READING BETWEEN THE LINES: PSYCHOLINGUISTIC INDICES OF PREDICTION AND FORMULAICITY IN LANGUAGE COMPREHENSION

BY

NYSSA BULKES

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Assistant Professor Darren Tanner, Chair
Professor Kiel Christianson
Associate Professor Emerita Susan Garnsey
Associate Professor Tania Ionin

Abstract

A comprehensive model of language processing must account for not only how people process literal language, but also how nonliteral language is processed. Further, of theoretical interest to psycholinguists is the role that prediction plays in language processing, namely the conditions under which anticipating linguistic forms and structures can facilitate language comprehension. L1 research has underscored prediction as facilitative; namely, the more informative the surrounding context, the more readers anticipate upcoming information. Research using the *transposed-letter (TL) effect* shows that a target with transposed letters (*cholocate*) are read faster than targets containing substitutions (*choeotate*), as letter position/identity are encoded separately (Perea & Lupker, 2003, 2004). Luke and Christianson (2012) demonstrated that higher semantic constraints lead to specific expectations for letter position/identity, showing that TL effects index prediction. While L2 research has investigated prediction in L2 processing, this research primarily addresses comprehension of literal language. In cases of semantically opaque—or idiomatic—language, it is unclear whether phrase literality affects predictive mechanisms in L1 or L2 processing. Finally, it is also unclear whether semantic opacity differentiates how expressions—literal or nonliteral—are stored and retrieved from the lexicon, namely in cases where dimension such as whole-string or substring frequency are controlled for. Results from three experiments in this dissertation support a dual-route model of language processing, where the mode of processing that is employed is ultimately determined by context.

Acknowledgements

I would like to thank my advisor, Dr. Darren Tanner, and my committee members, Dr. Kiel Christianson, Dr. Susan Garnsey, and Dr. Tania Ionin, for their feedback and commentary on the dissertation document. I would particularly like to thank Dr. Tanner for all of his guidance and support throughout my PhD program at Illinois and through the dissertation process. Without him, this work would not have been possible, and I am thankful to have been his first student. Darren, I promise to do good science and publish acceptable baselines.

I would like to thank the National Science Foundation for generously funding this research (BCS-1528701), as well as the Beckman Institute for Advanced Science & Technology at the University of Illinois at Urbana-Champaign, for funding me during the dissertation year.

I thank my participants for their time and effort, and for contributing to this work. Without them, this work would not have been possible.

Finally, I would like to thank my family and friends for their support and comradery during my time as a graduate student at Illinois. Connection is important.

Table of Contents

1. Introduction

The average literate adult English speaker uses up to 6 nonliteral expressions per minute (Pollio, Barlow, Fine, & Pollio, 1977). Native speakers are able to communicate at such a rapid and seemingly effortless rate, and research on multi-word expressions (MWEs) suggests this is in part due to the maximal use of configurations, or pre-stored expressions that are retrieved as chunks from the lexicon (e.g. Ellis, 2002; Goldberg, 2003; Wray, 2002). While productivity is at the heart of the definition of human language, language users are also incredibly well-versed in making the most out of what they already know works, namely *formulaic language*. Scholars in this domain have argued that, given the options, a language user will use the most common configuration to communicate a meaning, a choice informed by prior experience, namely co-occurrence knowledge (Ellis, 2002). If, in essence, the mental lexicon is a dictionary comprised of all the words in a language user's arsenal, then collocations and constructions are stored here too. Collocations and constructions transcend the boundaries of individual words, and are tapped for use when deemed the most efficient means of conveying a meaning given the context (e.g. Goldberg, 2003).

1.1 Formulaicity in language

In order to formulate a comprehensive model of language processing, we must also understand how we process both literal (i.e. collocations) and nonliteral (i.e. idioms) formulaic language. Psycholinguistic work demonstrates that language users are sensitive to what forms are frequent in their language, and that processing is facilitated for frequent items compared to less frequent ones (e.g. Diessel, 2007; Ellis, 1996, 2002; Hasher & Chromiak, 1977; Shapiro, 1969; see Bulkes & Tanner, 2017, and Libben & Titone, 2008 for discussions of subjective frequency in idioms). However, acquisition of these forms and knowledge of their frequency must accrue over time, through immersion in the language environment. With more experience, language learners become sensitive to co-occurrence information, or the knowledge that certain words "go together" more often than others. Although two expressions may convey a semantically

equivalent meaning, communication is expedited by using the form that is most conventional, or lexicalized (Pawley & Syder, 1983). By using a preconstructed expression to articulate a thought, speakers can economize on processing resources while also ensuring successful uptake. Learning formulae, or chunks, arises from the binding together of items that frequently co-occur and the subsequent recognition of these chunks as meaningful (Ellis, 2002). Conversely, it would be considered marked behavior to use a less familiar, uncommon expression when there is a more, expected canonical way of saying something.

A formulaic sequence, more broadly, is defined as: "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use" (Wray, 2002: 9). By attending to linguistic input over time, experience teaches the learner which expressions are common in a language. By producing these configurations during interaction, the learner is afforded more fluent production by nature of sounding like other people around her in the language environment. Seminal work in the field stipulates that the language processing task is optimally efficient when speakers master the retention and encoding of clauses rather than only individual lexical items (Pawley & Syder, 1983). For example, the use of preconstructed sequences and syntactic frames economizes the language interaction task, eliminating the need to generate exclusively novel utterances (i.e. Goldberg, 2003), ultimately freeing up cognitive resources to attend to other demands placed on a language user during fluent conversation.

Formulaic language is an umbrella term comprising both literal and nonliteral expressions that, over time, have established a rather direct form-meaning mapping within the minds of native speakers. Comprised of comparatively more or less canonical "sentence stems" (Pawley & Syder, 1983), this knowledge is a continuum along which certain expressions are more frozen than others (e.g. see Gibbs & Nayak, 1989; Nunberg, 1978, for discussions of how idioms vary), and along which more or less information is predetermined. For example, in the case of an idiom (i.e. *kick the bucket*), lexical items are specific, and the argument is that early activation relies heavily on recognition of the configuration as

meaningful. However, constructionist approaches (e.g. Goldberg, 1995, 2003, 2006) expand on this by stipulating that other constructions, such as those using a conventional syntactic structure (e.g. *Who did what to whom* relationships) also embody this form-meaning mapping indicative of formulaic language. Goldberg (2003) argues that language, as a whole, is built up entirely of constructions—"constructions all the way down," she says, borrowing from the popular metaphor (2003: 223). She argues that this theoretical departure from the traditional sense of grammar is required in order to account for the patterns apparent in everyday language use. Everything from the fully specified idiom to an entirely abstract phrasal pattern is accounted for by this approach, which defines language on the basis of expressions where the specific surface form(s) used to convey meaning are not entirely novel or conceptualized in the moment.

If we are to construct a comprehensive model of language processing, it must also explain formulaic language use, namely the preference for or dominance of one construction compared to another when other candidates, equivalent in meaning, are available. Rather than rely on an argument of subjective preference, a variety of research has outlined how co-occurrence knowledge modulates processing. Accounts such as these, which focus on how speakers make use of distributional information, must be considered. These models are discussed in the following section.

1.2 Theories accounting for frequency effects in language processing

Research on how frequency affects language processing demonstrates that language users retain information of varying grain sizes as part of their lexicon. As language is inherently built up of finite, discrete units, it is logical to suggest that language users retain a range of linguistic representations, from the smallest of grains (i.e. phonemes) to comparatively larger ones (i.e. words). Relatedly, there is a large body of work studying the psychological reality of transitional probabilities, for example, a speaker's knowledge of the likelihood of word N+1 given N (e.g. Smith & Levy, 2013). Findings from both language production and comprehension studies show more frequent strings are both produced as well as processed faster than less frequent ones in both adults and children (e.g. Arnon & Clark, 2011; Arnon &

Cohen-Priva, 2013, 2014; Arnon & Snider, 2010; Bannard & Matthews, 2008; Conklin & Schmitt, 2008; Jiang & Nekrasova, 2007; Sosa & MacFarlane, 2002; Siyanova-Chanturia, Conklin, & van Heuven, 2011; Tremblay & Baayen, 2010; Tremblay, Derwing, Libben & Westbury, 2011; Tremblay & Tucker, 2011). There is little work, however, that specifically tests how different kinds of formulaic language are represented in the lexicon (i.e. comparing literal collocations to nonliteral collocations), and whether semantic opacity differentiates processing of frequently co-occurring expressions. While there is a body of work that tests the status of larger-grained linguistic chunks in the lexicon (i.e. MWEs; e.g. Arnon & Snider, 2010), there is less work on how smaller subsets of formulaic language differ. Additionally, while there is work looking at how MWEs are represented, namely showing advantages for formulaic expressions used in their canonical form (i.e. "bread and butter", Siyanova-Chanturia, Conklin, & van Heuven, 2011), there is less work testing explicitly how phrase and part frequency affect processing in expressions with varying levels of semantic opacity (although see Jolsvai, McCauley, & Christiansen, 2013, described below, for an investigation in this domain).

A large body of work has demonstrated that language users tend to reuse the same types of recurrent clusters of sounds or words (e.g. Bybee, 2006; Cowie, 1998; Moon, 1998; Sinclair, 1991; Tomasello, 2003). For example, this research employs large corpora to determine which phrases in a language occur more often—and relatively, how often—compared to more novel strings, and the data is used to model differences in language-task performance (e.g. Biber, 2006; Biber, Conrad, & Reppen, 1998; Moon, 1998). Additionally, the argument from this domain is that using corpora to inform stimuli creation and data analysis facilitates a more descriptive approach to the study of language and the lexicon as opposed to a prescriptive one (e.g. Biber, Johansson, Leech, Conrad, & Finegan, 1999; Schmitt, Grandage, & Adolphs, 2004). Specifically, this research shows that frequency effects are not unique to highly frequent strings. As frequency is a continuous rather than a binary distinction (e.g. Bybee, 2006), results in this thread reveal frequency effects in cases where one token string is more frequent than other, highlighting that relative frequency—if a string is more or less frequent than another—also demonstrates graded advantages to processing. This efficiency is seen, for example, in the production of formulaic

sequences, as has been measured empirically by phonetic duration (e.g. Van Lancker, Canter, & Terbeek, 1981; Bybee & Scheibman, 1999). Arnon and Cohen-Priva (2013, 2014) show that, when controlling for part — that is, lexical — frequency as well as speech rate, higher frequency led to shorter duration during production of a target. In a study from their 2014 paper, Arnon and Cohen-Priva illustrated the influence of frequency on phonetic duration by measuring the duration of the middle word in a string rather than the final word to examine frequency independently of predictability. Results showed the duration of the middle word was shorter when the frequency of the word preceding it was higher, further illustrating the psychological separation of frequency and predictability.

Usage-based approaches are a larger class of viewpoints stipulating how speakers' knowledge is informed by the input (e.g. Bybee, 1995, 2002, 2010; Goldberg, 2006; Tomasello, 2003). Perspectives in this domain maintain that the more frequently lexical items co-occur—and are experienced in the input together—the more chunk-like the string of items can be represented and retrieved together from the lexicon. Scholars in this domain continue to question what constitutes a chunk and how holistic representation might affect processing. Namely, if a string has a meaning other than the sum of its parts and can be understood holistically, it is of interest to us how its meaning is retrieved independently of its part semantics or syntax. A true chunk is an expression whose meaning is retrieved once a particular configuration or syntax is recognized, and the more often a person is exposed to this ordering, the more familiar the expression becomes, both its meaning and the social scenarios in which its use would be appropriate. Usage-based approaches would predict, however, that even if a string is experienced frequently in the input, its processing would still be affected by the frequency of its component parts, ultimately ruling out truly holistic storage and retrieval (e.g. Arnon & Cohen-Priva, 2014). The effects of word-level frequency can be attenuated when the string is frequent, and conversely, the less frequent the string, the more its processing is affected by word-level, lexical frequency. The results described above from Arnon and Cohen-Priva (2014) illustrate this point empirically, namely demonstrating the interaction between the information provided by a string and the information provided by its component parts.

Usage-based approaches reside under the broader umbrella of *connectionism*, which relatedly suggests that all language input connects in networks to other language input; isolated units are irrelevant and meaningless (e.g. Elman, 1990; MacWhinney, 1998; Seidenberg, 1994). Connectionist models posit a system comprised entirely of units with dense connections to other units. The more input a person receives, the stronger the connections become between the nodes of the network, where connections are constantly revised and updated with more exposure. A learner becomes more proficient in a language the more opportunities she has to experience naturally occurring input and to use that information to forge stronger connections, for example, the links between frequently co-occurring lexical items. For example, despite similar periods of exposure to a second language, if one learner experiences a particular construction more often in their environment (i.e. field-specific jargon) then the connections in that learner's mind for that expression will be stronger and more robust compared to those of another learner with less experience with the expression.

Connectionist approaches contrast with a *words-and-rules* approach (e.g. Pinker, 1998, 1999; Pinker & Ullman, 2002). Proponents of this view would argue that multiword expressions or phrases are generated by rule, and not represented in the lexicon; only highly formulized expressions (i.e. idioms) are retained unitarily, specifying ruled-based concatenation for other regular forms (i.e. words, phrases, sentences). However, the more conventionalized a configuration—particularly in cases of fossilization of an expression, where it loses its literal meaning over time—the more it can be stored as a word. For collocations literal in nature, these might start out generated by the grammar in a rule-based fashion, but as they gain popularity of use as a construction, the more likely it would be that it be retained as a chunk in the lexicon. While some idioms would allow rule-based processes to modify tense or aspect (i.e. *tip the balance* in "*The balance was tipped in her favor*"), other more nondecomposable forms (i.e. *kick the bucket* to "*The bucket was kicked by John*") do not permit syntactic changes, and would thus be treated like irregulars, where compositional analysis would be blocked by the stored form. Within this framework, only irregular forms—or words (e.g. Pinker 1998; Pinker & Ullman, 2002)—are typically stored, leaving a variety of strings, including those of a larger grain size (i.e. binomials, complex

prepositions), to be governed by rules (for investigations of how expression frequency and generative linguistic knowledge interact, see Morgan & Levy, 2015, and Morgan & Levy, 2016). This suggests that, despite collocational frequency, phrases should not be subjected to whole-phrase frequency effects; frequency effects would be reserved for irregular, memorized forms, and not computed ones. At first glance, this may be permissible, as in the example of a binomial, where "short and sweet" can become "shorter and sweeter". However, what a rule-based approach cannot account for is why both native and proficient nonnative speakers respond to forms like "sweet and sour" faster than they respond to the inverse "sour and sweet" (Siyanova-Chanturia, Conklin, & van Heuven, 2011). A variety of other empirical work has likewise supported whole-phrase frequency as psychologically relevant (e.g. Arnon & Snider, 2010; Bannard & Matthews, 2008; Tomasello, 2003). If regular forms are computed based on rules alone, frequency information would not have the effect it does when comparing speakers' responses to nearly identical forms, where the only difference is the frequency of the order of the configuration.

In addition to frequency, there are other dimensions along which formulaic expressions vary, for example literality. Nonliteral language (e.g. idioms, metaphors, proverbs) are pervasive in the input, and are culturally specific. In a discussion of how frequency and formulaicity modulate processing, it is worthwhile to discuss the psychological reality of nonliteral configurations, and how semantic opacity affects comprehension. I review the relevant literature to this end in the following sections.

1.3 Idioms in language comprehension

Research on idioms has demonstrated empirically that a variety of factors contribute to differences between literal and nonliteral language comprehension, including, but not limited to, what kinds of information—bottom-up or top-down—are used when and to what extent (e.g. Nunberg, 1978; Rommers, Dijkstra, & Bastiaansen, 2013). Whether we compositionally or holistically analyze idioms has been debated for many years, and there are a number of arguments. *Noncompositional* models argue that idioms are stored and retrieved as chunks, entailing a processing advantage for idioms used figuratively in both comprehension as well as production (e.g. Bobrow & Bell, 1973; Gibbs, 1980; Gibbs & Gonzales,

1985; Holsinger, 2013; Swinney & Cutler, 1979). For example, in Swinney and Cutler's (1979) seminal work, the authors introduced the *lexical representation hypothesis*, which suggested that idioms are retrieved like long words from the lexicon. If an expression has both a plausible literal and idiomatic interpretation, both meanings are simultaneously activated and entertained as the context unfolds. This proposal stood in stark contrast to the earlier *idiom list hypothesis* posited by Bobrow and Bell (1973), which suggested that when processing idioms, speakers had to enter a special "idiom mode" where they could retrieve the idiomatic meaning from a longer list of semantically opaque expressions, but that this list was distinct from compositional phrasal processing. To test the psychological reality of a mode of processing unique to idioms, Swinney and Cutler's participants took part in a phrase classification task, where they read literal (i.e. *break the cup*) and nonliteral (i.e. *break the ice*) strings one at a time and were asked to decide whether the expressions were meaningful phrases of English. Participants were faster to indicate idiomatic stimuli were meaningful than they were to indicate literal controls were, a finding that was evidence against Bobrow and Bell's claim that idioms somehow required extra work to process. Swinney and Cutler ultimately used this finding to suggest that idioms provide a computational advantage in processing. They argued this was due to the idioms' long word-like representation in the lexicon; in contrast, the literal strings could not be represented as chunks in the lexicon, and their interpretation required compositional analysis.

Relatedly, in his 1980 work, Gibbs similarly demonstrated that when idioms were used idiomatically—conventionally, as he put it— processing was facilitated (Gibbs, 1980). Gibbs conducted three experiments to investigate how conventionality and presence (or absence) of context affects idiom comprehension. Namely, by presenting participants with idioms used either figuratively or literally, both with and without a surrounding context, he asked whether the use of an idiom in its conventional, figurative sense would take more or less time to read than the expression used literally, and how either the presence or absence of surrounding context would impact this. In the first experiment, Gibbs' participants read idioms embedded in longer sentence contexts (on average 6 lines of context each). Each line of the context was presented on its own on a computer screen, and reading times were collected for how long it

took a person to read each sentence presented on the screen. Once the entire passage was done, the person was asked to provide a true/false paraphrase judgment about a possible paraphrase of the final sentence they just read. Results from this experiment indicated that idioms took less time to read than the same expressions used literally, and there was no additional effect of context, meaning it did not matter whether the idiom was used within or without a surrounding context. Gibbs argued that the results from this experiment supported conventionality as beneficial in measures of overall processing time. In his second experiment, Gibbs was interested in how people would remember idiomatic expressions in conversation either when they were used idiomatically or when they were used literally. He hypothesized that idioms used literally should be remembered more easily because of the additional computation required to reject the idiomatic meaning, saying the configuration would be expected to be used figuratively. In a recall task, Gibbs presented people with the same stimuli from the first experiment. People were asked to come back 24 hours later and provide the last sentence of the stimuli they heard the day before, with correct responses considered those that contained all of the "important content words" and the same syntactic structure from the stimuli-final sentence (Gibbs, 1980, p. 152). Results showed that participants were better at recalling literal expressions compared to idiomatic expressions. The third experiment also tested recall of idiomatic expressions used either figuratively or literally, investigating what kinds of prompts—idiomatic or literal—would lead to proper recall of each the idiom used either idiomatically or literally. Results showed that when people were provided with literal paraphrases to help them recall the expressions they saw, these led to more successful recall than idiomatic paraphrases. While the scoring criteria are at best vague, Gibbs' work highlighted conventionality as facilitative early on in the literature on idiom processing, as impactful both in reading them and in the ability to recall them later on. Additionally, Gibbs often discusses familiarity to define conventionality, saying, for example, in the same paper that "…conventional uses of idioms are very familiar" (1980, p. 152). This paper is seminal, one that is often cited to discuss how familiarity affects comprehension of idioms, and this is frequently, as familiarity is one of the most discussed properties of how idioms vary.

In contrast, *compositional* models of idiom comprehension suggest that idioms are

compositionally analyzed, with each of the expression's component parts attended to in analysis of the string (e.g. Cacciari & Tabossi, 1988; Titone & Connine, 1994). Proponents of this view highlight that successful idiomatic meaning activation requires the recognition of an expression as a configuration, specifically a conventional construction conveying more than the sum of its parts. Cacciari and Tabossi (1988) presented the *configuration hypothesis*, which suggested that in every idiom is a key, specifically the point at which a person realizes the string is an idiom. This recognition point varies from idiom to idiom, but it is at this point that a person is able to recognize the construction as significant and successfully retrieve an idiomatic interpretation. Compositional analysis is employed at the start—the default processing mode initiated in the earliest stages of processing and also used in literal language comprehension—but as soon as the key is encountered and recognition occurs, the person no longer entertains the string's literal interpretation. Within this view, both the literal and figurative interpretations are pursued, and only when the idiomatic interpretation has reached sufficient activation, is pursuit of the literal meaning abandoned. Titone and Connine (1994) argued in favor of the configuration model, after demonstrating in a series of cross-modal priming experiments that idiom predictability—the likelihood of a phrase-final word—facilitated figurative meaning activation. In cases where an idiom also had a plausible literal interpretation, this meaning still showed priming despite what may have been stronger support for the idiomatic interpretation.

