COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12885

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12885

Meaningfulness Beats Frequency in Multiword Chunk Processing

Hajnal Jolsvai,[a] Stewart M. McCauley,[b] Morten H. Christiansen[c]

[a]*Oracle Cloud Infrastructure*
[b]*Department of Communication Sciences and Disorders, University of Iowa*
[c]*Department of Psychology, Cornell University*

Received 9 July 2019; received in revised form 7 July 2020; accepted 20 July 2020

Abstract

Whereas a growing bulk of work has demonstrated that both adults and children are sensitive to frequently occurring word sequences, little is known about the potential role of meaning in the processing of such multiword chunks. Here, we take a first step toward assessing the contribution of meaningfulness in the processing of multiword sequences, using items that varied in chunk meaningfulness. In a phrasal-decision study, we compared reaction times for triads of three-word sequences, corresponding to idiomatic expressions, compositional phrases, and phrasal fragments, while controlling for phrase and substring frequencies. Chunk meaningfulness, as assessed by a separate subjective rating study, was found to speed up decision times for all three types of strings: The more meaningful a multiword sequence was judged to be, the faster it was processed, independently of whether it was idiomatic, compositional in nature, or a phrasal fragment. These results highlight the importance of taking meaning into account when considering the processing of multiword chunks, consistent with predictions of construction-based approaches to language.

Keywords: Multiword sequences; Chunking; Distributional statistics; Usage-based approach; Meaningful chunks; Cognitive linguistics; Constructions; Idioms

1. Introduction

Words have long been recognized as building blocks of our linguistic abilities, but recently multiword sequences have also been proposed to play a similar key role in language acquisition and processing (see contributions in Arnon & Christiansen, 2017;

Correspondence should be sent to Morten H. Christiansen, Department of Psychology, Cornell University, 228 Uris Hall, Ithaca, NY 14853. E-mail: christiansen@cornell.edu

All three authors contributed equally to this study and its write-up.

A preliminary version of the paper was presented at the 35th Annual Conference of the Cognitive Science Society, Berlin, Germany.

Christiansen & Arnon, 2017, for a review). For example, developmental studies have shown that children as young as 2 and 3 years of age are sensitive to the frequency of multiword chunks when imitating four-word phrases (e.g., Bannard & Matthews, 2008). Four-year olds are more accurate in producing irregular plurals when they occur inside frequent multiword chunks (such as *Brush your teeth*; Arnon & Clark, 2011). In the same vein, adults show processing advantages for frequent multiword chunks across different experimental paradigms, including phrasal decisions (Arnon & Snider, 2010), self-paced reading (Reali & Christiansen, 2007), sentence recall (Tremblay, Derwing, Libben, & Westbury, 2011), and event-related potentials (Tremblay & Baayen, 2010). Adult language production is also affected by multiword chunk frequency, both onset latencies (e.g., Janssen & Barber, 2012) and duration (e.g., Arnon & Cohen-Priva, 2013). And, strikingly, similar to the Age-of-Acquisition effects observed for individual words, whereby early acquired words are retrieved faster (e.g., Ellis & Morrison, 1998), Arnon, McCauley, and Christiansen (2017) demonstrated that multiword chunks acquired early in life are processed faster in adulthood.

When it comes to single words, though, studies have shown that, in addition to frequency (e.g., Hall, 1954), meaning plays a key role in accessing and processing individual lexical items (e.g., Balota, 1990; Yap, Tan, Pexman, & Hargreaves, 2011; see Taylor, Duff, Woollams, Monaghan, & Ricketts, 2015, for a review). If multiword chunks are building blocks of language on par with individual words, we would expect that meaning should also affect their processing. Indeed, this is what would be predicted by construction-based approaches to language, not only by perspectives grounded in cognitive linguistics (e.g., Bybee, 2006; Goldberg, 2006; Wray, 2002, 2017) but also by those building on generative grammar (e.g., Culicover, Jackendoff, & Audring, 2017). Accordingly, along with distributional information, language users may also utilize the degree to which a multiword sequence conveys a coherent communicative meaning as a whole (Dąbrowska, 2014; Fillmore, Kay, & O'Connor, 1988; Langacker, 1987). Here, we therefore take a first step toward exploring the role of meaning in the processing of multiword sequences, including ones where meaning is normally viewed as being important: idiomatic expressions.

Traditionally, idioms, such as *chew the fat*, have been treated as a special case because their figurative meanings go beyond what would be expected from the semantics of their component words (Jackendoff, 1997). Indeed, idioms have been suggested to be retrieved directly from memory similar to individual words (Chomsky, 1980; Pinker, 1999) and therefore processed faster than compositional multiword phrases, which were thought to be generated "on the fly." This idea of idioms as stored units was supported by studies finding processing advantages for idiomatic expressions compared to similar phrases. For example, studies involving phrasal decision (Swinney & Cutler, 1979) and self-paced reading (Conklin & Schmitt, 2008) found faster processing for idioms compared to similar control strings (e.g., *break the ice* vs. *break the cup*). Similarly, eye-tracking studies have observed a processing advantage for idiomatic expressions (e.g., *at the end of the day*) over comparable control phrases (e.g., *at the end of the war*; Siyanova-Chanturia, Conklin, & Schmitt, 2011; Underwood, Schmitt, & Galplin, 2004).

