Meaning Overrides Frequency in Idiomatic and Compositional Multiword Chunks

Hajnal Jolsvai (jolsvai@cornell.edu)
Stewart M. McCauley (smm424@cornell.edu)
Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853 USA

Abstract

In line with usage-based accounts, recent psycholinguistic studies have confirmed that frequency of occurrence impacts processing latencies for multiword strings (e.g., Arnon & Snider, 2010). However, these studies have not been concerned with the meaning of the multiword chunks in question, which is central to accounts of formulaic language rooted in cognitive linguistics (e.g., Wray, 2002). Here, we address this issue by comparing processing latencies for three types of multiword chunks: idiomatic expressions, meaningful compositional phrases, and less meaningful fragments. All three chunk types were matched for whole- and sub-string frequency. Our results show that frequency facilitates processing for all three chunk types, but to a lesser extent than their "meaningfulness" (as assessed in a separate norming study), indicating that the meanings of multiword expressions may have implications for models of language processing which extend beyond those of frequency of occurrence.

Keywords: idiomatic phrases; chunks; distributional statistics; Construction Grammar (CG); usage-based approach; cognitive linguistics

Introduction

It has been recognized through a large number of corpus studies that everyday language involves a wide array of fixed co-occurring multiword sequences (such as *let it go*), or formulaic sequences (e.g., Sinclair, 1991). Usage-based theories suggest that strings of words can become fixed linguistic patterns through simple repetition of use (e.g., *strong tea* or *powerful computer*, as opposed to *powerful tea* or *strong computer*; Bybee, 2006; Bybee & Hopper, 2001; Tomasello, 2003). Indeed, recent psycholinguistic work has shown that the processing of a multiword sequence is affected by the frequency of the sequence as a whole (e.g., Arnon & Snider, 2010; Tremblay, Derwing, Libben, & Westbury, 2011).

But how much insight does frequency alone give us into formulaic language processing? Is it possible that some multiword chunks become conventionalized because they convey a highly specific meaning, in much the same way as do individual words (e.g., *let it be*)? In addition to their frequency of use, the "meaningfulness" of multiword chunks may be another factor affecting stored linguistic units, in line with what would be expected from a cognitive-grammar perspective (e.g., Fillmore, Kay, & O'Connor, 1988; Langacker, 1987).

Idioms are perhaps the most canonical form of formulaic language, involving a particular string of words which is linked to a specific meaning that does not follow compositionally from the meanings of the constituent words (e.g., *spill the beans*). Importantly, it has traditionally been assumed that idiomatic phrases are represented and retrieved from memory as single units and thus processed faster than compositional phrases (e.g., Swinney & Cutler, 1979). Whereas idioms are often treated as special cases, Construction Grammar approaches suggest that both idiomatic and high-frequency compositional multiword expressions are stored as conventionalized form-meaning pairs (Goldberg, 2003).

Supporting the lexicalized idioms assumption (e.g., Swinney & Cutler, 1979), an eye-tracking study by Underwood, Schmidt and Galpin (2004) found fewer and shorter fixations for the last words of idiomatic expressions (e.g., *met the deadline by the skin of his teeth*) as compared to the same words in non-idiomatic contexts (e.g., *the dentist looked at his teeth*). In a self-paced reading study, Conklin and Schmidt (2008) found that idiomatic expressions (e.g., *hit the nail on the head*) were read faster when compared to similar phrases (e.g., *hit his head on the nail*). Moreover, a recent eye-tracking study demonstrated a processing advantage for idiomatic expressions (e.g., *at the end of the day*) over similar phrases (e.g., *at the end of the war*) (Siyanova-Chanturia, Conklin, Schmitt, 2011). However, a shortcoming of these studies, which reveal faster processing for formulaic expressions, is that whole- and sub-string frequencies were not controlled for in a systematic way.