Finally, *hybrid* models underscore a person's experience with an idiom as a leading factor in determining the ease with which meaning can be retrieved. Specifically, these models argue that the more familiar a speaker is with an idiom, the more directly its figurative meaning can be activated and retrieved (e.g. Libben & Titone, 2008; Titone & Connine, 1999; Titone & Libben, 2014). For very familiar idioms, compositional analysis takes place after direct retrieval, where the expression's frequent use as nonliteral makes the literal interpretation unlikely; after the idiomatic meaning is retrieved, the literal meanings of the component parts become more available, but they do not interfere with the activation of the nonliteral sense. In cases of unfamiliar or infrequent idioms, compositional analysis takes place first, with the literal interpretation entertained first. If and only if this is infelicitous with the context, is a nonliteral

interpretation considered. Libben and Titone (2008) examined the dimensions along which idiom comprehension varies (also see Bulkes & Tanner, 2017, for a recent account broader in scope), homing in on factors such as decomposability and familiarity as central in determining the ease with which an idiomatic meaning can be retrieved. As one of the earlier proponents of a hybrid model, this paper demonstrated that decomposability—the degree to which an idiom's meaning can be deduced from the semantics of its component parts—is less influential in early stages of processing compared to other factors such as familiarity or predictability. This notion integrates insights from Cacciari and Tabossi's configuration hypothesis, such that the form an expression takes—namely its configuration—plays a key role in cluing in a reader to the fact that they are reading an idiom.

Questions on how idioms are analyzed dominated the field in the latter part of the 20th Century and into the 2000s. Similarly, this research led to the definition of a variety of key constructs involved in idiom processing, for example, familiarity, decomposability, and literality. A number of studies have published normative data to demonstrate both how idioms vary along key dimensions, as well as the extent to which speakers are knowledgeable about this information (e.g. Bulkes & Tanner, 2017; Cronk & Schweigert, 1992; Libben & Titone, 2008; Popiel & McRae, 1988; Schweigert & Cronk, 1992; Titone & Connine 1994). *Familiarity* refers to the degree of salience or subjective exposure a participant has to a particular expression. *Meaningfulness* refers to the degree to which a person is familiar with the actual meaning of the expression. *Literal plausibility* refers to an expression having a plausible literal interpretation in addition to the idiomatic one (e.g. *tie the knot*). *Decomposability* refers to the degree to which an idiom can be decomposed and interpreted based on lexical semantics. For example, *be on cloud nine* is nondecomposable; the meaning of *cloud* and the number *nine* have nothing to do with the notion of being elated. On the other hand, *hit a wall* is more decomposable, as the notion of a wall can indicate a barrier or obstacle, and hitting could refer to a sudden or difficult effort to accomplish something. *Predictability* is measured by an idiom's cloze probability, specifically the likelihood of a participant providing the final word in an idiom in a fill-in-the-blank task. *Frequency* of an idiom refers to the relative frequency of the expression in the language. While frequency is not typically included in norming

papers due to the inherent subjectivity of the norming task, it is typically quantified using data from large corpora. Frequency can be used as a more objective measure of an idiom's prevalence in the language, although the accuracy of this measure necessarily depends on the quality and size of the corpus used.

An important take-away from large-scale norming datasets and studies that produce them is that there are a number of properties of idioms that contribute to how heterogeneous the greater class of expressions is. Some of these properties contribute to the relative semantic opacity or transparency of an idiom, where idioms vary with respect to their degree of decomposability. Nondecomposable idioms, for example, require more experience with the input than decomposable idioms for the configuration to be successfully recognized as idiomatic. Whereas notions such as predictability also apply to literal language, it is unknown whether the predictive mechanisms underlying literal language and nonliteral language comprehension differ. Whereas compositional analysis is sufficient in literal language comprehension, nonliteral language comprehension requires an additional processing step where a person realizes what is meant resides above the phrasal level. First, realization that an expression's meaning is nonliteral requires recognition of the configuration, and second, it requires subsequent activation of the meaning. For nondecomposable idioms, this requires prior experience, which for some idioms can be relatively sparse or nonexistent entirely. While it would be reasonable to suggest that speakers actively predict upcoming stimuli when reading in informative environments, the notion of Cacciari and Tabossi's (1988) idiom key begs the question of whether prediction plays out differently in idioms. Once an idiom is recognized as a configuration, there may only be one or two felicitous completions; literal expressions more readily allow a variety of synonymous completions rather than one or two specific lexical items. If in every idiom there is a point at which a comprehender realizes she is reading an idiom, what is left for prediction after the key has been encountered? Is it the case that semantic opacity carves out a different role for prediction than is observed in semantically transparent, literal language? It is possible that once an expression is recognized as idiomatic, the parser proceeds in more of a "good enough" fashion, such that once a nonliteral meaning has been activated, interpreting the rest of the idiom comes computationally cheaper, requiring fewer attentional resources. This would manifest, for example, when

reading phrase-final words, potentially leading to greater skipping rates during natural reading for idioms compared to literals.

However, in order to ask questions of how prediction works in idiom comprehension, we must first define prediction. I move to this discussion next.

1.4 Predictive mechanisms in language comprehension

The role of prediction in language comprehension is far from fully defined, and the field has yet to reach consensus. For the past 50 years, the concept of prediction in language processing has been of great interest to psycholinguists (e.g. Miller & Isard, 1963; Tulving & Gold, 1963). Mounting evidence suggests a prominent role for predictive mechanisms at multiple levels of processing, including wordform, semantics, discourse, morphology, and syntax (e.g., Brothers, Swaab & Traxler, 2015; DeLong, Urbach & Kutas, 2005, 2014; Dikker et al., 2009, 2010; Farmer et al., 2006; Federmeier & Kutas, 1999; Federmeier et al., 2007, 2010; Fine et al., 2013; Kim & Lai, 2012; Kutas & Hillyard, 1984; Levy, 2008; Lew-Williams & Fernald, 2007; Van Berkum, Brown, Zwitserlood, Kooijman, & Hagoort, 2005; Wicha, Moreno & Kutas, 2004; Wlotko & Federmeier, 2012; see Huettig, 2015; Kaan, 2014; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2013; Van Petten & Luka, 2012, for recent reviews), and that global sentential constraint at the semantic level can encourage prediction (Federmeier et al., 2007; Luke & Christianson, 2012). A variety of experimental methodologies have been employed to better understand the time course with which predictive mechanisms come into play, as well as to observe the costs of disconfirmed predictions or processing of information that deviates from what was anticipated, both in relation to expectations for low-level bottom-up perceptual input as well as higher-level top-down contextual information and constraint (e.g., Federmeier et al., 2007; Kutas & Hillyard 1984; Wlotko & Federmeier, 2012, although see Luke & Christianson, 2016 for an account showing no evidence of costs from eye movement data). For example, research using event-related brain potentials (ERPs) has established there are a number of neurocognitive indices of prediction in language processing as a stimulus unfolds over time. Language comprehension research shows that words are integrated

incrementally into prior context, as opposed to only after all of the information has been fully accessed and processed (Kutas & Hillyard, 1983). Similarly, in the seminal Kutas & Hillyard (1984) paper, the authors introduced electrophysiological patterns—namely, the amplitude of the N400—that indexed the interaction between the cloze probability of a stimulus and the level of semantic activation and priming, suggesting the more likely a particular lexical item given a prior context, the greater its level of priming for integration once confirmed by bottom-up perceptual input. Work around this time, additionally, clarified semantic integration processes as being more immediate rather than delayed (e.g. Van Petten, Coulson, Rubin, Plante, & Parks, 1999). Broadly speaking, work in this thread has focused on what kinds of predictive mechanisms, if any, are used in processing linguistic input, and further, *which* mechanisms come into play *when*. Additionally, questions of whether prediction is an overt, committed process or more of a weaker variant—anticipation or expectation—are actively being debated (see Huettig, 2015; Kuperberg & Jaeger, 2016; Luke & Christianson, 2016; and Staub et al., 2015 for discussion on the terminology).

1.4.1 Empirical investigations in predictive processing

Despite the merits that behavioral methods have to offer, the time-sensitivity of online methods such as eyetracking and ERPs has allowed for more fine-grained analysis of when information is processed in what ways respective of its presentation to a participant. By studying when information becomes available in the brain for processing mechanisms, language researchers have been able to investigate under what conditions people are inclined to anticipate or expect upcoming input, and under what conditions this type of processing is less facilitated. For example, the *visual-world eyetracking paradigm*, in particular, has been successfully used to record participants' anticipatory eye movements after presentation of a stimulus (e.g. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Specifically, this method affords the ability to track a person's eye movements starting from when a stimulus is first presented to when the participant began to act on expectations for a particular outcome, or when contradictory information became available to motivate a saccade somewhere else. For example, Altmann

and Kamide (1999) employed the paradigm to demonstrate that speakers are sensitive to selectional restrictions on verbs, and that verbs with stronger selectional restrictions encourage more anticipatory eye movements than verbs with fewer selectional restrictions. In their study, speakers demonstrated earlier eye movements to a picture of something edible when listening to sentences like *The boy will eat the cake*, than while listening to sentences like *The boy will move the cake*. In this example, "eat" selectively restricts for something that can plausibly be eaten, while "move" allows a wider range of plausible objects. Results such as these suggest that speakers are sensitive to selectional restrictions and that the use of this information encourages efficient sentence processing.

Findings from studies such as these show us that as new input unfolds, speakers retain information provided by the preceding discourse to inform language processing behavior downstream. Further, we know from this work that speakers are also well-versed in attending to things like thematic role assignments and semantic constraint to generate an idea of what may be upcoming, incorporating each new piece of a stimulus into the discourse representation (e.g. thematic role assignments; Altmann & Kamide, 1999; Kamide, Altmann & Haywood, 2003; Boland, 2005). For example, in their 2005 study, DeLong and colleagues showed that, semantic constraint of a sentence informed participants' expectations of upcoming words, including phonological information provided by constrained determiners (i.e. *a/an*; although see Ito, Martin, & Nieuwland, 2016 and author response DeLong, Urbach, & Kutas, 2016 for more on the replicability of this effect). Research like this suggests that at a rudimentary level, speakers construct a representation that constrains upcoming expectations to include basic information such as phonology and part of speech. In a 2005 study, Van Berkum and colleagues demonstrated a similar level of prediction during comprehension using ERPs, where participants' responses were argued to be affected by their expectation for grammatical gender, as in Dutch, the gender on the article and noun must agree (for similar findings in Spanish, see Wicha, Bates, Moreno, & Kutas, 2003; Wicha, Moreno, & Kutas, 2004; and Foucart, Martin, Moreno, & Costa, 2014). Studies such as this further illustrate how, given a prior context, speakers can predict specific words as well as the features of those words (i.e. phonological information as in the *a/an* distinction described above). Proponents of

predictive language processing argue that as bottom-up information becomes available, this information is added to the existing framework composed of already-processed context and that this built-up information facilitates the formation of predictions downstream. As new information becomes available, the representation gets confirmed and updated in cases where revision is needed.

The appeal of a parser that predicts upcoming input is that it attenuates the problem of having to interpret in a noisy or impoverished environment (e.g. Stilp & Kluender, 2010; see Davis & Johnsrude, 2007 for a review) and accounts for how speakers overcome noise during comprehension. Likewise, it is also said that comprehenders covertly produce what is being processed in the input (Dell & Chang, 2014; Pickering & Garrod, 2013; Pickering & Garrod, 2007; Wilson & Knoblich, 2005). Wilson and Knoblich (2005) argue that by imitating covertly what is being perceived, predictions are made that enhance the perception of that input, making use of the covert production system to establish better memory of what is being perceived. By actively incorporating incoming input into the representation of what is being understood, some research suggests comprehenders pre-activate input they consider to be highly likely given the discourse, down to specific lexical items or semantic features (e.g. Federmeier & Kutas, 1999; Luke & Christianson, 2012; Van Berkum et al., 2005; Wicha et al., 2003, 2004; Wlotko & Federmeier, 2015). Pre-activation, however, remains a controversial notion in the language processing literature. I discuss the relevant literature on this distinction next.

1.4.2 Prediction, anticipation, and expectation – A war of words

What remains to be seen is whether comprehenders actively predict—that is, entertain one potential candidate as likely and go so far as to pre-activate a lexical item—or whether people more loosely anticipate what's coming. Recent work has made an effort to distinguish whether routine language processing constitutes overt *prediction*, as a purposeful and committed process, or whether this is too strong a claim. Where prediction requires expectation of a particular form, some research insists this claim is too strong in most instances (e.g. Huettig, 2015; Kuperberg & Jaeger, 2016; Luke & Christianson, 2016). Instead, unless it economizes processing, *anticipatory* processes are engaged, which

denote much less commitment to a particular form and, instead, the consideration of multiple possible candidates, some of which may be more probable than others. Probabilistic models of language processing use *prediction* as the sense more strongly tied to the likelihood of an item's occurrence given what preceded it (e.g. Bayesian approaches; Kleinschmidt & Jaeger, 2015; Levy 2008; Smith & Levy, 2013, and others). Approaches in this thread permit anticipation of multiple candidates with potentially different weights or beliefs.

While it is uncontroversial that prediction occurs to some extent in processing, more recent debates also (e.g. Kuperberg & Jaeger, 2016) grapple with whether prediction is serial or whether multiple candidates can be entertained in parallel. While serial prediction would allow for pre-activation of a highly likely candidate (e.g. "bucket" in *kick the bucket*), a parallel approach would allow pre-activation of several possible candidates when perhaps all share requisite semantic or orthographic features and many options are equally likely. Recent research (e.g. Kuperberg & Jaeger, 2016) takes issue with the presumption that input can be pre-activated at all. Rather, they maintain that pre-activating lexical items is too burdensome and rarely successful; in most language processing tasks, input is more weakly *anticipated* rather than overtly predicted and pre-activated. This view is also termed *graded prediction* (Luke & Christianson, 2016). For example, in a typical case of language processing, the preceding context is not so constraining such that only one or two candidates are viable. Additionally, studies showing highly predictive effects employ stimuli with a high level of constraint at varying levels of the linguistic representation (e.g. DeLong et al., 2005; Kim & Lai, 2012; Luke & Christianson, 2012), suggesting that results supporting overt prediction may be an artifact of experimentation and ultimately unlikely in natural discourse (Kuperberg, in-person communication; Luke & Christianson, 2016). Namely, in scenarios where the constraint is not so high, lexical prediction may not be appropriate or even all that helpful.

However, to ask whether we predict in language comprehension—on either a course- or fine-grained level—we must also identify the goal of language comprehension. Whatever processes are

employed in an efficient parser must also support the goals of a successful, cooperative language consumer. If a person uses language to convey or receive a message, then the best theory must describe how a language user most optimally navigates the task. Luke and Christianson (2012) demonstrated that predictions can be incredibly facilitative, providing highly specific cues for what information is likely coming next—although, in their case, the authors also showed that these highly specific predictions leave little room for deviation from those expectations, where the more specific the prediction, the greater the disruption to an anomaly (see Section 1.4.3 for in-depth discussion of these findings). In cases where the configuration of an expression leaves little room for deviations or alternate completions—for example, a familiar MWE—it may be more in the best interest of the comprehender to partially activate the phrase-final completion compared to when she comprehends a novel string. In this case, this would make the most out of prior language experience to economize processing in the present. In the case of an infrequent, completely novel construction, however, specific prediction for particular lexical items would be a waste of time. Instead, it might be easier—and computationally cheaper—to simply wait for the sentence to unfold and process it when it becomes available without explicitly predicting anything. However, Luke and Christianson (2016) found that people are good at predicting things like word category information, which, again, suggests that something weaker—like expectation or anticipation—may be more theoretically tenable than prediction. When discourse is a highly predictable, though, prediction can be more facilitative, as it both maximizes the likelihood of felicitous interpretation and frees up processing resources for other cognitive demands (i.e. planning the next utterance, further listening or reading). In the case of an idiom, particularly familiar ones, where the configuration requires specific lexical items, predictions to pre-activate those particular words could expedite processing.

For this reason, idioms are an interesting test case for examining predictive mechanisms in comprehension. Due to their predictability given a recognizable configuration, idioms are a prime locus for an investigation of how local and global constraints modulate prediction in cases of formulaic expressions, and similarly, how bottom-up and top-down information interact to inform language comprehension. Comparing idioms to formulaic strings that are literal in nature not only has the

opportunity to provide novel insight into how phrase and part frequency affect prediction, but also into the predictive underpinnings of nonliteral language comprehension. There are no studies to-date that control for part as well as phrase frequency while manipulating semantic opacity, an endeavor that would be uniquely insightful for teasing apart how literal and nonliteral language comprehension differ. Filling this gap in the literature is one of the primary goals of this dissertation.

What MWEs—both literal and nonliteral—have in common is the degree to which their configuration is recognizable; namely, once a configuration is identified as meaningful, a holistic representation is available. In their comprehension, we expect semantic opacity to play a leading role in the relative ease or difficulty of processing. Specifically, in cases where the expression requires higher-level processing—i.e. to retrieve a nondecomposable meaning—semantically opaque expressions should be harder to process than semantically transparent ones because they require that additional computation. Further, we know that other sources of information also contribute to comprehension of a text or passage, namely the appearance or order of expected letters in the string. In an investigation where reading is the method, word recognition mechanics come into play and, relatedly, can be manipulated to ask how literal and nonliteral MWEs differ in processing. In the next section, I discuss the existing literature on how bottom-up information (i.e. visual feature, orthography) informs linguistic prediction.

1.4.3 Prediction and the visual input

Eyetracking has widely been used to investigate predictive mechanisms in online sentence reading (e.g. Ehrlich & Rayner, 1981; Kliegl et al., 2004; Rayner & Well, 1996), as this method affords temporal precision as to when information becomes available for predictive inferencing. For example, work in the reading domain has shown that low-level information from upcoming wordforms (i.e. letter identity) is available in the parafovea, and that the availability of this information affects which words can be skipped in natural reading. Namely, word length, predictability, and frequency are all factors that determine which words are more likely to be skipped, with length and predictability being the most influential (e.g. Balota, Pollatsek, & Rayner, 1985; Blanchard et al., 1989; Drieghe, Rayner & Pollatsek, 2005; Rayner et

al., 1982; Rayner, 1975; Rayner & McConkie, 1976; Rayner & Well, 1996). For example, Rayner and McConkie (1976) found three-letter words were skipped about 67% of the time, whereas longer, seven- to eight-letter words were skipped much less frequently, just 20% of the time. This supports the benefit offered by the parafovea, where information to the right of fixation is available to the comprehender even though it is not explicitly attended to. Similarly, when matched for length, a word that is more predictable given the prior context is more likely to be skipped than one that is not as well supported (e.g. Ehrlich & Rayner, 1991; Rayner & Well, 1996).

Research in this domain argues that predictions made during processing are not vague, but fine-grained, precise enough to account for upcoming letter and sound information (Frisson et al., 2005; Morris, 1994). For example, studies using the *transposed-letter (TL) effect* show letter identity and letter position information are encoded separately. This would predict that a word containing a transposition (i.e. *cholocate*) should be processed almost as fast as the correctly spelled word (*chocolate*), and faster than a misspelling with substituted letters, despite retaining any visual similarity to the expected characters (i.e. *choeotate*; Duñabeitia et al., 2012; Grainger, 2008; Perea et al., 2008; Perea & Lupker, 2003a, 2003b, 2004; Rayner, White, Johnson, & Liversedge, 2006). What many of these studies have in common, however, is the use of masked priming as a paradigm, where the transposition or substitution is not consciously fixated, but rather flashed very briefly onscreen prior to a mask. In most studies using this effect, where the perturbation of the stimulus is not directly attended to, results widely suggest at least partial if not full facilitation for target activation given a transposed prime compared to a substituted prime (e.g. Forster, Davis, Schoknecht, & Carter, 1987; Forster, Mohan, & Hector, 2003; Perea & Lupker, 2003a, 2003b, 2004; Schoonbaert & Grainger, 2004). A prime with letter substitutions is less facilitative in priming a target, as the visual string is even less of a match to what a speaker's experience tells them is a word of the language—it is a mismatch both in terms of character identity as well as order. While a prime with a transposition would still be considered a nonword, the expected letters are in the string despite incorrect placement, and it still provides the processor with a good-enough match to activate the intended target. Additionally, Perea and Lupker (2004) demonstrated that this effect holds

when the transposed letters are not adjacent (i.e. *casino* to *caniso*), corroborating the finding that only the first and last positions in a string are privileged.

