is empirical evidence that multiword expressions are processed as units. Arnon and Snider (2010) asked participants to judge whether four-word expressions were possible in English or not (similar to a lexical decision task with single words). The stimuli were high-frequency expressions (don't have to worry, I don't know why), matched low-frequency expressions with equally frequent words (don't have to wait, I don't know who), and nonexisting expressions (I saw man the, jump during the pool). The authors observed that the high-frequency expressions were accepted faster than the low-frequency expressions (see also Senaldi et al., 2022; Siyanova-Chanturia et al., 2011). Arnon et al. (2017) used the same phrasal decision task and reported that early acquired three-word expressions (for the baby, in the trash) were accepted more rapidly than frequency-matched late acquired expressions (for the teacher, in the hills). This replicates the age-of-acquisition effect in single word recognition (Juhasz, 2005). Effects of multiword expressions have also been shown in eye movement data of participants reading sentences, wherein idiom and particle verb sentence forms facilitate processing compared to literal sentences (Titone et al., 2019; Tiv et al., 2019) and in priming, where idioms

✉ Emiko J. Muraki
ejmuraki@ucalgary.ca

[1] Department of Psychology, University of Calgary, 2500 University Drive, Calgary, AB T2N 1N4, Canada

[2] School of Languages, Linguistics, Literatures and Cultures, University of Calgary, Calgary, Canada

[3] Department of Experimental Psychology, Ghent University, Ghent, Belgium

with lower literal plausibility and more familiarity showed stronger priming effects (Titone & Libben, 2014).

Multiword expressions are also an important element in language acquisition, both in children learning their first language and in individuals acquiring a second language (Arnon, 2021; Boers et al., 2006). A language user without knowledge of multiword expressions faces serious challenges in everyday communication. As a result, multiword expressions are ideally included in text processing algorithms (Constant et al., 2017; Gamallo et al., 2018; Savary et al., 2015).

Scarcity of available resources

Researchers have started collecting norms for multiword expressions. In English, Titone and Connine (1994) collected ratings of familiarity, compositionality, predictability, and literality for 171 idiomatic expressions. Libben and Titone (2008) collected ratings of predictability, decomposability, familiarity, meaningfulness, and plausibility for 210 idiomatic expressions. Brysbaert et al. (2014) reported concreteness ratings for 2900 two-word expressions. Brysbaert and Biemiller (2017) validated test-based age-of-acquisition norms collected by Dale and O'Rouke (1981) for 3100 multiword expressions. Bulkes and Tanner (2017) published ratings of familiarity, meaningfulness, literal plausibility, global decomposability, and predictability for 870 idioms. Jolsvai et al. (2020) asked participants to rate plausibility, idiomaticity, and meaningfulness for 536 three-word expressions. Lindstromberg (2022) collected imageability and literality ratings on 150 phrasal verbs. In French, Bonin et al. (2022) collected norms of lexeme meaning dominance, semantic transparency, sensory experience, conceptual familiarity, imageability, age-of-acquisition, and subjective frequency for 506 hyphenated compound words. Finally, in Dutch, Hubers et al. (2019) collected norms of subjective frequency, subjective usage, familiarity, imageability, objective knowledge, and transparency for 374 idiomatic expressions and Sprenger et al. (2019) collected familiarity ratings for 189 idioms.

The above resources are important but rather limited when it comes to big data analysis. So, when Hills and Adelman (2015) investigated changes in the concreteness of English texts between the years 1800 and 2000, they had no other resource than the Brysbaert et al. (2014) ratings of 37,000 words and about 3000 two-word expressions (see also Snefjella et al., 2019). As a result, the concreteness of the expression "spill the beans" had to be estimated through the ratings of the words *spill* and *beans*. Similarly, authors looking at historical differences in the valence of texts (Hills et al., 2019), must look at the valence of *spill* and *beans*. The present paper is a first attempt to make big data available for multiword expressions.

Concreteness

Of the subjective norms collected for single words, concreteness is one of the most used, as can be concluded from citations to large-scale datasets of single-word ratings. Citation data suggest that affective ratings are used most, followed by concreteness, and age-of-acquisition. Concreteness refers to the degree to which a word's meaning is based on perception and action, or is conveyed through language. Examples of words scoring high on concreteness in the Brysbaert et al. (2014) ratings are "sled, daisy, peacock, bird". Examples of words scoring low are "ambivalent, belief, idealize, essentialness".

Looking at the references to Brysbaert et al. (2014), the following questions have been addressed with concreteness ratings:

- Selection of images for norming studies and the training of visual perception models (Hebart et al., 2019; Mahajan et al., 2018)
- Automated text analysis (Althoff et al., 2016; Humphreys & Wang, 2018)
- Estimating text sophistication (Kyle & Crossley, 2015)
- Selection of stimuli for behavioral, clinical, and brain imaging studies (Anderson et al., 2017; Bailey et al., 2020; Fini et al., 2022; Pereira et al., 2018; Ponari et al., 2018; Winter et al., 2017)
- Historical changes in word use (Hills & Adelman, 2015; Snefjella et al., 2019)
- Evaluating (computational) models of word semantics (Dubossarsky et al., 2017; Hollis & Westbury, 2016; Köper, & Im Walde, 2016; Vankrunkelsven et al., 2018; Villani et al., 2019)
- Examining the quality of word ratings (Pollock, 2018)
- Factors influencing word difficulty (Cervetti et al., 2015; Puimège & Peters, 2019; Yap et al., 2015)

Having concreteness norms for a good number of multiword expressions will expand the scope of these and other research topics. Below, we describe the compilation of a list of English multiword expressions. The resulting concreteness norms are the first measure for that multiword expression list.

