Behav Res
DOI 10.3758/s13428-016-0747-8

"Going to town": Large-scale norming and statistical analysis of 870 American English idioms

Nyssa Z. Bulkes[1,2] · **Darren Tanner**[1,2,3]

Abstract An *idiom* is classically defined as a formulaic sequence whose meaning is comprised of more than the sum of its parts. For this reason, idioms pose a unique problem for models of sentence processing, as researchers must take into account how idioms vary and along what dimensions, as these factors can modulate the ease with which an idiomatic interpretation can be activated. In order to help ensure external validity and comparability across studies, idiom research benefits from the availability of publicly available resources reporting ratings from a large number of native speakers. Resources such as the one outlined in the current paper facilitate opportunities for consensus across studies on idiom processing and help to further our goals as a research community. To this end, descriptive norms were obtained for 870 American English idioms from 2,100 participants along five dimensions: familiarity, meaningfulness, literal plausibility, global decomposability, and predictability. Idiom familiarity and meaningfulness strongly correlated with one another, whereas familiarity and meaningfulness were positively correlated with both global decomposability and predictability. Correlations with previous norming studies are also discussed.

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0747-8) contains supplementary material, which is available to authorized users.

✉ Nyssa Z. Bulkes
nyssabulkes@gmail.com

[1] Department of Linguistics, University of Illinois at Urbana-Champaign, 4080 Foreign Languages Building, MC-168, 707 S. Mathews Avenue, Urbana, Il 61801, USA

[2] Beckman Institute for Advanced Science and Technology, University of Illinois, Urbana, Il, USA

[3] Neuroscience Program, University of Illinois, Urbana, Il, USA

Keywords Idioms · Figurative language · Nonliteral language · Language processing · Language norms

Fluency—and subsequent mastery—of a language necessitates the successful processing of literal as well as nonliteral language (e.g., idioms, metaphors, etc.). Experience with the linguistic input enables comprehenders to recognize highly co-occurring phrases as chunks of words that "go together," ultimately reducing processing load and supporting computational economy (e.g., Ellis, 2002; Goldberg, 2007; Sinclair, 1991; Wray, 2002, among others). Although compositional analysis may suffice for literal language comprehension, nonliteral expressions require a comprehender's prior exposure to a particular construction, as well as contextual cues, to lead them to a felicitous interpretation, going beyond the sentence level to activate the nonliteral meaning of a particular configuration. Swinney and Cutler (1979) defined an idiom as "a string of two or more words for which meaning is not derived from the meanings of the individual words comprising that string" (p. 523). Whereas literal meaning can be gleaned at the sentence level by composing the meanings of the individual words with respect to some inferred syntactic structure, the meaning of an idiom is classically defined as one that is more than the sum of its component parts.

Within the greater class of idioms, however, there is more variation that ultimately affects processing. Mounting psycholinguistic evidence has shown that a variety of dimensions (i.e., familiarity, predictability) affect the speed and accuracy with which an idiomatic interpretation can be retrieved (e.g., Cacciari & Tabossi, 1988; Gibbs, 1980; Libben & Titone, 2008; Schweigert, 1986; Titone & Libben, 2014). For example, there is evidence that the figurative meaning of highly familiar, predictable idioms can be retrieved more quickly than those of less familiar, less predictable expressions

Published online: 05 August 2016

(Cacciari & Tabossi, 1988; Titone & Connine, 1999). Similarly, an idiom's internal structure varies with respect to how well it permits syntactic transformations (e.g., passivization) while still maintaining a possible figurative interpretation, and this variability distinguishes idioms as a heterogeneous class of expressions (Gibbs & Nayak, 1989; Nunberg, 1978). An additional level of complexity arises when idioms have a plausible literal interpretation (e.g., *It was a slap in the face*.). Here, contextual cues ultimately guide the listener to select the most contextually appropriate interpretation—either literal or nonliteral—as these expressions in isolation are ultimately ambiguous (Colombo, 1993). Finally, the predictability of an idiom has also been found to affect the speed with which an idiomatic interpretation can be retrieved from the lexicon, because the more likely the idiom-final word, the more unitary its representation in the lexicon (Cacciari & Tabossi, 1988; Titone & Connine, 1999). An idiom's predictability, however, hinges on its component parts, as the more unique its configuration is to a particular interpretation, the more predictable it is. For example, *a slap in the face* is fairly predictable, as an alternate completion is highly unlikely, and perhaps even ungrammatical (e.g., *a slap in the cheek*). *Hit the wall*, on the other hand, despite the lexical specificity of the verb *hit*, allows a variety of felicitous idiomatic completions (e.g., *hit the ceiling*, *hit the floor*, *hit the sack*), not to mention nonidiomatic continuations. Ultimately, the sheer number of felicitous completions, both idiomatic and literal, renders this expression fairly unpredictable without the help of a surrounding context.

Previous norming studies of English idioms (e.g., Libben & Titone, 2008; Schweigert & Cronk, 1992; Titone & Connine, 1994b) have reported ratings for different subsets of idioms along varying dimensions, with subjective familiarity most often discussed. For example, Schweigert and Cronk (1992) normed 390 idioms for familiarity, with a subset of their participants also rating items for the likelihood of encountering the phrase used literally. Titone and Connine (1994b) normed 171 expressions for familiarity, compositionality, literality, and predictability. Libben and Titone (2008) normed 219 idioms for familiarity, meaningfulness, literality, predictability, global decomposability, normal decomposability, abnormal decomposability, as well as for the semantic relatedness of an idiom-internal verb and/or noun with its phrasal meaning. In order to build a comprehensive model of language comprehension, however—one that also accounts for the processing of nonliteral expressions—research is needed to determine how idiom dimensions relate to one another, and specifically, whether any dimension is a good predictor of any other. Furthermore, in an effort to facilitate cross-study comparisons within the idiom literature and avoid the need to renorm a different subset of idioms for every new study, large-scale norming studies provide a valuable publicly available resource to the field, the use of which ultimately helps to ensure external validity. To accomplish this, not only are these publicly available resources needed, but breadth and scope are required, such that they are able to accommodate a similarly broad variety of research questions. It is in this spirit that we present the current work.