However, a common limitation of the studies supporting the "lexical representation hypothesis" for idiomatic expressions (Swinney & Cutler, 1979) is that whole-string and substring frequencies of the stimuli were not controlled in a systematic way. For example, Swinney and Cutler (1979) replaced one word in the idiom with a similar or higher frequency item to create controls, but the bigram and whole-phrase frequencies of the sequences were not controlled for. Moreover, meaningfulness was not a factor in these studies, making it possible that the idioms were inherently more meaningful than the controls. This raises the question of whether idioms will still have a processing advance over comparable multiword phrases, if distributional information is adequately controlled across items? If not, then perhaps the overall meaningfulness of a multiword sequence might be the primary factor in determining ease of processing?

To determine the potential role of meaning in the processing of multiword chunks, we employed a phrasal decision task (Arnon et al., 2017; Arnon & Snider, 2010; Swinney & Cutler, 1979) to examine reaction times for responses to three different kinds of three-word sequences: (a) idioms, which have traditionally been viewed as forming meaningful (possibly non-compositional; Swinney & Cutler, 1979) units (e.g., *of two minds*); (b) compositional phrases, whose meaning would generally have been seen as deriving from the particular combination of their parts (e.g., *a bad attitude*); and (c) phrasal fragments that did not align with syntactic phrase-level boundaries (e.g., *for all practical*). Accordingly, we created matched triads of idioms, phrases, and fragments, controlling for whole-string and substring frequency. Prior to the phrasal decision study, we conducted three subjective rating studies to further control the stimuli, assessing the plausibility, idiomaticity, and meaningfulness of our items. By design, all items were chosen to be equally plausible (idioms = phrases = fragments) but to differ in their idiomaticity (idioms > phrases > fragments) and meaningfulness (idioms = phrases > fragments).

Given the centrality of meaning to construction-based approaches to language (e.g., Dąbrowska, 2014; Goldberg, 2006; Wray, 2002), we hypothesized that the overall meaningfulness of a multiword sequence would be a key factor in determining how easy it is to process. The degree of meaningfulness should therefore predict reaction times for our three-word sequences across all three trigram types. Moreover, because we controlled for whole-string and substring frequency information in idioms and compositional phrases, we predicted no difference in reaction times between these two item types when they are equally meaningful. Finally, the prior results showing frequency effects in the processing of multiword sequences (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008) would further predict that we should obtain a main effect of whole-phrase frequency.

2. Methods

2.1. Participants

Four separate groups of monolingual American-English-speaking Cornell undergraduates participated for extra credit. We based our sample size on previous studies using

comparable statistical analysis (e.g., Snider & Arnon, 2012). Prior to data collection, we decided to test at least 30 participants in each subjective rating study and 40 participants in the phrasal decision study. Thus, we had 33 participants in the Plausibility Rating study, 46 participants in the Idiomaticity Rating study, and 33 participants in the Meaningfulness Rating study. We omitted data from any participant whose overall performance fell below 80% in a random memory recall task (see below; $n = 2$ in the Plausibility Rating study; $n = 13$ in the Idiomaticity Rating study; $n = 2$ in the Meaningfulness Rating study). After exclusions, 31 participants rated item Plausibility, 33 participants rated item Idiomaticity, and 31 participants rated item Meaningfulness. In all, 40 students participated in the Phrasal Decision Study. All experimental protocols were approved by the Cornell University Institutional Review Board for Human Participants.

2.2. Materials

We first extracted three-word sequences (trigrams) from a combination of the American national corpus (ANC; Reppen, Ide, & Sudeman, 2005) and the Fisher corpus (Cieri, Graff, Kimball, Miller, & Walker, 2004, 2005), containing a total of 39 million words of American English. The Fisher corpus consists of spoken language (telephone conversations), while the ANC consists of spoken as well as written texts. We then selected all three-word idiomatic expressions appearing in the following collections: McGraw-Hill's Essential American Idioms Dictionary (Spears, 2008), The Handbook of Commonly Used American Idioms (Makai, Boatner, & Gates, 1991), the IdiomQuest (http://www.idiomquest.com), and American Idioms (http://www.americanidioms.net) online idiom dictionaries. In all, 82 three-word idiomatic expressions from these collections appeared in the combined ANC/Fisher corpus.

Next, for each idiomatic expression (e.g., *play the field*), we picked frequency-matched compositional phrases (e.g., *nothing to wear*) and frequency-matched fragments (e.g., *without the primary*). Both the phrases and the fragments were frequency-matched to corresponding idioms such that the trigram frequency, first bigram, second bigram, first unigram, second unigram, and third unigram frequencies were within $\pm 10\%$ of the corresponding idiom's frequencies, respectively.[1] Log_2 transformation was applied to all raw phrase and substring frequencies prior to this selection process. The resulting preliminary set of items consisted of 82 idioms, 236 phrases, and 218 fragments. Table 1 shows the results of the individual ANOVA tests of the phrase and substring frequencies across idioms, phrases, and fragments. Any further minute differences between the frequencies of the tokens of each triad were controlled for through the linear mixed-effects (LME) analyses.

2.2.1. Subjective rating studies

To further inform the selection of item triads for the phrasal decision experiment, we conducted three separate subjective rating studies involving the previously selected 82 idioms, 236 phrases, and 218 fragments. In each rating study, a three-word sequence was presented on a computer screen one at a time. As a control, 90 ungrammatical three-word

Table 1
Individual ANOVA tests showing no differences between the averages of the six frequency measures across the three experimental conditions

	df	*F* Score	*p* Value
Phrase	2	0.027	.973
1st bigram	2	0.092	.912
2nd bigram	2	0.022	.978
1st unigram	2	0.845	.432
2nd unigram	2	0.582	.561
3rd unigram	2	0.051	.951

combinations were also included as foils. The ungrammatical tokens were created by scrambling matching phrases and fragment tokens that were not used in the experimental material.