Method

Participants

We recruited 2825 participants in 2020–2021 through three recruitment platforms: Amazon MTurk ($n = 1713$), Prolific

($n = 667$), and the University of Calgary's Research Participation System ($n = 435$). All participants were fluent English-speaking adults and all participants recruited through Amazon MTurk and Prolific were L1 speakers. Of the participants recruited from the University of Calgary, 130 reported that English was not their first language, however all of these participants reported being either completely fluent or very fluent in English. Participant demographic data were not collected for Amazon MTurk participants. Participants recruited through Prolific and the University of Calgary had a mean age of 25.74 ($SD = 9.55$), a mean of 14.73 years of education ($SD = 2.35$), and included 651 females, 437 males, six non-binary individuals, one gender-fluid individual, and one gender-questioning individual. The participants recruited through Amazon MTurk and Prolific were compensated between $3.50 to $11.53 USD, with variable compensation rates to account for different survey completion times. University of Calgary students received class credit in exchange for study participation. After our data cleaning procedures (reported in the Results section), we retained a sample of 1831 participants, including participants from Prolific and the University of Calgary with a mean age of 26.52 ($SD = 9.98$) and a mean of 14.97 years of education ($SD = 2.27$), and 94 participants from the University of Calgary who reported that English was not their first language. The final sample included 515 females, 339 males, six non-binary individuals, one gender-fluid individual, and one gender-questioning individual.

Stimuli

The multiword expression list was collected by MB over several years. It was based on an analysis of freely available dictionaries, lists of expressions recommended for language learning, Ngram frequency lists, and analysis of stimuli used in language studies. The list was reviewed to exclude any terms that were deemed inappropriate, transparently offensive, or very regionally specific, resulting in a final list of 66,458 expressions. The stimuli list was randomly divided into ten different surveys, with ~6646 expressions per survey, to accommodate limits of the survey software used for administration (Qualtrics). An additional 20 calibrator expressions and 20 control expressions were selected for inclusion in each of the surveys. The calibrator expressions were presented as practice and spanned the entire concreteness range (ten estimates were based on the Brysbaert et al. ratings; the others were judged by the authors). The control words were from the entire concreteness range as well (based on Brysbaert et al., 2014), and were used to detect noncompliance with the instructions (see below).

Procedure

The surveys were administered through Qualtrics. After providing informed consent, participants were asked to rate the expressions using a scale from 1 to 5, where 1 indicated that the expression was very abstract and 5 indicated that the expression was very concrete. Participants were instructed that if they were not familiar with the expression they were to select "I don't know the meaning of this expression". Participants were given 20 calibrator expressions at the beginning of the survey to practice making ratings, followed by a random subset of 440 expressions and the 20 control expressions, presented in a randomized order. The survey instructions are provided at https://osf.io/ksypa/ and were adapted from those used for the single-word and two-word expressions in Brysbaert et al. (2014). With each round of data collection, expressions that had acquired at least ten valid ratings or acquired "I don't know the meaning of this expression" responses eight or more times were removed from subsequent surveys. These numbers are based on previous experiences showing that increases in reliability of concreteness ratings tend to level off after 10 ratings (as also shown below). As the number of expressions in each survey decreased through the data collection processes, remaining expressions were consolidated into fewer surveys to retain enough expressions to randomly select a subset of 440 expressions per participant.

Results

The data were cleaned to screen for invalid responses (e.g., participants not attending to the task or responses from survey bots) using the methods described in Pexman et al. (2019) and consistent with the suggestions for online questionnaire data from Dupuis et al. (2019). Thus, we excluded participants who completed less than 33% of the ratings ($n = 49$) and participants who provided the same rating for 30 or more expressions in a row (i.e., an entire survey page; $n = 123$). As the last stage of data cleaning, each participant's ratings on the control[1] expressions were correlated with existing mean concreteness ratings for the same expressions from Brysbaert et al. (2014). Any participant whose correlation was less than $r = 0.20$ or whose correlation we were unable to calculate due to the same rating being provided for

[1] In the first round of data collection, a technical error resulted in six surveys that only had calibrator expressions and no control expressions. For these surveys, the calibrator expressions were used to assess individual participants' correlations to existing concreteness ratings from Brysbaert et al. (2014).

Table 1 Number of participants excluded as a function of exclusion criterion and platform (total number tested: $n = 2825$)

Platform	Exclusion criterion			Participants included in analysis
	Less than 33% of expressions rated	30 or more same ratings in a row	Correlation to control words < .20	
Amazon MTurk ($n = 1713$)	30	48	672	963
Prolific ($n = 677$)	0	19	102	556
University of Calgary RPS ($n = 435$)	19	56	48	312

Note. Sample sizes reported in parentheses are before any participants were excluded

every control expression was excluded ($n = 822$). The same criteria were used in Brysbaert et al. (2014).