When transpositions are unmasked, this impoverishment leads to more disruption in natural reading, as was the case in White, Johnson, Liversedge, & Rayner (2008), where the researchers used eyetracking to measure readers' eye movements in response to both letter position (e.g. *problem* v. *porblem*) and externality (i.e. *problem* v. *rpoblem*) manipulations. Similarly, stimuli varied with respect to frequency (i.e. frequent: *problem*; infrequent: *anagram*) to determine how differing degrees of frequency affected processing of targets with transpositions. Results further illustrated the privileged status of initial and final characters, as transpositions were most disruptive to reading when the first or final characters were not in place, as measured by longer fixation durations, with word-initial transpositions more disruptive than word-final transpositions. Also, results showed that higher frequency words with transpositions were fixated for shorter times than lower frequency words with transpositions, suggesting that higher frequency was facilitative in activating the base form. For lower frequency words, which were already more challenging for participants due to their infrequency in the language, transpositions provided an additional obstacle, more tangible in infrequent words than frequent words. These results are important findings in the domain of visual word recognition and sentence processing. Namely, these results illustrate how top-down knowledge can feed forward to aid in processes such as lexical access. They demonstrate the support that top-down cues, such as lexical frequency, can provide in overcoming impoverished bottom-up sensory cues, such as spelling and wordform appearance. Additionally, with respect to the TL effect, these results also demonstrate the effect that an impoverished visual stimulus has on natural reading, specifically when a stimulus is attended to and not appear only as a prime, as is found in a number of other previous studies. While a transposition is less disruptive when masked, it is more disruptive during foveal presentation; however, it is much less disruptive than a letter substitution (see also Stites, Federmeier, & Christianson, 2016, for an investigation on compound words using the TL effect). This suggests that, unless the target is skipped, natural reading does show disruptions from a

visual perturbation but this disruption is graded with respect to the degree of mismatch between the experienced and the expected target (e.g. Rayner et al, 2006).

Other work has examined the effects of letter information and parafoveal preview on natural reading. For example, Johnson, Perea, & Rayner (2007) embedded five-letter targets from Perea and Lupker (2003) in sentences for three silent reading experiments. Sentences were weakly constraining such that none of the targets were deemed predictable given the prior context. They found that participants were able to extract letter identity information from the region to the right of fixation and extract this independently of letter position information. Transpositions were more facilitative in natural reading than substitutions, a result the authors used to argue in favor of the ability to flexibly encode letter position and identity when a target is not directly fixated. Relatedly, Luke and Christianson (2012) found that, in highly constraining contexts, transpositions are just as disruptive as letter substitutions in non-biasing prior contexts. In a series of two experiments, the authors first replicated results from Johnson et al. (2007) by using a new experimental paradigm—self-paced reading with masked priming (SPaM)—where they tested effects of top-down processing in sentence reading. While Johnson et al. (2007) found that TL-medial (i.e. *jugde*) and TL-final (i.e. *judeg*) primes provided fairly equal facilitation when presented in the parafovea, Luke and Christianson's Experiment 1 found greater facilitation for TL-medial primes than for TL-final primes, an effect they argue replicates more closely the earlier work of Perea and Lupker (2003), where the authors found a similar pattern. This demonstrates again that, at least in weakly constraining contexts, word-internal transpositions still provide facilitation in lexical access, and that letter position and identity information are flexibly encoded. In Experiment 2, Luke and Christianson used the SPaM paradigm again, this time to present participants with highly constraining sentences (>75% expected completions in a cloze task) containing masked primes featuring either a transposition, substitution, or the identity target. They found that the TL priming effect observed in less constraining sentences disappeared when the targets appeared in high-constraint sentences. This finding suggests that the more informative the preceding context, the more specific predictions become, including specific predictions for both letter position as well as identity, illustrating an interesting interaction between top-

down and bottom-up cues in processing. These precise predictions led comprehenders to more specific predictions about the upcoming linguistic input, and letter transpositions were more disruptive to reading than if they were embedded in lowly constraining contexts. This suggests that although the required letters are visually available, high constraint of the prior context leads to more specific expectations for what should be upcoming, ultimately underscoring the mismatch between what is visually perceived and what was expected. Together with White at al.'s results, findings from Luke & Christianson (2012) demonstrate the effect that predictability has on processing, where both types of cues are top-down and show how top-down and bottom-up information work together to inform a person's processing of an item. Also, while both predictability and frequency affect processing, these notions are not synonymous. For example, a lexical item can be highly likely given a context (i.e. the completion to *Merry* _____), yet infrequent in the input. Taken together, these two studies show how skilled native speakers make use of the available cues to potentially overcome information that is misleading or somehow lacking during subconscious processing—as in priming studies, where the linguistic cue in question is not directly fixated—as well as in purposeful, explicit processing—as White and colleagues found in their investigation of natural reading.

For short, predictable words, there is a greater likelihood of skipping compared to longer words that are in less constraining environments. Further, we also know from studies such as Johnson et al (2007) that information about wordform (i.e. letter position, identity) is available in the parafovea and can affect real-time processing. However, it is possible that if a phrase-final completion is fairly predictable, less processing resources may be allocated to it, and a misspelling might not be noticed in such an environment. However, it is also possible that when low-level information available in the parafovea is processed as anomalous, this may make skipping less likely. MWEs are a perfect tool with which to examine this mechanism, as the lexical items in an MWE, by definition, are part of a configuration, where certain items are expected to appear in a particular order. If when reading an MWE the configuration is recognized as meaningful, this too, may encourage skipping, as the information available in the parafovea should match what the person's world knowledge is telling them should follow in the configuration. In

the case of reading in suboptimal conditions—namely, having words with transpositions or substitutions available in the parafovea—if at least part of the upcoming target matched what the person thought should come next in the configuration, this might mitigate even further the act of processing impoverished visual input. When reading MWEs, it is possible that letter order and identity encoding may be comparatively less critical than this type of encoding in reading novel language strings, thus encouraging the person to either attend to the stimulus less carefully or attend to it at all. If the person has enough information to realize they're reading a conventional, familiar expression, this may encourage them to rely less heavily on low-level visual cues, thus economizing on processing resources. If this is the case, we would expect that transpositions be less disruptive than substitutions, where as long as the expected letters were present in the string—as would be the case with a transposition—preview would still facilitate efficient processing, compared to a string with substituted, unexpected characters. On the other hand, as Luke and Christianson (2012) found, it may also be that when a specific phrasal completion is highly expected, these unexpected spellings might be even more disruptive compared to when reading a novel language string. If more specific predictions lead to stricter encoding for letter position and identity, this should incur a greater processing burden when an unexpected visual stimulus is encountered.

Findings from reading studies underscore how both bottom-up as well as top-down information affect sentence processing. When reading in configurational contexts, it may be that higher-level, top-down cues are more useful due to the canonicity of the syntax—as long as the expected pieces look as if they are in place, the expression's chunk-like representation may require less from bottom-up sensory input in processing. Further, considering the larger class of expressions that are MWEs, it is unclear whether the type of MWE makes a difference—namely, if we compared relatively transparent, compositional expressions like literal collocations to comparatively noncompositional expressions like idioms. Specifically, it is unclear how the degree of semantic opacity inherent in the expression would make a difference in how reading something like a chunk would be influenced by impoverished sensory input. What is interesting about the findings from Luke and Christianson (2012) is that they suggest that the level of constraint in a sentence directly impacts the strength of predictions that can be made in

processing, including predictions for visual features. While both bottom-up and top-down information both influence comprehension, it is also unclear whether the literality of an expression—or semantic opacity—impacts this any differently. For example, it is unclear whether the degree of semantic opacity in an MWE provides qualitatively different top-down information, and whether collocational information affects anticipatory mechanisms and how heavily bottom-up cues are relied upon. In cases of fixed expressions (i.e. idioms, collocations), co-occurrence information may act as an additional cue to either more highly constrain what upcoming information is deemed acceptable, or perhaps encourage more skipping, such that information to the right of fixation is relied on less. Reading a collocation—either literal or nonliteral—may actually help to mitigate any disruption wrought by visual anomalies, such that as long as nearby information appears to be intact in the parafovea, processing the rest of the MWE could come at a computational discount—i.e. less disruption when reading targets with unexpected orthography.

From a formulaic language perspective, it would be empirically interesting to test how literality impacts prediction. Idioms often contain at least one function word (e.g. *slap in the face*; *kick the bucket*), which are often less than three characters long. If idioms are accessed holistically, intra-word spaces in idioms might not be treated like intra-word spaces in literal collocations. Despite controlling for phrase frequency, the effect of recognizing a configuration as idiomatic may be so strong that comprehending the rest of the expression may come computationally cheaper, even automatically. As an index of this, if expressions like familiar idioms are stored and retrieved more unitarily, letter transpositions farther along within an idiom may be less disruptive when directly fixated than transpositions within literal collocations. Letter transpositions and eyetracking can thus be used to explore predictability effects, the manner in which idioms and collocations are retrieved and accessed from the lexicon, as well as the effects of semantic opacity when processing high-probability, locally constraining strings.

1.4.4 Prediction and top-down influences of frequency and context

Findings from previous research—particularly those from work like Luke & Christianson (2012)—motivate further inquiry on the influence of global versus local context on prediction in comprehension.

Context, understandably, is a leading cue in language processing—recall the last time a person asked for a translation, and was met with "What was the context?" While the notion of context is immense and far from simple, its inclusion in language processing models is necessary. Formulaic expressions vary along a number of dimensions—i.e. degree of literality, for starters—yet little work has examined the interaction between low-level sensory information and top-down phrase- or sentence-level constraint. While Luke and Christianson found global sentential constraint affected linguistic predictions, it is unclear whether predictions are differentially affected by a phrase's semantic opacity, and how in cases where expressions have equally high lexical co-occurrence, whether literality provides any additional advantage in processing. Some research has look at this, namely comparing frequently co-occurring strings to more novel strings and results show that the local constraint of an expression is highly influential in processing. Underwood, Schmitt, and Galpin (2004) used eyetracking to record fixations on a target word used in both formulaic and novel environments (e.g. *as a matter of fact* vs. *a well-known fact*), finding fewer fixations for the target *fact* when used in formulaic strings than when used in novel contexts. They argued that the meanings of formulaic sequences were retrieved more unitarily, underscoring co-occurrence information as predictive of reading behavior. This supports other findings showing that highly frequent strings are processed holistically, where the local context highlights the configuration needed for holistic retrieval (e.g. Cronk & Schweigert, 1993; Katz & Ferretti, 2001, 2003; Schweigert, 1986; Schweigert & Moates, 1988).

Eyetracking has similarly been used in the past to investigate online figurative meaning activation (e.g. Cronk & Schweigert, 1993; Frisson & Pickering, 1999, 2001; Lowder & Gordon, 2013; Schweigert, 1986; Schweigert & Moates, 1988; Titone & Connine, 1999). Among researchers, there is a consensus that the configuration is important for holistic meaning activation and retrieval, and that this is true for literal strings as well. For literal language, it is likely that the configuration is recognized through transitional probabilities, or the frequency of the n-gram, where more expected, likely completions or words N+1 are read more quickly or with greater ease (e.g. Smith & Levy, 2013). This contrasts with idioms, again, where recognition of the configuration triggers a subsequent mechanism, namely activation

of a meaning that resides above the sentence level. We could predict, then, that once an idiom's configuration is identified, this should generate stronger expectations for a particular idiom-final completion. While literal strings can felicitously be completed with synonyms, idioms require specific lexical completions and are arguably more locally constraining than literal collocations. Further, if stronger predictions are made about phrase-final words in idioms, deviations from those predictions in the visual stream might slow processing, comparatively more for idioms than for unexpected completions to literal collocates. Because the configuration of an idiom is so required for successful processing, any deviation from that configuration should incur a processing penalty. While literal collocations can also be recognized as having a meaningful configuration, arguably other words could appear in phrase-final position and be more or less felicitous; this is not the case with most idioms, as their semantic opacity necessitates that a particular completion to retain the figurative meaning.

However, this hypothesis only holds if semantic opacity distinguishes strings with high local co-occurrence, and there is work that argues this is not the case. Jolsvai, McCauley, and Christiansen (2013) argue that, when matched for phrase and part frequency, multiword expressions are part of a greater homogenous class. In their study, participants completed a phrase judgment task similar to that of the early Swinney & Culter (1979) work, where participants were asked to determine whether visually presented stimuli were English phrases. Participants saw either an idiom (e.g. *over the hill*), a collocation (e.g. *had a dream*)—idioms and collocations were frequency matched—or a random word string (e.g. *hear I isn't*). Participants were equally as fast to say an idiom was a phrase as they were to say the same of a collocation. The authors argue this result is evidence that, in native speakers, differences in frequency—and not semantic opacity—are what lead to an advantage in processing. Arnon and Snider (2010) also support this notion, where they argue that frequency effects are observable in all expressions along a continuum of frequent to infrequent expressions. They stipulate that frequency, specifically, is the only key differentiating factor determining how and when processing is facilitated. They also argue that there is little qualitative difference in processing idioms and literal language, though they do not test this directly. There does not appear to be any research to date that directly compares the recognition and

processing of literal and nonliteral collocations embedded in sentence contexts, specifically controlling for local co-occurrence. While Jolsvai et al.'s results support frequency as highly influential in successful identification of strings as possible, legal strings of the language, their task ultimately does not tap into whether participants actually accessed the meaning of the expressions, or whether their responses were based on perceptions of grammaticality. It remains an open question how collocation information gleaned from one's linguistic input affects predictive mechanisms, and whether top-down, semantic constraint and information about a phrase's literality both interact with bottom-up, sensory cues, and whether this informs prediction any differently when comparing different phrase types.

In sum, we know from prior studies that what a person knows of their linguistic input affects how they process language. Top-down cues, such as semantic constraint or frequency, affect how we process language. Similarly, we know that readers actively make use of bottom-up, sensory cues as they read a visual stimulus; the spelling must be in place for language communication to succeed. While we know that language users can anticipate upcoming input constraining, informative contexts, it is unclear how collocational information affects prediction, and whether semantic opacity, or literality, differentially impacts this. If experience with the native-language input informs a speaker about what things have a higher or lower likelihood of coming next, then it would be reasonable to suggest that collocations should receive facilitation in processing. However, we also know that MWEs can vary with respect to literality, and we do not yet have a firm theoretical understanding of how literality affects prediction. To investigate this, we would need to test how readers process unexpected visual cues in both literal and nonliteral contexts. Further, we would need to control for the variety of factors that we know to affect prediction—i.e. length, phrase frequency, part of speech, and the frequency of the individual words that make up the target phrases—in order to compare idioms and collocations' processing in a controlled, direct way. Further, if perceptions of literality are informed by experience with the input, doing all of this in a second language would also be of theoretical interest to scholars of second language acquisition. I will discuss the body of research in second-language processing in the next section.

1.5 Second-language sentence processing

Accounting for how linguistic knowledge is stored and accessed is one of the cornerstones of psycholinguistic research. A relevant view of this question lies in second-language processing research, namely looking at both qualitative and quantitative differences in reading in one's first (L1) compared to their second language (L2).

1.5.1 Models of second-language sentence processing

There have been a number of models proposed to explain how two languages can be maintained in the brain in adulthood, and to describe the mechanisms involved in accessing them during language comprehension. Research by Ullman (2001a, 2001b, 2001c) and others posit that differences in L1 and L2 processing are rooted in the cognitive architectures implicated in memory systems. Specifically, according to Ullman's *declarative/procedural* model, rule-based operations and lexical access rely on different neurological areas—i.e. procedural memory in the left frontal lobes and basal ganglia, and declarative memory in the temporal lobes. L1 processing, he argues, can be characterized as automatic and implicit, whereas L2 processing is comparatively more explicit and conscious. The reason for this, he says, is that computations reserved for the procedural system in the L1 are shifted to the declarative system in the L2, and that this shift is mostly affected by L2 age of exposure. For example, the model suggests that non-productive, noncompositional forms (i.e. *go-went*) reside in the L1 declarative system, whereas productive forms derived from morphological transformations (i.e. *walk-walked*) reside in the procedural system. Noncompositional strings in the model would include both irregular forms (i.e. *sing-sang*) as well as memorized expressions, such as idioms, where neither type of construction can be computed based on rules alone.

Other models have been proposed to illustrate differences in L1 and L2 processing. In the domain of complex syntax, the *shallow structure hypothesis* (Clahsen & Felser, 2006) suggests that the primary difference between L1 and L2 processing lies in the simplicity of the structural representations computed by L2 learners compared to native speakers. For example, in cases of complex filler-gap dependencies or

reduced relative clauses, where native speakers compute hierarchies to process the relationships in a sentence, L2 learners are restricted to computing shallower structures, all the while being adept at utilizing lexical-semantic and pragmatic information to arrive at an interpretation. The authors compare their proposal with other hypotheses from the sentence processing literature, for example "good enough" processing (i.e. Ferreira, Bailey, & Ferraro, 2002) and underspecification (i.e. Sanford & Sturt, 2002). They compare this type of processing to L1 comprehension, for example, when native speakers are misled by the meaning of content words in passive sentences, identifying them as plausible when they are not (i.e. *The dog was bitten by the man*). In cases where L2 learners are highly proficient, there is evidence to suggest these speakers employ the same processing mechanisms for understanding morphology as native speakers.

What these models share is a possible foundation from which to explain L2 acquisition and processing of idioms, yet both have their drawbacks. For example, for both irregular forms and idioms, Ullman's model would predict L2 users with greater proficiency and earlier age of exposure to be better at processing and productively using noncompositional forms. While this makes intuitive sense, this prediction is too simplistic, as it ignores the dimensions by which noncompositional strings vary (i.e. familiarity, decomposability). For example, where more decomposable idioms (i.e. "sign on the dotted line") allow compositional analysis and can be successfully processed without prior exposure, other expressions do not permit this (i.e. "go pear-shaped"). While nondecomposable strings may be memorized as chunks and stored unitarily, decomposable expressions still allow computation; this non-binary range of expressions throws a wrench in Ullman's declarative/procedural distinction when applied to idiom comprehension. Prior work illustrates the consensus that idioms, as a class of expressions, are heterogeneous (Bulkes & Tanner, 2017; Libben & Titone, 2008), and that the dimensions along which they vary affect processing ease. For example, some idioms may be easier for an L2 user to acquire due to item-level characteristics. Relatedly, a shallow structure account for idiom comprehension might predict a compositional-first approach, where lexical-semantic knowledge guides processing. This would suggest that in cases of both L1 and L2 comprehension, readers would be biased toward a literal interpretation of

an idiom. As a person gains more experience encountering a particular idiom, the configuration may more easily trigger idiomatic meaning retrieval, more easily than for an uncommon or unfamiliar expression. However, this would suggest literal interpretations would be entertained prior to nonliteral ones all of the time, as the meaning of the literal expression resides at the sentence level, and not above, as an idiomatic meaning does. This prediction, too, is complicated when considering a person's relative familiarity and frequency of exposure to an item (i.e. Cacciari & Tabossi, 1988; Titone & Libben, 2014), an idiom dimension we know to affect processing.

For the purposes of conceptualizing what types of information are housed where in the mind, both models may be valid, but they ultimately fail to both capture the complexity of the mechanism underlying comprehension of noncompositional expressions, as well as to explain differences in L1 and L2 processing.

1.5.2 Reading in a second language

There is arguably an increased demand on processing when reading in the L2, as a person recognizes and activates words and phrases in a language that is not their native tongue (e.g. Segalowitz & Segalowitz, 1993). However, a variety of studies have shown evidence for quantitative differences between L1 and L2 reading behavior, but not qualitative differences (e.g. Foucart & Frenck-Mestre, 2012; Frenck-Mestre & Pynte, 1997; Hoover & Dwivedi, 1998). For example, Hoover and Dwivedi (1998) found that both "fast" and "slow" L2 reader groups processed syntactic ambiguity, showing that reading speed was not predictive of whether readers noticed the ambiguity. Relatedly, Frenck-Mestre and Pynte (1997) found that highly proficient bilinguals were not only sensitive to ambiguity, but they were able to use verb subcategorization information (that is, idiosyncratic information about the co-occurrence of a particular verb and a particular syntactic frame, which must be acquired via language experience) to resolve the ambiguity in real time. Further, their processing was not differentiated by whether they were reading in their first or second language, suggesting that highly proficient L2 speakers utilize lexical-semantic cues in online L2 reading, even in cases where these lexical constraints differ between the L1 and L2. Further,

Frenck-Mestre (2002) showed some qualitative differences in reading, itself—for example, what was read when and how often it was revisited—and not differences in general processing. Specifically, she showed that skilled non-native readers experience the same challenges as native readers in first-pass reading of sentences (i.e. syntactic ambiguity). However, slower reading behavior in nonnative readers was attributed more specifically to more re-reading and regressions compared to skilled native readers. These three studies, together, provide evidence supporting the qualitative similarity between reading in one's L1 and L2, suggesting it may be other factors that drive differences in L1 and L2 reading (i.e. proficiency, age of acquisition). For example, the more experience a person has reading in the L2, the more likely it is the person will have had prior experience to a phrase—assuming it is not a novel expression. There is a breadth of empirical work supporting proficiency as a determining factor in predicting an L2 user's ability to use linguistic cues in real-time (e.g. Hahne & Friederici, 2001; Hopp, 2006; Jackson, 2008; Keating, 2009; McLaughlin, Osterhout, & Kim, 2004).