Participants were asked to rate items on a 1–7 scale by pressing a key between 1 and 7 (similar to Konopka & Block, 2009; Snider & Arnon, 2012). The items were randomized across participants. To ensure that participants read each sequence, a random memory recall test was included. In 10% of the trials for each condition (idioms, phrases, and fragments), participants were asked to type an English sentence that included the three-word sequence they had just seen. There was a different set of random items tested in the memory recall test for each participant.

In the Plausibility Rating study, participants were instructed to rate each trigram according to how plausible it was as part of an English sentence. The participants' task in the Idiomaticity Rating study was to rate each trigram according to how idiomatic they found each sequence of words. Finally, in the Meaningfulness Rating study, participants were asked to rate each sequence of words according to how meaningful they were as a unit.

Table 2 shows the examples of different frequency-matched triads of idioms, phrases, and fragments with Plausibility, Idiomaticity, and Meaningfulness ratings along with their $\log_2$ transformed trigram frequencies. The complete set of materials used in the phrasal decision study can be found in the Supporting Information (including frequency and rating information) on OSF: https://osf.io/zfm9d/.

There was no effect of trigram type on plausibility ratings (means of 6.84, 6.9, and 6.87 for Idioms, Phrases, and Fragments, respectively; $F[2, 117] = 0.798$, $p = .45$), while meaningfulness ratings differed significantly across the three conditions (means of 5.89, 5.88, and 1.97 for Idioms, Phrases, and Fragments, respectively; $F[2, 117] = 297.4$, $p < .0001$) as did Idiomaticity ratings (means of 4.34, 2.12, and 1.45 for Idioms, Phrases, and Fragments, respectively; $F[2, 117] = 100.6$, $p < .0001$).

2.2.2. Phrasal decision task

A final set of 40 triads (each comprising an idiom, a phrase, and a fragment) was selected by an algorithm which chose, from all possible sets of the same size, the one which differed the least according to plausibility ratings as well as whole-string and

Table 2
Examples of each trigram type with their plausibility, idiomaticity, and meaningfulness ratings, as well as whole-string frequency counts

Trigram Type	Trigram	Plausibility	Idiomaticity	Meaningfulness	Frequency ($\log_2$)
Idiom	*play the field*	6.4	6.4	6.5	1.6
Phrase	*nothing to wear*	7.0	2.6	5.9	1.6
Fragment	*without the primary*	6.9	1.3	2.6	1.6
Idiom	*on my mind*	7.0	4.5	6.6	6.5
Phrase	*is really nice*	7.0	1.2	3.6	6.7
Fragment	*know it gets*	6.8	1.3	1.5	6.5
Idiom	*up the creek*	7.0	4.5	6.1	2.6
Phrase	*get a certificate*	6.8	1.2	6.2	2.6
Fragment	*because it lets*	6.4	1.3	1.8	2.6

substring frequencies across the three trigram types, while differing as much as possible in meaningfulness ratings between fragments and the other two conditions (to ensure that the fragments remained low in meaningfulness). The resulting set of 40 idioms, 40 phrases, and 40 fragments did not differ significantly along the six frequency dimensions (trigram, first bigram, second bigram, first unigram, second unigram, and third unigram). There were no differences in forward or backward transitional probabilities between the three types of sequences. The $\log_2$ trigram frequencies of the final set of 40 triads in the phrasal decision study ranged between 1 and 10.4.

Importantly, the 40 idioms, 40 phrases, and 40 fragments did not differ according to the percentage of subjects rating items as 6 or 7 in the plausibility rating study, as illustrated in Fig. 1a. Additionally, as indicated by Fig. 1b, we ensured that idioms were rated as more idiomatic than both phrases and fragments. Finally, the items were constrained such that within a triad, idioms, and phrases did not differ in terms of their meaningfulness, whereas the fragments had the lowest meaningfulness scores possible, as shown in Fig. 1c. Besides the 40 experimental triads (totaling 120 items), 120 ungrammatical sequences (such as *hear I isn't*) were used as fillers.

2.3. Procedure

To determine whether the overall meaningfulness of a multiword sequence might facilitate its processing over and above mere frequency of use, and independently of whether the sequence is idiomatic or not, we conducted a phrasal decision task. In this task, participants are presented with a multiword sequence and have to decide as quickly and as accurately as possible whether the presented item could form part of an English sentence (Arnon & Snider, 2010; Arnon et al., 2017; Swinney & Cutler, 1979). This task is thus a phrasal version of the classic lexical decision task (Meyer & Schvaneveldt, 1971).

In our study, we presented participants with the three-word sequences (120 experimental and 120 ungrammatical filler tokens) one by one, in random order, on a computer screen, and asked them to judge (by quickly pressing one of two keys) whether or not