Nevertheless, recent experimental evidence has shown that language users are sensitive to the frequency of compositional multiword phrases. Typically, in these studies, pairs of high- and low-frequency phrases are compared, using corpus data to control for substring frequencies. Bannard and Matthews (2008) found that 3- and 4-years-olds repeat more frequent variants of compositional four-word phrases (such as *a drink of milk)* more easily than similar but less frequent four-word strings (such as *a drink of tea).* Similarly, high-frequency compositional multiword phrases are processed faster by adults than comparable sequences of lesser frequency (e.g., Arnon & Snider, 2010; Tremblay & Baayen, 2010).

In addition to findings of frequency effects for compositional multiword phrases, recent studies have investigated whether similar frequency effects can be

revealed for multiword sequences crossing syntactic boundaries. For instance, higher *n*-gram frequency chunks crossing syntactic boundaries (such as *in the middle of the)* are read faster compared to less frequent non-constituents (e.g., *in the front of the)* when embedded in a sentence (Tremblay, Derwing, Libben, & Westbury, 2011). Similarly, Arnon and Cohen-Priva (in press) found no effect of constituency when comparing production latencies for high- and low-frequency constituent and non-constituent pairs. The presence of frequency effects for not only phrases, but for sentence fragments as well, suggests that constituency may be a less important factor for theoretical accounts seeking to incorporate multiword chunks as linguistic units in their own right. Alternatively, it is possible that meaning may also need to be considered in order to reveal constituency effects for multiword sequences.

Although these findings make it clear that more frequent multiword phrases are processed faster, there is work to suggest that other factors may need to be taken in to consideration. For instance, Siyanova-Chanturia, Conklin and van Heuven (2011) found that readers are also sensitive to the fixedness of multiword chunks, above and beyond their frequency of occurrence. In this eye-tracking study, more frequent formulaic binominal phrases (such as *bride and groom*) were more easily read by participants as compared to their less preferred reversed forms, which nevertheless shared the same syntactic and semantic attributes (*groom and bride*). More importantly, even lower frequency formulaic chunks with fixed configurations (like *sweet and sour*), were read faster than higher frequency reversed phrases (such as *west and east*).

The above-mentioned studies are limited in that the "meaningfulness" of multiword chunks is not taken into account. Language users may also be sensitive to the relative meaningfulness of a multiword sequence as a whole, in line with predictions made by cognitive linguists (e.g., Langacker, 1987; Wray, 2002).

In the current study, we attempt to fill a gap left by previous psycholinguistic studies concerned with the processing of multiword sequences, which have tended to focus primarily on distributional properties, such as frequency of occurrence. We investigate the processing of multiword chunks that vary in the degree to which people find them meaningful as a whole unit, while controlling for frequency. Specifically, our goal is to determine whether the relative meaningfulness of multiword chunks may impact their processing over and above the well-known effects of frequency. Moreover, we aim to test the idea that highly meaningful chunks show processing advantages independently of their compositional status, in accordance with Construction Grammar approaches that treat both idiomatic and compositional sequences as pairings between form and meaning (e.g., Goldberg, 2003).

We address these issues by examining readers' processing of three different types of 3-word sequences; idiomatic expressions, highly meaningful compositional phrases. and less meaningful fragments which cross syntactic boundaries. All three types of trigrams were selected from the same corpora and matched for phrase (trigram) and substring (bigram, unigram) frequencies. Both idiomatic expressions and compositional phrases were rated as being equally meaningful in an initial norming study. Furthermore, participants in this norming study rated sentence fragments as being significantly less meaningful than their frequency-matched idiomatic and meaningful compositional phrases. A different set of participants found the three types of tokens to be equally plausible as part of an English sentence in a separate norming study (thus, our items differed only in their meaningfulness, and not plausibility as possible sequences of words in English sentences). Instead of comparing pairs of high and low *n*-gram frequency sequences differing by only one word (*don't have to worry* vs. *don't have to wait*), as in previous studies, we investigated sequences which varied in the whole-string frequency ($\log_2$ transformed) with which they appeared in a large corpus, along a continuum ranging from 1.0 to 10.4.