An influential factor in L1 and L2 processing is the frequency of a stimulus—or the relative amount of exposure a person has had to a particular form in their input, and psycholinguistic models underscore that frequency directly predicts the level of processing ease or difficulty (MacDonald, Pearlmutter & Seidenberg, 1994; MacWhinney, 2001). Whereas L1 users have a lifetime's worth of opportunities to internalize the frequency of linguistic tokens in the input, L2 users will necessarily have less, and this applies to the acquisition of expressions of all grain sizes. Namely, the more a person experiences a string of words together as an expression, the more likely the person is to recognize the configuration as meaningful. Additionally, whether an L2 learner's education is immersive or confined to a classroom with a textbook will also result in different frequencies for different types of forms (i.e. colloquialisms, formal structures). In cases of noncompositional or ambiguous, unfamiliar expressions—like some idioms—a reader must make use of top-down contextual cues and prior knowledge to arrive at the correct interpretation. When a person has less experience with the L2 compared to their L1, we might predict a transfer of ambiguity resolution strategies from the L1 to processing ambiguities in the L2 (e.g.

MacWhinney, 1997). However, it has also been shown that informative contexts can facilitate L2 processing in ambiguous environments, namely when reading cognates, although results are mixed as to whether this facilitation manifests in early or late processing measures (i.e. Libben & Titone, 2009; Van Assche, Drieghe, Duyck, Welvaert, & Hartsuiker, 2010).

1.6 Formulaic language processing in a second language

To those who argue frequency is the primary factor in differentiating processing, the distinction between literal (i.e. collocations) and nonliteral (i.e. idioms) formulaic expressions may be an arbitrary one. Specifically, if frequency is the only dimension along which these expressions vary in processing, then more exposure to input should mitigate any issues nonnative speakers have in acquiring and using formulaic strings appropriately. To native speakers, formulaic strings are pervasive in everyday language, as they are recognized as conventional, trademarks of what it means to sounds like a native speaker. It is this notion, though, that creates an obstacle: idioms, and other formulaic expressions, are culture specific (Wray, 2002). Alongside knowledge of the language, a speaker also needs cultural knowledge to inform herself which expressions are used to communicate which meanings when. To say frequency alone drives variation in processing is too simple. Instead, asking what modes of analysis (i.e. compositional, noncompositional) nonnative speakers utilize in processing would arguably provide more nuanced insight into second-language formulaic language processing, insights which would better inform an understanding of the status of these expressions in the lexicon—both native and nonnative.

In one school of thought, usage-based approaches highlight experience as the primary indicator of a learner's relative ease or difficulty in acquiring frequent, colloquial expressions in the L2. Simply put, a person's relative experience with the language and their exposure to colloquialisms are the primary means of accumulating the statistical information needed to successfully process and recognize formulaic language (Bod, 2006). Over time, a learner would eventually use the compiled memory representations from her experience and use this knowledge to inform which sequences or strings are appropriate when

and, additionally, which phrasal configurations are most expected in what scenarios to economize communication. Fluency is achieved when a learner successfully uses this statistical knowledge to understand meaning above the sentence level and to use formulaic strings in her own speech in culturally licensed environments.

Work on L2 formulaic language processing highlights acquisition of formulaic expressions as essential for near-native-like attainment and fluency (Cowie, 1998; Pawley & Syder, 1983; Sinclair, 1991; Tomasello, 2003; Wray, 2002). Prior research has underscored the importance not only of vocabulary acquisition in an L2, but also knowledge of how words fit together in an L2 (Wolter & Gyllstad, 2011; Wray, 2002). Although language learners undoubtedly understand which words co-occur in their first language (L1), this knowledge is not always helpful in a second language, since both collocations and idioms tend to be language-specific; this has been demonstrated by research reporting even proficient L2 learners struggle with collocations (Granger, 1998; Nesselhauf, 2005). As both types of expressions are highly predictable, L2 learners can learn both kinds of language like long words, likely experiencing them as chunks, which would directly support their mental representation of the regularity of certain patterns in the L2. Whereas experience supports the storage and retrieval of both idioms and collocations as chunks, L2 users encounter an obstacle arguably more so than native speakers, namely the inaccuracy of compositional analysis. Whereas the meaning of a collocation is derivable through the meaning of its component parts, nonliteral meaning resides above the sentence level, the key to whose interpretation resides in prior knowledge and exposure. If, however, these expressions are experienced roughly equally, then time with the input should mitigate any processing difficulties when comprehending nonliteral compared to literal language, which would suggest proficiency modulates the relative ease or difficulty with which nonliteral expressions can be comprehended in an L2. If, however, the mode of processing is key, specifically with compositional analysis as the default processing mode for L2 speakers, then this would suggest an advantage when reading collocations, as only literal language allows successful compositional analysis the majority of the time. In everyday language use, there are no overt cues to signal to a reader that an idiom or collocation has been encountered; recognition of a string as meaningful

only arises with sufficient frequency and experience, and there are theoretical arguments to be made for the role that co-occurrence information plays. Namely, others suggest that co-occurrence would be the dominating factor determining the ease with which language can be processed, including MWEs (e.g. Arnon & Snider, 2010, Jolsvai, McCauley, & Christiansen, 2013).

Whereas research shows an advantage for familiar nonliteral strings in processing for native speakers, the story becomes more complex when comparing native- and non-native speaker performance. In the latter case, some studies have found that nonnative speakers process idioms like novel language expressions (e.g. Siyanova-Chanturia et al., 2011; Underwood et al., 2004), interpreting the string as literal prior to a figurative one. Especially in the absence of a prior biasing context, these studies support the notion of compositional analysis first with little to no facilitation in processing due to a recognition point (Cieslicka, 2006; Matlock & Heredia, 2002). For example, Siyanova-Chanturia, Conklin, and Schmitt (2011) used eyetracking to investigate differences in native and nonnative processing of idioms. Participants read sentences featuring ambiguous idioms—used either figuratively or literally—and novel phrases. While native speakers showed faster reading times in conditions where idioms were used idiomatically, nonnative speakers did not show this advantage. In fact, nonnative speakers read idioms slower when used figuratively than when they were used literally. In a norming study conducted prior to the main task, participants completed a pre-test to show that the meanings of the idioms were familiar. Despite this, however, nonnative performance in the sentence processing task suggested that, despite prior experience with the targets, participants employed compositional analysis first.

Other studies, however, report similar processing behaviors across native and nonnative speakers when participants are presented with both literal and figurative uses of an idiom (Conklin & Schmitt, 2008; see Conklin & Schmitt, 2012 for a review). In their 2008 study, Conklin and Schmitt used a self-paced line-by-line reading paradigm to compare button push times across L1 and L2 groups when reading idioms embedded in longer passages. For both groups—native and proficient nonnative speakers—the authors found a facilitation effect for idioms over literal controls. A primary difference between this study and Siyanova-Chanturia, et al. (2011) is that stimuli in Conklin and Schmitt (2008) were longer, which

would be considered comparatively more informative environments than single sentences. It may be that, in the presence of a rich prior context, proficient nonnative speakers experience facilitation for idioms used figuratively much like native speakers do. Ultimately, however, the evidence is mixed, and more research is needed to determine what factors determine when processing is facilitated for idioms and when it is not, and how this manifests in the L1 and the L2.

Additionally, it remains unclear whether L2 speakers use context to anticipate a stimulus as it unfolds, and whether they predict at all. Some studies have reported that L2 users show reduced effects of lexical prediction compared to native speakers (e.g. Grüter et al., 2012; Lew-Williams & Fernald, 2010; Martin et al., 2013), though one recent report now shows ERP evidence of prediction in the form of anticipatory N400s and late frontal positivity (LFP) effects in highly constraining sentences (e.g. Foucart, Martin, Moreno, & Costa, 2014; see Kaan, 2014, for discussion). In Foucart, et al., (2014), Spanish native speakers, Spanish-Catalan early bilinguals—both as control groups—as well as French-Spanish late bilinguals were recruited to read highly constraining sentences ending in an NP either supported by the prior context or an unexpected NP; expected and unexpected noun targets were frequency-matched. The study found evidence of anticipation in all three groups, including the French-Spanish late bilingual group, which the authors interpret as evidence of linguistic anticipation in L2 speakers. They add that this effect may be modulated by the linguistic similarity between French and Spanish, citing a large lexical overlap between the languages as perhaps supporting these anticipatory processes.

All of these studies, including Foucart et al. (2014), focus on semantic constraints in literal sentences or on the use of morphosyntax as a predictive cue. Given strong effects for construction-based L2 processing (e.g. Ellis, 2012; Ellis et al., 2014), it is quite possible that lexical co-occurrence frequency may facilitate predictive processing in second-language populations. It remains to be seen, however, how semantic opacity interacts with co-occurrence-based predictions in non-native processing, and this is one of the key research questions of the current proposal.

1.7 Research questions

This dissertation is focused on three research questions:

1) When controlling for lexical and phrase frequency, when does information about local co-occurrence information become available in processing?
2) What role does semantic opacity play in discriminating between processing of different types of chunk-like expressions?
3) How do top-down and bottom-up sources of information interact in L2 reading of formulaic expressions
    a. When does information about co-occurrence become available in L2 reading?

1.8 Description of the experiments

1.8.1 Experiment 1

To answer Research Questions 1 and 2, Experiment 1 was designed to examine reading of idioms and literals in sentence contexts. By controlling for whole-string and substring variables we know affect processing—i.e. whole-string and substring frequency, length—we ensure that both idioms and literal expressions exhibit a comparable degree of local co-occurrence probability, such that native English speakers would recognize both types as meaningful, chunks of words that often appear together. By doing this, we are able to isolate semantic opacity as a potentially influential factor, allowing us to investigate how a phrase's relative semantic opacity or transparency impacts language comprehension. To investigate how bottom-up and top-down information interacts during processing, I incorporated a letter manipulation paradigm, as researchers have done previously using the transposed-letter effect. I did this to see how impoverished visual information affects recognition of phrases that can be characterized as chunks. Further, classic idiom models suggest there is something critical about an idiom's initial word that is needed to activate the idiomatic meaning. For this reason, the letter manipulation was incorporated

in two places in the phrases: In Experiment 1a and 2a, the letter manipulation occurred in the phrase-final word, and in Experiment 1b and 2b, the manipulation appeared in phrase-initial position. If there is something significant about an idiom's initial word that acts as a gateway to the nonliteral meaning, then manipulating the visual information about the first word should lead to processing difficulty. Also, if there is something unique about the configuration of a literal expression, we should see this manifest in a penalty for literals, as well, when encountering manipulated letter order and identity in the first word of the literal expression.

1.8.2 Experiment 2

To answer Research Questions 2 and 3, Experiment 2 was designed to test how information about semantic opacity and local co-occurrence affects processing in L2 reading. Namely, if there is something significant about idioms that requires prior experience in order to recognize a configuration as meaningful, then recruiting L2 learners for this task should provide the locus needed for this. As this group will have less overall exposure to these forms in the input, testing the same sentences from Experiment 1 in this group will provide an important comparison of how exposure to the input modulates the recognition of these configurations. Namely, if, as other studies have shown (i.e. Siyanova-Chanturia et al. 2011), L2 speakers employ compositional analysis first before entertaining other possible figurative interpretations, this should also manifest in natural reading, where we expect idioms to require more reading time compared to literal expressions. Further, when controlling stimuli for whole-string and substring frequency information, frequency information was obtained from English corpora, knowing full well that these types of resources are meant to represent frequency within a language. As a group of L2 speakers will have qualitatively different exposure to these forms than native English speakers, comparisons between L1 and L2 groups' reading behavior of the same stimuli may provide valuable insight into how knowledge of local co-occurrence information manifests in processing differences between idioms and literals. Finally, by incorporating the same letter manipulation paradigm employed in Experiment 1 in Experiment 2, we can study how L2 comprehension is impacted by the quality of

bottom-up compared to top-down information, something that would provide nuance to the notion that L2 speakers use compositional analysis. Namely, perhaps this strategy is modulated by the quality of the input available, where a compositional approach may lead to greater reliance on bottom-up, visual cues; Experiment 2 has been designed to test this.

1.8.3 Experiment 3

Finally, Experiment 3 was designed to help to answer Research Questions 1 and 2. While Experiment 1 is designed to investigate how idioms and literals are processed when embedded in contexts, this does not tap into how peoples' perceptions of relative plausibility or meaningfulness are affected by semantic opacity. Experiment 3 is designed to understand how these types of MWEs are processed without this supportive contextual environment. Further, this experiment is designed to study how semantic opacity differentiates recognition and processing when these expressions occur in isolation. Again, by controlling for whole-string and substring variables, we can isolate semantic opacity and examine its effect on initial recognition and lexical access of chunk-like meanings. Presenting these expressions in isolation has the ability to be important and insightful with respect to questions of how idioms are represented in the lexicon—unitarily or compositionally—isolated presentation will better help us to answer that.

1.9 Hypotheses

1.9.1 Experiment 1

As prior work using eyetracking shows, we should see greatest disruption to natural reading when the target word contains a letter substitution compared to a letter transposition, and both should be more disruptive to natural reading compared to Identity targets. Prior work shows that the visual system is sensitive to the degree of stimulus degradation, such that unexpected letters should inhibit lexical access and integration more than if the expected letters in the string appeared in a different order. Further, while masked priming work shows letter transpositions in primes facilitate lexical access nearly the same as Identity primes, eyetracking results would predict that, when in direct fixation, transpositions should

demonstrate a processing burden compared to Identity targets. Namely, if target words with transpositions are equally as facilitative in lexical access as Identity targets—as masked priming work would indicate—and semantic opacity distinguishes reading behavior, then we should see more skipping of phrase-final words when the word either appears as expected (Identity) or contains a transposition (TL). Consistent with prior work, targets with substitutions should be more disruptive to processing in any case, as these targets contain incorrect information rather than just letters in the wrong order.

If semantic opacity distinguishes processing, then we would expect less reliance on bottom-up foveal information when reading idioms, as the knowledge that one is reading an idiom should lessen the need for bottom-up information. If this is the case, this should lead to skipping of upcoming targets that appear as expected or targets that deviate minimally from what is expected (i.e. TLs). Targets containing substitutions should be skipped less, as although the substituted letters retain features of the expected letters, they are still not complete matches to how the word should appear. While ascenders and descenders may be retained, substitutions ultimately contain unexpected characters entirely, and this should decrease the likelihood of skipping. Further, if semantic opacity differentiates processing of formulaic chunks, there should be less disruption to natural reading of idioms when letter transpositions occur in the phrase-final word compared to literal collocations. When in direct fixation, we would expect greater disruption to reading from targets containing transpositions compared to Identity targets and even more disruption when the target contains a substitution. For literal targets, we would expect fewer skips overall. This would suggest a greater influence of bottom-up information in processing the phrase-final word of a literal collocation compared to the phrase-final word of an idiom. Namely, at the end of an idiom, people will have access to most of the phrase's information, arguably enough to judge that what they are reading is formulaic and semantically opaque, an expression that requires a specific completion. For literals, while there is a likely completion, potentially, there are more options for literals, which would suggest bottom-up information should be needed in literal collocation processing up until the final word. Further, in literal targets, we would also expect transpositions to be less disruptive than substitutions but

more disruptive than the expected target. The integrity of the visual cues should still make an impact regardless of the type of expression being processed.

However, if semantic opacity does not distinguish comprehension of idioms and literals, we should see comparable rates of skipping across conditions. Specifically, skipping in this case would be influenced by the quality of information available in the parafovea only. By controlling for a variety of factors in the stimuli, the only factor left to distinguish idiom and literal trials is the degree of semantic opacity of the target expression. With respect to skipping, we should see more skipping of the phrase-final word when the target is an Identity target, slightly less skipping for TL targets, and the least skipping for SUB targets, suggesting that the degree of mismatch in the parafovea should influence the planning of upcoming eye movements. In direct fixation, we would also expect that transpositions lead to longer reading times compared to Identity targets, but not as long as targets with substitutions, again, highlighting the degree of degradation as influential in the relative ease or difficulty of processing. Such patterns would highlight both idioms and literals as expressions that are commonly seen as chunks, and that as long as a person recognizes the configuration as meaningful—regardless of semantic opacity—this should facilitate economization of processing resources at the phrase-final word.

In Experiment 1b, when the letter manipulation is in phrase-initial position, there should be longer reading times overall for both idioms and literals. For both types of expression, they represent a configuration. By impoverishing the visual information at the first location where participants could start to represent a configuration, we should see longer reading times compared to Experiment 1a. However, if semantic opacity distinguishes idioms from literals, we should see longer overall reading times for idioms. If, as Cacciari and Tabossi would say, the first content word in an idiom is part of a gateway to the nonliteral meaning, manipulating the orthography in the first word should diminish any advantage idioms have. In such a case, a literal string might be read faster, as compositional analysis might facilitate filling in the gap, so to speak, of the impoverished first word. If semantic opacity does not differentiate processing of these expressions, we should see a comparable disadvantage for both types of expressions, as the first word in the configuration in either case would require looking to other cues to resolve the

representation. This would support idioms as being part of a larger class of expressions that frequently co-occur and would be evidence against there being anything special about idioms in processing.

1.9.2 Experiment 2

With respect to second-language processing, if compositional analysis is the default processing route, we should see a penalty for idiomatic expressions. As compositional analysis is all that is required for literal language comprehension, literal collocations should be easier for L2 speakers to comprehend compared to idioms, which we know require an additional computation above the phrasal level. Further, a compositional-first route might prioritize bottom-up cues over top-down information (i.e. prior knowledge). If this is the case, TL and SUB targets should both be disruptive to bilinguals in reading, as both forms will ultimately contain misspellings and be novel forms to these participants. However, if cross-script bilinguals flexibly encode letter position information, we might see an advantage for TL targets over SUB targets in processing. This would suggest an influence of prior exposure to the input on processing, where participants could use TL targets to better facilitate lexical access than they could SUB targets. Further, if semantic opacity distinguishes idioms and literals for bilinguals, we would likely see the advantage for literals. With presumably less exposure to the input, these participants may have less familiarity with identifying idioms as nonliteral, and less experience with assigning the nonliteral meaning to the configuration. With literals, this is not necessary, as compositional analysis, alone, will provide the meaning of the phrase.

When the letter manipulation appears in the phrase-initial word, if L2 learners are sensitive to phrase-level, co-occurrence information, we would expect longer reading times in Experiment 2b compared to 2a. Namely, if the knowledge that a person is reading a meaningful expression affects reading behavior, we might see faster reading times in Experiment 2a, where a person can anticipate more what might be upcoming, even as broadly as part of speech information. This would support a role for anticipation in L2 processing, namely the understanding that a particular kind of word should be coming

soon in a stimulus. If, however, phrasal knowledge does not impact sentence reading, then we should see comparable reading times across Experiments 2a and 2b. In both experiments, we should see slow-downs at the target word when it contains any kind of letter manipulation. If phrasal knowledge does not impact reading, then the position of the manipulation should not lead to differences in reading times.

1.9.3 Experiment 3

Finally, Experiment 3 was designed to help to answer Research Questions 1 and 2. While Experiment 1 is designed to investigate how idioms and literals are processed when embedded in contexts, this does not tap into how peoples' perceptions of relative plausibility or meaningfulness are affected by semantic opacity. Experiment 3 is designed to understand how these types of MWEs are processed without this supportive contextual environment. Further, this experiment is designed to study how semantic opacity differentiates recognition and processing when these expressions occur in isolation. Again, by controlling for whole-string and substring variables, we can isolate semantic opacity and examine its effect on initial recognition and lexical access of chunk-like meanings. Presenting these expressions in isolation has the ability to be important and insightful with respect to questions of how idioms are represented in the lexicon—unitarily or compositionally—isolated presentation will better help us to answer that.

2. Experiments

2.1 Experiment 1a

2.1.1 Method

Participants

Sixty-three monolingual speakers of American-English were recruited for participation in Experiment 1a; prior to analysis, data from three participants were excluded due to being exposed to another language in the home before age 6 (some of the participants were recruited from classes for course credit). Data from 60 participants were included in Experiment 1a (range: 18-29 years old; mean age=21 years; 45 female). All participants were either students at the University of Illinois or members of the surrounding Champaign-Urbana community. Participants reported having normal or corrected-to-normal vision and no history of dyslexia or developmental reading disorders. All participants completed a handedness questionnaire, and all reported being right-handed. All participants were compensated with cash for their time.

Materials

60 idioms and sixty literal collocations were collected for use in stimuli in a 2 (Phrase type) × 3 (Letter) design (see Table 1 for example stimuli; see Appendix A for stimuli used in Experiments 1 and 2). Idioms were selected from Bulkes & Tanner (2017), and only those that had a minimum average rating of 3.5 for familiarity (Likert scale of 1-5, 1=low; 5=high) were chosen for inclusion in this study. Sentences contained either an idiom or a literal collocation following a preamble where the context was felicitous with the target expression (i.e. a figuratively biasing context for the idiom, or a literal context for the collocation; see Table 1, where target expression is underlined). The first and last words of the idiom or literal expression were always content words, with any verbs being lexical verbs as opposed to auxiliary verbs. In Experiment 1a, the target word was the last word in the idiom or literal collocation; in Experiment 1b, the target word was the first word (underlined and italicized in Table 1).