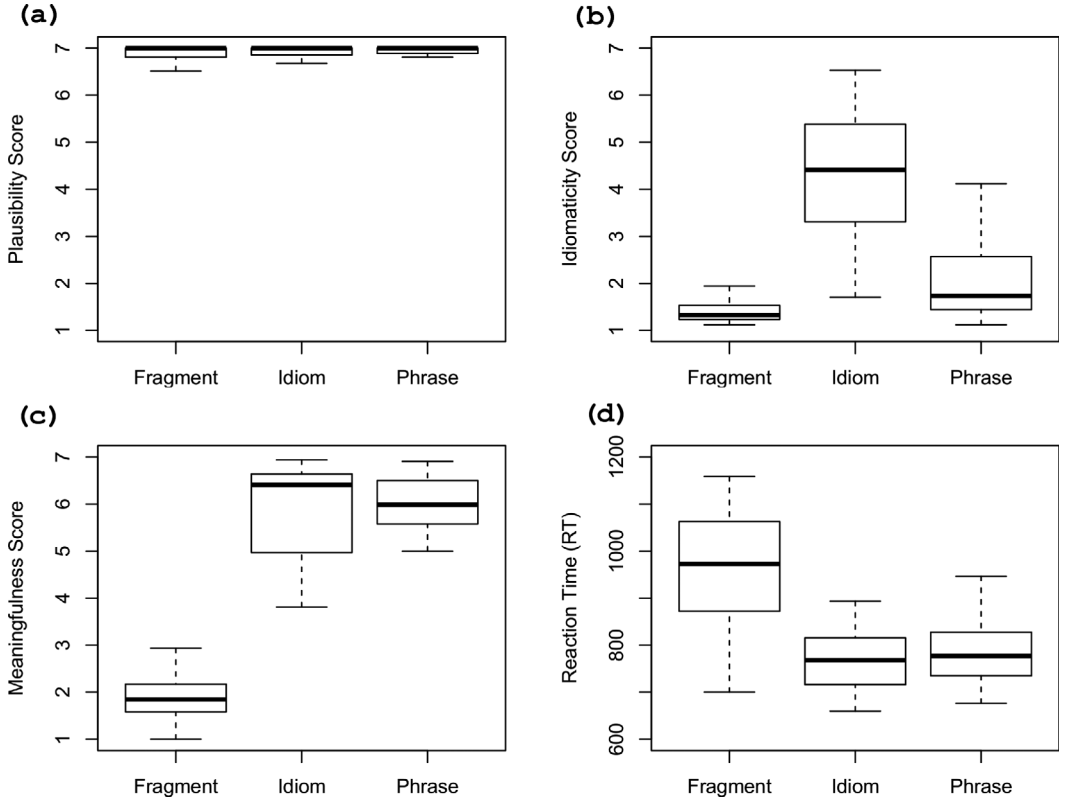

Fig. 1. Boxplots depicting (a) mean Plausibility Rating, (b) mean Idiomaticity Rating, (c) mean Meaningfulness Rating, and (d) mean reaction times for the Fragment, Idiom, and Phrase conditions in the phrasal decision task. Boxes depict the median (thick line), with upper and lower edges representing quartiles, while whiskers depict 1.5 interquartile range.

they formed possible word combinations in the context of English sentences. Participants were asked to make a decision as fast as possible but without sacrificing accuracy.

2.4. Data analysis

Responses of less than 200 ms as well as reaction times exceeding the upper quartile by more than three times the interquartile range were removed. This resulted in a 1.5% data loss. We analyzed the results using LME models in R version 3.4.0 (R Core Team, 2017), with packages LME4 version 1.1.19 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest version 2.0.33 (Kuznetsova, Brockhoff, & Christensen, 2016). We began using the *lme4* package to define the maximal model justified by the experimental design, which included Item and Subject as random effects, and the Idiomaticity ratings, the Meaningfulness ratings, Trigram Type (using Idiomatic Expressions as the base case), Frequency (whole-string), substring frequencies (including frequency predictors for First Bigram, Second Bigram, First Unigram, Second Unigram, and Third Unigram), and

Length in Characters as fixed effects. Moreover, we included all possible interaction terms involving the variables of interest: Trigram Type, Frequency (whole-string), Meaningfulness, and Idiomaticity. Finally, we included (by-subject) random slopes for Frequency (whole-string), Meaningfulness, and Idiomaticity. There was no substantial degree of multicollinearity: The condition number for the matrix of predictors was only 11.4 (cf. Belsley, Kuh, & Welsch, 1980).[2] Because the Meaningfulness and Idiomaticity scores ranged between 1 and 7, linear transformation was applied ($n/7$), and then the data were logit-transformed prior to entry in the model (cf. Warton & Hui, 2011). As the reaction times (RTs) were not normally distributed, they were logit-transformed prior to the analysis. All other numerical predictors were centered before inclusion in the model. The RT data as well as the associated item ratings are available from OSF along with the R analyses script: https://osf.io/zfm9d/.

3. Results

Fig. 1d shows that participants responded similarly to idioms (*M*: 766.4, *SD*: 248) and phrases (*M*: 789.4, *SD*: 260.3), but more slowly to the fragments (*M*: 949.9 ms, *SD*: 343.7). This overall pattern of results was corroborated by an LME analysis of the RT data. Using the above-described maximal model as a starting point (estimates and test statistics for this maximal model are shown in Appendix A), we employed a step-wise model comparison to reduce the number of fixed effects (in the interest of interpretability and identifying the best model fit, given the high number of non-significant effects). This was achieved using the *lmerTest* package in *R* (Kuznetsova et al., 2016). The resulting final model, shown in Table 3, included Trigram Type, Meaningfulness, the Third Unigram frequency, both Bigram frequencies, Length in Characters, and the interaction term between Trigram Type and Meaningfulness as fixed effects, with Subject and Item as random effects. The model also included a by-subject random slope of Meaningfulness.

The final model was compared to a version without the fixed effects ($\chi^2 = 224.22$, $p < .0001$) as well as a version of the model without the variables of interest (Meaningfulness and Trigram Type; $\chi^2 = 196.94$, $p < .0001$), indicating that the full version of the final model captured more of the variance in both cases. As an additional step, we carried out comparisons between the final model and reduced versions which removed only the fixed effect of Meaningfulness ($\chi^2 = 79.38$, $p < .0001$) and only the fixed effect of Trigram Type ($\chi^2 = 69.17$, $p < .0001$), finding that the full final model provided a significantly better fit in each case. For thoroughness, we also carried out these comparisons using the maximal model, finding that the removal of the fixed effect of Meaningfulness ($\chi^2 = 80.25$, $p < .0001$) as well as Trigram Type ($\chi^2 = 63.42$, $p < .0001$) damaged model fit to a significant degree.