We predicted that idiomatic expressions and compositional phrases would be processed faster than less meaningful fragments because they vary in the extent to which people find them meaningful as units. We also predicted that when controlled for phrase and substring frequencies, idiomatic expressions would not be processed faster than compositional meaningful sequences. This prediction is at odds with the traditional view that contrasts faster access to stored idiomatic expressions with slower processing of computed compositional phrases (e.g., Swinney and Cutler, 1979). However, our prediction is in line with a Construction Grammar perspective (e.g., Goldberg, 2003).

Methods

Three different types of trigrams were extracted from the American National Corpus (ANC; Reppen, Ide, & Suderman, 2005) and the Fisher corpus (Cieri, Graff, Kimball, Miller, & Walker, 2004, 2005). The two corpora were combined into a single corpus containing a total of 39 million words of American English. The Fisher corpus comprises spoken language (telephone conversations), while the ANC consists of spoken as well as written texts. From this combined corpus, we selected all 3-word idiomatic expressions using the following collections: McGraw-Hill's Essential American Idioms Dictionary (Spears, 2008); Handbook of Commonly Used American Idioms (Makai, Boatner, Gates 1991); and the IdiomQuest (http://www.idiomquest.com) and American Idioms (http://www.americanidioms.net) online idiom dictionaries. Eighty-two 3-word idiomatic expressions from these collections appeared in the combined corpus. For each idiomatic expression (e.g., *over the hill*) we selected frequency-matched compositional phrases (e.g., *had a dream*) and frequency-matched fragments (e.g., *by the postal*). Log_2 transformation was applied to all raw phrase and substring frequencies prior to the selection process (described below). Table 1 presents examples of the three

groups of items alongside their frequencies in the corpus. Each token belongs to a "triad" of whole- and sub-string frequency-matched items, with one from each of the three conditions (idiomatic expression, compositional phrase and fragment).

Table 1: Example triads of idiomatic expressions, compositional phrases, and fragments with their $\log_2$ transformed phrase frequencies

Idiomatic Expressions		Compositional Phrases		Fragments	
just the ticket	1	that's the agreement	1	if the global	1
for a song	3.2	it's a lie	3.2	of a heavy	3.5
over the hill	5.4	had a dream	5.6	by the postal	5.2
in the wind	5.8	to the edge	5.4	in the larger	5.6
on my mind	6.5	is really nice	6.7	know it gets	6.5

Both the compositional phrases and the fragments were frequency-matched to a corresponding idiomatic expression such that trigram frequency, first bigram, second bigram, first unigram, second unigram, and third unigram frequencies were within ±10% of the corresponding idiomatic phrase's frequencies, respectively. Table 2 shows the results of the individual ANOVA tests of the phrase and substring frequencies across idiomatic expressions, compositional phrases, and fragments.

Table 2: 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	0.0338	0.9668
1st bigram	2	0.092	0.9121
2nd bigram	2	0.0259	0.9745
1st unigram	2	0.8341	0.4368
2nd unigram	2	0.6037	0.5485
3rd unigram	2	0.05	0.9513

To ensure that the items differed only in the extent to which they would be judged as meaningful (i.e., that the fragments were less meaningful than the idiomatic expressions and compositional phrases, which in turn should not differ from one another) as opposed to plausibility of occurrence, we conducted two norming studies. In the first norming study, participants judged the plausibility of each trigram. In the second norming study, a different set of participants rated the tokens according to how meaningful they were as units.

Norming Study 1: Plausibility

The purpose of the first norming study was to collect judgments regarding the plausibility of the idiomatic expressions, compositional phrases, and fragments. This was necessary to ensure that stimuli from each condition were equally plausible as strings in American English, despite any differences in the extent to which items from each condition conveyed coherent meanings as units.

Participants Thirty-three native speakers of American English from the Cornell undergraduate population participated in the study for extra credit (mean age = 19.69; SD = 1.74). Data from 2 participants were omitted because their overall performance fell below 80% in a random memory recall task (see below).

