Is all formulaic language created equal? Unpacking the processing advantage for different types of formulaic sequences

Gareth Carrol, University of Birmingham

Kathy Conklin, University of Nottingham

Corresponding author:
Gareth Carrol
Department of English Language and Linguistics
University of Birmingham
Edgbaston
Birmingham
B15 2TT
+44 (0)121 414 9060
g.carrol@bham.ac.uk

Abstract

Research into recurrent, highly conventionalised 'formulaic' sequences has shown a processing advantage compared to 'novel' (non-formulaic) language. Studies of individual types of formulaic sequence often acknowledge the contribution of specific factors, but little work exists to compare the processing of different types of phrases with fundamentally different properties. We use eye-tracking to compare the processing of three types of formulaic phrases–idioms, binomials and collocations–and consider whether overall frequency can explain the advantage for all three, relative to control phrases. Results show an advantage, as evidenced through shorter reading times, for all three types. While overall phrase frequency contributes much of the processing advantage, different types of phrase do show additional effects according to the specific properties that are relevant to each type: frequency, familiarity and decomposability for idioms; predictability and semantic association for binomials; and mutual information for collocations. We discuss how the results contribute to our understanding of the representation and processing of multiword lexical units more broadly.

Keywords:

Formulaic language, lexical processing, idioms, collocations, binomials, eye-tracking

Introduction

Cheetahs and Ferraris are two examples of members of the category ‘things that are fast’. Beyond this broad similarity, there is not much that makes the two particularly comparable, and it is clear that the mechanisms that make each one fast are very different. In linguistics, formulaic language is an example of something that has sometimes been defined just as broadly. It encompasses a broad range of multiword sequences that fulfil a number of communicative functions (Wray, 2002, 2008), and knowledge of such sequences is an important part of how we use language. For example, in English we implicitly know to describe coffee as ‘strong’ not ‘powerful’, or to ask for ‘salt and pepper’ not ‘pepper and salt’.

Two features are common to all examples of what is generally considered under the heading of formulaic language. The first is that they must be recurrent, in the sense that they occur in natural language more frequently than comparable novel phrases. What counts as the threshold for “frequent” is an open question, but widespread evidence now supports the second feature: faster processing of recurrent sequences compared to “novel” control phrases, or, put another way, frequency effects at the multiword level, in line with usage-based accounts of how language develops and is organised (e.g., Bybee, 2006; Goldberg, 2006; Tomasello, 2003). Often, claims are made about the “holistic” nature of formulaic sequences, suggesting that all recurrent sequences are simply stored in the lexicon and retrieved directly, although the nature of what is meant by this may also be quite variable (c.f. Wray, 2012, p.234).

Beyond this very broad designation, what counts as “formulaic” can differ widely, and a view of “holistic” representation may not tell the whole story. Formulaic sequences vary along a number of important dimensions (Titone, et al., 2015), such as their degree of

fixedness/conventionalisation, their schematicity (whether they allow for internal variation through open "slots"), their semantic unity, the degree of compositionality, and the function they perform (see Buerki, 2016, for a very useful overview). As a result, when we look at the types of phrase that are usually included in the category of formulaic language, we may find that they are just as difficult to compare as cheetahs and Ferraris. For example, idioms are semantically opaque, self-contained figurative phrases, such as *kick the bucket*. Multiple studies attest the *idiom superiority effect*, whereby idioms are processed more quickly than matched control phrases (e.g., Gibbs, 1980; McGlone, Glucksberg & Cacciari, 1994; Rommers, Dijkstra & Bastiaansen, 2013; Swinney & Cutler, 1979; Tabossi, Fanari & Wolf, 2009). In comparison, collocations are very broadly defined as pairs of words that occur together more frequently than we would expect by chance, such as *strong coffee*. They are generally transparent and have a literal meaning that is the result of combining the component words, but these too have been shown to be processed more quickly than non-formulaic comparators (e.g., Bonk & Healy, 2005; Durrant & Doherty, 2010; Sonbul, 2015; Vilkaite, 2016; Wolter & Gyllstad, 2011). Other types of sequence have been shown to demonstrate a similar processing advantage: binomials (e.g., Arcara, et al., 2012; Siyanova-Chanturia, Conklin & van Heuven, 2011); phrasal verbs (e.g., Blais & Gonnerman, 2013; Kim & Kim, 2012; Matlock & Heredia, 2002; Paulmann, Ghareeb-Ali & Felser, 2015); and lexical bundles (e.g., Arnon & Snider, 2010; Bannard & Matthews, 2008; Bod, 2001; Hernández, Costa & Arnon, 2016; Tremblay & Baayen, 2010; Tremblay, Derwing, Libben & Westbury, 2011).

In this paper we set out to test how far the distributional properties of formulaic sequences can explain the processing advantage, regardless of specific features that may vary from type to type. In other words, is it simply that speakers register occurrences of frequently occurring phrases, or are additional properties important in why formulaic sequences seem to be

processed more efficiently than novel strings? Results from studies like those cited above for lexical bundles would suggest that frequency of occurrence, above all else, contributes to faster processing, hence any further distinctions may be fairly arbitrary and of little actual use. For example, Tabossi et al. (2009) showed that although idioms were judged to be meaningful more quickly than control phrases, the same was true for what they called clichés (entirely compositional frequently occurring phrases). They argued that the *idiom superiority effect* is a property of any frequently occurring phrase, regardless of other aspects such as semantics. On the other hand, Jolsvai, McCauley and Christiansen (2013) showed that the meaningfulness of a word sequence was an important factor in how it was processed in a phrasal decision task, over and above simply how frequently it occurs. Gyllstad and Wolter (2016) found that frequency but also transparency were important factors in how participants judged word combinations, with shorter reaction times for more transparent combinations (free combinations as opposed to restricted collocations).

In this paper we select three different types of formulaic sequence–idioms, binomials and collocations–which all qualify as "formulaic" from the point of view of being recurrent phrases. Beyond this, they differ markedly in terms of specific properties that contribute to their formulaic status. For example, idioms are broadly (but variably) non-decomposable, and in all cases the meaning of the whole phrase must be retrieved directly from the lexicon to some degree. In contrast, binomials–sequences of *x-and-y* where a specific word order is highly preferred, such as *salt and pepper*–can be literal or figurative and often constitute semantic associates, but are highly fixed in the sense that the reversed form is rarely if ever used. Collocations are very broadly defined as co-occurring word pairs, and here we define them as combinations of words that are entirely compositional and semantically "free", but which co-occur in conventional and recurrent patterns. Crucially, based on the established body of evidence, we expect all three types of formulaic language (idioms, collocations and

binomials) to show a processing advantage relative to control phrases. Our first aim is to see how far the broad distributional property of phrase frequency can explain the overall differences, and whether this alone can account for the observed processing effects. We then go on to explore how aspects of predictability, and type-specific constraints contribute to processing, over and above the effects of frequency. Note that throughout this paper, predictability refers to the expectancy for the final word of a formulaic sequence, once the initial word or words have been seen. Various factors, such as context, are likely to determine this, but we do not consider these in detail here.

Idioms

Idioms are amongst the most studied of all formulaic phrases, and have been described as "prototypically" formulaic (Siyanova-Chanturia & Martinez, 2014). Titone et al. (2015) suggest that this is because as a class, idioms can vary along all of the dimensions that are relevant to the study of formulaic language more generally, including frequency, familiarity, transparency, decomposability, literalness, etc. Importantly, as well as being clear examples of formulaic sequences, idioms are also often included in the class of figurative (non-literal) language, which may further have a bearing on how they are processed and understood. Early non-compositional models (e.g., Bobrow & Bell, 1973; Gibbs, 1980; Swinney & Cutler, 1979;) suggested that idioms are highly lexicalised entries that can be retrieved directly, but subsequent work has shown that idioms are not simply 'long words', and do undergo compositional analysis (Cacciari & Tabossi, 1988). More recent 'hybrid' models (e.g. Sprenger, Levelt and Kempen, 2006; Titone & Connine, 1999) see idioms as both single entries and compositional wordstrings, and are supported by widespread evidence demonstrating that idioms show internal syntax (Cutting & Bock, 1997; Konopka & Bock, 2009; Peterson, Burgess, Dell & Eberhard, 2001) and that the literal meanings of component

words are activated as an obligatory part of processing (Hamblin & Gibbs, 1999; Holsinger & Kaiser, 2013; Smolka, Rabanus & Rösler, 2007; Titone & Connine, 1994).

A robust finding is that idioms are generally recognised more quickly than matched 'novel' phrases. For example, Swinney and Cutler (1979) used a phrasal decision task to show that idioms (e.g., *break the ice*) are judged to be meaningful phrases more quickly than control phrases (e.g., *break the cup*). Recent eye-tracking research has also supported the fast processing of idioms compared to control phrases (Carrol & Conklin, 2017; Carrol, Conklin & Gyllstad, 2016), regardless of whether idioms are used in figurative or literal contexts (Siyanova-Chanturia, Conklin & Schmitt, 2011). The literature suggests that such effects stem from the fact that idioms are highly familiar (Schweigert, 1986, 1991; Schweigert and Moates, 1988), and predictable (Fanari, Cacciari & Tabossi, 2010; Libben & Titone, 2008), hence it seems likely that idioms are recognised quickly primarily because they are well-known phrases (Van Lancker Sidtis, 2012).

However, the role of decomposability (how much the figurative meaning of an idiom can be mapped onto the literal meanings of the component words) remains open to debate. For example, the idiom decomposition hypothesis (Gibbs & Nayak, 1989; Gibbs, Nayak & Cutting, 1989) proposed that only decomposable idioms should be processed quickly, since analysis of the phrase would be consistent with a literal reading. Results here have been mixed, with some studies showing support (Caillies and Butcher, 2007), and others showing the opposite pattern (Cieślicka, 2013; Titone & Libben, 2014) or no difference (Tabossi, Fanari & Wolf, 2008). Libben and Titone (2008) only found effects of decomposability on meaningfulness judgements for less familiar idioms, suggesting that familiarity may "trump" other aspects in how idioms are recognised, represented and understood (c.f. Abel, 2003; Carrol, Littlemore & Dowens, 2018).

For the purposes of the present study, it is useful to make a distinction between familiarity (which is a person-specific, subjective measure) and frequency, although these may often be highly correlated. Idioms as a class are widespread (Brenner, 2003, estimated that English contains over 10,000 idioms), but individual idioms occur relatively infrequently, at least based on corpus evidence (Moon, 1998). We therefore aim to assess how far frequency alone can explain the *idiom superiority effect*, before considering the additional contributions of specific properties such as decomposability and familiarity.

Binomials

Binomials have generated less interest in the literature than idioms, and represent an example where their formulaicity comes from an entirely different property. Here we define binomials as combinations of *x-and-y* where a reversal of the order is entirely possible, but where one word order is highly conventionalised.[1] Examples are most often noun-and-noun (e.g., *salt and pepper, king and queen*), and a complex set of variables have been identified that determine the order (e.g., Benor & Levy, 2006; Cooper & Ross, 1975; Lohmann, 2012; Mollin, 2012; Morgan & Levy, 2016). These include conceptual factors (e.g., general before specific, animate before inanimate), cultural restrictions (e.g., power relations) and phonological variables (e.g., length and stress patterns of each word), amongst others. However, for each constraint there are also frequent exceptions (e.g., for 'male before female': *man and wife, men and women* but *bride and groom*), and the overriding factor

[1] We exclude what have been called irreversible binomials (e.g., Arcara et al., 2012) – phrases such as *hit and run* where the order is not only iconic and logical, but also where the phrase itself has a meaning over and above the constituent parts, therefore effectively operating as an idiom.

seems to be conventionalisation of one order over the other. Morgan and Levy (2016) suggest that for any example there is a trade-off between abstract knowledge of these constraints and direct experience.

In a recent study, Conklin and Carrol (2016) showed that frequency effects emerge rapidly for novel binomials in a natural reading task. For non-attested binomials (e.g., *grass and leaves, plates and bowls*, where there is no highly conventionalised order) a processing advantage (incrementally faster reading times) was observed in as few as three or four presentations, confirming the importance of phrasal frequency in how binomials become fused in a particular order. Siyanova-Chanturia, Conklin and van Heuven (2011) argued that frequency only accounts for some of the effects observed for attested binomials. They showed faster reading times for binomials (e.g., *bride and groom*) compared to reversed forms (*groom and bride*), but their analysis suggested that phrasal frequency explained only part of the effect. They proposed that the configuration itself played a vital role, in that the preferred (binomial) form was privileged compared to the dispreferred (reversed) form, even when overall phrase frequency was accounted for.