Five dimensions included in previous norming endeavors were chosen for inclusion in the present work: familiarity, meaningfulness, global decomposability, literal plausibility, and predictability. *Familiarity* is defined here as how often a participant either hears or uses a given idiomatic expression. *Meaningfulness* is defined as how well an individual knows the meaning of an idiomatic expression. *Global decomposability* is defined as the degree to which an idiom's component parts contribute semantically to the overall meaning of the idiom (e.g., The idiom in *It was a done deal* would be considered relatively decomposable, whereas the idiom in *She was a flash in the pan* would be comparatively less decomposable). *Literal plausibility* is defined as the plausibility of the phrase's literal interpretation (e.g., *She was larger than life* could be considered relatively implausible, whereas *She hit the wall* could be rated as very literally plausible). *Predictability* is defined as the likelihood of providing the idiomatic completion in a fill-in-the-blank (cloze) task (e.g., *She was larger than* _____). These characteristics have all been demonstrated to ultimately affect the online processing of idioms (e.g., Cacciari & Glucksberg, 1991; Cacciari & Tabossi, 1988; Cronk, Lima, & Schweigert, 1993; Swinney & Cutler, 1979; Titone & Connine, 1994a, 1999; Titone & Libben, 2014), and have been included in previous idiom norming studies (e.g., Libben & Titone, 2008; Schweigert & Cronk, 1992; Titone & Connine, 1994b). It is for this reason that they have been chosen for inclusion in the present work. What is more, the goal of the present endeavor is not only to corroborate previously obtained norms, but to expand the scope of this work by obtaining a much larger norming set, both with more idioms than any previous endeavor, as well as more unique ratings from more participants. This will not only build upon the prior efforts, but will also lead to more generalizable norms through the inclusion of a wide range of raters, allow for latent variable analyses of the normed dimensions to better understand how these dimensions interrelate, and ultimately broaden the scope of information available for other researchers in this domain.

A further motivation of the present study is to provide a more contemporary set of norms. Although ratings are available in other published works, the need for recent information is crucial in order to properly control experimental materials, and other studies in the figurative language domain have similarly updated other publicly available norming data (e.g., Campbell & Raney, 2016, for an updated investigation on metaphors; and Brysbaert, Warriner, & Kuperman, 2014, for updated ratings of English lemma concreteness). Additionally, as specific research objectives vary across idiom studies, so do

the idioms appropriate for investigation. To this end, the current work includes ratings for the largest subset of idioms to date, in an effort to significantly expand the scope of the currently available data.

Method

Participants

Participants were recruited via Amazon Mechanical Turk (n = 2215). To obtain ratings specific to the North American context, participants' IP addresses were limited to include only those located in the United States and Canada. All participants were self-reported native speakers of English. Participants who reported a native language other than English were excluded from further analyses. All participants provided informed consent prior to the onset of a norming survey. Participants received a small amount of cash for their participation.

Materials

A total of 870 American English idioms were selected for norming from the *Longman Pocket Idioms Dictionary* (Lee, 2000). Idioms were chosen on the basis of modernity and the ability to be used in a simple sentence structure (e.g., *It was a slap in the face*). Carrier sentences were all of the minimal structure "She/It verbed X noun," in which each sentence began with a simple pronoun (i.e., *she*, *he*, *they*, *it*), followed by the verb. "X" could either be an adjective, determiner, and/or a preposition, as idioms of all lengths were included in an effort to diversify the norming data. Verbs were either in the present or past tense. No additional contexts were included, in order to ensure ratings described impressions of the idioms in isolation and were not influenced by other sentential or discourse context biases. Materials were randomized into four separate lists, with 217 or 218 idioms in each list. Four additional lists were created—for a total of eight lists per dimension—by rerandomizing the original set of 870 idioms. This was done to ensure ratings were not overtly affected by the presence of neighboring idioms.

Procedure

Participants were first asked to complete a brief language background survey prior to completing the norming. Each participant was then asked to provide responses for one list of idioms via the SurveyGizmo (www.surveygizmo.com) interface. No participant completed more than one list, and each participant only completed one list within one dimension, so that each list was rated by a unique set of participants. A minimum of 50 participants was recruited for each list; 50–58 participants completed each list. This meant that a minimum of 100 unique individuals rated each idiom on each dimension (see supplementary materials for number of participants per idiom within each dimension). For the dimensions of familiarity, meaningfulness, and literal plausibility, participants were asked to rate on a Likert scale of 1–5 (1 = *low*, 5 = *high*) each expression in their assigned list. For global decomposability, participants were asked to indicate whether each item was decomposable or nondecomposable with a checkbox response. For predictability, participants were asked to complete the simple carrier sentence with the first word that came to mind. Instructions for each dimension were adapted from those used in Libben and Titone (2008; see Appendices 1, 2, 3, 4, and 5 for the instruction texts given with the different tasks). To ensure that the directions were followed, eight catch trials were included in each list. For familiarity, meaningfulness, and literal plausibility, novel idioms were created (e.g., *She married the bench under the barn*) and included among the target items. For global decomposability, eight nonidiomatic expressions (e.g., *She used a hammer*) were included in each list, in addition to the novel idioms included in the lists for familiarity, meaningfulness, and literal plausibility, for a total of 16 catch trials. For the global decomposability lists, idiomatic meanings were presented alongside each item; for the catch trials, meanings unrelated to the component parts of the expression were created (e.g., *She married the bench under the barn. Idiomatic meaning: She used a computer*). For the literal catch trials, literal paraphrases were included alongside those items (e.g., *She did laundry. Idiomatic meaning: She washed her clothes*). Literal catch trials were included to doubly ensure that participants were following instructions, since piloting of the global decomposability task indicated that it was more challenging than the other tasks. Participants with an average rating of 3.5 or greater over all eight catch trials for familiarity, meaningfulness, and literal plausibility were excluded from the analyses. Global decomposability participants who rated the novel idioms as decomposable 80 % or more of the time were excluded from further analyses, and those who rated the literal catch trials as nondecomposable 80 % or more of the time were likewise excluded. For all dimensions, a total 115 participants, or 0.05 %, were excluded in this way. Additional participants were recruited to replace those participants. For predictability, eight expressions containing highly predictable, literal collocations were included in each list among the targets (e.g., *The American flag is red, white, and* ___). All participants supplied the expected and felicitous completions, so none were excluded due to this criterion. The final dataset over all five norming dimensions included ratings from 2,100 total participants. For each list, the idioms provided as example items were included in other lists, so that participants only rated idioms they were encountering for the first time during the task.