As predicted, RTs were affected by trigram type: It took longer for participants to respond to fragments ($\beta = 0.15$, $p < .001$). Furthermore, decision times for phrases were not significantly slower than for idioms ($\beta = 0.01$, $p = .75$). Indeed, there was no effect of Idiomaticity on responses ($\beta = -0.03$, $p = .38$ in the maximal model; Idiomaticity did

Table 3
Fixed effects for final model

Fixed Effect	Estimate	*SE*	df	*t* Value	*p* Value
(Intercept)	6.661	0.0374	75.9	178.03	0.000000***
Trigram type: Fragment	−0.155	0.0407	115.0	−3.806	0.000229***
Trigram type: Phrase	0.010	0.0320	107.2	0.314	0.753953
Meaningfulness	−0.025	0.0088	118.0	−2.773	0.006457**
Unigram 3	0.011	0.0032	110.0	3.497	0.000680***
Bigram 1	−0.011	0.0033	108.2	−3.31	0.001267**
Bigram 2	−0.008	0.0034	113.0	−2.399	0.018068*
Length in characters	0.014	0.0029	108.0	5.027	0.000002***
Trigram type: Fragment × Meaningfulness interaction	−0.280	0.0337	126.4	−8.313	0.000000***
Trigram type: Phrase × Meaningfulness interaction	0.000	0.0137	107.0	0.016	0.987376

Notes *$p < .05$, **$p < .01$, and ***$p < .001$.

not enter the final model). Rather, participants showed sensitivity to the overall meaningfulness of the trigrams: RTs were faster for more meaningful tokens, as revealed by a significant main effect of Meaningfulness ($\beta = -0.02$, $p < .01$). Moreover, there was a significant interaction between Meaningfulness and Trigram Type for Fragments ($\beta = -0.28$, $p < .0001$), reflecting heightened importance for coherent meaning in that condition.

Frequency also reached significance for the first ($\beta = -0.01$, $p < .01$) and second ($\beta = -0.008$, $p < .05$) bigrams, indicating that subjects responded faster to items with greater bigram substring frequency. Importantly, contrary to expectations from previous studies of multiword sequences (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008), whole-string frequency did not enter into the final model. Length in Characters ($\beta = 0.014$, $p < .001$) and Third Unigram ($\beta = 0.011$, $p < .001$) also reached significance.[3]

The Idiom and Phrase trigrams varied with respect to the type of phrase represented (e.g., noun phrases such as *sample of data*; verb phrases such as *had a dream*; prepositional phrases such as *in the mailbox*), and because such variation was not controlled in a systematic fashion, we conducted a follow-up analysis to ensure that this did not play an unexpected role in shaping the reaction times.[4] To this end, we tagged each trigram using the Stanford Parser (Manning et al., 2014) and included those phrase tags which were represented by at least 10 trigrams in our stimulus set. These included noun phrase (NP), verb phrase (VP), adverbial phrase (ADVP), and prepositional phrase (PP). We then included tag type as a fixed effect control variable in a version of the final LME model (see above). None of the phrase types (NP, VP, and PP) differed from the base case (ADVP) in terms of predicting RTs. Moreover, exclusion of phrase type as a fixed effect did not harm model fit ($\chi^2 = 1.18$, $p > .75$).

As the three conditions also differed slightly in terms of the mean number of function words per item (Phrases: 1.56; Idioms: 1.82, Fragments: 1.98), we carried out additional analyses to determine whether this difference affected RTs to a significant degree. To this

end, we included Function Word Count as a fixed effect in the final model described above. This led to no change in the pattern of results reported previously, and Function Word Count did not reach significance in the model ($\beta = -0.017$, $p > .18$). Moreover, removal of Function Word Count from the model did not damage fit to a significant degree ($\chi^2 = 1.9$, $p > .16$).

Finally, we analyzed the accuracy of participants' responses, independently of the RT data (recall that incorrect responses were excluded from the above analyses). Overall, participants achieved an accuracy rate of 98.5% for Idioms, an identical accuracy rate of 98.5% for Phrases, but a slightly lower accuracy rate of 87.5% for Fragments. A repeated-measures ANOVA with subject as a random factor confirmed a significant main effect of Condition $F[2, 78] = 83.94$, $p < .0001$. Thus, the participants' response accuracy mirrored the general pattern observed with RTs, with respect to the three trigram types.

4. Discussion

Similar to single words, prior research has observed compelling frequency effects for multiword sequences in language acquisition (e.g., Arnon & Clark, 2011; Bannard & Matthews, 2008) as well as in adult processing (e.g., Arnon & Snider, 2010; Tremblay et al., 2011) and production (e.g., Arnon & Cohen-Priva, 2013; Janssen & Barber, 2012). In the current study, we took a first step toward determining whether whole-sequence meaning also is a factor in the processing of multiword chunks, as has been shown for individual words (Balota, 1990; Yap et al., 2011). As predicted, when comparing the processing of three-word sequences varying in chunk meaningfulness, idiomaticity, and co-occurrence frequency, we found that language users strongly relied on the degree to which a multiword sequence conveys a coherent communicative meaning as a whole: The more meaningful a trigram was rated, the easier it was to process in the phrasal decision task. Moreover, decision times for compositional phrases were on par with idioms while processing times for fragments were significantly slower. Finally, whereas we observed frequency effect at the internal bigram level, we did not obtain a whole-phrase frequency effect, only a chunk meaningfulness effect. These results confirm the key role of chunk meaningfulness in language processing—a factor that has not been considered by previous studies focusing solely on the frequency of multiword phrases (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008). Taken together, our findings suggest that sequences that are perceived as highly meaningful as a unit leave traces in memory, regardless of idiomatic status or whole-string frequency.