An additional consideration with many binomials is that they often represent primary semantic associates (*knife-fork*, *king-queen*, *salt-pepper*), which may also contribute to how they are processed. That is, as well as frequency effects for the whole phrase, and direct representation of the configuration for the most common binomials, semantic priming between component words may contribute to faster processing. Carroll and Slowiaczek (1986) found within sentence priming for semantically related words (e.g., *author-book*) when these appeared within the same clause. In a similar study, Camblin, Gordon and Swaab (2007) found that association priming effects were only robust when the overall discourse context was impoverished or not cohesive. Given such results, it is likely that the semantic relatedness of binomial word pairs play at least some role in how they are processed,

although conventionalization is assumed to be the single biggest factor. Note that this does not necessarily equate to frequency, as a phrase may be fairly infrequent but have a highly fixed order, or may be very frequent but occur just as often in the reversed form. Both characteristics will be explored in this study.

Collocations

Collocations can be very broadly defined as any frequently co-occurring words, or, more accurately, words that co-occur more frequently than we might expect by chance (Biber, Johansson, Leech, Conrad & Finegan, 1999). They can often be at least partially figurative, and various classifications have been proposed that consider collocations from a phraseological point of view (e.g., Howarth, 1998). From a frequency-based perspective (e.g., Sinclair, 1991), collocations are defined according to certain corpus-derived metrics, such as t-scores (a test of the null hypothesis that there is no connection between two words) or mutual information (MI), which measures the strength of co-occurrence between two words that form a collocation. Typically, an MI score of 3 is taken as the threshold above which a word pair can be considered to be of linguistic interest (Hunston, 2002).

Importantly, various types of collocations have been shown to demonstrate the same processing advantage as the other formulaic types considered so far.[2] Wolter and Gyllstad (2011) showed that native speakers of English are faster and more accurate when responding to adjective + noun collocations (relative to non-formulaic baseline word combinations) in a primed lexical decision task. In a subsequent study, Wolter & Gyllstad (2013) showed

[2] Many of the studies on collocations discussed here also investigate non-native processing, but we concentrate here only on results for native speakers as these are more relevant to our study.

facilitation for verb + noun combinations on a phrase-level grammaticality judgement task. Durrant and Doherty (2010) used a primed lexical decision task for low, mid and high frequency collocations, as well as for high frequency collocations that were also semantic associates. In an unmasked priming task they found priming for only high frequency word pairs (MI higher than 6), and this effect was observed for both associated and non-associated word pairs. In a second study using a masked prime, facilitation was only observed for those word pairs that were high frequency collocations *and* semantic associates. They suggested that, firstly, the threshold at which collocations become psychologically 'real' may be much higher than that adopted in the corpus literature, and secondly, the mental representation of collocations may depend on both frequency of encounter and semantic association.

Sonbul (2015) used both offline measures (a rating task asking how typical a phrase is in English) and online measures (eye-tracking) to investigate responses to adjective + noun collocations. She compared synonymous word pairs of high frequency (e.g., *fatal mistake*), lower frequency (e.g., *awful mistake*) and no frequency (i.e., non-collocations, e.g., *extreme mistake*). There was a clear effect of frequency in the offline task (higher frequency collocations were rated as more typical). In the eye-tracking task, early measures showed an effect of frequency (first pass reading time was shorter for higher frequency collocations), but this disappeared in later measures (overall reading time and fixation count). This suggests that while higher frequency collocations may be more easily recognised, they are not necessarily easier to process and integrate into context than lower frequency synonymous word pairs, hence different factors may be important at different stages of processing.

Finally, Vilkaite (2016) used eye-tracking to show that verb + noun collocations (e.g., *provide information*) were read more quickly by native speakers than control phrases (e.g., *compare information*), both in their canonical adjacent configurations, and when they were separated by three words (e.g., *provide some of the information*). Vilkaite interpreted these

result as arguing against a 'holistic' hypothesis in how collocations are processed. Instead, they support more general probability-based models (e.g., Barlow & Kemmer, 2000; Kuperberg & Jaeger, 2016). The results are also in line with Hoey's theory of lexical priming (2005; also Pace-Sigge, 2013), and the notion of 'concgrams' described by Cheng, Greaves, Sinclair and Warren (2009), which are co-occurrences of words regardless of whether they are sequential or adjacent. Such approaches allow for much more flexibility in how multiword sequence are conceived.

Overall, there is clear evidence that frequently occurring collocations are processed quickly. Frequency alone may explain much of this, but other factors such as mutual information (a measure of strength of co-occurrence, rather than simply how frequent a collocation may be) are also important. Our study will enable us to explore how each of these contributes to overall processing patterns.

Comparing formulaic subtypes

Despite the widespread research into specific types of formulaic expressions, there remains a relative lack of work directly comparing subtypes. Columbus (2010) compared reading time for idioms, lexical bundles and restricted collocations. All three types were read more quickly than non-formulaic controls, and idioms were processed the most quickly overall. She concluded that these differences may not be the result of the different subtypes per se, but that different variables relevant to each type produce different effects. Columbus (2013) went on to show that both corpus data and human ratings can reliably distinguish between subtypes, using measures such as frequency, familiarity and perceived transparency. How these factors influence online processing remains to be explored, although as noted earlier, Gyllstad and Wolter (2016) showed that both transparency and frequency affected reaction times in their phrasal decision task. That is, restricted collocations were judged more slowly

than free combinations, but within each category more frequent phrases were judged more quickly. Other studies reviewed here have also shown that multiple factors may be at play in how formulaic language is processed, and aspects of conventionalisation (of which frequency is a reflection) may only tell us part of the story.

Importantly, little of the work discussed thus far has involved natural reading, as most tasks required overt responses or judgements. We therefore aim to explicitly compare the contribution of a range of factors to understand how different kinds of formulaic sequence are processed in more natural contexts. Our overall research questions are: 1) How far do distributional variables (frequency and predictability) explain the processing advantage for different kinds of formulaic phrase?; 2) how do different variables/constraints contribute to the processing of each type of phrase?

Experiment

Materials

Items were selected based on our definitions for each formulaic type introduced in the previous section, using a range of linguistic and distributional criteria. Frequency values were taken from the British National Corpus (BNC).

Idioms were selected from the Oxford Learner's Dictionary of Idioms (Warren, 1994) and previously published idiom studies. They were all of the form Verb-X-Noun (e.g., *kick the bucket, push your luck*) or Preposition-det-Noun (*behind the scenes, below the belt*). All items occurred at a phrase frequency of at least 11 per 100 million words (mean = 54, SD = 53). To ensure that they were generally well known to native speakers they were included in a rating task where native speaker participants (n = 21) rated each for how familiar it was on a seven-point scale. A final list of 45 items that scored highest for familiarity was created (mean = 5.9/7, SD = 0.7).

Binomials were of the form X-and-X, where both words were of the same lexical class. All phrases were from online lists and previously published studies. In order to trim this list down, we adopted a minimum phrase frequency of 20 per 100 million words (mean = 251, SD = 221), and a minimum ratio of 4:1 in favour of forward to reversed occurrences (i.e., a binomial must occur at least four times as often in the forward form than in the reversed form, mean = 9.3, SD = 0.7). All items were included in a rating study with native English speakers (n = 48) to assess how figurative/literal they were and to assess their reversibility (whether participants thought that the phrase could be reversed and the meaning retained). Figurative/Literalness was assessed on a scale from 1-3, with 1 being entirely figurative, 3 being entirely literal, and 2 being potentially both (mean = 2.7/3, SD = 0.36). Reversibility was assessed on a scale from 1-7 (mean = 4.6/7, SD = 1.1). There was a high correlation here, with more figurative idioms being seen as less reversible (r = .64, p < .001). A final list of 45 items was created, including only binomials that were considered to be broadly literal and reversible. The majority were noun-and-noun (e.g., *salt and pepper*, n = 31), with some verb-and-verb (e.g., *pick and choose*, n = 3), adjective-and-adjective (e.g., *sick and tired*, n = 10) and preposition-and-preposition (e.g., *out and about*, n = 1).

For collocations we extracted a list of commonly co-occurring adjective-noun combinations from the BNC. We only considered non-idiomatic (non-figurative) examples, and all items occurred at least 10 times per 100 million words (mean = 110, SD = 133), with a minimum MI score of 3 (mean = 6.7, SD = 2.2) and a minimum t-score of 2 (mean = 9, SD = 4.8). Forty-five items were selected based on these criteria.

For all items we created two control phrases: for Control Type 1 phrases, the first word was changed (e.g., *spill the beans* became *drop the beans*); and for Control Type 2 the second word was changed (e.g., *spill the beans* became *spill the chips*). This is a potentially important comparison as most studies have generally compared formulaic phrases to controls

where the final word is changed (e.g., *break the ice* vs. *break the cup*, Swinney & Cutler, 1979). By comparing two types of control phrase we will be able to see whether there is a formulaic advantage compared to one, both, or neither. There is also a key difference between the two types of control phrase. For the comparison of formulaic to control type 1 phrases, we are comparing a more frequent phrase with a less frequent one. We are also comparing a phrase where seeing word 1 may generate a strong expectancy for what word should follow (formulaic condition) compared to a phrase where this is not the case. For control type 2 phrases, we are again comparing a more frequent with a less frequent phrase, but additionally we are comparing phrases where the same expectation generated by word 1 is met (formulaic condition) or not met (control type 2 condition). A list of all formulaic and control items is provided in the supplementary materials.

Each control phrase was matched with its formulaic comparator for single word length and frequency. We then created short sentences for all items (see Table 1). These were designed to be as neutral a possible before the critical phrase was seen, in order to minimise any effects of context that might make a particular completion more or less predictable. The sentences were designed so that the formulaic phrase and both of its controls would make sense in the same context. The immediate post-context (the words following the critical phrase) was also the same to ensure that no contextual information could be extracted in the parafovea during processing of the final word of the phrase, then the final part of the sentence was tailored so that each version was completed in a coherent manner (n.b. in most cases the binomials and collocations fitted into the same context as the controls, so the same post-context was used for all three). We made sure that the number of words preceding (mean = 3.8, SD = 0.8) and following (mean = 11.9, SD = 2.1) the critical phrase was similar for all items. To ensure that all items (formulaic phrase and controls) were equally plausible sentences, we included all items in a rating study and asked native speakers (n = 25) to judge each sentence for how

acceptable it was on a scale from 1-5. There were no differences between formulaic units and either control (Formulaic, mean = 4.4, SD = 0.4; Control Type 1, mean = 4.4, SD = 0.4; Control Type 2, mean = 4.4, SD = 0.4; one way ANOVA by condition: $F = 0.13$, $p = .876$).

Table 1. Examples of context sentences for idioms and control phrases.

	Pre-context	Phrase	Post-context
Idiom	It was hard not to	spill the beans	when I heard such a juicy piece of gossip.
Control type 1	It was hard not to	drop the beans	when I burned my hand on the hot pan.
Control type 2	It was hard not to	spill the chips	when I stumbled on my way out of the kitchen.
Binomial	I heard that the	king and queen	will be visiting the city next week.
Control type 1	I heard that the	prince and queen	will be visiting the city next week.
Control type 2	I heard that the	king and prince	will be visiting the city next week.
Collocation	They were in	abject poverty	but they seemed to make the best of their situation.
Control type 1	They were in	total poverty	but they seemed to make the best of their situation.
Control type 2	They were in	abject agony	but they seemed to make the best of their situation.

We collected a range of other data for the experimental items (formulaic and control phrases). To make the frequency counts comparable between phrase types, we converted all raw values to the Zipf scale (Van Heuven, Mandera, Keuleers & Brysbaert, 2014). We then calculated two measures of predictability for the final word of each phrase. The first was a Cloze probability score for all formulaic and control items, which we considered to be a subjective measure of how predictable the phrase was in a short, neutral context. Phrases were presented as the first part of the sentence up to (but not including) the final word of the phrase, e.g., *It*

was hard not to spill the…, and we asked participants (native speakers of English, n = 69, with participants seeing one of four versions of the materials) to provide the first word that came to mind that could plausibly continue the sentence. It was stressed that these were sentence fragments, and that the word did not have to complete the sentence, simply to continue it in a reasonable way. Cloze probability was calculated as the percentage of participants who provided the correct (or intended) completion in each case. We also calculated a measure of transitional probability for each formulaic and control phrase. This is an objective measure of how likely it is that word 2 follows word 1 based on corpus frequencies. We follow the same formula used by McDonald and Shillcock (2003) and Frisson, Rayner and Pickering (2005), although this was adapted to calculate the likelihood of the final word following the first word and the determiner or conjunction in idioms and binomials. Transitional probability was calculated as Overall phrase frequency ÷ Frequency of Word 1 (+ determiner or conjunction) * 100, e.g., *spill the beans* (39) ÷ *spill the* (93) * 100 = 42%. Finally, for all items we obtained semantic association scores between the two content words using the Edinburgh Associative Thesaurus (EAT: Kiss, Armstrong, Milroy & Piper, 1973). A summary of the properties of the items is presented in Table 2.