Results

The descriptive norming data for each of the 870 idiomatic expressions are available for download. For familiarity, literal plausibility, and predictability, responses spanned the entire possible range (see Table 1 for descriptive statistics, and Fig. 1 for visualizations).

For global decomposability, the average rating expressed in Table 1 indicates the average rating of participants' selections of either 0, for a *nondecomposable idiom*, or 1, for a *decomposable idiom* (i.e., the closer to 0 the item's average, the more nondecomposable the item). For predictability, the average rating expressed in Table 1 indicates the proportion of participants' responses that matched the expected idiomatic completions for the items. All other responses (i.e., nontarget idiomatic completions) for each item are included in the accompanying spreadsheet. The average ratings for meaningfulness indicate that most participants were highly familiar with the subset of included idioms, suggesting that other ratings may be similarly reliable, since a high level of familiarity with the meaning of the expressions herein indicates a greater likelihood of accurately evaluating items along the subsequent dimensions, preventing responses based on conjecture alone (see Table 2 for examples of highly and lowly rated idioms from the present dataset).

As the average familiarity rating is slightly higher than the midpoint of the ratings scale, this, too, supports participants' overall familiarity with the items in the task and the subsequent reliability of the other ratings. No idioms were excluded from the database due to low familiarity or meaningfulness ratings, in an effort to provide as comprehensive a database as possible.

To examine the interrelationships among the dimensions, pairwise Spearman's rho correlations were calculated between each pair of the five dimensions (see Table 3).

Results show that familiarity and meaningfulness were highly positively correlated, with moderate positive correlations between familiarity and global decomposability, familiarity and predictability, meaningfulness and global decomposability, and meaningfulness and predictability. Literal plausibility did not correlate with any other dimension.

To compare the present normative data to previous works, Spearman's rho correlations were calculated comparing the subset of idioms in common with those in each of three previous studies: Schweigert and Cronk (1992), Titone and Connine (1994b), and Libben and Titone (2008) (see Table 4 for correlations). These studies were selected on the basis of comparability of the task and instruction type, as well as for their common focus on American English idioms (see Nordmann, Cleland, & Bull, 2014, for an investigation of British English idioms; Caillies, 2009, and Bonin, Méot, & Bugaiska, 2013, for normative data on French idioms; and Tabossi, Arduino, & Fanari, 2011, for Italian idioms). As Titone and Connine (1994b) used a 7-point Likert scale, raw scores were transformed to standard *z* scores prior to calculating correlations, in order to better compare across the studies (Colman, Norris, & Preston, 1997).

The negative correlations between the present dataset and the data from Schweigert and Cronk (1992) can be attributed to the latter study's used of a reversed scale (1 = *high*, 5 = *low*). Correlations with Schweigert and Cronk's familiarity ratings were calculated for both the present familiarity and meaningfulness dimensions, as their instructions asked participants to rate the expressions they read for how often they heard the phrase used figuratively. Although the inclusion of "often" in their instructions indicates that their task was an assessment of subjective frequency, the division in Titone and Connine (1994b) of familiarity into both familiarity and meaningfulness—and similarly in Libben and Titone's study—led the authors to question which meaning was represented in Schweigert and Cronk's data. As their familiarity ratings correlated moderately with both our familiarity and meaningfulness ratings, this suggests that, in their task, the notion of how well a participant knew the meaning of an idiomatic expression may have been at least partially conflated with how frequently one encountered the phrase. As the discussion of familiarity and meaningfulness in the literature is varied, this suggests that the two dimensions may reflect either one or separate underlying constructs. We return to test this distinction below.

Since the instructions for the present study were adapted from Titone and Connine (1994b), we assumed that the tasks employed in their study, as well as in Libben and Titone (2008)—which also used similar instructions—were comparable, and thus dimensions were only compared with their counterparts in the earlier studies. Our familiarity ratings

Table 1 Descriptive statistics for the normative data

	Familiarity	Meaningfulness	Global Decomposability	Literal Plausibility	Predictability
Average rating	3.4650	4.5628	.5427	3.1205	.5503
St. deviation	0.5718	0.3674	.2091	1.0406	.3496

Familiarity, Meaningfulness, and Literal Plausibility are measured on a Likert scale (1–5); the Global Decomposability average is based on participants' binary selections of 0 = *nondecomposable* or 1 = *decomposable* for each item; Predictability is expressed as the proportion of idiomatic completions.