Materials The materials consisted of 3-word sequences of 82 idiomatic phrases, 236 compositional phrases, and 218 fragments. The 3-word sequences were presented to participants on a computer screen (one 3-word sequence at a time). As a control, 90 impossible 3-word combinations were also included. The impossible tokens were created by scrambling matching compositional phrases and fragment tokens that were not introduced in the experimental material.

Procedure Participants' task was to rate each trigram according to how plausible the sequence of words was as part of an English sentence. Participants rated each token on a 1-7 scale by pressing a key. To ensure that participants read each sequence, a random memory recall test was included. In 10% of the trials for each condition (idiomatic expression, compositional phrase, and fragment), participants were asked to type an English sentence that included the 3-word sequence they had just seen.

Results Table 3 shows the mean scores for each condition in the plausibility norming study. As the trigrams were rated as equally plausible, we submitted them to a second norming study evaluating their meaningfulness.

Table 3: Mean ratings and standard deviations of plausibility (Norming Study 1) scores for idiomatic expressions, compositional phrases, and fragments.

	Mean	SD
Idiomatic Expressions	6.84	0.32
Compositional Phrases	6.91	0.22
Fragments	6.87	0.22

Norming Study 2: Meaningfulness

The purpose of the second norming study was to collect judgments regarding the meaningfulness of the idiomatic expressions, compositional phrases, and fragments as units.

Participants A different set of 33 native speakers of American English from the Cornell undergraduate population (mean age = 19.84; SD = 1.15) participated in the study for extra credit. Data from two subjects were removed because their overall performance on the random memory recall test (see below) was less than 80%.

Materials The materials consisted of the same 3-word sequences used in the first norming study (82 idiomatic phrases, 236 compositional phrases, and 218 fragments). The 3-word sequences as a whole were presented to participants one-by-one on a computer screen.

Procedure Participants' task was to rate each trigram according to how meaningful they found each sequence as a unit. Participants rated each token on a scale of 1-7 by pressing a key. To ensure that participants read each sequence, a memory recall test was included. In 10% of the trials for each condition (idiomatic expression, compositional phrase, and fragment), participants were asked to type an English sentence that included the last 3-word sequence they had seen.

Results Table 4 shows the mean scores for each condition in the meaningfulness norming study.

Table 4: Mean ratings and standard deviations of meaningfulness (Norming Study 2) scores for idiomatic expressions, compositional phrases, and fragments.

	Mean	SD
Idiomatic Expressions	5.90	1.02
Compositional Phrases	5.89	0.86
Fragments	1.98	0.48

In order to arrive at the final set of experimental items, we submitted the following to a selection algorithm (more details below): the arcsine-transformed proportion of subjects rating each item 6 or 7 (1 or 2 for fragments) in Norming Study 2, the arcsine transformed proportion of subjects rating each item as a 6 or 7 in Norming Study 1, and the whole- and sub-string frequencies of each item. The algorithm selected the set of 40 triads (each comprising an idiomatic expression, a compositional phrase, and a fragment) which differed least according to plausibility norming scores as well as phrase and substring frequencies across the three trigram types, while also differing maximally in Norming 2 scores between fragments and the other two conditions.

Reaction Time Study

Our prediction was that the overall meaningfulness of a sequence would facilitate processing over and above mere frequency of use, independently whether the sequence was an idiomatic phrase. Thus, we predicted that processing latencies for idiomatic phrases and compositional meaningful phrases would not differ from one another. However, both idiomatic phrases and compositional phrases should show processing advantages over fragments.

Participants An additional 40 native speakers of American English from the Cornell undergraduate population were recruited, none of which participated in either of the two norming studies (mean age = 20.5; SD = 1.58).