Based on the data cleaning procedure a total of 994 participants were removed from further analysis, leaving 1831 participants with valid ratings. Table 1 provides the exclusions by recruitment platform and exclusion reason. In particular, the data loss with Amazon MTurk was high, in line with concerns that have been raised about recent developments in the platform (Agley et al., 2022; Eyal et al., 2022).[2]

Fourteen duplicate expressions were identified in the original expression list, ten that had an extra space and four with extra punctuation or an extra letter. For these 14 expressions, the ratings from the duplicate entries were combined when calculating the mean concreteness rating. Two expressions were excluded from analysis because they did not include the properly accented letters and a version of the expressions with the appropriate accents was already on the list. Two further expressions were excluded because of formatting errors. Eight expressions were excluded because they were also included as control items.

The final dataset comprised 66,432 expressions and 803,479 observations. Of those observations, 691,689 were valid ratings and 111,790 were "I don't know the meaning of this expression" responses. On average, each expression received 10.4 valid ratings, with a maximum of 55 valid ratings. Only expressions with at least ten valid ratings ($n = 62,889$) are included in the subsequent analyses. The raw data and mean concreteness ratings are available at: https://osf.io/ksypa/.

To assess the reliability of the ratings, we calculated two intraclass correlation coefficients (ICC) using a multilevel model with a random intercept and the expression as a fixed effect (Brysbaert, 2019; Fletcher, 2015). The first ICC (representing the average correlation of expression ratings between participants) was 0.34 and the second ICC (representing the reliability of the average ratings) was 0.84, indicating good reliability. To further assess the reliability of the ratings, we selected the expressions that were present in our dataset and in Brysbaert et al., 2014 ($n = 2928$). The correlation between the two datasets was $r = 0.86$, confirming that the current ratings represent the same construct of concreteness as in Brysbaert et al. (2014).

The multiword expressions had a mean concreteness of 3.13 ($SD = 0.97$). A kernel density plot of mean concreteness values for all expressions (Fig. 1) showed a bimodal distribution, indicating two types of expressions, one more abstract and the other more concrete. We also investigated how the distribution of concreteness ratings varied by expression type. To do so, the expressions were coded as either compound nouns (e.g., *purchase price*; $n = 34,013$), particle verbs (e.g., *chime in*; $n = 5,115$), or idiom/fixed expressions (e.g., *don't count your chickens before they hatch*; $n = 23,761$).[3] The kernel density plots of each expression type (Fig. 1) indicate that idiom/fixed expressions and particle verb expressions are positively skewed and rated as more abstract ($M = 2.42$, $SD = 0.69$ and $M = 2.56$, $SD = 0.54$ respectively), whereas compound noun expressions are negatively skewed and rated as more concrete ($M = 3.71$, $SD = 0.79$).

We further examined the relationship between the mean concreteness ratings of the expressions, and the concreteness ratings of the individual words within each expression. We extracted concreteness ratings for individual words in each expression that were available in Brysbaert et al. (2014). We then calculated the mean concreteness of all individual words within an expression, the maximum concreteness of all individual words within an expression (i.e., the highest concreteness rating for the words within an expression), and the minimum concreteness of all individual words within an expression (i.e., the lowest concreteness rating for the words within an expression). This resulted in a list of 62,127 expressions with mean, maximum, and minimum concreteness values. When assessing these expressions, the mean

[2] A look at the distribution of correlations with control words revealed a bimodal distribution, with one mode close to $r = .8$ and another close to $r = .0$. Based on this observation, the criterion of $r = .2$ set by Brysbaert et al. (2014) remained sensible, even though it resulted in a high number of exclusions.

[3] This coding is very preliminary, based on the authors' intuitions, and we invite readers to work out a more refined, theoretically grounded categorization.