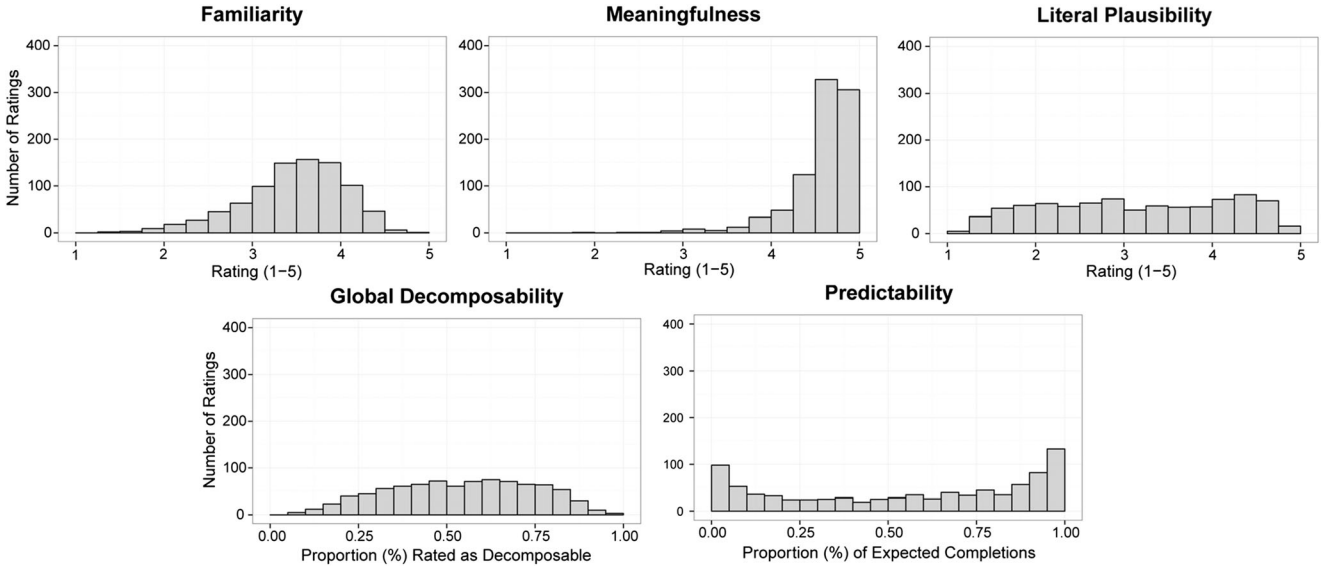

Fig. 1 Histogram plots showing the distribution of ratings for familiarity, meaningfulness, literal plausibility, global decomposability, and predictability dimensions

were strongly positively correlated with those from Titone and Connine (1994b), as were the present meaningfulness ratings with their data. A similar pattern was found when comparing the present data to the later Libben and Titone data. Since similar instructions and tasks were employed across the two studies, these findings show that within the subset of idioms shared across studies, the present ratings not only expand upon the scope of the available norming resources, but they also corroborate the ratings obtained in earlier works. The present literality ratings were very strongly positively correlated with the literal plausibility ratings in both Titone and Connine (1994b) and Libben and Titone (2008). This was expected, because although some idiomatic expressions have lost their literal plausibility over time (e.g., *kick the bucket*), this does not appear to be the case for the idioms common to the three datasets over the last two decades. Similarly, our predictability ratings were also positively correlated with those in the two earlier works, despite being somewhat lower. This may have been due to the much larger sample size of participants polled in our study, which may have allowed for more variability of responses. Although, overall, this suggests general corroboration among the dimensions between the present and previous studies, the implications of any differences in the ratings are discussed below.

After observing the relatively strong correlations between familiarity and meaningfulness, we questioned again the

Table 2 Examples of idioms rated as extremely low/high or predictable; actual ratings are in parentheses

	Low		High	
Familiarity	Go pear-shaped	(1.44)	Be no big deal	(4.77)
	Be as tough as boots	(1.50)	Drive someone crazy	(4.62)
	Upset the applecart	(1.60)	Be one of those days	(4.52)
Meaningfulness	Be all-singing, all-dancing	(2.81)	Be a piece of cake	(4.95)
	Be cannon fodder	(2.96)	Cost an arm and a leg	(4.94)
	Be a mutual admiration society	(2.97)	Burst someone's bubble	(4.92)
Global decomposability	Play chicken	(.08)	Sit up and take notice	(.98)
	Be a dark horse	(.09)	Go back to basics	(.97)
	Be the bomb	(.09)	Outstay your welcome	(.96)
Literal plausibility	Be the apple of someone's eye	(1.11)	Keep your eyes open	(4.83)
	Go bananas	(1.14)	Pat someone on the back	(4.82)
	Love someone to pieces	(1.25)	Get your hands dirty	(4.86)
Predictability	Go ballistic	(.00)	Vanish into thin air	(1.00)
	Be a riot	(.00)	Have a bad hair day	(1.00)
	Find something hard to stomach	(.00)	Be a couch potato	(1.00)

Table 3 Correlation matrix with Spearman's rho calculations comparing five dimensions

	Familiarity	Meaningfulness	Global Decomposability	Literal Plausibility	Predictability
Familiarity	1.0000	–	–	–	–
Meaningfulness	.8405	1.0000	–	–	–
Global decomposability	.3101	.2047	1.0000	–	–
Literal plausibility	–.0001	–.0586	.1185	1.0000	–
Predictability	.2842	.3183	.0480	–.0431	1.0000

varied treatments of these two constructs in the previous literature (e.g., Libben & Titone, 2008; Schweigert & Cronk, 1992; Titone & Connine, 1994b). Although it is perhaps a theoretical possibility that a participant's impression of an idiom's frequency in the language can be distinct from their knowledge of its meaning, it is quite possible that these similarly subjective metrics tap into the same underlying psychological construct—that of overall prior experience with an idiom. Specifically, it may be that when a person understands an idiom's meaning, this necessitates prior exposure to the idiom in a figurative context. Whether a person can truly judge the frequency of an idiom in their language, however, as compared to other idioms in the language, has yet to be seen (see, e.g., Deignan, 2006; Thibodeau & Durgin, 2008, for previous work suggesting a dissociation between subjective familiarity and corpus frequency in metaphor). To this end, we further investigated these two dimensions to determine whether they were tapping into the same underlying construct. Whereas familiarity and meaningfulness ratings might be seen as a subjective metric of our participants' experience with the present set of idioms, corpus data provide an objective means with which to compare subjective familiarity ratings with the actual frequency of the items in the language. If, in fact, participants' familiarity and/or meaningfulness ratings reflect their experience with the items in the language, we would expect to find a strong correlation between these ratings and frequencies from a corpus.

To this end, objective frequency information for each of the 870 idioms was obtained using the Corpus of Contemporary American English (COCA; Davies, 2008-). The data drawn from COCA are based on frequency estimates in written materials collected from 1990 onward. For our purposes, the COCA data provide an objective metric of frequency of occurrence in a large sample of American English, and they should provide a reasonable approximation of native speakers' experience with each idiom. More specifically, using corpus data can help us identify whether an idiom's meaningfulness aligns more with an individual's subjective impression of familiarity, or rather is more closely related to more objective corpus frequency metrics of the item.