Table 2. Example phrases and item characteristics for all stimuli. Phrase length is measured in total characters, including spaces; phrase frequency is expressed on the Zipf scale (1-7); Transitional Probability and Cloze Probability are scores out of 100; Association Strength is the strength of semantic association based on EAT scores and is also out of 100. Standard deviations are provided in brackets.

	Phrase	Phrase Length	Phrase Frequency	Cloze Probability	Transitional Probability	Association Strength
Idiom	*Spill the beans*	12.3 (1.6)	2.6 (0.3)	35.6 (26.1)	10.1 (13.3)	0.2 (0.6)
Control type 1	*Drop the beans*	12.2 (1.8)	1.7 (0.7)	3.2 (6.6)	0.5 (0.9)	0.2 (0.6)
Control type 2	*Spill the chips*	12.4 (1.5)	1.7 (0.6)	4.3 (8.2)	0.9 (1.5)	0.0 (0.2)
Binomial	*King and queen*	12.6 (1.8)	3.2 (0.4)	67.7 (33.8)	28.5 (19.0)	29.9 (26.1)
Control type 1	*Prince and queen*	13.3 (2.2)	1.5 (0.5)	12.9 (19.9)	0.9 (1.4)	3.6 (8.5)
Control type 2	*King and prince*	12.4 (1.8)	1.7 (0.5)	4.0 (13.9)	1.0 (1.4)	0.8 (2.0)
Collocations	*Abject poverty*	11.3 (2.2)	2.9 (0.4)	17.8 (26.6)	3.9 (7.2)	3.4 (7.0)
Control type 1	*Total poverty*	11.5 (2.4)	1.8 (0.4)	1.9 (6.8)	0.1 (0.1)	0.0 (0.0)
Control type 2	*Abject agony*	11.4 (2.1)	1.6 (0.4)	1.0 (5.6)	0.2 (0.4)	0.0 (0.0)

Participants

Thirty-six English native speaker undergraduate students took part in the experiment for course credit. They were randomly assigned to one of the three presentation lists.

Procedure

The experiment was administered on an Eyelink 1000+ eye-tracking system from SR Research. Stimuli were presented on a 1920 x 1080 computer monitor (refresh rate 60Hz). Eye movements (left-eye, monocular recording) were monitored using a desk-mounted camera (sample rate 500hz). Following initial setup, a nine-point calibration and validation procedure was used to verify accuracy, and was repeated at regular intervals throughout the experiment.

Participants were asked to read each sentence for comprehension and to press the spacebar when they had finished. Each sentence was preceded by a fixation cross to allow for trial by trial drift checking. One third of the sentences were followed by a simple yes/no comprehension question to ensure that participants paid attention throughout. Accuracy was high overall and comparable across all items (mean = 93%; scores ranged from 88% to 98% for individual subtypes/conditions). Participants saw a total of 180 sentences, including 45 filler sentences, with a short break and recalibration after every 60 items. Participants were randomly assigned to one of the three presentation lists (A, B and C).

Results

Two participants were removed because of technical problems, leaving data from 34 subjects (12 from List A, 11 from Lists B and C). As List was not a significant factor in any of the subsequent analysis, data were collapsed across lists. All eye tracking data was cleaned according to the default settings in the four-stage procedure within the Eyelink Data Viewer program. Here, very short fixations are first merged with neighbouring fixations within a specified distance (first stage = fixations below 40ms and within 0.5 degrees; second stage = 40ms and within 1.25 degrees), then any instances of three consecutive fixations below 140ms are merged into one fixation. Finally, any remaining fixations below 80ms or longer than 800ms are removed entirely, as these are assumed to represent, respectively, minor location errors rather than true fixations, and momentary losses of concentration. Data were also visually inspected and any individual trials where data was clearly unusable due to poor calibration, track loss had occurred, or where the whole phrase was skipped (received no fixations at all), were discounted. In total 4.6% of data were excluded based on these criteria, leaving 4379 data points for analysis.

We concentrate our analysis on both whole phrases and the final words (see Carrol & Conklin, 2014, for a rationale of this with formulaic sequences). Phrase level effects might reflect the processing and integration of the phrase as a whole, while the final word is likely to be the locus of any formulaic advantage (Columbus, 2010). We consider a range of early and late measures in our analysis. Broadly speaking, early measures are thought to reflect immediate lexical processing during an initial parse of a sentence, while late measures reflect post-lexical processes and integration of meaning into the overall sentence context (Altarriba, Kroll, Scholl & Rayner, 1996; Conklin, Pellicer-Sánchez & Carrol, 2018; Inhoff, 1984; Paterson, Liversedge & Underwood, 1999; Staub & Rayner, 2007). For whole phrases and final words we consider first pass reading time (early measure: the sum duration of all fixations on the phrase the first time it is encountered in the sentence, before gaze exits to the left or right) and total reading time (late measure: the sum of all fixations on the phrase over the course of the trial, including re-reading time). For final words we also considered likelihood of skipping (also an early measure: how likely is it that the final word receives no fixations at all during first pass reading).

We constructed linear mixed effects models using the lme4 package (version 1.1-13, Bates, Maechler, Bolker, Walker, Christensen, Singmann & Dai, 2014) in R (version 3.5.1, R Development Core Team, 2018). Each eye-tracking measure was considered in its own model. We included random intercepts for subject and item, and adopted the maximal random effects structure warranted by the dataset (Barr, Levy, Scheepers and Tily, 2013). All duration measures were log-transformed to reduce skewing. For analysis of likelihood of skipping we used a logistic linear model, and skipped items were discounted from any subsequent word-level durational analysis. In all models we included single word length and frequency (on the Zipf scale) for both content words as covariates to control for any word level differences.

Omnibus analysis of distributional variables

Table 3 reports the overall reading patterns (mean and SD) for each measure, for the data as a whole, and according to subtype. For each measure we constructed a model including fixed effects of phrase type and condition, and compared the difference between levels of Condition (formulaic vs. control 1 and formulaic vs. control 2) for each phrase type using the difflsmeans function in the lmerTest package (version 3.0-1; Kuznetsova, Brockhoff & Christensen, 2016) in R. Significant differences are indicated in Table 3, and the full output of this model is provided in the Supplementary Materials. Note that the data suggests that fixation durations in general are quite low, compared to averages reported in the wider literature. In fact, the average fixation duration during the experiment as a whole (considering all fixations made) was 197ms (SD = 76), which is somewhat lower than the mean fixation duration of 225ms for silent reading reported in Rayner (1998). Crucially, the data for the regions of interest are not markedly different than reading patterns for the sentences in general. In Table 3, whole phrase data includes items where the final word was skipped (but where at least one fixation was made elsewhere in the phrase). The data for the final word reports the proportion of phrases where final word was skipped entirely, then duration measures are reported for only those items that were not skipped.

Table 3. Mean phrase and word level measures for Idioms, Binomials and Collocations, and for Formulaic, Control 1 and Control 2 variants. Duration measures are reported in milliseconds; skipping is reported as a probability; duration values for the final word exclude zero values for skipped items. Values in brackets are standard deviations. Significant differences between formulaic and control conditions (based on the model reported in the supplementary materials – Table S1) are indicated with the convention of *, p < .05; **, p <.01; ***, p < .001

	Whole Phrase				Final Word					
	First pass RT		Total RT		Skipping rate		First pass RT		Total RT	
All types	335 (164)		428 (227)		.22 (0.42)		199 (76)		229 (118)	
Control 1	355 (180)	*	490 (285)	***	.14 (0.35)	***	212 (81)	***	248 (134)	***
Control 2	357 (176)	*	485 (261)	***	.14 (0.35)	***	212 (80)	***	251 (135)	***
Idioms	351 (149)		441 (215)		.29 (0.46)		196 (66)		218 (101)	
Control 1	392 (186)	**	539 (299)	***	.16 (0.37)	***	212 (82)	***	241 (122)	**
Control 2	371 (173)		491 (230)	***	.17 (0.38)	***	203 (66)	*	225 (90)	
Binomials	340 (180)		430 (234)		.16 (0.37)		204 (78)		230 (119)	
Control 1	364 (183)		484 (289)	*	.11 (0.31)	*	219 (85)	*	259 (148)	**
Control 2	359 (172)	*	473 (257)	***	.12 (0.33)	*	226 (93)	***	263 (141)	***
Collocations	315 (161)		413 (231)		.21 (0.41)		197 (82)		239 (131)	
Control 1	310 (162)		445 (257)		.15 (0.36)	*	205 (77)		244 (129)	
Control 2	339 (182)		491 (294)	***	.13 (0.34)	**	205 (77)	*	265 (159)	***

Table 3 shows the expected formulaic advantage, relative to control phrases, although this is inconsistent across the three phrase types. For idioms there is a clear advantage across all measures for control type 1 phrases, and for phrase level total RT, final word skipping and final word first pass RT for control type 2 phrases. For binomials, there are differences relative to control type 2 phrases across all measures, and relative to control type 1 phrases for all measures except phrase level first pass RT. For collocations there are significant differences relative to control type 2 phrases for all measures except phrase level first pass RT, but compared to control type 1 phrases, only final word skipping showed an advantage.

The different measures reflect different processes, and the one where all three sub-types showed an advantage compared to both types of control was final word skipping. Skipped words are assumed to have been at least partially recognised and processed in the parafovea, hence words that are part of a known sequence may be more likely to be skipped entirely, or may subsequently receive shorter fixations (as seen for idioms and binomials relative to both types of control phrase, and collocations relative to control type 2 phrase). Consistent effects were also seen for phrase-level total RT, which is a later measure that may reflect overall effort required to integrate the meaning of the phrase into the sentence. This suggests that the overall meaning of the formulaic phrases was easier to understand, while the non-formulaic controls require relatively more consideration (even when the meaning of the formulaic phrase was entirely literal, as in the case of binomials and collocations).

To address our first research question, we added phrase frequency and predictability into the model for each measure, to see firstly whether this improved the fit, and secondly whether it accounted for the formulaic advantage. In other words, were there still between-condition differences once these were included? For all duration measures, phrase frequency made a

significant improvement: phrase-level first pass RT: $\chi^2 = 18.55$, $p < .001$ and total RT: $\chi^2 = 2.95$, $p < .001$; final word first pass RT: $\chi^2 = 15.79$, $p < .001$ and total RT: $\chi^2 = 17.29$, $p < .001$. There was no further improvement for any model by adding phrase frequency as an interaction rather than a fixed effect. In all cases, higher phrase frequency led to shorter durations (all *t*s > 2.0, all *p*s < .05). However, phrase frequency did not significantly improve the model for likelihood of skipping as a fixed effect ($\chi^2 = 5.80$, p = .446) or as an interaction with type and condition.

We next considered whether predictability adds anything, over and above the effects of phrase frequency. As might be expected, phrase frequency, Cloze probability and transitional probability are highly correlated (all *r*s > .50, all *p*s < .001), suggesting that broadly they reflect similar properties. We ran a Principal Component Analysis that confirmed this: all three predictors loaded onto the first component in a similar way, and this accounted for 72% of the variance in these variables. Cloze probability and transitional probability loaded onto the second component (accounting for 15% of the variance) in the same way, while phrase frequency operated in a different direction. Based on this, and to avoid issues of collinearity, we removed transitional probability from any further analysis and instead included only Cloze probability as a measure of predictability.