Table 1. Example stimuli in a 2 (Phrase type) × 3 (Letter) design (target expression underlined, word italicized)

Experiment 1a	
Idiom	After Alyssa's success, no one could rain on her *parade / pardae / parebe* and upset her.
Literal	To catch his flight, Trent had to leave for the *airport / airprot / ainqort* to avoid being late.

All idioms and literal collocations were matched across lists for content word frequency, frequency of the literal or idiomatic expression, length of the target, and part of speech (i.e. verb + function word + noun) across conditions. Frequency information was obtained from the Corpus of Contemporary American English (Davies, 2008). Cloze probability of the target following the sentence preamble was assessed during norming prior to the study, where a separate set of participants (n=60) on Amazon Mechanical Turk read the preambles and were asked to supply the most likely completion in a fill-in-the-blank task. Cloze probability up to the target word was then matched across idiom and literal lists (see Table 2 for information on the average constraint in the sentences used in Experiments 1 and 2, and see Table 3 for t-tests on item parameters, such as whole-string and substring frequencies).

Table 2. Constraint of stimuli used in Experiments 1 and 2, up to target word when used in phrase-final position

	Idioms	Literals
Average constraint	0.83	0.80
Range	0.26-1.00	0.06-1.00
Std. Deviation	0.16	0.21

Table 3. Paired t-tests for stimuli parameters

Comparison	t	df	p-value
Cloze	1.06	112	.29
Phrase length	0.28	124	.78
Phrase frequency (COCA)	0.05	124	.96
First word length	0.69	111	.49
First word frequency (COCA)	-1.11	115	.27
First word frequency (SUBTLEX)	-1.36	114	.18
Last word length	-1.57	108	.12
Last word frequency (COCA)	-1.38	124	.17
Last word frequency (SUBTLEX)	-0.55	112	.58

N.B. Cloze probabilities determined in norming prior to the study, see Appendix A for cloze probabilities of stimuli

In the main study, participants saw one of six experimental lists, where they saw one version of each item: the sentence with the target word as expected (Identity condition), the target with two internal characters

transposed (Transposition—or TL—condition), or the target with two internal characters substituted (Substitution—or SUB—condition). Transposed characters were always word-internal. Substituted characters always retained visual similarity to the letters they were replacing (i.e. retaining ascenders or descenders, letter shape). Sentences were distributed across 3 experimental lists using a Latin Square design, such that participants saw 20 items per condition. The three lists were then presented in the reverse order as 3 new lists for a total of 6 lists.

Procedure

At the start of the experimental session, participants provided informed consent and completed a language background and handedness questionnaire. Participants were then seated comfortably at a table in front of a desk-mounted SR-Research, Ltd. EyeLink 1000 eyetracker and a computer screen, where they would complete the sentence processing task. Each session began with a practice block of 10 sentences, where participants were guided through the instructions and sample items to get them comfortable with the procedure. After the practice, participants saw one version of six experimental lists (three main lists and three additional lists with the trial order reversed). Each list contained 120 experimental items—60 idiomatic sentences and 60 literal sentences—and 120 filler sentences, which included a subset of garden-path sentences adapted from Christianson, Hollingworth, Halliwell, & Ferreira (2001), as well as sentences containing fake idioms (i.e. *She married the bench under the barn*). Participants were asked to respond to comprehension questions ("Yes/No") following 1/3 of all sentences, and these questions only followed filler items.

Prior to each experimental block, camera accuracy was assessed using a 9-point calibration (acceptability threshold: ≤0.8 degrees). Calibrations were repeated as necessary throughout the experiment. Participants self-paced through the experiment, triggering the onset of each sentence themselves by fixating a dot on the left-hand side of the screen and pressing the space bar. Fixating this dot served as a calibration check before each trial, where the trial would only began if the camera detected the person's eye within the threshold of less than, or equal to, 0.8 degrees of the center of the point. To

minimized movements during the experiment, participants anchored their heads using the chin rest, creating a viewing distance of 98cm. Sentences were presented on a ViewSonic VX2268WM computer screen using fixed-width (Courier New) font, and there were approximately 5 characters per degree of visual angle. Sentences all fit on one vertically-centered line on the monitor and were presented in off-white, "chalk" font on a black background. Data were sampled at 1000 Hz. The experiment was broken into six blocks of 40 sentences each, where the participant was free to take a break between the blocks and allowed to move about freely and stretch. After the sentence processing task, participants completed a lexical decision task, 50 fill-in-the-blank questions taken from the Michigan English Language Institute College English Test (MELICET) and the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007) as language proficiency measures; an operation span task and a letter-number sequencing task as measures of working memory; the author recognition task (Acheson et al., 2008) as a measure of print exposure; and the Nelson-Denny reading test as a measure of reading speed. At the end of the session, participants provided subjective familiarity ratings for all of the idioms they saw in the sentence processing task.

Data Processing & Analysis

Data from both Experiments 1 and 2 were handled in the same way. Fixations less than 80ms in duration were merged with nearby fixations that were either within one character prior or after the fixation in question (1.6% of fixations). Further, any trials with track loss on the target word were eliminated from further analysis (4.1% of trials). Outlier fixations were moved to the closest interest area. After this, fixations less than 80ms or greater than 800ms were deleted. For trials where the region of interest was skipped on the first pass, but fixated during second- or later-passes through the sentence, these were treated as zeroes for first-pass measures (e.g., first pass skipping, first fixation duration, gaze duration; see below).

For both Experiments 1 and 2, I analyzed two main regions in the data: the target word—or, the phrase-final word—and the target expression, and four eye-movement measures were used in analysis. Four eye movement measures (Rayner, Pollatsek, Ashby, & Clifton, 2012) were examined. For analyses

of the target word, *first pass proportion skipped* was calculated, or the rate with which the target word was skipped on the first pass through the region. This measure was chosen as an early measure of graded prediction and anticipation, where the more a person is anticipating the next lexical item to come, the more they should be inclined to skip the word. For the region containing the target word, I measured *first fixation duration* (the total time spent during the initial first-pass fixation on the target word) and *total duration* (total time spent reading in a region, including time spent reading after re-entering the region from either the right or the left). I chose first fixation duration as an index of prediction and anticipation of an expected stimulus, where the more expected a target is—particularly for shorter words—the shorter the first fixation duration measure should be (e.g. Rayner & Duffy, 1986), as less resources must be allocated to its processing early on. Further, *total duration* was used, as it is an index of the relative ease or difficulty with which the information in the region can be integrated and incorporated into the representation of the surrounding text (Liversedge, Paterson, & Pickering, 1998). For the region containing the whole expression, I measured *first pass time* (the total of all fixation time on the first pass through the region before the eyes exited either to the right or to the left), and again total duration. First pass time was selected as a measure of lexical access as well as an index of the initial relative ease or difficulty of processing the expression as a whole. Total duration was also used for this region.

I first analyzed the data using an omnibus linear mixed effects model (Baayen, Davidson, & Bates, 2008), using the lme4 package (version 1.1-12; Bates, Maechler, Bolker, & Walker, 2015) in R (version 3.3.1, R Core Team, 2016). I included sum-coded main effects of Phrase Type (idioms v. literals) and Letter (Identity, TL, SUB). Random intercepts for participants and items were included in the model, and both of the experimental factors, Letter and Phrase Type, were included in the model as fixed effects (Barr, Levy, Scheepers, & Tily, 2013). For the skipping data, I performed a logistic mixed effects regression, with subjects and items as random effects, and the factors of experiment, Letter and Phrase Type, were entered into the model as fixed effects (Jaeger, 2008). Separate models were fitted for each eye movement measurement within each experiment. Fits for all models were constructed using the *mixed* function from the *afex* package in R (Singman, Bolker, Westfall, & Aust, 2017). For all models, the

likelihood ratio test (LRT) method was used for calculating p-values, a method which produces a chi-square statistic instead of an F statistic. This method is appropriate when there are more than 50 observations per subject (*mixed* documentation; Singman et al, 2017). Pairwise comparisons were made for significant main effects and interactions using the *lsmeans* package in R (Lenth, 2016). For pairwise comparisons, the Tukey method was used for p-value adjustment due to multiple comparisons.

2.1.2 Results

Target word measures

Means per condition are available in Table 4 and show that Identity targets were skipped more, on average, than TL targets, which were skipped more than SUB targets. The model output, available in Table 5, shows a significant main effect of Letter but no main effect of Phrase Type. Skipping rates were comparable to skipping rates found in other studies (i.e. Rayner & Well, 1996). This suggests that the amount of target-word skipping was predicted by the quality of information available in the parafovea, and not by the type of phrase people were reading.

Table 4: Proportion skipped for the region containing the phrase-final target word

	Identity	TL	SUB
Idiom	0.22 (0.01)	0.21 (0.01)	0.15 (0.01)
Literal	0.22 (0.01)	0.19 (0.01)	0.13 (0.01)

Values are averages expressed as a probability. Standard error of the mean indicated in parentheses.

Table 5. Native model for proportion skipped data from Experiment 1a

Effect	df	X^2	p-value
Letter	2	55.63	<.0001
Phrase Type	1	2.45	.12
Letter×Phrase Type	2	1.00	.61

Output constructed using mixed function with "afex" package in R.

Contrast	β	Std. Error	z	p-value
Identity – SUB	-0.59	0.08	-7.36	<.001
Identity – TL	-0.19	0.07	-2.50	<.05
SUB – TL	0.40	0.08	4.92	<.001

Output constructed using "lsmeans" package in R.

Pairwise comparisons showed that Identity targets were skipped more than SUB targets and TL targets, and SUB targets were skipped less than TL targets. In this study, it is possible that knowledge of lexical co-occurrence combined with expected information in the parafovea is what led to the differences in skipping rates. For example, given that the target manipulation was in the last word of the collocation—both idiomatic and literal—participants had access to the beginning and middle of each chunk. With this information, participants could anticipate a particular phrasal outcome, and speakers may have used the information in the parafovea to plan upcoming eye movements with respect to this. Further, we see evidence that the degree of visual degradation is what drives the effect of Letter; namely, when only letter position information is manipulated, there is a difference between Identity and TL targets. However, when both letter position and identity information are manipulated, the difference between TL and SUB targets is greater than the Identity-TL difference. This data are an example of the visual system's sensitivity to what, and how much, information is in place and what information is anomalous in a string.

Means and standard errors for first fixation duration are presented in Table 6. Identity targets, in general, had shorter first fixation durations than TL targets, which also had shorter first fixation durations than SUB targets. The model fits can be found in Table 7, which show a significant main effect of Letter, no main effect of Phrase Type, and a significant interaction. Pairwise comparisons indicated that Identity targets had shorter first fixation durations than SUB targets and TL targets, and SUB targets had longer first fixation durations than TL targets (see Figure 1 for visualization). There was also a Letter × Phrase Type interaction, where Identity targets had longer first fixation durations when the target word was in an idiom, but the reverse pattern was true for TL and SUB targets.

Table 6. Reading measures from English native speakers on phrase-final word in Experiment 1a

First fixation duration			
	Identity	TL	SUB
Idiom	206.73 (2.79)	212.20 (3.56)	244.87 (4.15)
Literal	195.77 (2.85)	218.64 (3.29)	250.32 (3.92)
Total duration			
	Identity	TL	SUB
Idiom	282.45 (5.29)	375.85 (8.46)	620.77 (13.27)
Literal	282.45 (5.68)	381.57 (7.88)	578.78 (13.50)

Values indicate mean durations in milliseconds. Standard error of the mean indicated in parentheses.

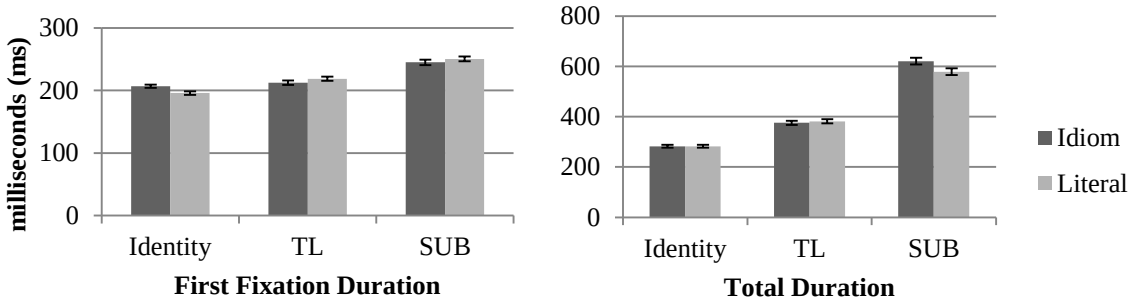

Figure 1. First fixation (left) and total duration (right) measures for Experiment 1a. Error bars reflect the standard error of each mean represented.

Table 7. Model for first fixation duration on the target word in Experiment 1a

Effect	df	X^2	p-value
Letter	2	203.12	<.001
Phrase Type	1	0.02	.88
Letter×Phrase Type	2	7.92	<.05

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-46.59	3.34	6284	-13.93	<.001
Identity – TL	-14.28	3.38	6283	-4.22	<.001
SUB – TL	32.31	3.30	6279	9.79	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	10.34	5.73	447	1.81	.46
TL, Idiom – TL, Literal	-7.08	5.62	417	-1.26	.81
SUB, Idiom – SUB, Literal	-5.12	5.53	391	-0.93	.94

Output constructed using "lsmeans" package in R.

Means and standard errors for total duration can be found in Table 6. Generally, Identity targets had shorter overall reading times, and TL targets had shorter total durations than SUB targets. Further, while

the difference between idioms and literals is smaller in Identity and TL conditions, idioms appear to incur slightly longer reading times than literals in the SUB condition. The model outputs can be found in Table 8 and show no effect of Phrase Type, but a significant main effect of Letter, as well as a Letter × Phrase Type interaction (Figure 1 for visualization). Pairwise comparisons revealed that Identity targets were read for less time overall than SUB targets and TL targets, and SUB targets took more overall time to read than TL targets. These results, taken together with those from first fixation duration, indicate that, when reading in a collocation—idiom or literal—any kind of deviation from an expected target incurs a processing burden. In the case of TLs, however, where letter identity information is retained, these targets more closely resemble the expected target. For SUB targets, manipulating both letter position and identity leads to greater processing difficulty, both in early, initial lexical access as well as in downstream, integrative processing. Further, this burden seems to be emphasized by the type of phrase. The interaction suggests that while Phrase Type was not a significant predictor of total reading times, idioms incurred a slight penalty in total time when the phrase-final word contained a substitution, even though the pairwise difference between SUB targets in idioms and SUB targets in literals was not significant after adjusting *p*-values for multiple comparisons (β =42.54, t=2.39, p=.16). The more impoverished the target word—comparing TLs to SUBs—the greater difficulty becomes to integrate the target into the expression, and this is particularly the case within an idiom. So far, these results suggest that the visual system is sensitive to the degree of impoverishment of a visual stimulus, and that semantic opacity may lead to stronger predictions about how the phrase-final word should appear and that this difference manifests in later, integrative stages of processing.

Table 8. Model for total duration on the target word

Effect	df	X^2	p-value
Letter	2	1126.64	<.001
Phrase Type	1	0.67	.41
Letter×Phrase Type	2	8.42	.01

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-316.55	9.29	6376	-34.07	<.001
Identity – TL	-96.08	9.40	6375	-10.22	<.001
SUB – TL	220.47	9.21	6374	23.95	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	-0.99	18.19	297	-0.05	1.00
TL, Idiom – TL, Literal	-5.94	18.02	287	-0.33	1.00
SUB, Idiom – SUB, Literal	42.54	17.79	273	2.39	.16

Output constructed using "lsmeans" package in R.

Whole phrase measures

Means and standard errors for first pass time can be found in Table 9. Phrases with Identity targets had shorter first pass times than phrases with TL targets, which in turn had shorter first pass times than SUB targets. Further, there seems to be a slight difference between idioms and literals throughout, where idioms seem to have slightly longer first pass times than literals. The model outputs are shown in Table 10, however, and show a significant main effect of Letter, but no effect of Phrase type, and no interaction (see Figure 2 for visualization). Pairwise comparisons among the levels of the factor Letter revealed that phrases with Identity targets were read for less time on the first pass through the region compared to TL targets and SUB targets. Phrases with SUB targets had longer first pass times than phrases with TL targets. However, the lack of an effect of or interaction with Phrase type indicates that, although there was a numerical reading time advantage for literal targets, this was not significant enough to reach statistical significance.

Table 9. Reading measures for the whole phrase in Experiment 1a

First pass time			
	Identity	TL	SUB
Idiom	616.03 (8.09)	648.90 (8.92)	771.89 (12.02)
Literal	579.04 (8.24)	638.93 (8.94)	740.64 (11.37)
Total duration			
	Identity	TL	SUB
Idiom	815 (13.15)	899.67 (14.19)	1213.66 (20.15)
Literal	787.55 (13.99)	885.95 (13.79)	1099.41 (18.91)

Values indicate mean durations in milliseconds. Standard error of the mean indicated in parentheses.

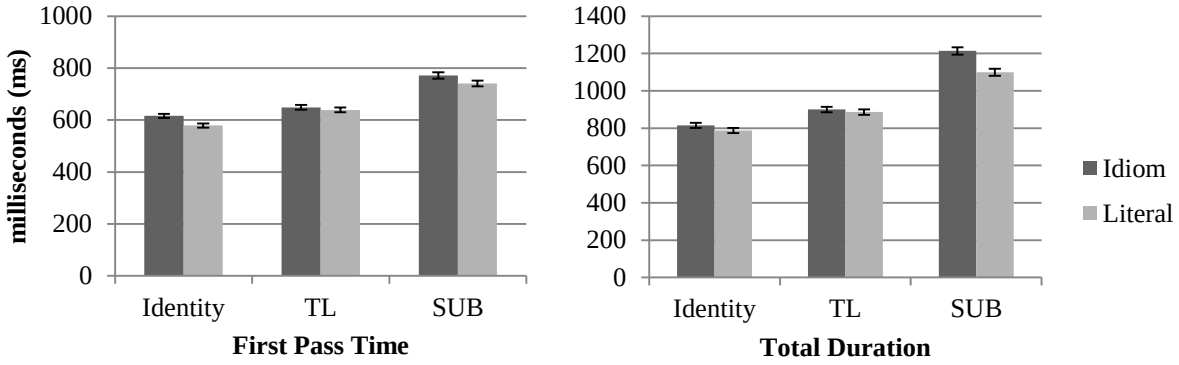

Figure 2. First pass reading time (left) and total reading time (right) for the whole expression. Error bars indicate standard error of each mean represented.

Table 10. Model for first pass reading time on the whole phrase from Experiment 1a

Effect	df	X^2	p-value
Letter	2	362.43	<.001
Phrase Type	1	1.47	.22
Letter×Phrase Type	2	2.84	.24

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Identity – SUB	-158.71	8.46	6964	-18.76	<.001
Identity – TL	-46.51	8.46	6964	-5.50	<.001
SUB – TL	112.20	8.46	6964	13.26	<.001

Output constructed using "lsmeans" package in R.

Means and standard errors for total duration measures can be found in Table 9. The means suggest that phrases with Identity targets had shorter total durations than phrases with TL targets, and that phrases with TL targets had shorter total durations than phrases with SUB targets. Further, there appears to be a slight advantage for literals, and this difference looks the largest in the SUB condition. The model fits are available in Table 11 and show a significant main effect of Letter, no significant main effect of Phrase type, and a Letter × Phrase Type interaction (see Figure 2 for visualization). Pairwise comparisons

among the levels of Letter showed that phrases with Identity targets were read for overall less time than phrases with TL targets and phrases with SUB targets, and phrases with SUB targets took longer to read than phrases with TL targets. The interaction suggests a similar pattern to what was seen in measures of reading on the target word: When the target contains a substitution, this manipulation incurs the largest penalty when appearing in an idiom.

Table 11. Native model for total duration on the whole phrase in Experiment 1a

Effect	df	X^2	p-value
Letter	2	653.27	<.0001
Phrase Type	1	2.46	.12
Letter×Phrase Type	2	14.63	.01

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-355.59	14.11	6918	-25.20	<.001
Identity – TL	-91.82	14.11	6918	-6.51	<.001
SUB – TL	263.77	14.11	6918	18.70	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	27.47	36.80	185	0.75	.98
TL, Idiom – TL, Literal	14.55	36.79	185	0.40	.99
SUB, Idiom – SUB, Literal	113.85	36.79	185	3.09	<.05

Output constructed using "lsmeans" package in R.

While this pattern first seems to appear in first pass time measures, the significant interaction in total duration suggests that this difference becomes more apparent in later processing. This is consistent with prior work that shows that idioms incur longer overall reading times in measures of late processing due to the additional effort that comes with integration (i.e. Titone & Libben, 2014). These results suggest that while idioms and literals do not differ significantly when the input is expected, idioms are penalized more in cases of degraded input and that this manifests mostly in later processing.