All versions of each idiom were included in the raw count data, such that no specific pronouns or tenses of the verbs limited the search. By utilizing COCA's speech tags in the syntax of our searches, the raw count data obtained were as comprehensive as possible. For example, for the idiom *hang someone out to dry*, the query syntax used was "[hang].[v*] [ppho1] out to dry" to permit inclusion of all phrasings of the idiom. This syntax provides raw counts of the number of times this idiom appears in the corpus, starting with any form of the specified verb ("hang"), followed by a third-person singular pronoun (i.e., "him," "her"), followed by the exact completion "out to dry" (see Table 5 for the correlations between the COCA frequencies and the five dimensions collected in the norming). Similarly, by replacing [ppho1] with [ppho2] in the search, we included instances using third-person plural pronouns (i.e., "them") in the raw count data as well. Additionally, literal uses of the idioms were not separated from idiomatic uses. Because the goal was to extract the frequency information of the *n*-grams, we included all uses of the target expressions in our count data. These raw counts for each idiom obtained from COCA were later log-transformed; all further analyses here were conducted using the log-transformed data.[1]

The correlations between the COCA frequencies and the familiarity and meaningfulness data were weaker than the Spearman's correlation calculated between the present familiarity and meaningfulness data (Spearman's correlation = .8405), which supports the notion that familiarity and meaningfulness may tap into the same underlying construct, which is distinct from corpus frequency.

To further investigate whether the high correlations between meaningfulness and familiarity are indicative of them being related to one underlying construct, and additionally whether these two variables were merely reflections of corpus frequency, we conducted a pair of exploratory principal component analyses (PCAs). We first fit a three-component PCA using varimax rotation on the familiarity, meaningfulness, and COCA frequency data, to better characterize the underlying structure of the data. First, all Pearson correlations were above .3, suggesting reasonable factorability (see Table 6 for the correlations used in the PCA).[2]

[1] Note that the log transformation would not impact the Spearman correlations reported here, as they work on ranked data. However, the log transformation was performed to reduce the impact of outliers in the principal component analyses reported below.

[2] These denote the Pearson correlations used in the PCA, not the Spearman's correlations conducted earlier to investigate relationships among the norming dimensions.

Table 4 Spearman's rho correlations between the previous and present norming datasets

	n	Familiarity	Meaningfulness	Global Decom.	Lit. Plausibility	Predictability
Schweigert & Cronk (1992)						
Familiarity	92	–.5859	–.5280	–	–	–
Titone & Connine (1994b)						
Familiarity	50	.7723	–	–	–	–
Meaningfulness		–	.6104	–	–	–
Literality		–	–	–	.9010	–
Predictability		–	–	–	–	.6729
Libben & Titone (2008)						
Familiarity	46	.7454	–	–	–	–
Meaningfulness		–	.6840	–	–	–
Global decom.				.8642		
Literality		–	–	–	.9246	–
Predictability		–	–	–	–	.6361

n = Number of idioms in common between the previous and present datasets, included in the correlations

Next, the Kaiser–Meyer–Olkin measure of sampling adequacy was .551, and Bartlett's test of sphericity was significant [$\chi^2(3) = 1{,}244.31$, $p < .001$], further suggesting reasonable factorability of the data. Three components were extracted to test for latent relationships among familiarity, meaningfulness, and corpus frequency, and whether these three dimensions could be reduced to either one or two latent constructs.

The initial eigenvalues showed that the first component accounted for 69.33 % of the variance, the second component for an additional 25.73 %, and the third component for the remaining 4.95 % of the variance. The eigenvalue for the first component was 2.080, well over a threshold of 1, whereas the eigenvalues for Components 2 and 3 were 0.772 and 0.148, respectively. Inspection of the component loadings (see Table 7) shows that each measure correlated strongly with separate components. However, meaningfulness and familiarity each cross-loaded moderately onto the other's component.

On the one hand, this suggests some independence of meaningfulness and familiarity as separable, but related, constructs. However, the amount of residual variance accounted for by Component 3 (i.e., familiarity) was very small. This suggests that a two-component solution may provide a better fit to the data. In the two-component solution, Component 1 accounted for 60.29 % of the variance, and Component 2 accounted for an additional 34.83 % of variance (see Table 8).

As can be seen in Table 8, both meaningfulness and familiarity showed strong correlations with Component 1, whereas corpus frequency showed a low correlation with this component. Instead, COCA frequency showed a strong correlation with Component 2, which has smaller correlations with both meaningfulness and familiarity. This suggests two things. First, the meaningfulness and familiarity ratings that we (and others) obtained may tap a single latent construct—namely, subjective familiarity. Second, the latent subjective familiarity construct is not necessarily a direct reflection of the frequency of occurrence of an idiom in a language (cf. Deignan, 2006; Thibodeau & Durgin, 2008).

Discussion

In the present work, we conducted large-scale norming on the largest subset of idioms in the literature to date, to expand upon the scope of prior norming studies and provide a large-scale database of norms for 870 idioms. This dataset will provide a valuable resource to the research community, specifically to others interested in investigating psycholinguistic aspects of idiom comprehension.

The correlations found in this study show that, among the five dimensions—familiarity, meaningfulness, global decomposability, literal plausibility, and predictability—an idiom's familiarity and meaningfulness are highly interrelated.

Table 5 Correlation matrix with Spearman's rho calculations comparing five dimensions to objective frequencies obtained from COCA

	Familiarity	Meaningfulness	Global Decomposability	Literal Plausibility	Predictability
COCA	.3920	.2759	.2388	.1442	–.1280

Table 6 Correlation matrix showing Pearson correlations ($N = 870$)

	Familiarity	Meaningfulness	COCA
Familiarity	1.000	.842	.415
Meaningfulness	.842	1.000	.300
COCA	.415	.300	1.000

Although both of these rating dimensions relate to different aspects of subjective familiarity, our PCA analysis showed that they may reflect a single underlying latent construct, which, interestingly, is largely separable from measures of corpus frequency. This dissociation between the frequency of occurrence of an idiom and raters' subjective familiarities with that idiom suggests that a further mediating variable, not measured here, may link frequency with meaningfulness and familiarity in language users' minds. Note that we fit additional exploratory PCAs that included all of our measured variables, to test for a possible additional related metric in our dataset, in the hope that we might explain this discrepancy between frequency and subjective familiarity (as well as to see whether the rating dimensions could be further reduced to a smaller set of constructs). Inspection of the eigenvalues and scree plots suggested either a two- or three-component solution for the dataset; however, these components were unstable, and models sometimes did not converge. Thus, it is likely that variables that went unmeasured here may provide the link between frequency and subjective familiarity. For example, it is likely that, given the diversity represented in the Mechanical Turk user base, our rater group represented a broad range of language experience, print exposure, and dialectal variation. Furthermore, many idioms in our dataset are arguably more colloquial than others (i.e., compare *love someone to bits* to *stay on track*), so that some forms may be more suited to written presentation than others. Future researchers may wish to quantify participant-level traits, such as print exposure, dialect, and language experience, in order to identify whether any of these factors may mediate the lack of a clear association that we observed between corpus frequency and the subjective familiarity of an idiom.