We added Cloze probability to the models including phrase frequency to determine whether this made any further improvement. This made no improvement to the model for phrase-level first pass RT ($\chi^2 = 0.71$, $p = .397$) or skipping rates ($\chi^2 = 1.10$, $p = .295$), but did for phrase-level total RT ($\chi^2 = 11.86$, $p < .001$), final word first pass RT ($\chi^2 = 4.12$, $p = .042$) and final word total RT ($\chi^2 = 8.80$, $p = .003$). Table 4 reports the differences between condition once phrase frequency and phrase frequency and Cloze probability are included in the models. With the exception of final word skipping for idioms, inclusion of these two variables eliminates the formulaic advantage for all subtypes on all measures.

Table 4. Contribution of phrase frequency and Cloze probability, with subsequent differences between conditions when these are included in the model. Note that with the exception of final word skipping in idioms, any remaining between-condition differences are in favour of the control phrases.

	Whole Phrase				Final Word					
	First Pass RT		Total RT		Skipping		First Pass RT		Total RT	
	t	*p*	*t*	*p*	*z*	*p*	*t*	*p*	*t*	*p*
Idioms										
Control 1	0.39	.670	1.75	.080	-3.46	.001***	0.42	.678	-0.33	.739
Control 2	-1.15	.252	-0.23	.820	-2.80	.005**	-0.53	.595	-1.25	.211
Binomials										
Control 1	-2.02	.044*	-2.85	.004**	-0.50	.615	-1.79	.074	-1.99	.046*
Control 2	-1.58	.114	-2.23	.026*	-0.45	.656	-0.64	.519	-1.18	.236
Collocations										
Control 1	-3.14	.002**	-1.43	.154	-1.03	.303	-0.89	.376	-1.82	.070
Control 2	-1.19	.233	0.73	.466	-1.58	.115	-0.76	.448	-0.11	.910
Phrase Freq	-4.37	.000***	-4.07	.000***	1.06	.291	-3.55	.000***	-3.57	.000***
Cloze Prob	-0.83	.407	-3.46	.001**	1.07	.285	-2.02	.044*	-2.96	.003**

Individual constraints for each subtype

We next considered each phrase type separately, in order to consider how type-specific constraints may influence the processing of the different subtypes of formulaic language. For example, mutual information (MI) is a measure of the strength of a collocation, and therefore is only relevant to this type of phrase. Similarly, idiom specific variables such as decomposability are not relevant to binomials or collocations. We included type specific variables and the main variables of Condition, phrase frequency and Cloze probability.

For each subtype we again constructed separate models for each eye-tracking measure. We began with a model including a fixed effect of Condition, and word-level length and frequency for both content words as covariates. We then added in each of phrase frequency,

Cloze probability and the type-specific predictors outlined below individually, then used log-likelihood tests to see whether any of these improved the model, either as fixed effects or interactions with condition. Next, we constructed cumulative models using forward model selection, where if any variable made a significant improvement this was retained, then subsequent variables were added to this (as fixed effects and interactions with Condition), and the models were compared using log likelihood tests. This gave us an indication of the combination of predictors that exerted an effect for each subtype. Finally, we constructed a maximal model for each measure, where all variables were included as fixed effects and interactions with Condition. We compared this to the best fitting cumulative model to ensure that the pattern of significant variables was the same and found no notable differences in terms of significant effects. Below, we report first the results of the individual models, then the results of the cumulative models, for each subtype and each eye-tracking measure. We include coefficient values from the models to give an indication of the size of any significant effects. (Note: coefficients relate to log-RTs for duration measures, and log-odds for skipping rates). At the end of the section an overall summary of predictor variables for each subtype (based on the best-fitting cumulative model) is presented in Table 5.

Idioms

For idioms, as well phrase frequency and Cloze probability, we also included familiarity and decomposability to assess their effect. Familiarity was rated by a set of native speakers of English (n = 21) as indicated in the methodology section. Mean rating was 5.9 (SD = 0.73, range = 3.5-7). Decomposability was rated by a separate set of native speakers (n = 19), who were asked to judge how much the component words contributed to the figurative meaning which was provided for them (e.g., If you make peace with someone you *bury the hatchet*). Mean rating was 4.1/7 (SD = 1.3, range 1.8-6.2).

At the phrase level, the model for first pass RT was improved by the addition of fixed effects for phrase frequency (χ^2 = 16.43, p < .001) and decomposability (χ^2 = 6.30, p = .012). In an additive model both remained significant, hence higher phrase frequency (β = -0.11, t = -4.01, p < .001) led to shorter reading times and higher decomposability (β = 0.03, t = 2.44, p = .019) led to longer reading times. In this model, differences between idioms and control type 1 (t = -0.22, p = .827) and control type 2 (t = -1.75, p = .082) were not significant. For total RT, the addition of phrase frequency (χ^2 = 16.29, p < .001) and an interaction between condition and familiarity (χ^2 = 10.89, p = .012) improved the initial model. In the additive model, this meant that more frequent phrases were read more quickly (β = -0.11, t = -3.68, p < .001). The interaction between condition and familiarity for control type 2 phrases (t = 2.80, p = .005) meant that for control type 2 phrases only, greater familiarity with the corresponding idiom led to marginally longer RTs (β = 0.08, t = -1.89, p = .061), while there was no effect on idioms or control type 1 phrases. In this model, there was a marginal difference between idioms and control type 1 phrases (t = 1.90, p = .059) but no difference compared to control type 2 phrases (t = 0.33, p = .745).

For final word skipping, the initial model was improved by the addition of an interaction between condition and familiarity (χ^2 = 13.05, p = .005) and condition and decomposability (χ^2 = 10.17, p = .017), but neither phrase frequency nor Cloze probability made any improvement. In the additive model, the best fit included the interaction of condition and familiarity, and a fixed effect of decomposability (addition of this as an interaction with condition made no further improvement once familiarity was also included in the model). This meant that skipping rates were higher for more familiar idioms (β = 0.46, z = 2.69, p = .007), while higher decomposability led to less skipping for all phrases (β = -0.17, z = -2.56, p = .010). In this model there were significant differences between conditions for idioms

compared to control type 1 phrases (z = -4.61, p < .001) and compared to control type 2 phrases (z = -4.12, p < .001).

For final word reading times, phrase frequency (χ^2 = 10.17, p = .017) and decomposability marginally (χ^2 = 5.74, p = .057) improved the model for first pass RT as fixed effects. In the additive model, once phrase frequency was included decomposability made no further improvement (χ^2 = 0.45, p = .501). Here, higher phrase frequency led to shorter reading times (β = -0.05, t = -2.30, p = .022) and there were no differences between idioms and control type 1 (t = 0.96, p = .338) or control type 2 (t = 0.10, p = .922) phrases. For total RT, only phrase frequency improved the initial model as a fixed effect (χ^2 = 8.54, p = .003). This meant that higher frequency led to shorter reading (β = -0.08, t = -2.92, p = .004), and there were no differences between idioms and control type 1 (t = 0.25, p = .806) or control type 2 (t = -0.56, p = .578) phrases.

Binomials

For binomials, as well as the main variables, we included semantic association strength and the ratio of forward to backward occurrences in the BNC.

At the phrase level, the model for first pass RT was improved by the addition of fixed effects for each of phrase frequency (χ^2 = 4.02, p = .045), association strength (χ^2 = 7.68, p = .006), and marginally Cloze probability (χ^2 = 3.05, p = .081) and ratio (χ^2 = 3.19, p = .074). In an additive model, however, the inclusion of association strength removed any effects of phrase frequency or Cloze probability. The best fitting model included a fixed effect of semantic association (β = -0.02, t = -2.78, p = .006) and a marginal effect of ratio (β = -0.08, t = -1.77, p = .077), where both variables contributed to shorter reading times. In this model, no between condition differences were observed for binomials compared to control type 1 phrases (t = -0.76, p = .448) or control type 2 phrase (t = -0.82, p = .414). For total RT the

initial model was improved by the addition of fixed effects for phrase frequency (χ^2 = 9.20, p = .002) and Cloze probability (χ^2 = 7.91, p = .005), and an interaction of condition and association strength (χ^2 = 14.23, p = .003). In an additive model, combinations of both phrase frequency and Cloze, and phrase frequency and association strength were significant, but only phrase frequency remained as a significant predictor when all three were included in the model. These models suggested that phrase frequency was always facilitative (led to lower overall RTs), and Cloze was also facilitative, but only when semantic association strength was not included. When association strength was included, the effects on binomials and control type 1 phrases were non-significant, but there was a significant facilitative effect for control type 2 phrases, whereby more strongly associated phrases had overall less reading time. Comparison of the two models (phrase frequency + Cloze and phrase frequency + association) showed that they were very similar in terms of their fit, suggesting that Cloze and association strength may be in part reflecting a similar property for binomials. When all three variables were included, the difference between binomials and control type 1 phrases disappeared *(*t = -0.92, p = .358) but there remained a significant difference compared to control type 2 phrases (t = -2.83, p = .005).

For final word skipping, addition of an interaction with condition made an improvement to the initial model for association strength (χ^2 = 12.44, p = .006), and marginally for Cloze probability (χ^2 = 7.39, p = .060). Ratio also made a marginal improvement as a fixed effect (χ^2 = 2.91, p = .088). The additive model showed that once an interaction with association strength was included, no other variables made any further improvement. In this model a higher level of association between the component words led to higher rates of skipping in binomials (β = 0.13, z = 2.57, p = .010), and there were differences between binomials and control type 2 phrases (z = 2.01, p = .043) but not control type 1 (z = -1.18, p = .237).

For final word reading, the initial model for first pass RT was improved by the addition of a fixed effect for Cloze probability only (χ^2 = 9.22, p = .002). In an additive model, no other variable further improved this, so the final model showed a facilitative effect of Cloze probability for all phrases (β = -0.01, t = -3.03, p = .003). In this model there were no differences between binomials and either control type 1 (t = -0.45, p = .653) or control type 2 phrases (t = 0.48, p = .635). For total RT, the initial model was improved by the addition of fixed effects for phrase frequency (χ^2 = 34.21, p <.001) and Cloze probability (χ^2 = 38.22, p <.001), and by the addition of interactions with condition for association strength (χ^2 = 38.44, p <.001) and ratio (χ^2 = 35.68, p <.001). The additive model suggested that a combination of factors were important here. Once phrase frequency was included, Cloze probability made an additional improvement as a fixed effect (χ^2 = 7.32, p = .007), and association strength made a marginal improvement as an interaction with condition (χ^2 = 6.75, p = .080). In the model including Cloze probability, phrase frequency made a marginal improvement as a fixed effect (χ^2 = 3.31, p = .069) and ratio made a marginal improvement as an interaction with condition (χ^2 = 7.64, p = .054). Association strength made no contribution once other factors were included. The final model included fixed effects of phrase frequency (β = -0.08, t = -2.09 p = .038) and Cloze probability (β = -0.02, t = -2.65, p = .008), where both led to shorter overall RTs, and a marginal interaction of condition and ratio, whereby for control type 1 phrases (but not binomials or control type 2) a higher ratio (therefore a more strongly conventionalised order) for the corresponding binomial led to longer reading times (β = 0.16, t = 2.67, p = .008). In this model, there remained overall marginal differences between binomials and control type 1 (t = -1.83, p = .068) and control type 2 phrases (t = -1.94, p = .052).

Collocations

For collocations we included semantic association strength and mutual information (MI) score as predictors in all models.

At the phrase level, none of the predictors made any improvement to the model for first pass RT, either on their own or in combination with other variables. In the model including only Condition and the word-level covariates, there were marginal differences between formulaic and control type 2 phrases (t = 1.75, p = .083) but not control type 1 phrases (t = -0.80, p = .430). The model for total RT was improved by the addition of Cloze probability (χ^2 = 15.19, p < .001), MI (χ^2 = 9.84, p = .002) and semantic association (χ^2 = 9.24, p = .002) as fixed effects. An additive model included fixed effects of both Cloze probability (β = -0.03, t = -3.49, p < .001) and MI (β = -0.02, t = -2.63, p = .009). The addition of association strength made no further improvement to this, but the addition of an interaction between condition and phrase frequency did make a marginal improvement (χ^2 = 7.11, p = .068). Here, phrase frequency did not have an effect on collocations themselves (t = 0.48, p = .631), but was inhibitory for both control type 1 (t = 2.37, p = .018) and control type 2 phrases (t = 2.11, p = .035). In this model, there was a significant difference between collocations and control type 1 phrases (t = -1.97, p = .049) but not control type 2 phrases (t = -1.60, p = .111).

For final word skipping, no variable made an improvement to the initial model on its own, either as a fixed effect or interaction term, and no combination of predictors made any improvement in an additive model. This meant that in the basic model including only Condition and word-level covariates, the second word of collocations was skipped more often than in control type 2 phrases (z = -2.98, p = .003) and marginally in control type 1 phrases (z = -1.74, p = .081).