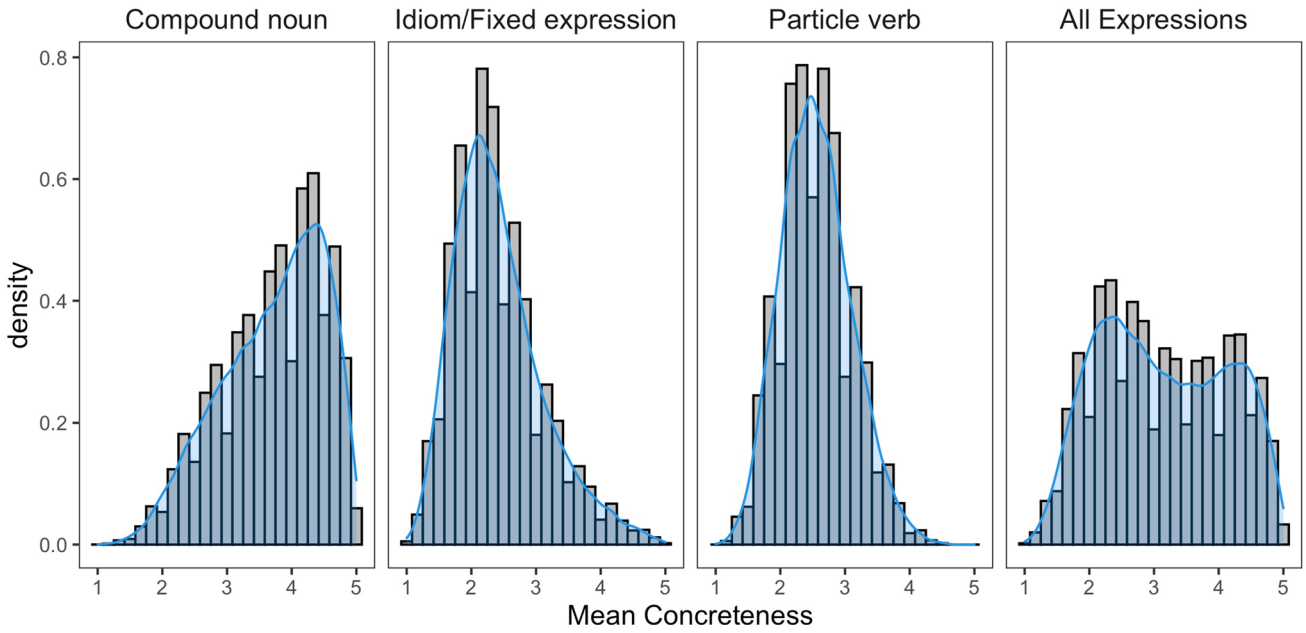

Fig. 1 Distributions of concreteness ratings for the different types of multiword expressions (1 = very abstract, 5 = very concrete)

concreteness rating for the entire expression correlated $r = 0.67$ to the mean concreteness ratings of the individual words within the expression and $r = 0.67$ to the minimum concreteness rating of the individual words within the expression. The concreteness rating for the entire expression was less related to the maximum concreteness rating of individual words within the expression ($r = 0.45$). When these relationships were examined by expression type, we observed that concreteness ratings of compound noun expressions were more strongly related to the mean, maximum, and minimum concreteness ratings of individual words within the expression ($r = 0.63$, $r = 0.58$, and $r = 0.57$ respectively). In contrast, the concreteness ratings of both idiom/fixed expressions and particle verb expressions were less related to the mean, maximum, and minimum concreteness ratings of individual words within the expression ($r = 0.30$, $r = 0.14$, $r = 0.27$ and $r = 0.26$, $r = 0.20$, $r = 0.23$ respectively). These relationships are shown in Fig. 2 and the correlations are reported in Table 2.

To examine to what extent the above correlations were influenced by function words, we repeated the analyses limited to content words. Function words are some 300 high-frequency words signaling the syntactic structure of the phrase. They consist of determiners, prepositions, pronouns, particles, and conjunctions. An expression like "drop in the bucket" contains two function words (in, the) and two content words (drop, bucket). For the particle verbs, the pruning of particles meant that they were stripped to the verb (give in, give up, … were all limited to give). Some multiword expressions consist entirely of function words (e.g., near to, next to, on top of, one another). These were excluded from the analysis using only content words, resulting in a list of 60,707 expressions included in this analysis.

In the analysis limited to content words the concreteness ratings for the expression were less related to the mean concreteness ratings of the content words within the expression than in the first analysis including all words: mean ($r = 0.52$), maximum ($r = 0.47$), minimum ($r = 0.48$). When these relationships were examined by expression type, we observed no great changes for the compound noun expressions, as these include very few function words ($r = 0.62$, $r = 0.58$, and $r = 0.56$ mean, maximum, and minimum respectively). In contrast, the concreteness ratings of the idiom/fixed expressions were less related to the mean concreteness ratings ($r = 0.25$) when using only content words and more related to the maximum concreteness ratings ($r = 0.18$; although the relationship is still weak), with a smaller change in the relationship between the concreteness ratings of the expression and the minimum concreteness rating of content words within the expression ($r = 0.26$). Particle verb expressions were less related to the mean concreteness ratings of individual content words within the expression ($r = 0.21$). The change in relationship between the concreteness rating of the expression and the maximum and minimum concreteness ratings when using only content words was modest ($r = 0.22$ and $r = 0.20$, respectively). These relationships are shown in Fig. 3 and the correlations are reported in Table 2.