Aside from the correlation between familiarity and meaningfulness, other intercorrelations among the dimensions were only moderate; this underscores the necessity of properly norming experimental items along these dimensions to ensure that all variables are sufficiently controlled for. Global decomposability correlated positively with both familiarity and meaningfulness. This relationship may suggest that when an idiomatic expression is frequently encountered and its meaning is known to the reader, the level of semantic relatedness between an idiom's component parts and the meaning of the string may support comprehension and subsequent retention of the string in the lexicon. Alternatively, this relationship may also suggest that idioms whose meanings are relatively

Table 7 Rotated component correlations based on a principal component analysis using varimax rotation ($N = 870$)

	Component 1	Component 2	Component 3
Meaningfulness	.91	.13	.39
COCA	.12	.98	.16
Familiarity	.55	.23	.81

semantically transparent are used more frequently, due to their more explicit semantic relatedness. On the other hand, because the correlations are only moderate, this suggests, too, that while familiar idioms whose meanings are understood and well-known can be decomposable, this is not always the case. To exemplify, from the present dataset, *cross your fingers* has an average familiarity rating of 4.33, a meaningfulness rating of 4.81, and a global decomposability rating of just .45.

Predictability correlated positively with both familiarity and meaningfulness. This relationship may suggest that when an idiomatic expression is frequent and its meaning is known to the reader, its completion is more easily predicted than in other when a phrase is infrequent or its meaning opaque. It may also be the case, however, that although a variety of expressions in this study were rated as frequent and meaningful, an idiom that is considered infrequent may still be considered predictable, and its idiomatic completion may still be easily supplied; *be a balancing act* is an example of such an idiom from the present dataset. Although participants provided the correct completion for this idiom 96 % of the time, its average familiarity rating was just 3.18. Although this is a moderate familiarity rating, it was very highly predictable, and is perhaps illustrative of the potential relationship between an idiom's familiarity and its predictability. To further investigate the contributions to an idiom's predictability, Spearman's rho correlations were calculated between each of the five main norming dimensions and idiom length, as measured by the number of words in the idiomatic string (see Table 9 for the correlations between idiom length and each of the five normed dimensions).

Although idiom length did not correlate with any other norming dimension, we did observe a moderate correlation of .4736 between idiom length and predictability, or its cloze probability, as defined above. This suggests that

Table 8 Rotated component correlations based on a principal component analysis using varimax rotation ($N = 870$)

	Component 1	Component 2
Meaningfulness	.96	.11
Familiarity	.92	.26
COCA	.19	.98

Table 9 Spearman's rho correlations between idiom length and each of five norming dimensions

	Familiarity	Meaningfulness	Decomposability	Lit. Plausibility	Predictability
Length	.0203	.0474	–.0230	.0211	.4736

although longer idioms are not more familiar, meaningful, decomposable, or literally plausible, they are more predictable. This additionally suggests that an idiom's length accounts for a moderate amount of variance in its predictability, over and above the other dimensions we investigated. We interpret this as support for the relationship between context and the degree to which upcoming information can be predicted. Specifically, the moderate correlation reported here suggests that the more information is available prior to the idiomatic completion, the earlier it can be preactivated from the lexicon. This is in line with other work in the broader domain of literal sentence comprehension—specifically, findings from work using event-related brain potentials (ERPs), showing that linear position within a sentence is highly predictive of a word's N400 amplitude. That is, with more preceding context, words become more predictable, and easier to access and integrate into a context (e.g., Payne, Lee, & Federmeier, 2015; Van Petten & Kutas, 1991). Our study shows that length and linear position are important for lexical prediction in idiom comprehension, as well.

Literal plausibility correlated with none of the other dimensions. This is perhaps due to the fact that literal plausibility, relative to the other dimensions, is not as subjective. Although varying amounts of prior experience may explain a person's low rating of *be off one's rocker*, this does not affect the plausibility of the literal interpretation. Literal plausibility did strongly correlate with previous norming studies, and this was expected. This supports the idea that, despite the diachronicity of language, it is unlikely that expressions would become opaque in just 20 years; at least for the idioms in common between the present dataset and the earlier data, this appears to be the case.

Our norms additionally corroborate and extend the previous norms obtained in earlier works. Schweigert and Cronk's (1992) familiarity norms correlated moderately with both the present familiarity and meaningfulness norms. The present familiarity and meaningfulness norms were also highly positively correlated with those of Titone and Connine (1994b), and even more highly with those of Libben and Titone (2008). The moderate correlations between the present norms and these earlier data for some of the dimensions are potentially explained by the differences in location: The norms obtained in Titone and Connine (1994b) were collected in New York, Libben and Titone collected their dataset in Montreal, Canada, and the present norming comprises ratings from participants all over North America. What is more, Libben and Titone's work was conducted in a region where the majority of the participants were bilingual.[3] It is very possible that any differences in the ratings obtained by each of these works are due to the various cultural influences of the diverse groups of the participants.

Finally, the presently obtained predictability ratings correlated positively with the previous norms, albeit less positively than did the other dimensions discussed above. A closer look at the data suggests that some of the common expressions were completed more often with the expected completion in the present data than in previous data. For example, in the present norms, *to smell a rat* was completed as expected 51 % of the time, whereas in Libben and Titone (2008), the expression was completed as expected only 11 % of the time. Similarly, comparing the same two datasets, from the present norms, *to let off steam* was completed idiomatically 86 % of the time, and in the 2008 dataset, just 14 % of the time. Although other expressions had similar cloze probabilities between the two studies, this suggests that perhaps time or participant variability may have had effects on the predictability of the targets. Indeed, as was mentioned above, the context of the 2008 norming study in Montreal, where bilingualism is the norm, may have led to slightly disparate ratings, while still maintaining a positive correlation.