For final words that were not skipped, the model for first pass RT was improved by the addition of both phrase frequency (χ^2 = 4.11, p = .042) and MI (χ^2 = 6.42, p = .011) as fixed

effects. The additive model suggested that addition of both variables caused a confound, as neither was significant when they were both included, hence we looked at separate models for phrase frequency and MI. Phrase frequency was not improved by the addition of either Cloze probability or association strength, and in the final model led to shorter first pass RTs (β = -0.06, t = -2.05, p = .041), with no differences between collocations and control type 1 (t = -0.47, p = .642) or control type 2 phrases (t = -0.47, p = .636). The model for MI similarly included a facilitative fixed effect (β = -0.01, t = -2.54, p = .011), with no differences between collocations and control type 1 (t = -0.47, p = .639) or control type 2 phrases (t = -0.52, p = .604). The model for total RT was improved by the addition of fixed effects for each of Cloze probability (χ^2 = 5.31, p = .021), MI (χ^2 = 5.29, p = .021) and association strength (χ^2 = 4.19, p = .040). An additive model showed that once Cloze and MI were included, association strength made no further improvement, hence the final model included fixed effects of Cloze probability (β = -0.02, t = -1.96, p = .051) and MI (β = -0.02, t = -1.98, p = .048), with no differences between collocations and control type 1 (t = -1.24, p = .216) or control type 2 phrases (t = -0.11, p = .914). When Cloze probability was not included, both MI (β = -0.02, t = -2.14, p = .033) and marginally association strength (β = -0.06, t = -1. 58, p = .065) had a facilitative effect on all phrases.

Table 5 summarises the constraints that are relevant for each of the formulaic subtypes. We indicate whether each variable has a facilitative effect (speeds up processing / leads to shorter reading) or inhibitory effect (slows down processing / leads to longer reading times). Effects in brackets are only significant when other predictors are not included. We also indicate which between-condition differences remain when significant predictors were included in an additive model.

Table 5. Summary of predictor variables for each subtype (+ = facilitative effect / shorter reading times; - = inhibitory effect / longer reading times; effects in brackets are significant only as individual predictors and not in a cumulative model). Advantage columns indicate whether the formulaic advantage (relative to each control type) remains once all significant variables are included in the model (n.s. = not significant, otherwise *p*-values are reported as $^{+}$, p < .10; *, p < .05; **, p < .01; ***, p <.0001).

Idioms	Variables				Advantage	
	PhraseZipf	Cloze	Familiarity	Decomp	Control 1	Control 2
Phrase first pass RT	+			-	n.s.	n.s.
Phrase total RT	+				$^{+}$	n.s.
Word skip			+	-	***	***
Word first pass RT	+				n.s.	n.s.
Word total RT	+				n.s.	n.s.
Binomials						
	PhraseZipf	Cloze	Semantic Association	Ratio		
Phrase first pass RT			+	+	n.s.	n.s.
Phrase total RT	+	(+)	(+)		n.s.	**
Word skip			+		n.s	*
Word first pass RT		+			n.s.	n.s.
Word total RT	+	+			$^{+}$	$^{+}$
Collocations						
	PhraseZipf	Cloze	Semantic association	MI		
Phrase first pass RT					n.s.	$^{+}$
Phrase total RT		+		+	*	n.s.
Word skip					$^{+}$	**
Word first pass RT	(+)			(+)	n.s.	n.s.
Word total RT		+	(+)	+	n.s.	n.s.

As Table 5 shows, the advantage for idioms was generally explained by phrase frequency, and when this was included the between-condition differences disappeared for all duration measures. Increased frequency led to shorter first pass RT (initial recognition of the

sequence) and total RT (overall integration of the meaning). Decomposability, which is often seen as an important factor in how idioms are processed, had a limited a role whereby less decomposable (i.e. more opaque) idioms were read more quickly during the first pass, and were more likely to have the final word skipped. Familiarity (but not frequency) also made an important contribution to skipping rates, but Cloze probability was not a significant predictor for any measure, either on its own or when phrase frequency was included. The only measure where the included variables did not account for the formulaic advantage was skipping rates, where in the best fitting model the differences between idioms and both control conditions remained.

For binomials, a wider spread of variables was implicated. Phrase frequency was important for later measures (phrase-level and final word total RT). Cloze probability contributed to final word reading patterns on early and late measures, and was important at the phrase level only when semantic association was not accounted for. Similarly, semantic association contributed to phrase-level reading times (for total RT, and only when Cloze was not included), and was the only variable that contributed to final word skipping. Ratio, which might be seen as a measure of fixedness (in terms of how much more often the binomial is seen compared to its reversed form) contributed to facilitated reading only during initial recognition of the whole phrase.

For collocations, on two measures our variables failed to explain the formulaic advantage at all, although for first pass reading time this advantage was not strongly apparent in our data. For final word skipping there was a clear advantage compared to control type 2 phrases (and a marginal advantage compared to control type 1) that was not explained by any of our variables. Both Cloze probability and MI contributed to shorter overall reading times, suggesting that these might be a good indicator of how easily a collocation can be integrated into the overall sentence, rather than simply how quickly it is recognised. Phrase frequency

had a minimal effect (and when it did seemed to reflect the same property as MI), and semantic association was also minimal in its contribution.

Discussion

We compared reading patterns for three distinct types of formulaic sequence (idioms, binomials, collocations), in three conditions (conventional formulaic form, control phrase where the first content word was changed, control phrase where the second content word was changed). We found evidence that for all three types there was an advantage for the formulaic phrases compared to both types of control phrase, across a range of eye-tracking measures. Based on the previous research indicating a processing advantage for formulaic language in general, this was expected. Our primary questions were: How far can phrase frequency and predictability alone account for this advantage? And what contribution do type-specific variables make for each subtype?

Distributional variables: omnibus analysis

The omnibus analysis suggested that both frequency and predictability jointly explain why formulaic expressions are processed (recognised during initial reading and then integrated into the surrounding context) more quickly than non-formulaic control phrases. Put simply, formulaic language is processed quickly largely because it is known, hence phrases that have been encountered more often (as measured by their overall frequency) are assumed to be more strongly encoded in the lexicon, independently of their component words. Note that what this means could cover a multitude of things, from a truly "holistic" entry for some phrases (such as non-decomposable idioms), to something more akin to a lexical priming mechanism whereby links between co-occurring words become strengthened through experience (e.g. Wray, 2012; Hoey, 2005). Our results suggest that broad distributional properties do a fairly good job of explaining the formulaic processing advantage, and when

these are accounted for, differences between formulaic and control conditions largely disappear.

If these findings seem straightforward, it is worth remembering that some formulaic phrases–and in particular idioms–are actually not particularly frequent, at least compared to other sequences. In our study, the control phrase "see the film" had a higher phrase frequency (84) than the majority of idioms (only 7 idioms had a higher frequency), hence frequency alone does not equate to formulaicity. Predictability clearly contributes as well, and adding this to the analysis improved the models for three of the five measures we included here. We considered two measures of predictability – Cloze probability and transitional probability – but our initial analysis suggested that these are highly correlated, they are likely to reflect very similar properties (Frisson, Rayner & Pickering, 2005; Janssen & Barber, 2012). Where these may differ is in the sensitivity to context: Cloze probability may vary according to the strength of bias provided by a preceding context, whereas transitional probability will not. Taken together, we can consider frequency and both measures of predictability to reflect the overall experience with each phrase, and it is clear that in broad terms, this does a good job of explaining how and why formulaic language is processed in the efficient way that it is.

The question of interest then becomes, to what extent do additional, phrase-specific factors such as those we have considered here represent something over and above experience-based effects of frequency and predictability? If so, are such features unique to formulaic language, or simply manifestations of features of the language processing system that are brought into focus by the particular subtypes we have looked at? We consider these questions for each of the subtypes and their respective variables in turn.

Idioms

Despite being relatively infrequent (at least compared to the binomials and collocations in our study), idioms were the sub-group where phrase frequency had the most consistent effect. On all duration measures, phrase frequency had a main effect, and when this was included in the analysis any between-condition differences all but disappeared. There were additionally effects of decomposability, whereby greater decomposability was inhibitory (led to longer reading times and less likelihood of skipping the final word). As we addressed in the literature review, the role of decomposability has been variable in previous studies: in terms of activation of idiom meaning, some find that greater decomposability leads to faster activation than for less decomposable phrases (e.g. Caillies & Butcher, 2007), whereas other find that greater decomposability interferes with the activation of figurative meaning (e.g. Cieślicka, 2013; Titone & Libben, 2014). One proposal is that all idioms are to some degree represented as unitary entries, at least in terms of the meaning of the phrase as a whole (c.f. Caillies & Declerq, 2011; Titone & Libben, 2014;), and other aspects such as decomposability or literal plausibility dictate the extent to which the literal meaning interferes with retrieval of the figurative. In this sense, "knowing" an idiom is the key driver of how it will be processed (like all fixed phrases, e.g. Tabossi et al., 2009), and we highlighted in the introduction that a crucial fact about idioms is that they are both formulaic and figurative. If an idiom is not known (i.e. has never been encountered before), aspects such as transparency or decomposability (along with context) will be essential in determining whether the meaning can be inferred. However, once an idiom is known, these properties may serve only to modulate the ease or difficulty with which the figurative meaning is selected, relative to the competing activation of the literal meaning. Importantly, these do not drive or over-ride the overall advantage for idioms compared to non-formulaic phrases, since this is based primarily on the recognition of a known combination of words.

The key difference between frequency and familiarity was demonstrated in our results only in terms of final word skipping, where greater familiarity led to a higher rate of skipping (but did not entirely account for the between-condition differences). Our results are in line with studies such as Carrol and Conklin (2017), where native speakers skipped the final words of idioms 31% of the time compared to 9% for control phrases. In eye-tracking research, word skipping is known to be affected by both visual and linguistic factors. Very short words are often skipped (Rayner & McConkie, 1976), as are function words (Carpenter & Just, 1983) and very high frequency content words (Rayner, Sereno & Raney, 1996). Of relevance here, words that are highly predictable are also skipped more often (Drieghe, Rayner & Pollatsek, 2005; Rayner, Slattery, Drieghe & Liversedge, 2011; Rayner & Well, 1996). Cloze probability in itself did not contribute to skipping, but we could posit that this is subsumed within the variable of familiarity, in that once an idiom is well known, it is highly predictable. In turn, for all idioms, but especially very familiar ones (and note that all of the idioms in this study were selected to ensure that they were generally familiar), recognition of a "known" configuration may increase the chances that the final word is skipped entirely, although the negative effect of decomposability reinforces the degree to which other variables might interfere with this.

Binomials

Binomials represent a very different case to idioms, since successful comprehension was always a simple case of combining the component words, rather than recognising an additional meaning at the level of the whole phrase. The items we used are highly frequent (on average much more frequent than idioms) and are in theory entirely reversible (in its non-idiomatic sense, *black and white* has the same propositional meaning as *white and black;* note: all of the items in our study were non-idiomatic). As we summarised previously, a range of linguistic factors have been proposed for how this order is determined, but

convention / experience seems to be the most important in processing terms (Morgan & Levy, 2016). Our results confirmed this, with overall phrase frequency and Cloze probability accounting for most of the phrase and word level reading patterns. Ratio, which we might interpret as an index of how conventionalised or fixed a word order is, was important only for initial recognition of the phrase as a whole (first pass RT at the phrase level). Siyanova-Chanturia, Conklin and van Heuven (2011) found a consistent advantage for how binomials were read, and their analysis suggested that something over and above phrase frequency contributed to this. In a follow up study using EEG, Siyanova-Chanturia, Conlin, Caffarra, Kaan and van Heuven (2017) found evidence to suggest that the configuration itself was an important part of processing, in that some element of pattern recognition as well as semantic expectancy was involved in how they were processed. Since conventionalisation might be seen as the result of very high levels of exposure (that is, repeated encounters lead to the formation of something akin to a "template" for very high frequency items), these results all point to the same conclusion, whereby frequency plays a very significant role in how binomials are assimilated into the lexicon, and how they are subsequently processed.