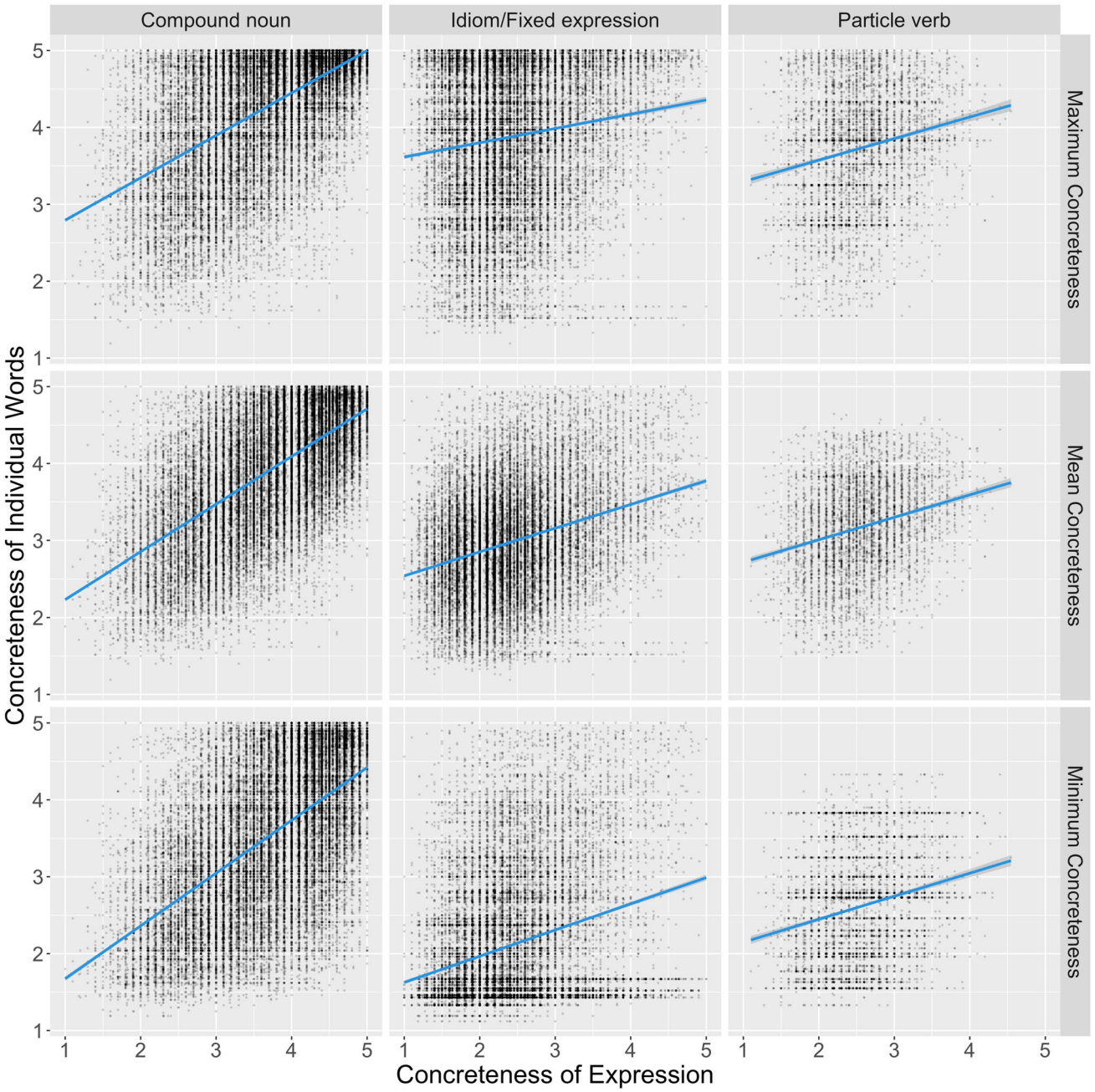

Fig. 2 Correlations between concreteness ratings of multiword expressions and concreteness ratings of the words in the expressions (all words included)

Discussion

In the present article, we present concreteness ratings for 62,889 English multiword expressions. These ratings will allow researchers to expand their research from single words to familiar sequences of words and will support efforts to study language processing in context (see Barsalou, 2020, and Murgiano et al., 2021, for a more thorough discussion of situated language processing, and Sidhu & Pexman, 2021, on the limits of single word recognition tasks).

The concreteness ratings have an overall reliability of .84, meaning that they give a good estimate of how much participants perceive their meaning as based on perception/action or on language. The ratings also provide us with the first curated list of multiword expressions that can be built upon for future norming studies. In addition, the study includes 3543 expressions that were not known to 59% of the

Table 2 Correlations between concreteness ratings of multiword expressions and concreteness ratings of the words in the expressions (all words and content words only)

Expression type	All Words (n = 62,127)			Content Words Only (n = 60,707)		
	Max Conc	*M Conc*	*Min Conc*	*Max Conc*	*M Conc*	*Min Conc*
All expressions	0.45	0.67	0.67	0.47	0.52	0.48
Compound nouns	0.58	0.63	0.57	0.58	0.62	0.56
Idiom/Fixed expressions	0.14	0.30	0.27	0.18	0.25	0.26
Particle verbs	0.20	0.26	0.23	0.22	0.21	0.20

Note. Max Conc = the highest concreteness rating of words within an expression; M Conc = the mean of concreteness ratings of words within an expression; Min Conc = the lowest concreteness rating of words within an expression. All concreteness ratings for individual words are derived from Brysbaert et al. (2014)

participants who received those expressions and that may be useful for learning studies.