A potential limitation of the present work is the manner of presentation of the items to participants. Although it was our intention to present the idioms in neutral contexts, no context is truly neutral. For example, items like *mean the world to someone* are less acceptable without the pronoun completion, requiring an additional lexical item as context. In isolation, *mean the world* has an implicit reading of meaning the world to a particular person or group. In the norming, expressions like this and others were presented to participants as *It meant the world to* _____, where pronoun completions were to be expected. Although this does not entail an entirely neutral context, it is minimal, and it is likely that for other items a richer context would have made the expressions sound more familiar to the participants, perhaps altering their ratings. Users of the available norms are cautioned about this point, since the predictability of items such as these may be higher than for idioms whose expected completion is something more specific (e.g., *be a riot*).

[3] We thank an anonymous reviewer for pointing this out.

While it was a primary goal to include as many expressions as possible, there are still other idioms that have not been normed, either in this study or in previous efforts. As language is always changing and new expressions are coined daily, it is beyond the scope of this work to include every possible idiomatic expression. Further, many of the expressions adapted from Lee (2000) were of a similar structure (e.g., *She hit the wall*, *He smelled a rat*), while others were longer in length (e.g., *She had eyes in the back of her head*). We felt it prudent to include a variety of structures in the present dataset, since idioms as a class do not conform to a short structure. Furthermore, in the interest of providing as comprehensive a dataset as possible, we did not exclude idioms on the basis of sentence length. With that in mind, however, longer expressions like these would certainly be more predictable, as the more words in an expression, the fewer potential candidates for a felicitous completion. There are undoubtedly implications of including idioms of varying lengths in a single experiment, and we caution users of this work to bear this in mind. What the present dataset will be of use for, however, is the availability of these measures for a variety of types of idioms. Different labs will have different research questions, and will ultimately require idioms with different characteristics. The scope of this project will facilitate this, and we hope that its availability will encourage further endeavors in research on idiom processing. Although different groups of participants will differ with respect to their exposure to any given set of idioms, the diverse set of ratings presented here, we hope, will be of use to a broad range of researchers for use in stimulus construction, or for use in multivariate statistical models when analyzing data on idiom processing. Whereas previous analyses using ANOVA approaches had no place for such metrics, statistical techniques such as hierarchical linear models are becoming increasingly popular in the field, allowing the inclusion of more continuous covariates and interactive predictors, such as the dimensions normed within the present study as well as others, including work investigating idiom comprehension (e.g., Titone & Libben, 2014).

Ultimately, idioms are multifaceted expressions, as has been demonstrated by this work and by other studies that have come before it. Although they present an interesting focus for psycholinguistic research, idioms also present researchers with the challenge of accounting not only for traditional concerns of character count and sentence length, but also for the dimensions described here. It is our hope that the public dissemination of these data will facilitate future endeavors within the psycholinguistics of idiom processing.

Author note This research was supported in part by a University of Illinois Campus Research Board Grant RB14158, awarded to D.T.. We thank the members of the Electrophysiology and Language Processing Lab at the University of Illinois at Urbana-Champaign for assistance with data entry and quantification, and Joseph Roy for helpful discussions about data analysis. We also thank the audience of the Illinois Language and Linguistics Society 7 conference for helpful comments on an earlier version of this work.

Appendix 1: Instructions for Familiarity rating task (adapted from Libben & Titone, 2008)

In this task, you'll read a series of short sentences, each of which contains an idiom (i.e., a fixed expression). For each idiom, you will be asked to rate how frequently you have seen, heard or used the idiom. Make your decision using the 1-to-5 scale provided, with a 5 signifying that you see or hear the idiom very frequently, and 1 signifying you have never or almost never heard or seen the idiom. A rating of 3 at the midpoint of the scale would indicate that you have come across the idiom moderately often. Please use the full range of the scale in making your decisions.

For each idiom you read, please make a judgment on the idiom itself, not on the entire sentence. For example, in judging "He was on cloud nine," make your judgment about the frequency of the idiom "to be on cloud nine."

Appendix 2: Instructions for Meaningfulness rating task (adapted from Libben & Titone, 2008)

On the following pages you will read a series of sentences containing idioms. For each of the idioms, you will rate them on a scale of 1 to 5, depending on how well you know the figurative, non-literal meaning of the idiom. A rating of 1 would mean that you have absolutely no idea what the idiom means. A 3 means that you are moderately certain of what it means, and a 5 would indicate that you are 100 % certain of the idiom's meaning and could easily put it into your own words.

Some of the idioms' meanings may not stem from the meanings of the idiom's individual words. For example, to say, "He was on cloud nine," does not literally mean that someone was on a numbered cloud, but rather than he was happy or elated. If you were familiar with this non-literal meaning of the phrase, you'd give it a high rating in this survey, like 4 or 5. If, however, you didn't know this non-literal meaning, you'd give it a low rating, like 1 or 2. If you were somewhat familiar, you would give it a rating of 3.

Some of the idioms in the list may have a plausible literal meaning in addition to the figurative meaning. For example, "to smell a rat" could either mean literally smelling a rat or it could figuratively mean to detect a lie. You should judge

whether you know the figurative meaning of the idiom or not, not the literal meaning, if there is one.

Please try to use the full scale when making your judgments.

Appendix 3: Instructions for Global Decomposability rating task (Libben & Titone, 2008)

In this task, you'll be asked to make a judgment as to the "decomposability" of various idiomatic phrases.

Decomposable idioms are defined as phrases whose individual components contribute to their overall meanings. Idioms whose individual words do not make such a contribution are called nondecomposable.

An example of a decomposable idiom would be the phrase "cover up your tracks," which has two words whose meanings are related, in their literal sense, to the idiomatic meaning. The word "cover" is closely related to the idea hide, while the word "tracks" refers to evidence of your actions. Another example of a decomposable idiom is "can't believe my ears," where again it is apparent how the individual words map onto the figurative meaning (unable to believe what is being said).

An example of a nondecomposable idiom would be "be the cat's whiskers." This idiom means to "be the best" and would be called nondecomposable because the individual word meanings do not directly relate to the overall meaning of the idiom.