The effect of semantic association may represent something additional to these distributional characteristics. Binomials are very often also primary semantic associates, and low-level semantic priming between words in natural reading has been demonstrated previously (e.g. Carrol & Slowiaczek, 1986; Camblin, Gordon & Swaab, 2007), in particular in the early stages of processing. We saw similar effects, whereby stronger association between component words led to shorter first pass RTs for the phrase, and made skipping of the final word more likely. These between-word priming effects are a part of language processing more generally (reflecting well-established properties such as spreading activation, e.g. Collins & Loftus, 1975), with automatic sematic priming thought to be driven by a combination of properties, such as feature overlap and association strength (e.g. Hutchison,

2003; Lucas, 2000). In binomials, these properties may serve to reinforce the one-way relationship between words, although this is driven primarily by having encountered the word combination multiple times in one particular configuration (which may be determined in the first place by a range of linguistic constraints discussed in the introduction). As with idioms, for the most frequent (and fixed) examples, something akin to a "lexical entry" may eventually form, or at least connections between words may become so automatized that the end result (highly speeded, activation of the second coordinate word) is the same.

Collocations

The collocations in the current study showed some evidence of a formulaic advantage, but even in the initial omnibus analysis this was less marked and consistent than for idioms or binomials. Phrase frequency did not do a good job of explaining these effects in the individual analysis, and was largely superseded by MI, which is a measure not simply of occurrence but of co-occurrence for two given words. MI and phrase frequency were highly correlated, and hence may be reflecting the same broad patterns. However, it may be that MI is both more nuanced, and more specific than phrase frequency. To take an example, *strong X* may be a combination that occurs highly frequently, but where the noun slot can be filled with several plausible options (many of which may in themselves be frequent, such as *strong tea, strong coffee, strong feeling, etc.*). In contrast, *abject X* is more restricted in that only a small number of words (*poverty*, *failure*) are likely candidates. Although the overall frequency may be low, the likelihood of co-occurrence is high, hence MI reflects a more nuanced knowledge of language experience than the coarser measure of phrase frequency. In line with this, Ellis, Simpson-Vlach and Maynard (2008) found that for lexical bundles derived from academic corpora, MI rather than phrase frequency determined speed of processing for native speakers, while language learners were more sensitive to overall frequency of occurrence. Our results suggest that both Cloze probability and MI (both

reflective of the expectation created by seeing the first word) are both better explanatory variables for collocations than phrase frequency, although ultimately all are derived from experience and therefore support the broad conclusion that distributional factors are the primary drivers of collocational processing. Interestingly, on two measures (first pass RT for the phrase and likelihood of skipping the second word altogether), none of the variables considered here explained the differences between collocations and control phrases (although these were marginal). Vilkaite (2016) also found no specific effects of phrase frequency or MI in her analysis of verb-noun collocations (which did show an overall advantage), and suggested that both variables were subsumed within the overall status of collocation.

We found little evidence that semantic association affected the processing of collocations, in contrast to, e.g. Durrant and Doherty (2010), who found a clear difference for collocations that were also semantic associates. Hutchison (2003) concluded that there is evidence of pure associative priming in the absence of any semantic overlap, but that the strongest effects of priming are often seen when both criteria are met (semantic overlap and pure association). In our stimuli, semantic association was in general low (much lower than for the binomials), and was routinely close to zero for both sets of control items, so it may be that we did not have enough variability here to see any effects. As above, any evidence of semantic priming for associated words would be consistent with language processing more generally, and would not necessarily reflect any aspect of the formulaic nature of the items themselves.

Overall conclusions

Our data support the view that formulaic expressions, regardless of fundamental differences in the properties that constitute categories like idioms, binomials and collocations, are all processed quickly, primarily because they are known phrases that have been encountered multiple times as part of the language experience of native speakers. Note that processed here

refers to two aspects: both the recognition of a "known" combination of words, and the analysis and integration of that sequence into the surrounding sentence context. Both aspects seem to be easier for formulaic expressions than literal expressions, even when the task of deriving propositional meaning is ostensibly straightforward (as in the case of binomials and collocations). This frequency of past occurrence may, in the most frequent or most fixed examples, lead to the formation of something akin to a "template", and evidence from the EEG literature supports this for a range of different types of expression: idioms (Vespignani et al., 2011; Zhang, Yang, Gu & Ji, 2013); figurative collocations (Molinaro & Carreiras, 2010); and binomials (Siyanova-Chnaturia et al., 2017).

As Siyanova-Chanturia and Martinez (2014) and Siyanova-Chanturia (2015) argue, much of the evidence that claims to show "holistic" processing in formulaic sequences doesn't actually speak directly to this claim, but instead simply shows a consistent speed advantage for formulaic phrases. A better way to conceive of formulaic sequences might be as distributed representations at a lexical level, with multiple connections both between words and with other levels (e.g., the lexical conceptual level postulated by the superlemma theory – Sprenger et al., 2006; see also the construction-integration account of idioms described in Caillies & Butcher, 2007). Experience and frequency of past encounter are the primary drivers here, as with language processing in general, and this view is not incompatible with a lexical priming account (Hoey, 2005; Pace-Sigge, 2013), whereby all examples of a word combination either serve to reinforce an existing link, or dilute it.

Our data suggest that distributional characteristics do account for most of the formulaic advantage, while other aspects serve to modulate the ease or difficulty with which the phrase as a whole might be interpreted. For example, semantic links provide low-level facilitation where these exist, and since binomials happen to be primary associates more often than not, these present an example of a more general phenomenon. Similarly, if idioms are not

familiar, there can be no formulaic advantage, and variables such as decomposability serve to make the process of working out the figurative meaning more or less straightforward (e.g. Carrol et al., 2018). When idioms are well-known, they are recognised quickly and easily, and decomposability has little effect on the most familiar examples (Libben & Titone, 2008). There are obvious linguistic differences between disparate subtypes (e.g. the difference between idioms and lexical bundles), but within the broad class of formulaic language, aspects of conventionalisation (frequency of past occurrence, and predictability of the sequence based on co-occurrence probabilities) are the main driver of the faster processing reported in the literature. Specific aspects of different phrase "types" (e.g. idioms, which are inherently ambiguous), serve to underpin processing in various subtle ways, which do not differ markedly from how language is processed more generally.

References

Abel, B. (2003). English idioms in the first language and second language lexicon: a dual representation approach. *Second Language Research, 19*(4), 329–358. doi: 10.1191/0267658303sr226oa

Altarriba, J., Kroll, J., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory and Cognition, 24*(4), 477-492.

Arcara, G., Lacaita, G., Mattaloni, E., Passarini, L., Mondini, S., Beninca, P., & Semenza, C. (2012). Is "hit and run" a single word? The processing of irreversible binomials in neglect dyslexia. *Frontiers in Psychology, 3*(11), 1-11. doi: 10.3389/fpsyg.2012.00011

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67-82. doi: 10.1016/j.jml.2009.09.005

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: the effect of familiarity on children's repetition of four-word combinations. *Psychol Sci, 19*(3), 241-248. doi: 10.1111/j.1467-9280.2008.02075.x

Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage-based models of language.* Stanford, CA: The Center for the Study of Language and Information Publications.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278. doi: 10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., & Dai, B. (2014). *lme4: Linear mixed-effects models using Eigen and S4*.

Benor, S., & Levy, R. (2006). The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Linguistic Society of America, 82*(2), 233-278.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Blais, M.-J., & Gonnerman, L. M. (2013). Explicit and implicit semantic processing of verb–particle constructions by French–English bilinguals. *Bilingualism: Language and Cognition, 16*(4), 829-846. doi: 10.1017/s1366728912000673

Bobrow, S., & Bell, S. (1973). On catching on to idiomatic expressions. *Memory and Cognition, 1*, 343-346.

Bod, L. (2001). *Sentence memory: Storage vs. computation of frequent sentences*. Paper presented at the CUNY 2001.

Bonk, W., & Healy, A. (2005). *The company words keep: Priming effects without semantic or associative links through collocation*. Paper presented at the 46th Annual Meeting of the Psychonomics Society, Toronto.

Brenner, G. (2003). *Webster's New World American Idioms Handbook*. New York: Houghton Mifflin Harcourt.

Buerki, A. (2016). Formulaic sequences: a drop in the ocean of constructions or something more significant? *European Journal of English Studies 20*(1), 15-34.

Bybee, J. L. (2006). From Usage to Grammar: The Mind's Response to Repetition. *Language, 82*(4), 711-733. doi: 10.1353/lan.2006.0186

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language, 27*(6), 668-683.

Caillies, S., & Butcher, K. (2007). Processing of Idiomatic Expressions: Evidence for a New Hybrid View. *Metaphor and Symbol, 22*(1), 79-108.

Caillies, S. & Declerq, C. (2011). Kill the song—steal the show: what does distinguish predicative metaphors from decomposable idioms? *Journal of Psycholinguistic Research, 40*(3), 205-223. doi: 10.1007/s10936-010-9165-8.

Camblin, C. C., Gordon, P. C., & Swaab, T. Y. (2007). The interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. *Journal of Memory and Language, 56*(1), 103-128. doi: 10.1016/j.jml.2006.07.005

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275-307). San Diego, CA: Academic Press.

Carrol, G., & Conklin, K. (2014). Eye-tracking multi-word units: some methodological questions. *Journal of Eye Movement Research, 7*(5), 5, 1-11.

Carrol, G., & Conklin, K. (2017). Cross language priming extends to formulaic units: Evidence from eye-tracking suggests that this idea "has legs". *Bilingualism: Language and Cognition, 20*(2), 299-317. doi: 10.1017/S1366728915000103

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in Translation. *Studies in Second Language Acquisition, 38*(03), 403-443. doi: 10.1017/s0272263115000492

Carrol, G., Littlemore, J. & Dowens, M. G. (2018). Of false friends and familiar foes: Comparing native and non-native understanding of figurative phrases. *Lingua, 204*, 21-44.

Carroll, P., & Slowiaczek, M. (1986). Constraints on semantic priming in reading: A fixation time analysis. *Memory and Cognition, 14*(6), 509-522.

Cheng, W., Greaves, C., Sinclair, J. M., & Warren, M. (2009). Uncovering the Extent of the Phraseological Tendency: Towards a Systematic Analysis of Concgrams. *Applied Linguistics, 30*(2), 236-252. doi: 10.1093/applin/amn039

Cieślicka, A. B. (2013). Do nonnative language speakers chew the fat and spill the beans with different brain hemispheres? Investigating idiom decomposability with the divided visual field paradigm. *Journal of Psycholinguistic Research, 42*(6), 475-503. doi: 10.1007/s10936-012-9232-4

Collins, A. & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407–428. doi:10.1037/0033-295X.82.6.407

Columbus, G. (2010). Processing MWUs: Are MWU Subtypes Psycholinguistically Real? In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 194-210). London: Continuum.

Columbus, G. (2013). In support of multiword unit classifications: Corpus and human rating data validate phraseological classifications of three different multiword unit types. *Yearbook of Phraseology, 4*, 23-43. doi: 10.1515/phras-2013-0003

Conklin, K. & Carrol, G. (2016). *Native and non-native sensitivity to word order in novel multiword sequences*. Paper presented at BAAL Vocabulary SIG, Nottingham UK, July 2016.

Conklin, K., Pellicer-Sánchez, A. and Carrol, G. (2018). Eye-Tracking: A Guide for Applied Linguistics Research. Cambridge: Cambridge University Press.

Cooper, W. E. and Ross, J. R. (1975). World order. In R. E. Grossman, L. J. San and T. J. Vance (Eds.), *Papers from the parasession on functionalism*, (pp. 63–111). Chicago: Chicago Linguistic Society.

Cutting, J., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory and Cognition, 25*(1), 57-71.

Drieghe, D., Rayner, K., & Pollatsek, A. (2005). Eye movements and word skipping during reading revisited. *Journal of Experimental Psychology: Human Perception and Performance, 31*(5), 954-959. doi: 10.1037/0096-1523.31.5.954

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory, 6*(2), 125-155.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *Tesol Quarterly, 42*(3), 375-396.

Fanari, R. Cacciari , C. & Tabossi, T. (2010). The role of idiom length and context in spoken idiom. *European Journal of Cognitive Psychology*, *22*(3), 321-334.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*(5), 862-877. doi: 10.1037/0278-7393.31.5.862

Gibbs, J. R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition, 8*, 149–156.

Gibbs, R. W., & Nayak, N. P. (1989). Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology, 21*, 100-138.

Gibbs, R. W., Nayak, N. P., & Cutting, C. (1989). How to Kick the Bucket and Not Decompose: Analyzability and Idiom Processing. *Journal of Memory and Language, 28*, 576–593.