Although we found an overall effect of meaningfulness across item types, a possible concern may be that perhaps it was constituency that drove the effect of meaningfulness given that both idioms and phrases were full constituents, whereas fragments were not. Arnon and Cohen-Priva (2013) is the only prior study to investigate whether constituency might affect multiword sequence processing when adequately controlling for frequency. They found similar frequency effects when comparing high- and low-frequency variants of constituents (e.g., *a lot of work* vs. *a lot of years*) as well as non-constituents crossing

syntactic boundaries (e.g., *as far as I* vs. *as far as you*). In the present study, the interaction between meaningfulness score and trigram type for fragments demonstrates that the effect of meaningfulness, rather than being driven by constituency, is actually heightened in the case of Fragments: Higher meaningfulness scores predict decreased RTs to an even higher extent for Fragments than for the other two conditions. So even though constituency and semantic coherence are intertwined, the more meaningful a fragment was rated to be, the faster it was processed, despite the fact that none of the fragments were full constituents. This suggests that constituency may be less important to multiword sequence processing than meaningfulness (and frequency).

These findings dovetail with construction-based theories of language that emphasize the role of meaning in language processing, and which treat idioms and compositional phrases merely as variants of stored form-meaning pairings (e.g., Goldberg, 2006; Langacker, 1987). By contrast, because our results blur the distinction between vocabulary and grammar, they add to the growing set of challenges facing accounts of language that rely on single words and rules for combining them as the primary means of explaining linguistic productivity (e.g., Chomsky, 1980; Pinker, 1999). Within this perspective, multiword sequences are viewed as constructed by rules on the fly, except for peripheral exceptions such as idioms. However, a growing number of corpus studies have shown that multiword sequences are by no means marginal (e.g., Jackendoff, 1997) but comprise up to 50% of normal written and spoken language use (DeCock, Granger, Leech, & McEnery, 1998; see Conklin & Schmitt, 2012, for a review). Whereas the meaningfulness of a linguistic structure is secondary in generative theories (e.g., Chomsky, 1980; Pinker, 1999), construction-based approaches argue for a more central role for meaning in language acquisition and use (e.g., Dąbrowska, 2014; Goldberg, 2006; Tomasello, 2009).

The implications of our findings reach beyond cognitive science and linguistics to computer science (see also Christiansen & Arnon, 2017). Indeed, the prevalence of multiword sequences and their heterogeneity has famously been flagged as "a pain in the neck" (Sag, Baldwin, Bond, Copestake, & Flickinger, 2002), a major obstacle for computer science approaches to language (and the focus of special issues in computational linguistics journals: e.g., Ramisch, Villavicencio, & Kordoni, 2013; Villavicencio, Bond, Korhonen, & McCarthy, 2005). Our study suggests that going beyond distributional information to incorporate the meaningfulness of multiword sequences into corpus analyses might provide a useful way forward, not only for computational natural language processing but also for usage-based approaches to language (for the latter, see also Pijpops & Van de Velde, 2016).

More generally, the results of our study might be best understood when considering that everyday language use is communicative and contextual; language elements that express a meaning together as a chunk, to describe a particular event, such as "nothing to wear" somehow become "glued" together. Previous work has explored this idea within usage-based (e.g., Bybee, 2006; Tomasello, 2009) and construction-based (e.g., Dąbrowska, 2014; Goldberg, 2006) perspectives on language acquisition and use, where commonly used multiword sequences develop into stored units, based on the fact that the words often occur together (see Arnon & Christiansen, 2017, for a review). Recently,

Christiansen and Chater (2016) argued that the use of such multiunit chunks is necessary to deal with the onslaught of input during real-time language acquisition and processing, given the severe memory and attentional constraints inherent to the language system (the so-called Now-or-Never bottleneck). This perspective emphasizes the role of "shallow" parsing in normal language processing (e.g., Pijpops & Van de Velde, 2016), whereby the input is chunked into larger units, and where the focus of processing is on arriving at a "good enough" interpretation of the utterance (e.g., Ferreira & Patson, 2007), rather than a full syntactic parse. Aspects of this theory were implemented in a cross-linguistic computational model capturing aspects of comprehension (shallow parsing) and production (word-chunk sequencing) in language acquisition (McCauley & Christiansen, 2019). Our study extends this work by highlighting the key role of chunk meaningfulness in language processing over and above frequency of use.

An important remaining issue pertains to how the meaningful multiword chunks may be represented by the language system. Our findings suggest that the meaningfulness of a multiword sequence has a direct influence on processing speed: The more a multiword sequence conveys a coherent meaning, the more likely it is that it will be processed as a linguistic unit in its own right. This may suggest that multiword sequences are stored as unanalyzed wholes (e.g., Pijpops & Van de Velde, 2016), possibly somewhat similar to single words. Indeed, a large-scale corpus analysis by Williams et al. (2015) showed that multiword sequences generally provide a better fit with a Zipfian distribution than single words (Zipf, 1935). However, research on idioms suggests that even within these multiword chunks, individual words may still be accessible (Libben & Titone, 2008). In a similar vein, Dąbrowska (2014) argues that most multiword sequences are not stored as unanalyzed wholes, but rather that the component words form a co-activated network of representations. Our findings suggest that meaningfulness may play an important role in the co-activation of elements within such a network.