Please read each phrase and its associated literal meaning (presented in parentheses) and then indicate, by checking the correct box, whether you feel the phrase is decomposable or nondecomposable. For example, you would check the box for "decomposable" for the phrase "cover up your tracks," as it is very decomposable, and you would check the box for "nondecomposable" for "be the cat's whiskers," as it is completely nondecomposable.

Appendix 4: Instructions for Literal Plausibility rating task (Libben & Titone, 2008)

For each of the idioms on the following pages you will need to make a literality judgment. While all of the idioms have a meaningful idiomatic or figurative interpretation, only some of them have a well-formed literal meaning. For example, the idiom "let the cat out of the bag" figuratively means, to reveal a secret and literally means to release a cat. However, the idiom "to give the cold shoulder," which figuratively means to ignore, does not have a clear literal meaning (if any at all) as compared to let the cat out of the bag (e.g., it is unclear what it means to literally give a cold shoulder).

Your task in rating these idioms is to decide if there is a possible literal interpretation, and if so, how plausible it is on a 5-point scale. That is, rate the idioms based on how likely the literal meaning of the phrase is (if you believe one exists). A rating of 1 would indicate that an idiom definitely does not have any possible literal interpretation and therefore is completely implausible literally. A rating of 5 would indicate that the idiom definitely has a clear and well-formed literal interpretation that is very plausible. Intermediate values of the scale should reflect your judgments of the plausibility of these phrases interpreted literally.

Please try to use the entire scale when doing your ratings.

Appendix 5: Instructions for Predictability rating task (expanded from Libben & Titone, 2008)

On the following pages you will read a series of sentence fragments. Your task is to complete these sentences with the first word that comes to mind and to write your answer in the box provided after each fragment. For example, you may get an incomplete sentence such as *The boy swung the* _____. In this case, the first word that might come to mind is bat. If this were so, you would write the word bat in the box provided. This would complete the phrase as *The boy swung the bat*.

You should complete the fragment with just a single word (e.g., table) or a compound noun (e.g., greenhouse).

References

Bonin, P., Méot, A., & Bugaiska, A. (2013). Norms and comprehension times for 305 French idiomatic expressions. *Behavior Research Methods, 45,* 1259–1271. doi:10.3758/s13428-013-0331-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904–911. doi:10.3758/s13428-013-0403-5

Cacciari, C., & Glucksberg, S. (1991). Understanding idiomatic expression: The contribution of word meanings. In G. Simpson (Ed.), *Understanding word and sentence* (pp. 217–240). Amsterdam, The Netherlands: Elsevier Science.

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language, 27,* 668–683. doi:10.1016/0749-596X(88)90014-9

Caillies, S. (2009). Descriptions de 300 expressions idiomatiques: Familiarité, connaissance de leur signification, plausibilité littérale, «décomposabilité» et «prédictibilité». *L'Année Psychologique, 109,* 463–508. doi:10.4074/s0003503309003054

Campbell, S. J., & Raney, G. E. (2016). A 25-year replication of Katz et al'.s (1988) metaphor norms. *Behavior Research Methods, 48,* 330–340. doi:10.3758/s13428-015-0575-2

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports, 80,* 355–362.

Colombo, L. (1993). The comprehension of ambiguous idioms in context. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 163–200). Hillsdale, NJ: Erlbaum.

Cronk, B. C., Lima, S. D., & Schweigert, W. A. (1993). Idioms in sentences: Effects of frequency, literalness, and familiarity. *Journal of Psycholinguistic Research, 22,* 59–82.

Davies, M. (2008). The Corpus of Contemporary American English: 520 million words, 1990–present [Database]. Available at http://corpus.byu.edu/coca/.

Deignan, A. (2006). The grammar of linguistic metaphors. In A. Stefanovich & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 106–122). Berlin, Germany: Mouton.

Ellis, N. C. (2002). Frequency effects in language acquisition: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24,* 143–188.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition, 8,* 149–156. doi:10.3758/BF03213418

Gibbs, R. W., Jr., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21,* 100–138. doi:10.1016/0010-0285(89)90004-2

Goldberg, A. (2007). Learning linguistic patterns. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 47, pp. 33–63). San Diego, CA: Elsevier Academic Press.

Lee, W. (Ed.). (2000). *Longman pocket idioms dictionary.* Essex, UK: Pearson Education.

Libben, M. R., & Titone, D. A. (2008). The multidetermined nature of idiom processing. *Memory & Cognition, 36,* 1103–1121. doi:10.3758/MC.36.7.1103

Nordmann, E., Cleland, A. A., & Bull, R. (2014). Familiarity breeds dissent: Reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability. *Acta Psychologica, 149,* 87–95. doi:10.1016/j.actpsy.2014.03.009

Nunberg, G. (1978). *The pragmatics of reference*. Bloomington, IN: Indiana University Linguistics.

Payne, B. R., Lee, C., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52,* 1456–1469. doi:10.1111/psyp.12515

Schweigert, W. A. (1986). The comprehension of familiar and less familiar idioms. *Journal of Psycholinguistic Research, 15,* 33–45.

Schweigert, W. A., & Cronk, B. C. (1992). Figurative meanings and the likelihood of literal meanings among U.S. college stuents. *Current Psychology: Research & Reviews, 11,* 325–345.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18,* 523–534. doi:10.1016/S0022-5371(79)90284-6

Tabossi, P., Arduino, L., & Fanari, R. (2011). Descriptive norms for 245 Italian idiomatic expressions. *Behavior Research Methods, 43,* 110–123. doi:10.3758/s13428-010-0018-z

Thibodeau, P., & Durgin, F. H. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language, 58,* 521–540. doi:10.1016/j.jml.2007.05.001

Titone, D. A., & Connine, C. M. (1994a). Comprehension of idiomatic expressions: Effects of predictability and literality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1126–1138. doi:10.1037/0278-7393.20.5.1126

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

Titone, D. A., & Connine, C. M. (1999). On the compositional and noncompositional nature of idiomatic expressions. *Journal of Pragmatics, 31,* 1655–1674.

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

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition, 19,* 95–112. doi:10.3758/BF03198500

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.