We hope that the current list will be a step forward in language research as multiword expressions now make their way into the big data language enterprise. For example, studies investigating text difficulty (Hills & Adelman, 2015; Hills et al., 2019; Kyle & Crossley, 2015; Snefjella et al., 2019) can now use concreteness of multiword expressions, rather than estimates based on the constituent words. Similarly, researchers interested in language learning can now add multiword expressions to their stimulus list.

At the same time, it is good to keep in mind some limitations of our enterprise. A first limitation of concreteness ratings is that they are largely influenced by vision (Brysbaert et al., 2014; Connell & Lynott, 2012). More information can be obtained by gathering multidimensional measures of perceptual and action strength (Lynott et al., 2020; Speed & Brybaert, 2022). So, for the expression "fun run", participants can be asked how much the expression is related to vision, hearing, touch, taste, olfaction, interoception, foot movements, hand movements, head movements, mouth movements, and torso movements. Collection of such multidimensional information requires many more resources than we have at our disposal and could be a focus for future norming work.

Our concreteness ratings are also limited because they do not make a distinction between various senses of an expression. For instance, the word "uniform" can be considered concrete (a garment) or abstract (unchanging in form or character). The additional context provided by multiword expressions in some cases clarifies words with multiple senses, such as "bank", which in the present dataset occurs in both "bank account" and "river bank". However, some ambiguity can still be seen in the variability of the ratings given by the participants. In the present dataset, we see large standard deviations for some of the idioms in our list (e.g., kid glove, a finger in every pie, it will all come out in the wash, dragon mouth, the tail wagging the dog), indicating large individual differences in how they were interpreted. Interestingly, recent studies have shown that words with highly variable concreteness ratings tend to be recalled better in memory studies (Brainerd et al., 2021). It will be interesting to see whether the same is true for multiword expressions.

Recently, suggestions have been made for better ways to collect subjective information about language stimuli (Hollis, 2020; Hollis & Westbury, 2018) or to calculate descriptive statistics from rating studies (Bürkner & Vuorre, 2018; Liddell, & Kruschke, 2018; Taylor et al., 2021). We have stuck to the traditional mean and SD to facilitate comparison with previous data (notably Brysbaert et al., 2014), but we are much interested to see if better measures are possible for the concreteness ratings we have collected. Therefore, we make not only the summary data but also the raw data available for reanalysis.

Despite the limitations of our effort, we are convinced that the present list will be a major step forward in language research, because multiword expressions now enter the big data language enterprise.

Availability

The concreteness ratings are available in two main files at https://osf.io/ksypa/. The first file contains all raw data collected, including those of participants that had to be excluded (indicated by the column Filter = 0), and answers indicating that the participant did not know the expression (ratings of 6). This file may be of interest to colleagues investigating online data gathering and response patterns across participants.

The second file is a summary file, containing processed information for the 66,432 English multiword expressions we presented. Expressions not known to the participants only include information about the number of selected participants who responded to the expression and the number of participants indicating they did not know the expression. This is the only information we think valid for these