To conclude, our results provide new insights into the representation and processing of multiword sequences, suggesting that the meaningfulness of such strings affects processing, independently of their frequency and idiomaticity. These findings are consistent with construction-based approaches that treat idioms and phrases as comparable form-meaning mappings. Furthermore, our findings are relevant to usage-based approaches to language more generally (e.g., Bybee, 2006) because they suggest that the meaningfulness of multiword sequences provides an additional dimension, apart from co-occurrence frequency, that such approaches must take into account.

Acknowledgments

We would like to thank Jaclyn Jeffrey-Wilensky, Julia Krasnow, Klejda Bejleri, and László Kálmán for help with materials and data collection as well as Haim Bar for statistical guidance.

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/zfm9d/.

Notes

1. Because of the complexity of matching stimulus triads for frequency, there were two exceptions: for the idiom *over a barrel*, both its phrase and fragment pairs' frequencies differed 13.9% from the idiom's trigram frequency. Also, for the idiom *take the case*, the phrase pair frequency differed 36.9% from the idiom's trigram frequency. Note, however, that these deviations were controlled for in the statistical analyses.
2. Note that we did not include the plausibility ratings in our analyses because they are largely redundant, as we found no significant differences across the three groups in terms of Plausibility. We ran the plausibility norming study merely as a control and did not have any a priori interest in plausibility per se, but rather wished to ensure tightly controlled stimuli.
3. The third unigram frequency may have reached significance because the last unigram was likely to be a content word while most of the first and second words were function words.
4. Because fragments by their very nature do not correspond to taggable phrases, they were excluded from this analysis.

References

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1–L2 differences. *Topics in Cognitive Science*, *9*, 621–636. https://doi.org/10.1111/tops.12271

Arnon, I., & Clark, E. V. (2011). When "on your feet" is better than "feet": Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, *7*, 107–129. https://doi.org/10.1080/15475441.2010.505489

Arnon, I., & Cohen-Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*, *56*, 349–371. https://doi.org/10.1177/0023830913484891

Arnon, I., McCauley, S. M., & Christiansen, M. H. (2017). Digging up the building blocks of language: Age-of-acquisition effect for multiword phrases. *Journal of Memory and Language*, *92*, 265–280. https://doi.org/10.1016/j.jml.2016.07.004

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Balota, D. A. (1990). The role of meaning in word recognition. In D. A. Balota, G. F. D'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 9–32). Hillsdale, NJ: Erlbaum.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*, 241–248. https://doi.org/10.1111/j.1467-9280.2008.02075.x

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Belsley, D. A.,Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 711–733.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Christiansen, M. H., & Arnon, I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*, 542–551. https://doi.org/10.1111/tops.12274

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2004). *Fisher English training speech (part 1) transcripts*. Philadelphia, PA: Linguistic Data Consortium.

Cieri, C., Graff, D., Kimball, O., Miller, D., & Walker, K. (2005). *Fisher English training speech (part 2) transcripts*. Philadelphia, PA: Linguistic Data Consortium.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and non-native speakers? *Applied Linguistics*, *29*, 72–89.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science*, *9*, 552–568. https://doi.org/10.1111/tops.12255

Dąbrowska, E. (2014). Recycling utterances: A speaker's guide to sentence processing. *Cognitive Linguistics*, *25*, 617–653. https://doi.org/10.1515/cog-2014-0057

DeCock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learning English on computer* (pp. 67–79). London: Addison, Wesley, Longman.

Ellis, A. W., & Morrison, C. M. (1998). Real age of acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 515–523. https://doi.org/10.1037/0278-7393.24.2.515

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*, 71–83. https://doi.org/10.1111/j.1749-818X.2007.00007.x

Fillmore, C., Kay, P., & O'Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, *64*, 501–538.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Hall, J. (1954). Learning as a function of word-frequency. *American Journal of Psychology*, *67*, 138–140.

Jackendoff, R. S. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS ONE*, *7*, 0033202. https://doi.org/10.1371/journal.pone.0033202

Konopka, A. E., & Bock, J. K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*, 68–101. https://doi.org/10.1016/j.cogpsych.2008.05.002

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0-33. Available at: https://CRAN.R-project.org/package=lmerTest

Langacker, R. W. (1987). *Foundations of cognitive grammar, Vol. 1: Theoretical prerequisites*. Palo Alto, CA: Stanford University Press.

Libben, M., & Titone, D. (2008). The multidetermined nature of idiomatic expressions. *Memory & Cognition*, *36*, 1103–1131. https://doi.org/10.3758/mc.36.6.1103

Makai, A., Boatner, M. T., & Gates, J. E. (1991). *Handbook of commonly used American idioms*. New York: Barron's Educational Series Inc.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In K. Bontcheva & J. Zhu (Eds.), *Proceedings of 52nd annual meeting of the Association for Computational Linguistics: System demonstrations* (pp. 55–60). Baltimore, MD: Association for Computational Linguistics.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*, 1–51. https://doi.org/10.1037/rev0000126

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234. https://doi.org/10.1037/h0031564

Pijpops, D., & Van de Velde, F. (2016). Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*, *50*, 543–581. https://doi.org/10.1515/flin-2016-0020

Pinker, S. (1999). *Words and rules*. New York: Harper.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/

Ramisch, C., Villavicencio, A., & Kordoni, V. (2013). Introduction to the special issue on multiword expressions: From theory to practice and use. *ACM Transactions on Speech and Language Processing (TSLP)*, *10*(2), 3. https://doi.org/10.1145/2483691.2483692

Reali, F., & Christiansen, M. H. (2007). Word-chunk frequencies affect the processing of pronominal object-relative clauses. *Quarterly Journal of Experimental Psychology*, *60*, 161–170. https://doi.org/10.1080/17470210600971469

Reppen, R., Ide, N., & Suderman, K. (2005). *American national corpus (ANC): Second release*. Philadelphia, PA: Linguistic Data Consortium.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Proceedings of the third international conference on intelligent text processing and computational linguistics* (pp. 1–15). Berlin: Springer.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 251–272. https://doi.org/10.1177/0267658310382068

Snider, N., & Arnon, I. (2012). A unified lexicon and grammar? Compositional and noncompositional phrases in the lexicon. In S. Gries & D. Divjak (Eds.), *Frequency effects in language* (pp. 127–163). Berlin: Mouton de Gruyter.