Goldberg, A. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press.

Gyllstad, H., & Wolter, B. (2016). Collocational processing in the light of a phraseological continuum model: Does semantic transparency matter? *Language Learning, 66*(2), 296-323.

Hamblin, J., & Gibbs, J. R. W. (1999). Why You Can't Kick the Bucket as You Slowly Die: Verbs in Idiom Comprehension. *Journal of Psycholinguistic Research, 28*(1), 25–39.

Hernández, M., Costa, A. and Arnon, I. (2016). More than words: multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience, 31*(6), 785-800.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Holsinger, E., & Kaiser, E. (2013). Processing (Non)Compositional Expressions: Mistakes and Recovery. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39*(3), 866-878. doi: 10.1037/a0030410

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*, 24-44.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: : Cambridge University Press.

Hutchison, K. (2003). Is semantic priming due to association strength or feature overlap? A *micro*analytic review. *Psychonomic Bulletin & Review, 10*(4), 785-813.

Inhoff, A. (1984). Two Stages of Word Processing during Eye Fixations in the Reading of Prose. *Journal of Verbal Learning and Verbal Behavior, 23*, 612-624.

Janssen, N., & Barber, H. A. (2012). Phrase frequency effects in language production. *PLoS One, 7*(3), e33202. doi: 10.1371/journal.pone.0033202

Jolsvai, H., McCauley, S. M., & Christiansen, M. H. (2013). *Meaning overrides frequency in idiomatic and compositional multiword chunks.* Paper presented at the 35th Annual Conference of the Cognitive Science Society, Austin, Texas.

Kim, S. H., & Kim, J. H. (2012). Frequency Effects in L2 Multiword Unit Processing: Evidence From Self-Paced Reading. *Tesol Quarterly, 46*, 831-841. doi: 10.1002/tesq.66

Kiss, G., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. from Edinburgh University Press

Konopka, A. E., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology, 58*(1), 68-101. doi: 10.1016/j.cogpsych.2008.05.002

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition and Neuroscience, 31*(1), 32-59. doi: 10.1080/23273798.2015.1102299

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2016). *lmerTest: Tests in Linear Mixed Effects Models*.

Libben, M., & Titone, D. (2008). The multidetermined nature of idiom processing. *Memory and Cognition, 36*(6), 1103-1121.

Lohmann, A. (2012). A processing view on order in reversible and irreversible binomials. *Vienna English Working Papers, 21*(1), 25-50.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7*, 618-630.

Matlock, T., & Heredia, R. R. (2002). Understanding phrasal verbs in monolinguals and bilinguals. In R. R. Heredia & J. Altarriba (Eds.), *Bilingual Sentence Processing* (pp. 251-274). Amsterdam: Elsevier.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research, 43*(16), 1735-1751. doi: 10.1016/s0042-6989(03)00237-2

McGlone, M., Glucksberg, S., & Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes, 17*, 167-190.

Molinaro, N., & Carreiras, M. (2010). Electrophysiological evidence of interaction between contextual expectation and semantic integration during the processing of collocations. *Biol Psychol, 83*(3), 176-190. doi: 10.1016/j.biopsycho.2009.12.006r

Mollin, S. (2012). Revisiting binomial order in English: ordering constraints and reversibility. *English Language and Linguistics, 16*(01), 81-103. doi: 10.1017/s1360674311000293

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition, 157*, 384-402. doi: 10.1016/j.cognition.2016.09.011

Pace-Sigge, M. (2013). The concept of lexical priming in the context of language use. *ICAME Journal, 37*, 149-174.

Paterson, K., Liversedge, S., & Underwood, G. (1999). The influence of focus operators on syntactic processing of short relative clause sentences. *The Quarterly Journal of Experimental Psychology, 52A*, 717-737.

Paulmann, S., Ghareeb-Ali, Z., & Felser, C. (2015). Neurophysiological markers of phrasal verb processing: Evidence from L1 and L2 speakers. In R. R. Heredia & A. B. Cieślicka (Eds.), *Bilingual Figurative Language Processing* (pp. 245-267). Cambridge: Cambridge University Press.

Peterson, R., Burgess, C., Dell, G., & Eberhard, K. (2001). Dissociation Between Syntactic and Semantic Processing During Idiom Comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27*(5), 1223-1237.

R Development Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin¸124*(3), 372-422.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research, 16,* 829-837.

Rayner, K., Sereno, S. C., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance 22,* 1188-1200.

Rayner, K., Slattery, T. J., Drieghe, D., & Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance, 37*(2), 514-528. doi: 10.1037/a0020990

Rayner, K., & Well, A. (1996). Effects of contextual constraint on eye movements in reading: a further examination. *Psychonomic Bulletin and Review, 3,* 504-509.

Rommers, J., Dijkstra, T., & Bastiaansen, M. (2013). Context-dependent semantic processing in the human brain: evidence from idiom comprehension. *Journal of Cognitive Neuroscience, 25*(5), 762-776. doi: 10.1162/jocn_a_00337

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

Schweigert, W. (1991). The Muddy Waters of Idiom Comprehension. *Journal of Psycholinguistic Research, 20*(4), 305-314.

Schweigert, W., & Moates, D. (1988). Familiar Idiom Comprehension. *Journal of Psycholinguistic Research, 17*(4), 281-296.

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

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory, 11*(2). doi: 10.1515/cllt-2014-0016

Siyanova-Chanturia, A., & Martinez, R. (2014). The Idiom Principle Revisited. *Applied Linguistics*. doi: 10.1093/applin/amt054

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*(2), 251–272. doi: 10.1177/0267658310382068

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

Siyanova-Chanturia, A., Conklin, K. Caffarra, S., Kaan, E., & van Heuven, W.J.B. (2017). Representation and processing of multi-word expressions in the brain. *Brain and Language, 175,* 111-122. [DOI: 10.1016/j.bandl.2017.10.004]

Smolka, E., Rabanus, S., & Rösler, F. (2007). Processing Verbs in German Idioms: Evidence Against the Configuration Hypothesis. *Metaphor and Symbol*, *22*(3), 213-231.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition, 18*(03), 419-437. doi: 10.1017/s1366728914000674

Sprenger, S., Levelt, W., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language, 54*(2), 161–184. doi: 10.1016/j.jml.2005.11.001

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford Handbook of Psycholinguistics* (pp. 327-342). Oxford: Oxford University Press.

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

Tabossi, P., Fanari, R., & Wolf, K. (2008). Processing idiomatic expressions: effects of semantic compositionality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 313-327. doi: 10.1037/0278-7393.34.2.313

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory and Cognition, 37*(4), 529-540. doi: 10.3758/MC.37.4.529

Titone, D., & Connine, C. (1994). Comprehension of Idiomatic Expressions: Effects of Predictability and Literality. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(5), 1126-1138.

Titone, D., & Connine, C. (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. *The Mental Lexicon, 9*(3), 473-496. doi: 10.1075/ml.9.3.05tit

Titone, D., Columbus, G., Whitford, V., Mercier, J., & Libben, M. (2015). Contrasting Bilingual and Monolingual Idiom Processing. In R. R. Heredia & A. B. Cieślicka (Eds.), *Bilingual Figurative Language Processing* (pp. 171-207). New York: Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: A behavioural and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on Formulaic Language: Acquisition and Communication* (pp. 151–173). London: Continuum.

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*(2), 569-613. doi: 10.1111/j.1467-9922.2010.00622.x

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*, 1176-1190. doi: 10.1080/17470218.2013.850521

Van Lancker Sidtis, D. (2012). Two-track mind: Formulaic and novel language support a dual-process model. In M. Faust (Ed.), *The Handbook of the Neuropsychology of Language* (pp. 342–367). Chichester: Wiley-Blackwell.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2009). Predictive Mechanisms in Idiom Comprehension. *Journal of Cognitive Neuroscience, 22*(8), 1682-1700.

Vilkaite, L. (2016). Are nonadjacent collocations processed faster? *J Exp Psychol Learn Mem Cognition, 42*(10), 1632-1642. doi: 10.1037/xlm0000259

Warren, H. (1994). *Oxford Learner's Dictionary of English Idioms*. Oxford: Oxford University Press.

Wolter, B., & Gyllstad, H. (2011). Collocational Links in the L2 Mental Lexicon and the Influence of L1 Intralexical Knowledge. *Applied Linguistics, 32*(4), 430-449.

Wolter, B., & Gyllstad, H. (2013). Frequency of Input and L2 Collocational Processing: A Comparison of Congruent and Incongruent Collocations. *Studies in Second Language Acquisition, 35*(3), 451-482.

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

Wray, A. (2008). *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Wray, A. (2012). What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play. *Annual Review of Applied Linguistics, 32*, 231-254. doi: 10.1017/S026719051200013X

Zhang, H., Yang, Y., Gu, J., & Ji, F. (2013). ERP correlates of compositionality in Chinese idiom comprehension. *Journal of Neurolinguistics, 26*(1), 89–112. doi: 10.1016/j.jneuroling.2012.05.002

Supplementary Materials for Carrol & Conklin – Is all formulaic language created equal? Unpicking the processing advantage for different types of formulaic sequences

These materials present the materials used in the study (formulaic phrases and control phrases), as well as the full output of mixed effects models reported in the main text, for the comparison of all formulaic subtypes compared to control items, followed by the same analysis including phrase frequency (Zipf) and residulaised Cloze probability.

Stimuli used in the eye-tracking experiment

Idiom			Control 1			Control 2		
behind	the	scenes	between	the	scenes	behind	the	bushes
below	the	belt	about	the	belt	below	the	line
bite	the	bullet	load	the	bullet	bite	the	packet
break	the	bank	hurt	the	bank	break	the	wall
break	the	ice	crack	the	ice	break	the	lock
bury	the	hatchet	find	the	hatchet	bury	the	cable
caught	the	sun	seen	the	sun	caught	the	flu
chewing	the	fat	using	the	fat	chewing	the	rind
dropped	the	ball	stopped	the	ball	dropped	the	plate
fit	the	bill	see	the	bill	fit	the	role
hold	the	fort	take	the	fort	hold	the	door
jump	the	gun	take	the	gun	jump	the	wall
jump	the	queue	join	the	queue	jump	the	fence
look	the	part	get	the	part	look	the	best
missed	the	boat	cracked	the	boat	missed	the	train
pass	the	time	use	the	time	pass	the	house
popped	the	question	shouted	the	question	popped	the	balloon
rock	the	boat	crash	the	boat	rock	the	table
runs	the	show	saw	the	show	runs	the	shop
saved	the	day	ruined	the	day	saved	the	cash
seen	the	light	found	the	light	seen	the	film
set	the	scene	paint	the	scene	set	the	clock
spill	the	beans	drop	the	beans	spill	the	chips
stole	the	show	liked	the	show	stole	the	phone
turn	the	tables	move	the	tables	turn	the	wheels
changed	your	tune	learned	your	tune	changed	your	tyre
eat	your	words	know	your	words	eat	your	beans
found	your	feet	hurt	your	feet	found	your	ring
hang	your	head	mind	your	head	hang	your	shirt
hold	your	horses	lead	your	horses	hold	your	drinks
lose	your	marbles	count	your	marbles	lose	your	memories
make	your	mark	show	your	mark	make	your	sign
mark	your	words	hear	your	words	mark	your	work
pick	a	fight	have	a	fight	pick	a	shirt
pick	your	brains	use	your	brains	pick	your	gift
playing	with	fire	cooking	with	fire	playing	with	dolls
pull	your	leg	grab	your	leg	pull	your	arm
push	your	luck	make	your	luck	push	your	body
smell	a	rat	hear	a	rat	smell	a	fire
stood	your	ground	kept	your	ground	stood	your	child
stretch	your	legs	rest	your	legs	stretch	your	back
tighten	your	belt	changed	your	belt	tighten	your	hands
twist	your	arm	hold	your	arm	twist	your	leg

wasting	your	breath	losing	your	breath	wasting	your	lives
watch	your	step	clean	your	step	watch	your	child