Experiment 1a summary

Results from Experiment 1a suggest that, for native English speakers, idioms and literal collocations do not differ substantially from one another in measures of early processing or prediction. The overall lack of an effect of Phrase Type suggests that, once the factors we know to affect prediction

are controlled for, these expressions both embody how readers formulate expectations for how chunk-like expressions should be completed, and how these expectations are aided by parafoveal preview during natural reading. Namely, prior experience with the language does not seem to discriminate based on the degree of semantic opacity in early processing. Differences between the two phrase types only seem to manifest in later processing and in cases where the visual input is particularly degraded. While targets containing transpositions did not enjoy the degree of facilitation in lexical access that expected targets did, they were more facilitative than targets containing substitutions. While both are technically misspellings, manipulating both the letter position and identity in SUB targets incurred greater reading difficulty across the board, demonstrating how the visual system seems to be sensitive to how anomalous an experienced target is to what was anticipated given the global context of the sentence and the local context of the chunk. This is evident in both target word and phrase measures, suggesting that this difference manifests in late processing. Further, while both measures show this difference, it only becomes apparent in the SUB condition, suggesting that more degradation is needed to see this effect.

For models of idiom comprehension, this suggests that idioms are harder to process than literals for native speakers when the visual input is largely degraded. Minor anomalies in spelling do not elicit this effect, but when more visual anomalies are apparent, idioms require more time to resolve the ambiguity and integrate the target within the surrounding context compared to literals. For models of prediction, these results do not support either idioms or literals as more predictive environments than the other. Particularly as evidenced by the skipping data, the type of phrase is not a significant predictor in whether people predict more or less phrase-final completions. However, when the bottom-up input is degraded, in both types of phrases, processing slows down, evident in measures for both the region containing the word and the region containing the phrase, suggesting that the relative quality of the information being fixated impacts relative ease or difficulty in reading. For idioms, this difference is more pronounced when the target contains substituted letters rather than just transposed letters, suggesting that integrative mechanisms are more heavily impacted in idiom comprehension when processing unusual visual cues.

2.2 Experiment 1b

2.2.1 Method

Participants

Thirty-six monolingual speakers of American-English were recruited for participation in Experiment 1b (range: 18-25 years old; mean age= 20.36 years; 23 female). All participants were either students at the University of Illinois or members of the surrounding Champaign-Urbana community. Participants reported having normal or corrected-to-normal vision and no history of dyslexia or developmental reading disorders. All participants completed a handedness questionnaire, and all reported being right-handed. All participants were compensated with cash for their time.

Materials

The idioms and literal collocations used in Experiment 1b were the same as were used in Experiment 1a. The same 2 (Phrase type) × 3 (Letter) design was used; the only change in Experiment 1b was that the letter manipulation (Identity v. TL v. SUB target) was in phrase-initial position (see Table 12 for example stimuli; see Appendix A for stimuli). Sentences still contained either an idiom or a literal collocation following a preamble where the context was felicitous with the target expression.

Table 12. Example stimuli from Experiment 1b (target expression underlined, word italicized)

Experiment 1b, 2b	
Idiom	After Alyssa's success, no one could *rain/rian/rejn* on her parade and upset her.
Literal	To catch his flight, Trent had to *leave/laeve/lcawe* for the airport to avoid being late.

In the main study, participants saw one of six experimental lists, where they saw one version of each item: the sentence with the target word as expected (Identity condition), the target with two internal characters transposed (Transposition—or TL—condition), or the target with two internal characters substituted (Substitution—or SUB—condition). Transposed characters were always word-internal. Substituted characters always retained visual similarity to the letters they were replacing (i.e. retaining ascenders or descenders, letter shape). Sentences were distributed across 3 experimental lists using a Latin Square

design, such that participants saw 20 items per condition. The three lists were then presented in the reverse order as 3 new lists for a total of 6 lists.

Procedure

The procedure in Experiment 1b was identical to the procedure in Experiment 1a.

Data Processing & Analysis

Data from Experiment 1b were processed exactly in the exact same way as in Experiment 1a. For Experiment 1b, I analyzed two main regions in the data: the target word—or, the phrase-initial word—and the target expression, and the same four eye-movement measures that were used in the analysis of Experiment 1a were analyzed for Experiment 1b: first pass proportion skipped, first fixation duration, and total duration for the phrase-initial word; and first pass time and total duration for the whole expression.

2.2.2 Results

Target word measures

Means per condition, available in Table 13, show that Identity targets were skipped the most compared to both TL and SUB targets, and also show no difference between skipping ratings of TL and SUB targets. The model output, available in Table 14, shows no significant main effect of either Letter or Phrase Type.

Table 13: Proportion skipped for the region containing the phrase-final target word

	Identity	TL	SUB
Idiom	0.23 (0.02)	0.19 (0.01)	0.19 (0.01)
Literal	0.21 (0.02)	0.22 (0.02)	0.22 (0.02)

Values are averages expressed as a probability. Standard error of the mean indicated in parentheses.

Table 14. Native model for proportion skipped data from Experiment 1b

Effect	df	X^2	p-value
Letter	2	1.21	.55
Phrase Type	1	1.30	.25
Letter×Phrase Type	2	2.77	.25

Output constructed using mixed function with "afex" package in R.

The only difference between Experiment 1a and Experiment 1b was the position of the target word. Comparing the two experiments (see Table 4 for Experiment 1a proportion skipped means), the data suggest there is something about reading an expression that affects skipping behavior. Namely, in Experiment 1a, where the phrase-final word was impoverished, people were already in the processing of reading an expression. By virtue of having experience with the language, a reader would know what kinds of completions would be felicitous when reading a phrase—perhaps as broad as knowing what part of speech should come next. In Experiment 1b, when the manipulation of the target was in phrase-initial position, this is arguably less constraining than a phrasal completion, which is perhaps what led to diminished skipped rates across conditions. This hypothesis is supported by the absence of a penalty when reading SUB targets, which should elicit the greatest penalty in reading and, in theory, be skipped the least. No differences between the levels of the factor Letter suggest these targets were all read with the same degree of expectation.

Means and standard errors per condition for first fixation duration are available in Table 15. These results show the shortest first fixation durations for Identity targets, and little difference between TL and SUB targets. The model output is available in Table 16. Results show a significant main effect of Letter but not of Phrase Type, and no interaction (see Figure 3 for visualization). Pairwise comparisons within the factor Letter revealed that Identity targets had shorter first fixation durations than TL targets and SUB targets, but that the difference between TL and SUB targets was not significant.

Table 15. Reading measures for the target word in Experiment 1b

First fixation duration			
	Identity	TL	SUB
Idiom	227.75 (4.00)	249.64 (4.86)	248.08 (5.50)
Literal	233.32 (3.76)	241.93 (4.75)	244.27 (5.67)
Total duration			
	Identity	TL	SUB
Idiom	259.07 (7.08)	360.26 (12.10)	526.14 (20.04)
Literal	262.48 (7.62)	333.98 (10.57)	472.73 (15.49)

Values indicate mean durations in milliseconds. Standard error of the mean indicated in parentheses.

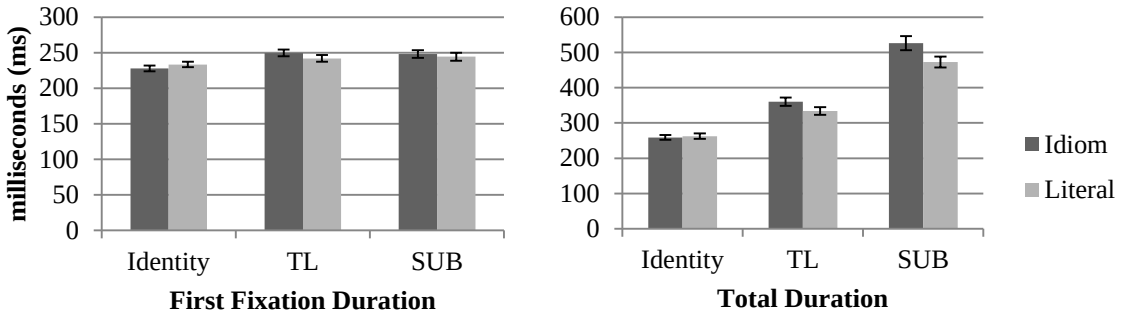

Figure 3. First fixation duration (left) and total reading time (right) for the phrase-initial word. Error bars indicate standard error of each mean represented.

Table 16. Model for first fixation duration for phrase-initial word from Experiment 1b

Effect	df	X^2	p-value
Letter	2	16.34	<.01
Phrase Type	1	0.47	.49
Letter×Phrase Type	2	2.37	.31

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	T	p-value
Identity – SUB	-16.69	4.61	3622	-3.62	<.01
Identity – TL	-15.92	4.66	3624	-3.41	<.01
SUB – TL	0.77	4.57	3622	0.17	0.98

Output constructed using "lsmeans" package in R.

These data suggest that, in the absence of a surrounding context with which to anticipate a particular kind of word, letter transpositions and letter substitutions are equally disruptive in early processing. In both cases, the targets are misspellings. Without a surrounding context with which to create an expectation for what should come next, these data suggest that the degree of perturbation of the stimulus does not matter. Neither target would match anything within a person's lexicon, creating problems for both targets in initial stages of lexical access.

Means and standard errors per condition for total duration on the phrase-initial word are available in Table 15. Results show comparable total durations for Identity targets across phrase type, longer total durations for TL targets, and even longer times for SUB targets. Further, while the difference between the phrase types does not appear to be significantly different within Identity targets, total duration times seem to diverge the more degraded the stimulus is, where idioms seem to incur a larger reading time penalty when the phrase contains a substitution. The model output, which is available in Table 17, shows that total duration measures on the phrase-initial word showed a significant main effect of Letter, such that the more impoverished the visual input, the longer the reading times for the phrase-initial word become. Further, there was additionally a main effect of Phrase Type, such that idioms were read on average more slowly than literal targets. The fact that this manifests in total time, and not in early measures of processing, suggests that it is driven by later re-reading processes. Finally, there was a marginally significant interaction between Letter and Phrase Type, which suggests that the more degraded the visual input, the more this results in longer reading times, as is evident in late measures. Namely, idioms incur a greater penalty in total reading time when the target word contains a transposition and an even greater penalty when the target contains substituted letters, suggesting that the more degraded the input, the harder it becomes to integrate the target into an idiom compared to a literal expression. Pairwise comparisons showed that Identity targets had shorter overall reading times than TL targets and SUB targets, and that SUB targets had longer overall reading times than TL targets (see Figure 3 for visualization).

This data confirms the pattern shown so far, where the visual system seems to be sensitive to the degree of input degradation. Similarly, targets containing substitutions were read significantly longer when they occurred at the onset of an idiom. This suggests that it was harder to overcome degraded bottom-up when reading an idiom and this is consistent with theories of idiom comprehension, specifically the hypothesis that idioms are configurations and must be recognized as such in order to activate the idiom meaning. In the presence of severely degraded bottom-up input, it makes sense that idioms demonstrate a penalty when the target word contains a substitution. While literal collocations are

also recognizable as configurations, this is perhaps more constraining in idioms. When the initial word is degraded at the start of an idiom, results indicate it is harder to integrate a target with misspellings into an idiom, and that the more anomalous the target, the harder this becomes. This is demonstrated by the presence of this effect in late measures and not in early measures, which suggests this effect indexes more so re-reading and regressions rather than initial activation difficulty.

Table 17. Model for total duration on phrase-initial word from Experiment 1b

Effect	df	X^2	p-value
Letter	2	415.76	<.001
Phrase Type	1	7.23	<.01
Letter×Phrase Type	2	6.06	.05

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-238.61	11.56	4229	-20.64	<.001
Identity – TL	-86.25	11.56	4229	-7.46	<.001
SUB – TL	152.36	11.56	4229	13.18	<.001
Phrase Type					
Idiom - Literal	25.40	9.44	4229	2.69	<.01
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	-3.41	16.35	4229	-0.21	1.00
TL, Idiom – TL, Literal	26.09	16.36	4229	1.60	.60
SUB, Idiom – SUB, Literal	53.52	16.35	4229	3.27	<.05

Output constructed using "lsmeans" package in R.

Whole phrase measures

Means and standard error rates per condition for first pass times are shown in Table 18. The data show that phrases with Identity targets have shorter first pass times than phrases with TL targets, and phrases with TL targets have shorter first pass times than phrases with SUB targets. Further, there seems to be an effect of phrase, where idioms, in general, have longer first pass times than literals. The model output is shown in Table 19. There was a significant main effect of Letter, such that the more degraded a stimulus, the longer the first pass times on the phrase became. There was no main effect of Phrase Type or interaction (see Figure 4 for visualization). Pairwise comparisons showed that phrases with Identity targets elicited shorter first pass times than phrases with TL targets and phrases with SUB targets, and that

phrases with SUB targets elicited longer first pass times than phrases with TL targets. This shows that, when the phrase-initial word was somehow degraded, this significantly affected the amount of time needed to read the entire expression, even on the first pass through the region. Although first fixation durations on the phrase-initial word did not reveal a difference between TL and SUB targets, first pass times suggest this difference was evident within the first pass through the region containing the phrase.

Table 18. Reading measures for the whole phrase in Experiment 1b

First pass time			
	Identity	TL	SUB
Idiom	609.53 (11.98)	692.20 (14.82)	860.95 (19.94)
Literal	573.91 (11.16)	678.33 (14.78)	807.94 (18.55)
Total duration			
	Identity	TL	SUB
Idiom	746.73 (14.80)	924.98 (21.85)	1219.91 (31.01)
Literal	743.82 (16.58)	866.27 (19.40)	1118.65 (25.62)

Values indicate mean durations in milliseconds. Standard error of the mean indicated in parentheses.

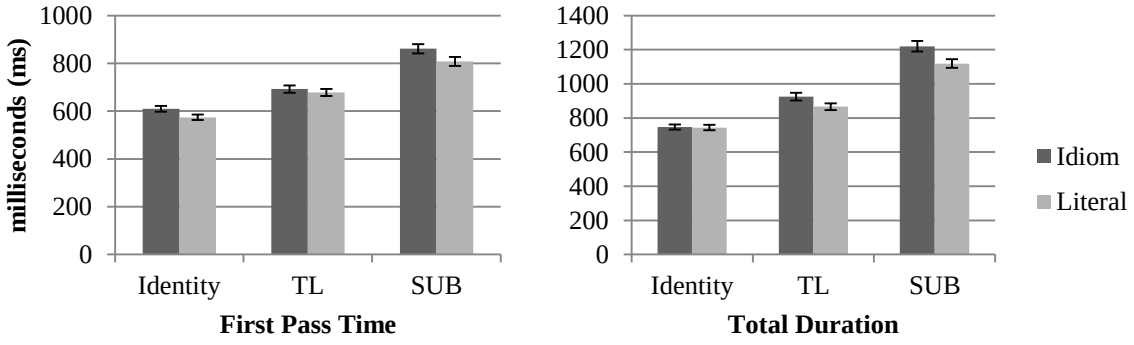

Figure 4. First pass time (left) and total duration (right) on the whole phrase in Experiment 1b. Error bars indicate the standard error of each mean represented.

Table 19. Model for first pass time for whole expression from Experiment 1b

Effect	df	X^2	p-value
Letter	2	319.42	<.001
Phrase Type	1	1.77	.18
Letter×Phrase Type	2	2.10	.35

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-242.72	13.44	4149	-18.07	<.001
Identity – TL	-93.68	13.44	4149	-6.97	<.001
SUB – TL	149.04	13.44	4149	11.09	<.001

Output constructed using "lsmeans" package in R.

Table 18 shows the means and standard errors per condition for total duration on the phrase. The data show that phrases with Identity targets had the shortest overall reading times compared to phrases with TL and SUB targets, and that phrases with TL targets had shorter total durations than phrases with SUB targets. Further, while there does not appear to be a difference between the phrase types within Identity targets, the more degraded the input gets, the more idioms seem to require longer reading times. The model output is available in Table 20. Results showed a significant main effect of Letter, which suggests that the more impoverished the visual input, the longer it took to read the entire phrase. There was no main effect of Phrase Type, which shows that reading times, overall, were not predicted by the type of phrase being read. There was also an interaction, which suggests that while phrase type did not make a difference when the target was intact, more degraded input led to longer reading times, and this was most apparent for idioms (see Figure 4 for visualization). Pairwise comparisons showed that phrases with Identity targets had shorter overall reading times than phrases with TL targets and phrases with SUB targets. Further, phrases with SUB targets had longer overall reading times than phrases with TL targets

Table 20. Model for total duration for whole expression from Experiment 1b

Effect	df	X^2	p-value
Letter	2	533.11	<.001
Phrase Type	1	2.41	.12
Letter×Phrase Type	2	7.48	<.05

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-424.01	18.03	4102	-23.52	<.001
Identity – TL	-150.31	18.03	4102	-8.34	<.001
SUB – TL	273.69	18.03	4102	15.18	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	2.92	40.54	220	0.07	1.00
TL, Idiom – TL, Literal	58.63	40.55	220	1.45	.70
SUB, Idiom – SUB, Literal	101.27	40.54	220	2.50	.13

Output constructed using "lsmeans" package in R.

Experiment 1 summary

Taken together, the results from Experiment 1b highlight the importance of recognizing a configuration as meaningful in idiom comprehension. Specifically, target word measures demonstrate the importance of the phrase-initial word in an idiom; when the word in this position contains unadulterated lexical items, these targets elicited comparable reading behaviors across phrase type. The more unusual the orthography—either the position of expected characters, or the identity of substituted characters—the more overall processing time is required, and this seems to be particularly the case when reading an idiom. The interaction in total reading time of the phrase-initial target suggests that when a person tries to integrate a misspelled target into a context, this misspelling can make it more difficult to integrate an idiom-initial word into the expression, culminating in longer total reading time of the target. For idioms, the phrasal configuration requires specific lexical items, which is perhaps what led to this greater difficulty downstream compared to literals. Early measures from the region containing the whole phrase confirm that, when all else is controlled for, there are few differences between idioms and literals. This suggests that idioms and literals do not show differences in processes related to prediction or lexical access. The differences that do exist, however, manifest most clearly in later reading times. Namely, when the phrase-initial word in the configuration is severely degraded, this is problematic for literal expression integration, but it is comparably worse for idiomatic expression integration. Further, this same pattern of results is found when looking at reading times of the entire phrase of interest, which supports the difference between the phrase types appears in late processing.

2.3 Experiment 2a

2.3.1 Method

Participants

Sixty L1 Mandarin Chinese-L2 English bilinguals were recruited for participation in Experiment 2a (range: 18-31; mean=22 years; 49 female), and 36 L1 Mandarin Chinese-L2 English bilinguals were

recruited for Experiment 2b (range: 18-33 years; mean= 22.7 years; 23 female; see Table 21 for a summary of the L1 Mandarin speakers' language background and proficiency measures).

Table 21. L1 Mandarin participants' language background, proficiency from Experiment 2

	Years studying English	Reading	Writing	Speaking	Listening	MELICET	PPVT	Length of residence (months)
Experiment 2a								
Mean	14.23	7.77	7.1	6.82	7.37	36	154	27.19
Range	6-22	4-10	3-10	3-10	4-10	26-49	86-216	1-89
St.Dev.	3.67	1.51	1.49	1.62	1.45	5.50	25.37	22.18
Experiment 2b								
Mean	14.32	7.62	7.19	7.30	7.92	35.69	160.14	23.19
Range	2-22	4-10	4-10	4-10	4-10	28-45	108-207	5-66
St.Dev.	4.50	1.40	1.22	1.27	1.14	4.70	21.98	17.06

Note: Reading, writing, speaking, and listening are subjective ratings provided by participants (1=poor; 10=high). The maximum score on the MELICET is 50 possible points; the maximum score on the PPVT is 220.

Paired t-tests showed no difference between the PPVT scores from Experiment 1b and 2b ($t=-1.38$, $p=.17$) and also no difference between the MELICET scores ($t=0.53$, $p=0.60$). All participants were either students at the University of Illinois or members of the surrounding Champaign-Urbana area. Participants reported having normal or corrected-to-normal vision and no history of dyslexia or developmental reading disorders. All participants completed a handedness questionnaire, and all reported being right-handed. All participants were compensated with cash for their time.

Materials

The materials for Experiment 2 were the same stimuli used in Experiment 1. To recap, 60 idioms and sixty literal collocations were used in a 2 (Phrase type) × 3 (Letter) design (see Table 1 for example items). Idioms were those with a minimum average familiarity rating of 3.5 from the norming study in Bulkes & Tanner (2017). Sentences started with a preamble that felicitously led to either an idiom or a literal collocation (i.e. a figuratively biasing for the idiom, or a literal context for the collocation). The first and last words of the idiom or literal expression were always content words, with any verbs being lexical verbs as opposed to auxiliary verbs. In Experiment 2a, the target word was the last word in the expression of interest; in Experiment 2b, the target word was the first word.