Materials A final set of frequency-matched tokens from the trigrams rated in the two norming studies was selected for the reaction time study (using a selection algorithm which sought to minimize differences along the frequency dimensions as well as the norming scores between the conditions). This set consisted of 40 triads, each comprising an idiomatic phrase, a compositional phrase and a fragment. The resulting set of 40 idiomatic expressions, 40 compositional phrases and 40 fragments did not differ significantly along the 6 frequency dimensions (trigram, first bigram, second bigram, first unigram, second unigram, third unigram) or according to the percentage of subjects rating items as 6 or 7 in the first plausibility norming study. Additionally, the items were constrained such that the idiomatic and compositional phrases did not differ in terms of their meaningfulness ratings from second norming study, whereas the fragments were chosen to have the lowest meaningfulness scores possible. All comparisons: $p > 0.4$.

The $\log_2$ phrase frequencies of the final set of 40 triads introduced in the behavioral study ranged between 1 and 10.4. Besides the 40 experimental triads, 120 impossible sequences (such as *hear I isn't*) were used as fillers.

Procedure We based our reaction time study on Arnon and Snider's (2010) phrasal decision task (which in turn is based on the classic lexical decision task). Participants were presented with the three-word sequences (120 experimental and 120 impossible filler tokens) separately, in random order, on a computer screen, and asked to judge (by quickly hitting one of two keys) whether they formed possible word combinations in the context of English sentences.

Data Analysis Data points corresponding to reaction times of less than 200 ms were removed, along with extreme outliers (defined as those reaction times exceeding the upper quartile by more than three times the inter-quartile range), resulting in a 1.5% data loss. The data were then submitted to a linear mixed-effects (LME) analysis, with Item and Subject as random effects, and the scores from the second norming study (hereafter referred to as the Meaningfulness Scores), 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), Length in Characters, and the Trigram Type x Frequency interaction term as fixed effects. Because the Meaningfulness scores had a fixed range of 1 to 7, they were converted to proportions (n/7), which were then logit-transformed prior to entry in the model (cf. Armitage & Berry, 1984). As the reaction times were not normally distributed, they were log-transformed prior to the analysis.

The LME resulted in no significant interaction between Frequency and Trigram Type and was therefore simplified to involve Item and Subject as random effects, and Meaningfulness scores, Trigram Type (using Idiomatic Expressions as the base case), Length in Characters, Frequency (whole-string), and the substring frequency predictors as fixed effects. There did not appear to be substantial multicollinearity between the fixed effects: The condition number for the matrix of predictors (cf. Belsley, Kuh, & Welsch, 1980) was only 6.6.

The final model was compared to a version without the fixed effects (χ^2 = 140.63, $p < 0.0001$) as well as a version of the model without the variables of interest (Meaningfulness Score, Frequency, Trigram Type; χ^2 = 111.07, $p < 0.0001$), indicating that the full model captured more of the variance in both cases.

Results

As predicted, participants showed sensitivity to the meaningfulness of the trigrams. Decision times were faster for more meaningful tokens, as revealed by a highly significant main effect of Meaningfulness Score (β = -0.037, $p < 0.0001$). Also in line with our predictions, reaction times were affected by trigram type: It took longer for participants to decide whether fragments were possible strings in English (β = 0.089, $p < 0.05$). However, decision times for compositional meaningful phrases were no slower than for idiomatic expressions (β = 0.003, $p > 0.3$). Table 5 presents the mean RTs and standard deviations for each condition.

Table 5: Mean RTs and standard deviations for idiomatic expressions, compositional phrases, and fragments.

	Mean	SD
Idiomatic Expressions	766.4	247.9
Compositional Phrases	789.4	260.3
Fragments	949.9	343.7

Frequency (whole-string) also reached significance (β = -0.014, $p < 0.05$) indicating that subjects responded faster to trigrams with greater whole-string frequency. Importantly, the effect of Frequency was less than that of Meaningfulness Score.

Of the fixed effects included to control for substring frequency and character length, only Length in Characters (β = 0.08, $p < 0.05$) and Third Unigram (β = 0.014, $p < 0.05$) reached significance.

General Discussion

In this study, we compared the processing latencies for triads consisting of idiomatic expressions, compositional phrases and less meaningful fragments. In each triad, both the meaningful compositional phrases and the fragments were frequency-matched to the idiomatic phrases. The aim of the study was to investigate (i) whether the relative meaningfulness of multiword chunks affects processing latencies in addition to their frequency, and (ii) whether idiomatic and meaningful compositional phrases are processed similarly, once frequency is adequately controlled for.