Binomial			Control 1			Control 2		
aches	and	pains	spasms	and	pains	aches	and	spasms
arms	and	legs	hands	and	legs	arms	and	feet
art	and	design	music	and	design	art	and	music
black	and	white	green	and	white	black	and	green
boys	and	girls	men	and	girls	boys	and	men
bread	and	butter	cheese	and	butter	bread	and	meat
brother	and	sister	cousin	and	sister	brother	and	cousin
deaf	and	dumb	blind	and	dumb	deaf	and	blind
doctors	and	nurses	surgeons	and	nurses	doctors	and	surgeons
fish	and	chips	beans	and	chips	fish	and	rice
food	and	drink	cups	and	drink	food	and	plates
gold	and	silver	diamond	and	silver	gold	and	diamond
goods	and	services	items	and	services	goods	and	items
horse	and	rider	pony	and	rider	horse	and	pony
husbands	and	wives	mothers	and	wives	husbands	and	sons
iron	and	steel	gold	and	steel	iron	and	gold
king	and	queen	prince	and	queen	king	and	prince
knife	and	fork	spoon	and	fork	knife	and	spoon
ladies	and	gentlemen	children	and	gentlemen	ladies	and	children
law	and	order	rules	and	order	law	and	rules
left	and	right	back	and	right	left	and	back
live	and	learn	think	and	learn	live	and	think
live	and	work	move	and	work	live	and	write
male	and	female	mixed	and	female	male	and	mixed
mum	and	dad	son	and	dad	mum	and	son
name	and	address	number	and	address	name	and	number
nice	and	easy	slow	and	easy	nice	and	slow
north	and	south	east	and	south	north	and	east
nuts	and	bolts	screws	and	bolts	nuts	and	screws
oil	and	gas	coal	and	gas	oil	and	coal
out	and	about	here	and	about	out	and	busy
peace	and	quiet	calm	and	quiet	peace	and	calm
pick	and	choose	select	and	choose	pick	and	select
plain	and	simple	easy	and	simple	plain	and	easy
read	and	write	spell	and	write	read	and	spell
rich	and	poor	sick	and	poor	rich	and	noble
salt	and	pepper	spices	and	pepper	salt	and	spices
sick	and	tired	bored	and	tired	sick	and	bored
soap	and	water	towels	and	water	soap	and	towels
son	and	daughter	friend	and	daughter	son	and	friend

tea	and	coffee	juice	and	coffee	tea	and	juice
time	and	money	people	and	money	time	and	people
trial	and	error	bias	and	error	trial	and	appeal
warm	and	dry	safe	and	dry	warm	and	safe
wind	and	rain	snow	and	rain	wind	and	snow

Collocation		Control 1		Control 2	
abject	poverty	total	poverty	abject	agony
ancestral	homes	traditional	homes	ancestral	house
ancient	history	distant	history	ancient	stories
anecdotal	evidence	additional	evidence	anecdotal	account
angry	mob	large	mob	angry	gang
classic	example	decent	example	classic	version
clean	clothes	fresh	clothes	clean	things
clear	sky	pretty	sky	clear	sea
complex	series	diverse	series	complex	string
cosmic	rays	stellar	rays	cosmic	dust
cruel	joke	nasty	joke	cruel	trick
current	affairs	modern	affairs	current	actions
daily	paper	regular	paper	daily	update
direct	result	straight	result	direct	change
final	exam	last	exam	final	task
foreign	debt	overseas	debt	foreign	plan
former	student	previous	student	former	neighbour
full	text	new	text	full	book
great	concern	large	concern	great	worry
heavy	rain	steady	rain	heavy	snow
human	health	animal	health	human	growth
inner	self	ideal	self	inner	dreams
likely	effects	normal	effects	likely	results
low	risk	small	risk	low	chance
luxury	items	special	items	luxury	things
married	couple	lovely	couple	married	person
menial	task	boring	task	menial	role
mental	picture	abstract	picture	mental	portrait
narrow	range	better	range	narrow	piece
parallel	lines	equal	lines	parallel	strips
pretty	girl	elegant	girl	pretty	view
private	homes	modern	homes	private	grounds
public	opinion	general	opinion	public	thought
quick	break	small	break	quick	rest
real	impact	huge	impact	real	result
rough	surface	poor	surface	rough	coating
separate	occasions	earlier	occasions	separate	attempts

serious	injury	nasty	injury	serious	outcome
shallow	water	normal	water	shallow	ground
short	stay	brief	stay	short	tour
special	unit	specific	unit	special	team
stone	floor	new	floor	stone	surface
tragic	death	awful	death	tragic	finish
trusted	friend	caring	friend	trusted	ally
wild	horses	crazy	horses	wild	ponies

Table S1. Omnibus linear mixed effects models comparing effects of Phrase Type (baseline = Idioms) and Condition (baseline = Formulaic) for each eye-tracking measure (summary reported in Table 3 in main text).

	Whole Phrase						Final Word								
	First Pass RT			Total RT			Skipping Rate			First Pass RT			Total RT		
	β	SE	t	β	SE	t	β	SE	z	β	SE	t	β	SE	t
Intercept	5.60	0.12	46.94	5.73	0.14	40.45	0.17	0.62	0.27	5.14	0.09	57.63	5.12	0.18	43.52
Type: Binomial	-0.06	0.03	-1.77	-0.04	0.04	-0.99	-0.73	0.19	-3.83***	0.04	0.03	1.47	0.05	0.04	1.27
Type: Collocation	-0.18	0.04	-4.77***	-0.15	0.04	-3.49***	-0.27	0.22	-1.22	-0.04	0.03	-1.34	0.02	0.04	0.63
Control Type 1	0.10	0.03	3.23**	0.19	0.03	6.17***	-0.90	0.11	-5.16***	0.08	0.02	3.36***	0.09	0.03	3.03**
Control Type 2	0.04	0.03	1.21	0.11	0.03	3.47***	-0.78	0.19	-4.22***	0.05	0.02	2.15*	0.05	0.03	1.75
Bin * Control 1	-0.05	0.04	-1.13	-0.11	0.04	-2.70**	0.43	0.26	1.65	-0.02	0.03	-0.79	-0.01	0.04	-0.08
Coll * Control 1	-0.13	0.04	-3.05**	-0.11	0.04	-3.22**	0.50	0.25	2.00*	-0.04	0.03	1.28	-0.06	0.04	-1.54
Bin * Control 2	0.03	0.04	0.64	0.00	0.04	-0.02	0.33	0.26	1.29	-0.04	0.03	1.42	0.07	0.04	1.79
Coll * Control 2	0.02	0.04	0.39	0.04	0.04	1.08	0.21	0.25	0.84	0.00	0.03	0.03	0.04	0.04	1.09
W1 Length	0.03	0.01	3.26**	0.04	0.01	4.10***	-0.10	0.04	-2.46*	0.02	0.01	3.34***	0.02	0.01	2.37*
W1 Freq (Zipf)	-0.01	0.01	-0.28	-0.01	0.02	-0.90	0.09	0.08	1.13	-0.01	0.01	-1.27	-0.01	0.01	-0.59
W2 Length	0.02	0.01	2.20*	0.01	0.01	1.49	-0.26	0.04	-6.47***	0.00	0.01	0.61	0.01	0.01	0.65
W2 Freq (Zipf)	-0.00	0.02	-0.15	0.02	0.02	0.88	-0.01	0.09	0.10	0.01	0.01	0.80	0.02	0.02	1.40
Random effects	Variance		SD	Variance		SD	Variance		SD	Variance		SD	Variance		SD
Item	0.006		0.08	0.017		0.13	0.037		0.19	0.004		0.06	0.008		0.09
Subject	0.033		0.18	0.042		0.21	0.346		0.59	0.014		0.12	0.014		0.12
Subject | Binomial	0.001		0.03	0.001		0.02	0.130		0.36	0.001		0.04	0.004		0.06
Subject | Collocation	0.003		0.06	0.004		0.06	0.568		0.75	0.003		0.05	0.006		0.08
Subject | Control 1	0.003		0.06	0.004		0.06	0.056		0.24	0.002		0.04	0.001		0.02
Subject | Control 2	0.003		0.06	0.006		0.07	0.208		0.46	0.001		0.02	0.000		0.02
Residual	0.207		0.45	0.206		0.45	-		-	0.092		0.30	0.156		0.40

Note: *p*-values are estimated using the lmerTest package in R (version 2.0-33; Kuznetsova, Brockhoff & Christensen, 2016). * < .05; ** < .01; *** < .001.

Table S2. Omnibus linear mixed effects models comparing effects of Phrase Type (baseline = Idioms) and Condition (baseline = Formulaic) for each eye-tracking measure. Phrase frequency and cloze probability are included in all models (summary reported in Table 4 in main text).

	Whole Phrase						Final Word								
	First Pass RT			Total RT			Skipping Rate			First Pass RT			Total RT		
	β	SE	t	β	SE	t	β	SE	z	β	SE	t	β	SE	t
Intercept	5.64	0.12	47.02	5.82	0.14	40.98	-0.00	0.63	-0.00	5.18	0.09	57.78	5.19	0.12	43.80
Type: Binomial	0.00	0.04	0.01	0.06	0.04	1.38	-0.88	0.21	-4.15***	0.09	0.03	3.12**	0.13	0.04	3.23**
Type: Collocation	-0.16	0.04	-4.05***	-0.15	0.05	-3.31**	-0.26	0.23	-1.15	-0.03	0.03	-1.10	0.03	0.04	0.65
Control Type 1	0.01	0.04	0.39	0.07	0.04	1.74	-0.72	0.21	-3.46***	0.01	0.03	0.42	-0.01	0.04	-0.33
Control Type 2	-0.04	0.04	-1.15	-0.01	0.04	-0.24	-0.60	0.22	-2.80**	-0.01	0.03	-0.53	-0.04	0.03	-1.25
Bin * Control 1	-0.11	0.04	-2.45*	-0.21	0.04	-4.59***	0.57	0.28	2.06*	-0.07	0.03	-2.23*	-0.08	0.04	-1.79
Coll * Control 1	-0.13	0.04	-3.14**	-0.12	0.04	-2.88**	0.48	0.25	1.90	-0.04	0.03	-1.15	-0.05	0.04	-1.27
Bin * Control 2	-0.03	0.04	-0.72	-0.10	0.04	-2.25*	0.47	0.27	1.75	-0.01	0.03	-0.23	0.01	0.04	-0.21
Coll * Control 2	-0.00	0.04	-0.11	0.04	0.04	0.89	0.21	0.26	0.81	-0.01	0.03	-0.22	0.04	0.04	0.94
W1 Length	0.02	0.01	3.23**	0.04	0.01	3.94***	-0.10	0.04	-2.46*	0.02	0.01	3.29**	0.02	0.01	2.27*
W1 Freq (Zipf)	0.02	0.02	0.99	0.00	0.02	0.25	0.07	0.09	0.80	-0.01	0.01	-0.20	0.01	0.02	0.37
W2 Length	0.02	0.01	2.23*	0.01	0.01	1.59	-0.26	0.04	-6.42***	0.01	0.01	0.63	0.01	0.01	0.69
W2 Freq (Zipf)	0.01	0.02	0.70	0.03	0.02	1.73	-0.01	0.09	-0.11	0.02	0.01	1.57	0.03	0.02	2.17*
Phrase Freq (Zipf)	-0.08	0.02	-4.37***	-0.08	0.02	-4.08***	0.11	0.11	1.06	-0.05	0.01	-3.55***	-0.06	0.02	-3.57**
Cloze probability	-0.03	0.04	-0.83	-0.01	0.00	-3.46***	0.02	0.02	1.07	-0.01	0.00	-2.02*	-0.12	0.04	-2.96*
Random effects	Variance	SD		Variance	SD		Variance	SD		Variance	SD		Variance	SD	
Item	0.006	0.08		0.016	0.13		0.043	0.21		0.004	0.06		0.008	0.09	
Subject	0.033	0.18		0.042	0.20		0.346	0.59		0.014	0.12		0.015	0.12	
Subject | Binomial	0.001	0.03		0.000	0.02		0.129	0.36		0.001	0.04		0.003	0.06	
Subject | Collocation	0.003	0.06		0.005	0.07		0.561	0.75		0.003	0.05		0.006	0.08	
Subject | Control 1	0.003	0.05		0.004	0.06		0.057	0.24		0.002	0.04		0.001	0.03	
Subject | Control 2	0.003	0.06		0.005	0.07		0.203	0.45		0.000	0.02		0.000	0.02	
Residual	0.206	0.45		0.204	0.45		-	-		0.092	0.30		0.155	0.40	

Note: *p*-values are estimated using the lmerTest package in R (version 2.0-33; Kuznetsova, Brockhoff & Christensen, 2016). * < .05; ** < .01; *** < .001.