Procedure

The procedure for Experiment 2 was the same as for Experiment 1.

Data Processing & Analysis

Data for Experiment 2a were processed and analyzed in the exact same manner as in Experiment 1, specifically in Experiment 1a, where the target manipulation occurred in the phrase-final word.

2.3.2 Results

Target word measures

Mean proportion skipped per condition are available in Table 22. These data suggest there seem to be few, if any differences between conditions with respect to skipping. Specifically, there seem to be no differences comparing the levels of Letter or Phrase Type. The model output is in Table 23 and confirms this: There was no significant main effect of either Letter or Phrase Type.

Table 22: Proportion skipped for the region containing the phrase-final target word in Experiment 2a

	Identity	TL	SUB
Idiom	0.02 (0.004)	0.03 (0.005)	0.02 (0.004)
Literal	0.02 (0.004)	0.01 (0.003)	0.02 (0.004)

Values are averages expressed as a probability. Standard error of the mean indicated in parentheses.

Table 23. Model for proportion skipped data from Experiment 2a

Effect	df	X^2	p-value
Letter	2	3.15	.21
Phrase Type	1	1.92	.17
Letter×Phrase Type	2	2.74	.25

Output constructed using mixed function with "afex" package in R.

This suggests the bilinguals' use of parafoveal preview information was qualitatively different in this task from that of the English native speakers in Experiment 1a. For example, it may suggest that the Mandarin natives, while privy to the information in the parafovea, did not use it to plan their eye movements in the same way that English natives did. Further, this data suggest that the bilinguals seldomly skipped the phrase-final word. This may be due to a high level of attention paid in foveal fixation when reading the phrase-final word, as this would be a crucial part of being able to understand the meaning of the

expression. As these were proficient bilinguals, participants would have been well aware that what was coming was a needed element to put together a representation of the phrase. This may have led to greater attention—and thus less skipping—on the phrase-final word.

Means and standard errors per condition for first fixation duration are available in Table 24. They show that Identity targets had shorter first fixation durations than TL targets, which had shorter first fixation durations than SUB targets. Further, there appears to be no difference between the levels of phrase. The model output is shown in Table 25. Results showed no significant main effect of Letter, such that first fixation durations were not significantly affected by the quality of visual information in phrase-final targets. Further, there was no main effect of Phrase type and no interaction. This suggests that the type of phrase was not a meaningful predictor of how long first fixation durations would be (see Figure 5 for visualization).

Table 24. Reading measures for the phrase-final word in Experiment 2a

First fixation duration			
	Identity	TL	SUB
Idiom	278.95 (3.21)	278.12 (3.28)	284.35 (3.39)
Literal	275.73 (3.14)	271.23 (2.96)	275.82 (3.07)
Total duration			
	Identity	TL	SUB
Idiom	1449.13 (27.67)	1654.89 (33.68)	1822.38 (34.23)
Literal	1183.47 (23.69)	1295.76 (25.33)	1598.25 (31.96)

Values indicate mean durations in milliseconds. Standard error of the mean represented in parentheses.

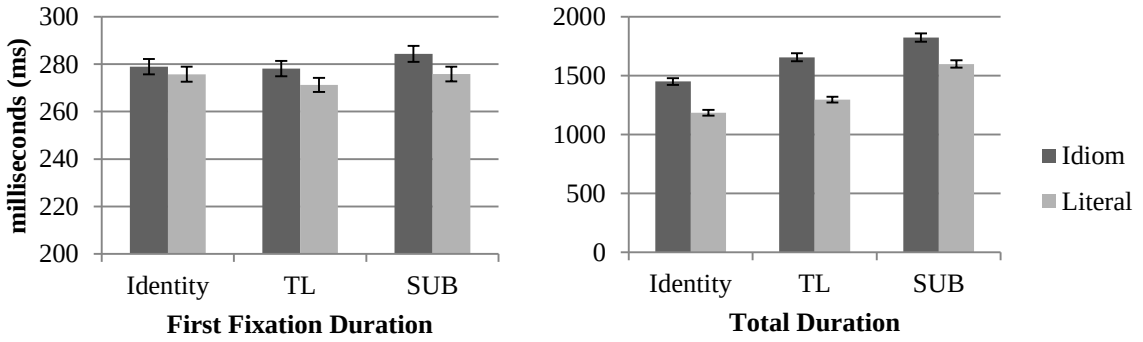

Figure 5. First fixation duration (left) and total duration (right) measures for Experiment 2a. Error bars indicate standard error of each mean represented.

Table 25. Model for first fixation duration for phrase-final word for Experiment 2a

Effect	df	X^2	p-value
Letter	2	3.22	.20
Phrase Type	1	2.65	.10
Letter×Phrase Type	2	0.84	.66

Output constructed using mixed function with "afex" package in R.

Means and standard error rates for total duration are shown in Table 24. They indicate show that Identity targets had shorter total reading times than TL targets, which had shorter total reading times than SUB targets. There also appears to be an advantage for literals, where targets within literals, overall, appear to have shorter reading times, comparing within the levels of Letter. The model output is shown in Table 26, and shows a significant main effect of Letter, such that the more degraded the visual input, the longer total reading times became. Further, there was a significant main effect of Phrase Type, such that targets within literals, overall, were read for less time than targets in idioms. Also, there was a Letter × Phrase Type interaction, such that the difference in total reading times for Identity and TL targets within literals is smaller than the same comparison within idioms (see Figure 5 for visualization).

Table 26. Model for total duration for phrase-final word from Experiment 2a

Effect	df	X^2	p-value
Letter	2	292.26	<.001
Phrase Type	1	22.45	<.001
Letter×Phrase Type	2	8.47	<.01

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-392.68	22.86	6662	-17.18	<.001
Identity – TL	-160.07	22.86	6662	-7.00	<.001
SUB – TL	232.61	22.86	6662	10.18	<.001
Phrase Type					
Idiom v. Literal	282.45	56.83	120	4.97	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	264.88	62.66	177	4.23	<.001
TL, Idiom – TL, Literal	356.04	62.66	177	5.68	<.001
SUB, Idiom – SUB, Literal	226.43	62.66	177	3.61	<.01

Output constructed using "lsmeans" package in R.

Pairwise comparisons showed Identity targets had shorter first fixation durations than TL targets and SUB targets, and SUB targets had longer first fixation durations than TL targets. This suggests that idioms were harder to read, overall, and that this difficulty manifested in downstream, integrative processes. Further, the degree of degradation also contributed to increased reading times for TL targets and SUB targets, as indexed by the difference between the two, suggesting that the visual system in cross-script bilinguals is also keenly sensitive to how much unexpected orthography is in a stimulus. Finally, the interaction suggests that letter transpositions within literals were easier to recover from—and took less time to read—than transpositions within idioms, suggesting that the type of phrase impacted the relative ease or difficulty with which unexpected orthography could be processed. This is further demonstrated by the smaller difference between the total time for idioms and literals when the target contained a substitution, where outright substituted characters were harder to recover from than transpositions, and that this difference was most pronounced in literal expressions.

Whole phrase measures

Means and standard errors per condition are shown in

Table 27 for first pass times. The data show phrases with Identity targets had shorter first pass times than phrases with TL targets, and that phrases with TL targets had shorter first pass reading times than phrases with SUB targets. Further, literal phrases, overall, seem to have the shortest first pass reading times. The model output is included in Table 28. Results show a significant main effect of Letter, showing that first pass times are the longest on phrases with SUB targets, shorter on phrases with TL targets, and shortest on phrases with Identity targets. There was a main effect of Phrase Type, such that first pass times were shorter, overall, when reading literals compared to idioms. Finally, there was a significant Letter × Phrase Type interaction (see Figure 6 for visualization). Pairwise comparisons showed that phrases with Identity targets had shorter first pass times than phrases with TL targets and phrases with SUB targets. Phrases with SUB targets had longer first pass times than phrases with TL targets.

Table 27. Reading measures for the whole expression in Experiment 2a

First pass time			
	Identity	TL	SUB
Idiom	937.30 (14.43)	1028.69 (17.28)	1088.33 (16.74)
Literal	838.19 (12.78)	872.39 (13.65)	994.15 (16.69)
Total duration			
	Identity	TL	SUB
Idiom	1523.42 (26.67)	1729.99 (32.54)	1913.58 (33.21)
Literal	1240.85 (22.77)	1355.67 (24.75)	1673.77 (31.16)

Values indicate mean durations in milliseconds. Standard error of the mean represented in parentheses.

These data also suggest that idioms were harder to process, and that this difficulty manifested in early measures of reading through the region. The interaction term, however, interestingly shows that the TL targets were no different from Identity targets when reading literals (β =-35.64, t=-1.83), but that this difference was significant when reading idioms (β =-91.98, t=-4.70). For the bilinguals, TL targets enjoyed more facilitation lexical access when embedded in a literal expression, suggesting a kind of "good enough" processing early on when only letter position information was manipulated. When both letter position and identity information were affected, this incurred extra difficulty in initial processing compared to TL targets, and this was true both when reading idioms (β =58.40, t=2.98) and literals (β =119.58, t=6.12).

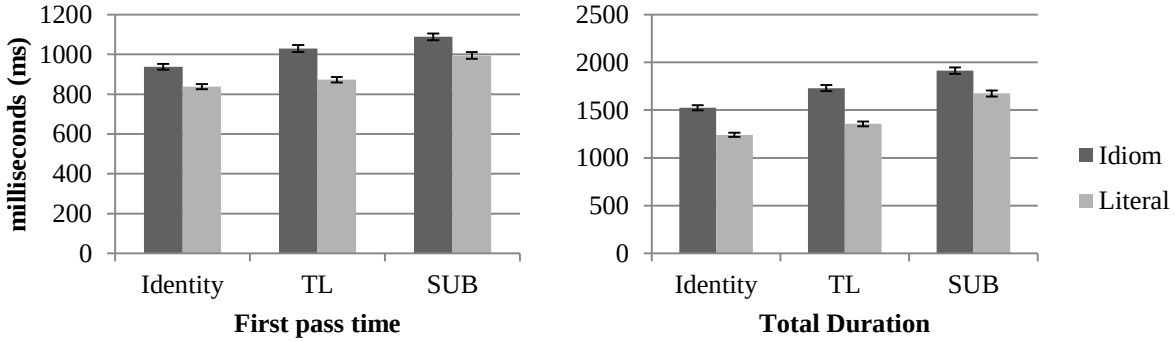

Figure 6. First pass time (left) and total duration (right) for the region containing the whole expression in Experiment 2a. Error bars represent the standard error of each mean represented.

Table 28. Model for first pass time for the whole expression from Experiment 2a

Effect	df	X^2	p-value
Letter	2	122.10	<.001
Phrase Type	1	104.21	<.001
Letter×Phrase Type	2	6.05	.05

Contrasts	β	St. Error	df	T	p-value
Letter					
Identity – SUB	-152.80	13.83	7007	-11.05	<.001
Identity – TL	-63.81	13.82	7007	-4.62	<.001
SUB – TL	88.99	13.83	7007	6.43	<.001
Phrase Type					
Idiom v. Literal	115.65	11.29	7089	10.25	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	98.48	19.54	7089	5.04	<.001
TL, Idiom – TL, Literal	154.82	19.55	7089	7.92	<.001
SUB, Idiom – SUB, Literal	93.64	19.56	7089	4.79	<.001

Output constructed using "lsmeans" package in R.

Means and standard errors per condition for total duration on the whole phrase are shown in Table 27. The data show that phrases with Identity targets had shorter overall reading times than phrases with TL targets, and that phrases with TL targets had shorter overall reading times than phrases with SUB targets. Further, literal phrases, overall, seem to have the shortest overall reading times. The model output is shown in Table 29. Results show a significant main effect of Letter, such that more degraded visual input led to longer overall reading times. There was a main effect of Phrase Type suggest, which suggests that idioms, overall, incurred a processing penalty in total reading times. Finally, there was a significant Letter × Phrase Type interaction (Figure 6 for visualization). Pairwise comparisons showed Identity targets had shorter overall reading times than TL targets and SUB targets, and that SUB targets had longer overall reading times than TL targets.

Table 29. Model for total duration for the whole expression from Experiment 2a

Effect	df	X^2	p-value
Letter	2	300.52	<.001
Phrase Type	1	240.28	<.001
Letter×Phrase Type	2	8.14	<.05

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-407.61	23.42	7272	-17.40	<.001
Identity – TL	-162.42	23.40	7272	-6.94	<.001
SUB – TL	245.18	23.42	7272	10.47	<.001
Phrase Type					
Idiom v. Literal	298.80	19.11	7033	15.63	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	282.40	33.09	7033	8.53	<.001
TL, Idiom – TL, Literal	372.38	33.09	7033	-12.92	<.001
SUB, Idiom – SUB, Literal	241.70	33.13	7033	7.30	<.001

Output constructed using "lsmeans" package in R.

Measures of total reading time further demonstrate the pattern found so far that idioms are harder to read, this time demonstrating that this difference is apparent throughout both early and late stages of processing. Additionally, these results show that any kind of misspelling in the target incurs a processing cost in downstream processing.

Experiment 2a summary

In sum, the results from Experiment 2a largely confirm the difficulty that idioms present to L2 speakers compared to literal expressions, which is consistent with prior work (e.g. Cieslicka, 2006). As the participants in this study have comparatively less exposure to the input than participants from Experiment 1a, this difference is expected, as exposure and experience are prerequisites for successful idiom comprehension (although in more globally decomposable idioms, this is perhaps less strict). In analyses of the target word, early measures of prediction and lexical access (skipping rate, first fixation duration) showed no differences of any kind among the levels of either factor, suggesting that any differences that arose in later measures were a result of trying to incorporate the target (and its presence or absence of unexpected spelling) into a surrounding phrase. This task was markedly more difficult when

the locus of integration was an idiom, where literals showed comparably less difficulty, as was indexed by less reading time. However, what is particularly interesting is that despite both TL and SUB targets containing what are ultimately misspellings, the bilinguals were sensitive to the similarity of these targets to actual items in the English lexicon, as was indexed by the difference in reading times for TL and SUB conditions. While targets from these conditions are all ultimately nonwords, L2 speakers demonstrated a keen sensitivity to the minor misspelling in TL targets compared to the more deviant spelling in SUB targets, by demonstrating shorter reading times in cases of a TL target compared to a SUB target. Similarly, bilinguals demonstrated more flexible letter position encoding when reading literals. Only in later measures, do we see clearer differences between Identity and TL targets, and this is an implication I will discuss more in the general discussion.

2.4 Experiment 2b

2.4.1 Method

Participants

Thirty-six L1 Mandarin Chinese-L2 English bilinguals were recruited for Experiment 2b (range: 18-33 years; mean= 22.7 years; 23 female; see Table 21 for a summary of the L1 Mandarin speakers' language background and proficiency measures). All participants were either students at the University of Illinois or members of the surrounding Champaign-Urbana area. Participants reported having normal or corrected-to-normal vision and no history of dyslexia or developmental reading disorders. All participants completed a handedness questionnaire, and all reported being right-handed. All participants were compensated with cash for their time.

Materials

The materials for Experiment 2b were the exact same stimuli used for Experiment 1b, namely sentences leading up to either idioms or literals, but with the target letter manipulation in phrase-initial position (see Table 12 for example stimuli).

2.4.2 Procedure

The procedure for Experiment 2a was identical to the procedure for Experiments 1a, 1b, and 2a.

Data Processing & Analysis

The same data processing and analysis steps used for Experiments 1a, 1b, and 2a were used for the data from Experiment 2b.

2.4.3 Results

Target word measures

Mean proportions skipped and standard error rates are shown in Table 30. These data show that while participants in Experiment 2b are skipping more than participants did in 2a, there do not seem to be differences in skipping rates when comparing across the levels of either factor. The model output is shown in Table 31, and there is no significant main effect of Letter or Phrase Type. These findings suggest that the type of letter information in the parafovea did not influence skipping behavior, and neither did the type of phrase the bilinguals were reading. Further, the skipping rate observed here is less than the skipping rate observed for monolinguals in Experiment 1, which shows bilinguals skipped phrase-initial targets, overall, less than monolinguals.

Table 30: Proportion skipped for the region containing the phrase-initial word in Experiment 2b

	Identity	TL	SUB
Idiom	0.11 (0.01)	0.11 (0.01)	0.10 (0.01)
Literal	0.10 (0.01)	0.12 (0.01)	0.11 (0.01)

Values are averages expressed as a probability. Standard error of the mean indicated in parentheses.

Table 31. Model for proportion skipped data from Experiment 2b

Effect	df	X^2	p-value
Letter	2	0.89	.64
Phrase Type	1	0.06	.81
Letter×Phrase Type	2	0.26	.88

Output constructed using mixed function with "afex" package in R.

Means per condition and standard error rates for first fixation duration on the phrase-initial word are shown in Table 32. These data show slight differences in first fixation duration, where Identity targets appear to have the shortest durations, followed by TL targets, where SUB targets appear to have the

longest first fixation durations. There does not appear to be a difference based on the type of phrase being read. The model output is provided in Table 33, which shows a significant main effect of Letter, suggesting that the more degraded the visual input, the longer first fixation durations became. There was no main effect of Phrase Type, suggesting that first fixation duration was not predicted based on the type of phrase being read. Pairwise comparisons showed that Identity targets had shorter first fixation durations than TL targets and SUB targets, and that SUB targets had longer first fixation durations than TL targets (see Figure 7 for visualization). This suggests that early measures of lexical access were impacted by the degree of degradation in the stimulus, where manipulating letter position led to less disruption in reading than when both letter position and identity were manipulated. Further, the interaction between Letter and Phrase Type neared significance, suggesting a trend where TL targets seemed less disruptive when occurring in a literal expression, but that this difference between the phrase types went away the more degraded the bottom-up cues became.

Table 32. Reading measures for the target word in Experiment 2b

First fixation duration			
	Identity	TL	SUB
Idiom	263.87 (4.58)	296.36 (5.68)	304.80 (6.15)
Literal	272.90 (4.35)	281.54 (5.40)	302.61 (6.10)
Total duration			
	Identity	TL	SUB
Idiom	476.11 (13.17)	626.29 (17.31)	814.74 (22.94)
Literal	413.26 (10.26)	550.64 (18.11)	804.71 (23.62)

Values indicate mean durations in milliseconds. Standard error of the mean represented in parentheses.

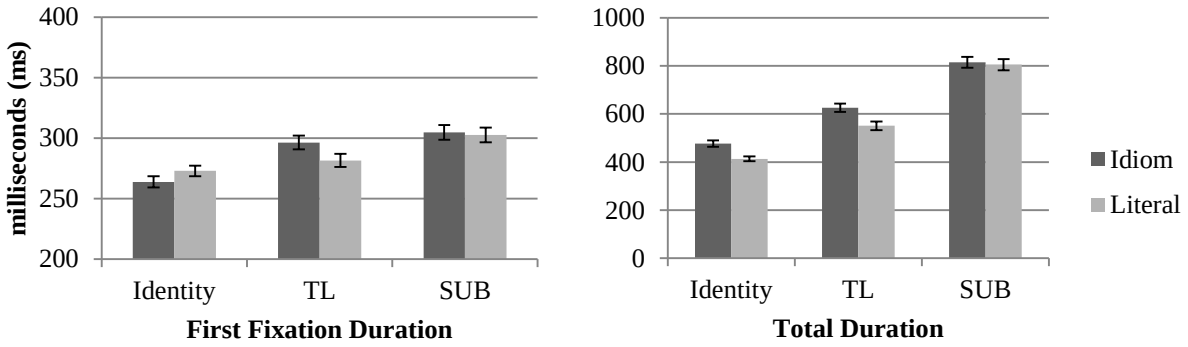

Figure 7. First fixation duration (left) and total duration (right) for the phrase-initial word in Experiment 2b. Error bars indicate standard error of each mean represented.

Table 33. Model for first fixation duration on phrase-initial word from Experiment 2b

Effect	df	X^2	p-value
Letter	2	47.91	<.001
Phrase Type	1	0.40	.53
Letter×Phrase Type	2	5.24	.07

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-35.89	5.19	4000	-6.92	<.001
Identity – TL	-20.78	5.21	4000	-3.99	<.01
SUB – TL	15.11	5.19	4000	2.91	<.05

Output constructed using "lsmeans" package in R.

Table 32 shows the means and standard errors per condition for total duration measures. These data show that total duration increased the more misspelled and anomalous the target word became. This difference also seems to be differentiated by phrase type only when the target appears without unexpected orthography—the more misspelled the target, the more problematic this seems to be for both types of phrases, as indicated by a similarity in reading times in the SUB condition. The model output is shown in Table 34. Results showed a significant main effect of Letter, where Identity targets led to the shortest overall reading times, TL targets had slightly longer reading times, and SUB targets had the longest overall reading times. There was a main effect of Phrase Type, suggesting that idioms, overall, took longer to read. There was no interaction between Letter and Phrase Type (see Figure 7 for visualization). These results suggest that idioms were harder to read in later integrative measures when the target contained a misspelling compared to when the misspelling appeared in a literal collocation. Further, the effect of Letter shows the same pattern from first fixation duration, where the degree of incorrect letter information also affects total reading times on the target word.