Our results indicate that participants were sensitive to the meaningfulness of the chunks. The meaningfulness of a given trigram—as indicated by the second norming study—successfully predicted reaction times in the final experiment. Participants' decision times for more meaningful trigrams were faster than for less meaningful ones. Furthermore, our findings show that the whole-string frequency of the tokens predicted the processing latencies, but to a lesser extent than the meaningfulness of the different chunks. These results thus suggest that in addition to frequency of occurrence, as emphasized by usage-based theories, the relative meaningfulness of multiword chunks should also be considered in accounts of language processing.

Our findings also showed that processing latencies for idiomatic phrases did not differ from frequency-matched compositional meaningful phrases, while processing latencies for less meaningful fragments were significantly greater. This suggests that meaningful compositional sequences may be represented and processed similarly to idiomatic phrases, despite their compositional nature. This is at odds with traditional distinctions between stored idioms and compositional phrases (Swinney & Cutler, 1979). However, these results are in line with Construction Grammar approaches, suggesting that there are no fundamental differences between the representation and processing of idiomatic constructions and compositional phrases; they are both instances of conventionalized form-meaning mappings (Goldberg, 2003).

Additionally, higher meaningfulness scores were associated with reduced processing latencies. These results, when viewed alongside the weaker frequency effect, provide a step forward for studies of formulaic language, suggesting that that the meaningfulness of multiword chunks may be as important to their processing as their distributional properties.

It is possible that the processing latencies for compositional phrases in the current study were affected by their status as constituents, whereas fragments crossed syntactic boundaries. There exists only one study that has investigated whether constituency affects multi-word sequence processing while controlling for frequency. In a recent production study, Arnon and Cohen-Priva (in press) found that constituency did not affect processing. In their study, similar frequency effects were found when

comparing high- and low-frequency variants of constituents (*a lot of work* vs. *a lot of years*) as well as non-constituents crossing syntactic boundaries (*as far as I* vs. *as far as you*). Similar frequency effects for both phrases and fragments suggest that constituency may be a less important feature of multiword chunk processing. On the other hand, it is possible that the lack of a constituency effect in the Arnon and Cohen-Priva study stems from not taking chunk meanings into account. Further studies are needed to examine the exact nature of the relationship between constituency and chunk meaningfulness.

To conclude, our results provide new insights into the representation and processing of formulaic expressions; they suggest that multiword compositional phrases that people find highly meaningful are likely to be processed similarly to idiomatic phrases, as a linguistic unit in its own right. Our findings are thus relevant for usage-based approaches to language, indicating that meaning provides an additional dimension that such approaches must take into account, in line with a number of expectations derived from cognitive linguistics (e.g., Langacker, 1987).

Acknowledgments

We would like to thank Jaclyn Jeffrey-Wilensky and Julia Krasnow for their assistance with the reaction time study. Thanks also go to Jaclyn Jeffrey-Wilensky and László Kálmán for their help during material selection. Finally, we are grateful to Haim Bar for his valuable comments on our statistical analyses.

References

Armitage, P. & Berry, G. (1994). Statistical methods in medical research (3rd ed.). Oxford: Blackwell.

Arnon, I. & Cohen-Priva, U. (in press). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech*.

Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*, 67-82.

Bannard, C. & Matthews, D. (2008). Stored word sequences in language learning. *Psychological Science*, *19*, 241-248.

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

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

Bybee, J., & Hopper, P. (2001). *Frequency and the emergence of linguistic structure.* Amsterdam: John Benjamins.

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

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

Fillmore, C., Kay, P., & O'Connor, M. K. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, *64*, 501-38.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*, 219-224.

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

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

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

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

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.

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

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

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behaviour*, *18*, 523-34.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge: Harvard 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 (pp. 151-167). In D. Wood (Ed.), *Perspectives on formulaic language*, NY: 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.

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.

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