Spears, R. A. (2008). *Essential American idioms dictionary*. Columbus, OH: McGraw-Hill.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, *18*, 523–534.

Taylor, J. S. H., Duff, F. J., Woollams, A. M., Monaghan, P., & Ricketts, J. (2015). How word meaning influences word reading. *Current Directions in Psychological Science*, *24*, 322–328. https://doi.org/10.1177/0963721415574980

Tomasello, M. (2009). The usage based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 69–87). Cambridge, UK: Cambridge University Press.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 151–167). New York: Continuum International Publishing.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 153–172). Amsterdam: John Benjamins.

Villavicencio, A., Bond, F., Korhonen, A., & McCarthy, D. (2005). Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, *19*, 365–377.

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology*, *92*, 3–10. https://doi.org/10.1890/10-0340.1

Williams, J. R., Lessard, P. R., Desu, S., Clark, E. M., Bagrow, J. P., Danforth, C. M., & Dodds, P. S. (2015). Zipf's law holds for phrases, not words. *Scientific Reports*, *5*, 12209. https://doi.org/10.1038/srep12209

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

Wray, A. (2017). Formulaic sequences as a regulatory mechanism for cognitive perturbations during the achievement of social goals. *Topics in Cognitive Science*, *9*, 569–587. https://doi.org/10.1111/tops.12257

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*, 742–750. https://doi.org/10.3758/s13423-011-0092-y

Zipf, G. K. (1935). *The psycho-biology of language*. New York: Houghton, Mifflin.

Appendix A: Estimates and test statistics for the maximal model

Fixed Effect	Estimate	*SE*	df	*t* Value	Pr(>|*t*|)
(Intercept)	6.665886	0.039356	87.92809	169.372	0.00000***
Trigram type: Fragment	−0.10889	0.113925	90.42113	−0.956	0.34172
Trigram type: Phrase	0.044717	0.062193	86.77151	0.719	0.47407
Trigram frequency	−0.00448	0.012597	87.39103	−0.356	0.72289
Meaningfulness score	−0.02937	0.012319	90.51409	−2.384	0.01922*
Idiomaticity score	−0.02945	0.033717	86.91654	−0.873	0.38489
Unigram 1 frequency	−0.00524	0.007429	90.4985	−0.705	0.48257
Unigram 2 frequency	−0.00181	0.009756	91.28403	−0.185	0.85337
Unigram 3 frequency	0.014199	0.004788	88.21909	2.966	0.00388**
Bigram 1 frequency	−0.00466	0.007378	88.60128	−0.632	0.5292
Bigram 2 frequency	−0.00823	0.006437	87.43743	−1.279	0.20421
Length in characters	0.014378	0.003233	87.76193	4.448	0.00003***
Trigram type: Fragment × Tri. freq.	−0.05585	0.108157	95.43802	−0.516	0.60682
Trigram type: Phrase × Tri. freq.	0.054977	0.041351	86.76808	1.33	0.18716
Trigram type: Fragment × Meaningfulness score	−0.24725	0.197913	95.13665	−1.249	0.21462
Trigram type: Phrase × Meaningfulness score	−0.01459	0.02863	86.39811	−0.509	0.61175
Tri. freq. × Meaningfulness score	0.002197	0.004456	86.63567	0.493	0.62315
Trigram type: Fragment × Idiomaticity score	0.093432	0.084937	90.90906	1.1	0.27423
Trigram type: Phrase × Idiomaticity score	0.057395	0.05552	86.90166	1.034	0.30411
Tri. freq. × Idiomaticity score	−0.00802	0.018245	87.03398	−0.439	0.66143
Meaningfulness score × Idiomaticity score	0.008933	0.013738	86.2614	0.65	0.51726
Trigram type: Fragment × Tri. freq. × Meaningfulness score	−0.11344	0.139278	100.5769	−0.814	0.41729
Trigram type: Phrase × Tri. freq. × Meaningfulness score	−0.02943	0.018919	86.46863	−1.555	0.12351
Trigram type: Fragment × Tri. freq. × Idiomaticity score	−0.01338	0.078685	94.26905	−0.17	0.86534

(continued)

Appendix A (continued)

Fixed Effect	Estimate	*SE*	df	*t* Value	Pr(>|*t*|)
Trigram type: Phrase × Tri. freq. × Idiomaticity score	0.035453	0.03479	87.0391	1.019	0.31099
Trigram type: Fragment × Meaningfulness score × Idiomaticity score	0.036741	0.132537	95.91334	0.277	0.78222
Trigram type: Phrase × Meaningfulness score × Idiomaticity score	−0.02156	0.026395	86.60832	−0.817	0.41622
Tri. freq. × Meaningfulness score × Idiomaticity score	−0.00102	0.007596	86.90308	−0.134	0.8937
Trigram type: Fragment × Tri. freq. × Meaningfulness score × idiomaticity score	−0.05928	0.094812	99.63791	−0.625	0.53322
Trigram type: Phrase × Tri. freq. × Meaningfulness score × Idiomaticity score	−0.01168	0.016792	86.8983	−0.696	0.48847

Notes *$p < .05$, **$p < .01$, and ***$p < .001$.