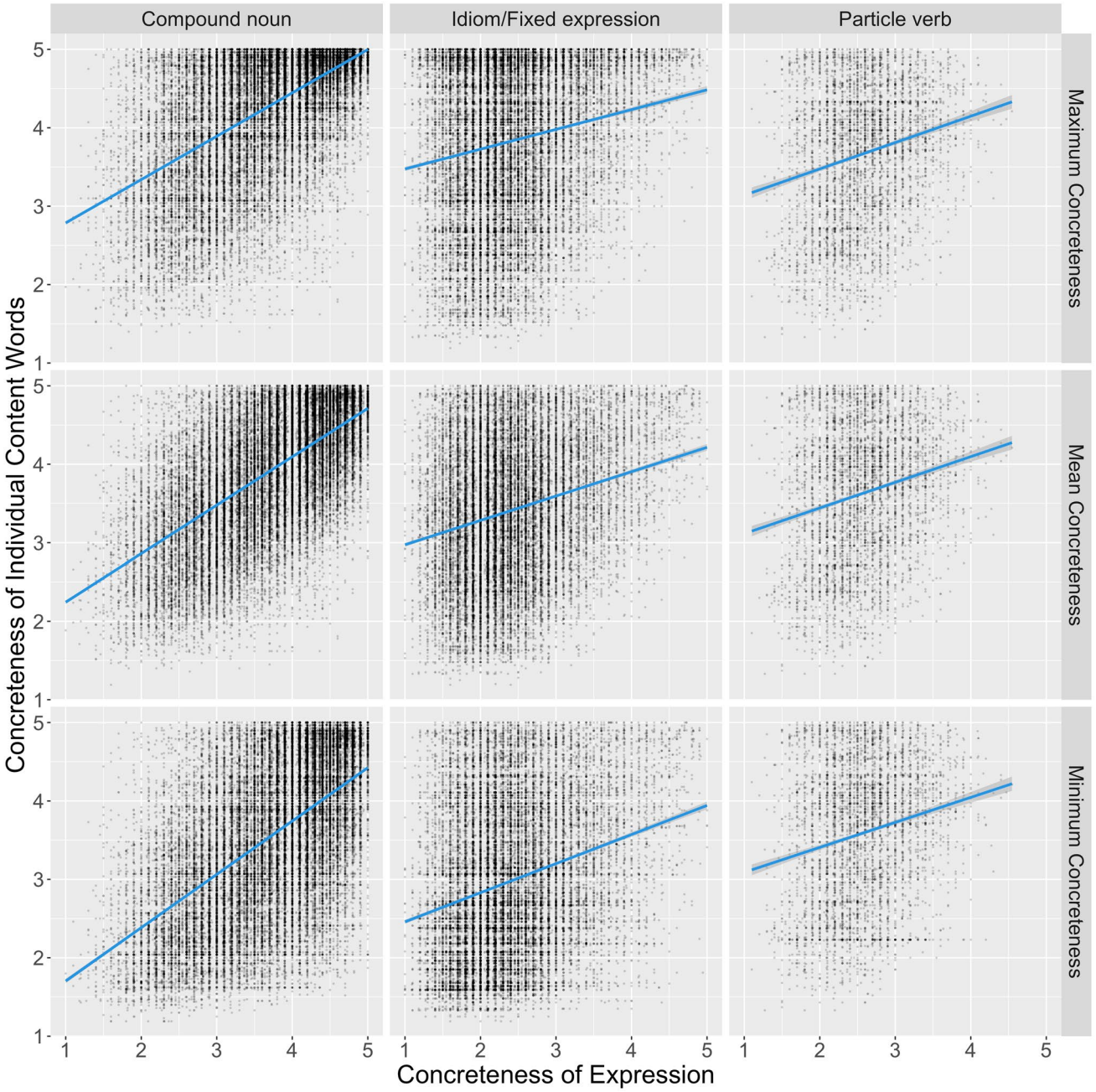

Fig. 3 Correlations between concreteness ratings of multiword expressions and concreteness ratings of the words in the expressions (content words only)

expressions. They may be of interest to colleagues wanting to teach unfamiliar expressions.

Expressions familiar to the participants include additional information about the mean rating, the standard deviation of the ratings, and the number of responses going from 1 (very abstract) to 5 (very concrete). This provides users with all the information they need to select stimuli for various types of research.

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01665-8

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale analysis of counseling conversations: An application of natural language

processing to mental health. *Transactions of the Association for Computational Linguistics, 4*, 463–476.
Anderson, A. J., Kiela, D., Clark, S., & Poesio, M. (2017). Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics, 5*, 17–30.
Arnon, I. (2021). The Starting Big approach to language learning. *Journal of Child Language, 48*(5), 937–958.
Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82.
Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effects for multiword phrases. *Journal of Memory and Language, 92*, 265–280.
Bailey, D. J., Nessler, C., Berggren, K. N., & Wambaugh, J. L. (2020). An aphasia treatment for verbs with low concreteness: a pilot study. *American Journal of Speech-Language Pathology, 29*(1), 299–318.
Barsalou, L. W. (2020). Challenges and opportunities for grounding cognition. *Journal of Cognition, 3*(1), 31. https://doi.org/10.5334/joc.116
Boers, F., Eyckmans, J., & Stengers, H. (2006). Motivating multiword units: Rationale, mnemonic benefits, and cognitive style variables. *EUROSLA Yearbook, 6*(1), 169–190.
Bonin, P., Laroche, B., & Méot, A. (2022). Psycholinguistic norms for a set of 506 French compound words. *Behavior Research Methods, 54*(1), 393–413.
Brainerd, C. J., Chang, M., Bialer, D. M., & Toglia, M. P. (2021). Semantic ambiguity and memory. *Journal of Memory and Language, 121*, 104286.
Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. https://doi.org/10.5334/joc.72
Brysbaert, M., & Biemiller, A. (2017). Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavior Research Methods, 49*(4), 1520–1523.
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.
Bulkes, N. Z., & Tanner, D. (2017). "Going to town": Large-scale norming and statistical analysis of 870 American English idioms. *Behavior Research Methods, 49*(2), 772–783.
Bürkner, P.C., & Vuorre, M. (2018). Ordinal regression models in psychological research: A tutorial. https://files.osf.io/v1/resources/x8swp/providers/osfstorage/5a973e25218b7b000f13bc0d
Cervetti, G. N., Hiebert, E. H., Pearson, P. D., & McClung, N. A. (2015). Factors that influence the difficulty of science words. *Journal of Literacy Research, 47*(2), 153–185.
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
Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics, 43*(4), 837–892.
Contreras Kallens, P., & Christiansen, M. H. (2022). Models of language and multiword expressions. *Frontiers in Artificial Intelligence, 5*, 781962. https://doi.org/10.3389/frai.2022.781962
Dale, E., & O'Rourke, J. (1981). *The living word vocabulary, the words we know: A national vocabulary inventory*. World Book.
Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology, 53*(8), 1560.
Dupuis, M., Meier, E., & Cuneo, F. (2019). Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices. *Behavior Research Methods, 51*(5), 2228–2237. https://doi.org/10.3758/s13428-018-1103-y
Eyal, P., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01694-3
Fini, C., Zannino, G. D., Orsoni, M., Carlesimo, G. A., Benassi, M., & Borghi, A. M. (2022). Articulatory suppression delays processing of abstract words: The role of inner speech. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218211053623
Fletcher, T. D. (2015). *Package 'psychometric'*. Available at https://cran.r-project.org/web/packages/psychometric/psychometric.pdf
Gamallo, P., Garcia, M., Pineiro, C., Martinez-Castano, R., & Pichel, J. C. (2018, October). LinguaKit: a big data-based multilingual tool for linguistic analysis and information extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 239–244). IEEE.
Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One, 14*(10), e0223792.
Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition, 143*, 87–92.
Hills, T. T., Proto, E., Sgroi, D., & Seresinhe, C. I. (2019). Historical analysis of national subjective wellbeing using millions of digitized books. *Nature Human Behaviour, 3*(12), 1271–1275.
Hollis, G. (2020). The role of number of items per trial in best–worst scaling experiments. *Behavior Research Methods, 52*(2), 694–722.
Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review, 23*(6), 1744–1756.
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.
Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative data of Dutch idiomatic expressions: Subjective judgments you can bank on. *Frontiers in Psychology, 10*, 1075.
Humphreys, A., & Wang, R. J. H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research, 44*(6), 1274–1306.
Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2020). Meaningfulness beats frequency in multiword chunk processing. *Cognitive Science, 44*(10), e12885.
Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin, 131*(5), 684–712.
Köper, M., & Im Walde, S. S. (2016, May). Automatically generated affective norms of abstractness, arousal, imageability and valence for 350 000 German lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2595–2598).
Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757–786.
Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36*(6), 1103–1121.
Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348.
Lindstromberg, S. (2022). The compositionality of English phrasal verbs in terms of imageability. *Lingua, 103373*. https://doi.org/10.1016/j.lingua.2022.103373

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods, 52*(3), 1271–1291.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., ... Van Der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 181–196).

Murgiano, M., Motamedi, Y., & Vigliocco, G. (2021). Situating Language in the Real-World: The Role of Multimodal Iconicity and Indexicality. *Journal of Cognition, 4*(1), 38. https://doi.org/10.5334/joc.113

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., et al. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications, 9*(1), 1–13.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body-object interaction ratings for more than 9,000 English words. *Behavior Research Methods, 51*(2), 453–466. https://doi.org/10.3758/s13428-018-1171-z

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.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science, 21*(2), e12549.

Puimège, E., & Peters, E. (2019). Learning L2 vocabulary from audiovisual input: an exploratory study into incidental learning of single words and formulaic sequences. *The Language Learning Journal, 47*(4), 424–438.

Savary, A., Sailer, M., Parmentier, Y., Rosner, M., Rosén, V., Przepiórkowski, A., ... Sangati, F. (2015, November). PARSEME–PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*.

Senaldi, M. S., Titone, D. A., & Johns, B. T. (2022). Determining the importance of frequency and contextual diversity in the lexical organization of multiword expressions. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 76*, 87–98.

Sidhu, D. M., & Pexman, P. M. (2021). Implications of the "Language as Situated" view for written iconicity. *Journal of Cognition, 40*, 1–4.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(3), 776–784.

Snefjella, B., Généreux, M., & Kuperman, V. (2019). Historical evolution of concrete and abstract language revisited. *Behavior Research Methods, 51*(4), 1693–1705.

Speed, L. J., & Brybaert, M. (2022). Dutch sensory modality norms. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01656-9

Sprenger, S. A., la Roi, A., & van Rij, J. (2019) The development of idiom knowledge across the lifespan. *Frontiers in Communication, 4*, 1–29. https://doi.org/10.3389/fcomm.2019.00029

Taylor, J. E., Rousselet, G. A., Scheepers, C., & Sereno, S. C. (2021, August 3). Rating norms should be calculated from cumulative link mixed effects models. https://doi.org/10.31234/osf.io/3vgwk

Titone, D. A., & Connine, C. M. (1994). Descriptive norms for 171 idiomatic expressions: Familiarity, compositionality, predictability, and literality. *Metaphor and Symbol, 9*(4), 247–270.

Titone, D., & Libben, M. (2014). Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon, 9*(3), 473–496.

Titone, D., Lovseth, K., Kasparian, K., & Tiv, M. (2019). Are figurative interpretations of idioms directly retrieved, compositionally built, or both? Evidence from eye movement measures of reading. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 73*(4), 216.

Tiv, M., Gonnerman, L., Whitford, V., Friesen, D., Jared, D., & Titone, D. (2019). Figuring out how verb–particle constructions are understood during L1 and L2 reading. *Frontiers in Psychology, 1733*.

Vankrunkelsven, H., Verheyen, S., Storms, G., & De Deyne, S. (2018). Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *Journal of Cognition, 1*(1).

Villani, C., Lugli, L., Liuzza, M. T., & Borghi, A. M. (2019). Varieties of abstract concepts and their multiple dimensions. *Language and Cognition, 11*(3), 403–430.

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic?: Iconicity in English sensory words. *Interaction Studies, 18*(3), 443–464.

Yap, M. J., Lim, G. Y., & Pexman, P. M. (2015). Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition, 43*(8), 1148–1167.