Table 34. Model for total duration on phrase-initial word from Experiment 2b

Effect	df	X^2	p-value
Letter	2	490.25	<.001
Phrase Type	1	14.11	<.001
Letter×Phrase Type	2	4.65	.10

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Letter					
Identity – SUB	-365.04	16.13	4267	-22.63	<.001
Identity – TL	-143.78	16.13	4267	-8.91	<.001
SUB – TL	221.26	16.13	4267	13.72	<.001
Phrase Type					
Idiom v. Literal	49.51	13.17	4267	3.76	<.001

Output constructed using "lsmeans" package in R.

Whole phrase measures

Means and standard errors per condition for first pass times are shown in Table 35. The data show longer overall first pass times for SUB targets compared to TL targets, and shorter overall first past times for Identity targets. Further, there appears to be a slight advantage in first pass time for literal expressions. The model output is included in Table 36. Results showed a significant main effect of Letter, where more degradation in the visual input led to longer first pass times, and no effect of Phrase Type, suggesting that this first pass time behavior was not significantly impacted by the type of phrase being read. Pairwise comparisons indicated that phrases with Identity targets had shorter first pass times than phrases containing TL targets and phrases containing SUB targets, and that SUB targets had longer first pass times than phrases with TL targets (Figure 8 for visualization). This suggests that, during the first pass through the region, when the target manipulation was in the first word of the expression, the type of expression did not influence reading behavior.

Table 35. Reading measures for the target word in Experiment 2b

First pass time			
	Identity	TL	SUB
Idiom	937.22 (19.66)	1063.71 (24.48)	1214.48 (26.49)
Literal	867.61 (17.09)	1019.17 (23.78)	1188.60 (26.41)
Total duration			
	Identity	TL	SUB
Idiom	1465.03 (32.54)	1690.25 (37.09)	1938.16 (44.11)
Literal	1186.12 (26.46)	1412.86 (37.65)	1791.64 (40.21)

Values indicate mean durations in milliseconds. Standard error of the mean represented in parentheses.

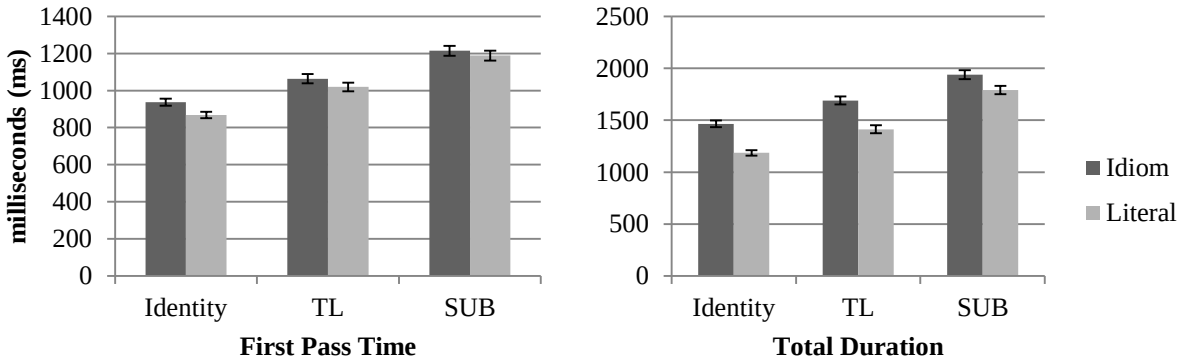

Figure 8. First pass time (left) and total duration (right) on the region containing the whole phrase in Experiment 2b. Error bars indicate standard error of each mean represented.

Table 36. Model for first pass time for the whole expression in Experiment 2b

Effect	df	X^2	p-value
Letter	2	213.50	<.001
Phrase Type	1	1.55	.21
Letter×Phrase Type	2	1.18	.56

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	T	p-value
Letter					
Identity – SUB	-299.12	20.23	4132	-14.79	<.001
Identity – TL	-139.03	20.23	4132	-6.87	<.001
SUB – TL	160.09	20.23	4132	7.92	<.001

Output constructed using "lsmeans" package in R.

Means and standard error rates for total duration of the whole phrase are shown in Table 35. The data show the same pattern from first pass times, namely where SUB targets seem to generate the longest total reading time for the phrase, TL targets yield a slightly shorter total reading time compared to SUB targets, and phrases with Identity targets have the shortest total reading time. Further, there seems to be a slight advantage in reading times when the phrase in question is literal. The model output is shown in

Table 37. Results showed a significant main effect of Letter, where total reading times increased the more misspelled the phrase-initial word was. There was a main effect of Phrase Type, which suggests that reading times for the phrase were shorter, overall, if the phrase was literal. There was also a significant Letter × Phrase Type interaction. Pairwise comparisons revealed that phrases with Identity targets had shorter overall reading times than phrases with TL targets and phrases with SUB targets, and that phrases with SUB targets had longer overall reading times than phrases with TL targets (see Figure 8 for visualization). This suggests that, in integrative processing, idioms were harder to read overall than literals, and that this difference was further affected by the quality of visual cues in the phrase-initial word. The interaction term illustrates that the more degraded the appearance of the phrase-initial word, the more this penalized reading times in both types of expressions. When the target manipulation was in the phrase-initial word, idioms incurred a greater overall processing difficulty than literals, suggesting that letter position manipulation only still facilitated, to an extent, activation of the phrase-initial word in literals and, to a lesser extent, in idioms. When both letter position and identity information were manipulated, this affected reading behavior for both types of expressions equally.

Table 37. Model for total duration for the whole expression in Experiment 2b

Effect	df	X^2	p-value
Letter	2	325.78	<.001
Phrase Type	1	11.80	<.001
Letter×Phrase Type	2	6.67	<.05

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	T	p-value
Letter					
Identity – SUB	-539.32	29.43	3928	-18.33	<.001
Identity – TL	-225.98	29.43	3928	-7.68	<.001
SUB – TL	313.34	29.43	3928	10.65	<.001
Phrase Type					
Idiom v. Literal	234.27	66.53	120	3.52	<.001
Letter×Phrase Type					
Identity, Idiom – Identity, Literal	278.91	74.70	190	3.73	<.01
TL, Idiom – TL, Literal	277.39	74.70	190	3.71	<.01
SUB, Idiom – SUB, Literal	146.52	74.70	190	1.96	.37

Output constructed using "lsmeans" package in R.

Experiment 2 summary

Overall, the proportion skipped data from Experiment 2, as a whole, suggests that the type of information available in the parafovea was not influential to bilinguals in informing the planning of saccades. Despite a diminished use of information in the parafovea for TL and SUB targets, participants did not show an effect of Letter, suggesting that upcoming information in the visual stream did not influence reading behavior. Further, bilinguals skipped less in Experiment 2a compared to Experiment 2b. In Experiment 2a, participants may have been more aware that they were reading a meaningful expression, perhaps encouraging them to skip the phrase-final word less in order to ensure they understood the meaning of the phrase. In Experiment 2b, there was more skipping, suggesting perhaps less expectation, overall, for any type of word. This may have led to less attention at the position of the target word, compared to phrase-final position, where they may have had more of an expectation for particular word properties (i.e. part of speech, or maybe features that fit with the interpretation of the expression they were reading).

Bilinguals did show an effect of Letter, however, the Identity < TL < SUB pattern observed in native speakers was not quite replicated in bilinguals. Namely, in literals, bilinguals demonstrated more similarity in reading Identity and TL targets when the target was embedded in a literal expression. In an idiom expression, TL targets showed more similarity to SUB targets in reading measures. This suggests that, for bilinguals, unexpected letter position information was easier to reconcile and recover from when they were reading a literal expression. In an idiom, the greater similarity in difficult between TL and SUB targets suggests unexpected orthography, in general, was harder to process when the target was embedded in an idiom. This pattern manifested most clearly in measures of the region containing the whole phrase, both in early and in late measures, suggesting that this process had something to do with the ease of accessing this target and integrating it into an already established context. This pattern was not evident in Experiment 2b, which suggests that people required some prior scaffolding of the phrase-level meaning in order to demonstrate this greater flexibility for TL targets compared to SUB targets in literals. Broadly

speaking, patterns such as these demonstrated the benefit that literal collocations enjoyed in bilingual processing, and how this benefit is driven by how the available top-down information interacts with the available bottom-up information, and how the larger presence of the former helps to mitigate processing of impoverished bottom-up cues.

2.4.4 Experiments 1 and 2 Discussion

Experiments 1 and 2 were designed to investigate how top-down information, like semantic opacity, interacts with bottom-up information in natural reading. Further the collective endeavor was intended to explore differences in L1 and L2 processing of collocations, both literal and idiomatic. When controlling for frequency variables, I was interested in how differences in literal and nonliteral processing manifested in native and nonnative speakers of a language, and how complex methods, like eyetracking, could demonstrate how differences manifest at different time points for these two populations. Contrary to the perspective that idioms bring with them a processing advantage over literal strings, the results of these two studies show that, if there is a difference, idioms seem to incur a processing penalty, rather than an advantage, and that this difference is most evident in late measures, signifying a greater difficulty in integrative processes, compared to initial stages of lexical access and word identification. Further, when we see this penalty in later processing, it was when anomalous visual input occurred within the idiom.

For example, in Experiment 1, target word measures showed no main effect of Phrase Type in first fixation duration or total duration, but a significant interaction in total duration. A closer look at pairwise comparisons revealed that the interaction was driven by the difference between idioms and literals when the phrase-final word contained a letter substitution. While both TL and SUB targets were more disruptive to natural reading than Identity targets, experiencing unexpected characters—rather than just the expected characters in the wrong order—incurred a greater penalty when it occurred in an idiom. However, when the target word was spelled correctly or contained a TL, there were no differences comparing the two phrase types. This pattern is also shown in measures taken from the whole phrase, where again, there was no interaction in first pass time, but the interaction in total duration was

significant. The difference between reading times in the SUB condition comparing idioms to literals illustrates the relationship between top-down and bottom-up information during processing, and specifically, how the difference in semantic opacity of a phrase can lead to more downstream difficulty in reading. This possibility goes back to Cacciari and Tabossi's notion of the configuration as particularly significant for idioms. For this reason, literals may be computationally easier than idioms; for idioms, not only do the lexical items need to be processed but the figurative meaning has to be computed on top of that. Further, in order to complete the configuration needed to signal an idiom, specific lexical items are required. For a literal collocation, while co-occurrence information may lead to greater expectations for a particular word, synonyms with the same critical properties or traits of that expected item will also lead to perfectly fine, felicitous completions and accomplish the same linguistic goal. In idioms, because a specific lexical entry is required, a reader will be on the lookout not only for lexical traits of the word but of visual characteristics of the wordform in the string, explaining the penalty observed most apparently for SUB targets but less so for TL targets. The reason for this is the amount of information manipulated or misplaced in the string. While TL targets ultimately retain all of the required characters, SUB targets have outright unexpected information, not just information that is in an unexpected order. It is possible that idioms may be slightly more predictive in this regard, where a person would be more likely to incur a penalty when reading unexpected orthography for an idiom compared to a collocation. The open-endedness of collocations would lead to more flexible encoding for letter position and identity information, which is why we see the greatest penalty for idioms when reading targets containing substitutions.

In Experiment 1b, when the target manipulation appeared in phrase-initial position, participants showed no differences in skipping behavior based on either the letter manipulation or the type of phrase. This is a departure from the pattern observed in Experiment 1a, and a potential reason for this is that people may have realized they were reading a meaningful expression or chunk in the first experiment, which would have more so driven orthographic expectations about upcoming words. By the time people got to the phrase-final word with the misspelling, they already had a rough idea of what should come

next. In Experiment 1b, by the time people saw the unexpected orthography, they had not yet experienced the local context of the MWE—either idiomatic or literal. At phrase-initial position, people presumably had fewer expectations for what should be coming next, perhaps leading to more flexible bottom-up processing, as would be encouraged by looser anticipation rather than looking for a word with particular features. This explanation is supported by results such as the similarity between TL and SUB targets in first fixation duration. The effect seems to be driven by the fact that, in both conditions, targets were misspelled; regardless of how much more degraded SUB targets were compared to TL targets, both are ultimately bad. Without any expectation yet from the context for what should appear next—both types of words or specific visual features of words—the longer reading times derived by TL and SUB targets were possibly simply a result of unexpected orthography, a purely bottom-up result. Without anything to compare it to, both TLs and SUBs are nonwords. This hypothesis contrasts with those of Luke and Christianson (2012), who might say that when readers are more flexible in reading, they show a difference between processing transpositions and substitutions. However the SPaM method used in the paper works by incorporating subconscious priming of words with self-paced reading, a method which differs from natural reading processes observed in eyetracking. In the current design, participants were able to regress and fixate for as long as they wanted to on these anomalous targets without any sort of prime. In seeing these targets with unexpected orthography for the first time, it is reasonable to suggest that TLs and SUBs elicit comparable reading times—longer times than Identity targets—in contexts where there is little to go off in the way of anticipation or prediction. When we do see the graded effect of substitutions being worse than transpositions is in later, integrative processing stages, when top-down information has had a chance to make an impact. For example, in total duration, Identity targets did not elicit a significant difference across the levels of Phrase Type, but targets containing letter substitutions did. This suggests that while phrase type was not influential on its own, when the bottom-up input was more degraded, this led to greater difficulty when reading the idioms compared to the literal collocations. This result is consistent with the fact that semantic opacity is a top-down cue, and as idioms seem to incur

more of a penalty than a benefit in this study, it makes sense that we see this difference manifest in late processing.

The data from Experiment 1 contribute a different perspective to the idiom literature. The great debate in idiom processing has focused on whether idioms are harder to process in literals, and when these differences manifest and in what conditions. This is the first study to examine how bottom-up input affects processing of things like semantic opacity, and the results show that, in natural reading, a more degraded visual input has greater implications for idioms than for literals. These results highlight an idiom as an environment that requires specific pieces for the idiomatic interpretation to arise. Swinney & Cutler (1979), for example, found an idiom benefit, but found it in a phrase classification task. In their study participants were faster to say idioms were phrases than other types of strings. The primary difference between their seminal study and this one is the method of presentation. It is possible that we see an idiom penalty here because items are embedded in contexts, which require compositional analysis. Switching from compositional analysis to more holistic, whole-form processing is also what may have driven the advantage for literals here, namely when the bottom-up information was of poorer quality—i.e. substitutions. In both phrase types, the degraded input must ultimately be reconciled, but for idioms, the nonliteral meaning needs to be computed as well, which can only happen after the unusual orthography was dealt with and resolved. What is unclear between this study and Swinney and Cutler's, however, is what role co-occurrence information plays. In particular, the authors do not discuss the relative frequency or infrequency of their control stimuli. Another study that looks at this question is Jolsvai, McCauley, & Christiansen (2013), where the authors conducted a phrase classification task on frequency-matched idioms and literals. However, the instructions in the later study conflate judgments of grammaticality with meaning, which ultimately leaves this question open: How does semantic opacity affect perceptions of meaning and grammaticality, namely when phrases like idioms and literals are presented in isolation? The design of the Experiment, too, leaves this question open, and I will return to this topic in Experiment 3.

Experiment 2 was conducted to determine how bottom-up and top-down information interact when reading in a second language, namely looking at the influence of semantic opacity or transparency

and bottom-up orthographic cues on second language processing. In Experiment 2a, when the letter manipulation occurred in phrase-final position, L1 Mandarin-L2 English bilinguals did not show any differences in skipping behavior with respect to either factor Letter or Phrase Type. In general, proportions skipped per condition were low across the board, suggesting that, although the letter information was available in the parafovea, the bilinguals were more incremental in processing and seemed to rely more on bottom-up information. First fixation duration did not reveal any main effects, suggesting that neither type of expression nor the quality of bottom-up information affected how long bilinguals fixated the target for the first time. This result, taken together with the skipping data, suggests that differences across stimuli conditions arose in later measures only. In total duration of the phrase-final word, for example, we see both a main effect of Letter and of Phrase Type as well as an interaction, showing that idioms were harder both in the Identity and TL condition. When phrase contained a substitution, however, bilinguals demonstrated comparable processing times across the levels of phrase, suggesting that processing differences between literals and semantically opaque expressions are most evident when the bottom-up input is discernable. After a certain level of degradation, however, it seems both phrase types incur processing difficulties. This difference, in particular, manifested in total duration, which suggests that the impoverished bottom-up information affected downstream, integrative mechanisms the most. Further, total duration measures for the phrase-final word were nuanced with respect to how the different levels of Letter influenced reading. Namely, the greater similarity between Identity and TL targets in literals compared to the comparison within idioms is evidence of how these cross-script bilinguals demonstrated flexible letter position encoding in late processing measures and how this was supported when the target occurred in a literal context. While it is uncontroversial to suggest that idioms would be harder to read, this result demonstrates how processing of degraded bottom-up input can be influenced by the top-down contextual environment. TLs within idioms were more disruptive than TLs in literals. While we know idioms require specific lexical items in certain configurations, this result suggests that orthographic encoding may also be more constraining for idioms, because the content in an idiom may be more likely to have been memorized as part of a chunk. Whole phrase measures also show

the slight advantage a literal environment provided bilinguals when resolving unexpected orthography in reading.

Experiment 2b showed a slightly different pattern than Experiment 2a. For example, while there were no differences in skipping based on the conditions, there was more skipping, overall, than in Experiment 2a. It is possible that this is due to the difference in sample size (compare 60 participants in 2a with 36 in 2b). However, it is also possible that this difference is driven by the context. In Experiment 2a, regardless of the type of expression being read, participants may have been more aware that they were reading a meaningful chunk, leading them to more carefully look for particular features in phrasal completions, ultimately leading to less skipping. In Experiment 2b, the point in the sentence where the target word was located—phrase-initial position of the expression—would not have warranted this, perhaps leading to more flexible reading and less anticipation. While this hypothesis suggests that the environment in which bilinguals would skip more is the opposite of what native speakers demonstrated in Experiment 1, if bottom-up information is relied on more than contextual information, this is a reasonable possibility.

Bilinguals in this experiment also demonstrated nuance in the degree to which letter manipulations influenced reading time. For example, early measures on the phrase-initial word in Experiment 2b revealed a main effect of Letter, where degraded bottom-up input led to longer first fixation durations. While the magnitude of the differences between conditions was not equal (see Table 33 for a review of the model output), bilinguals were keenly sensitive to the degree of letter manipulation in the string in Experiment 2b. This is also found in total duration measures, where there is a difference by Phrase Type when reading TL targets, but this difference is less pronounced the more degraded the input becomes. When orthography of the target word was manipulated along one dimension, bilinguals were better at resolving this unexpected input when the surrounding context is literal, and specifically, they seem to be better at overcoming unexpected letter position information than the native English speakers from Experiment 1, as is demonstrated by the significant interactions in many of the later measures in Experiment 2. However, contrast this with the results from first fixation duration in

Experiment 2a, where bilinguals showed no differences among the levels of letter, suggesting that the duration of the first time people looked at the target containing the manipulation, this same level of sensitivity was missing. The fact that bilinguals did not fixate targets any differently with respect to the levels of letter in early measures suggest that when people were in the middle of an MWE, the constraints on the phrase-final word may have been more relaxed during initial stages of lexical access and identification. In later measures, bilinguals were sensitive to the degree of degradation in the stimulus, suggesting that when the manipulation was in the phrase-final word, the letter manipulation only manifested in differences in late processing, suggesting the challenge was in integration, not access.

Taken together, the findings from Experiments 1 and 2 offer interesting and novel insights to the domains of figurative language processing as well as visual word recognition in first- and second-language reading. For example, for monolinguals, the data were influenced most by the degree of unexpectedness in the orthography. For the most part, when all else is controlled for, the type of phrase was not predictive of reading behavior for monolinguals. Only in later processing measures, such as total duration, when the input was severely degraded (i.e. SUB targets) did idioms show a processing difference, and interestingly, it was a penalty and not the benefit that classic idiom studies would predict. By controlling for many factors that affect prediction and lexical processing, these data demonstrate in a novel way how the relative degree of semantic opacity or transparency affects a person's ability to integrate top-down information with bottom-up cues of varying degrees of quality. Further, for bilinguals, the effects mostly bear out in later measures, and when they do, we see a clear benefit for literal language, consistent with prior work (e.g., Cieslicka, 2006). Knowing that idioms require explicit prior exposure and more computational work in processing, it is reasonable that these expressions incur a penalty when reading in one's second language. However, what is novel about this set of data is that they demonstrate how bottom-up and top-down information interact in bilingual sentence processing, and further under what conditions bilinguals tend to anticipate, more or less, and in what environments they seemingly do not predict, as is evidenced by the patterns observed in native reading of the same sentences. Namely, in early measures, in cases where a person has access to phrase-level information (i.e. Experiment 2a), the

processing penalty incurred by poor quality bottom-up information seems to arise in late processing, as is evidenced by the difference between the levels of Letter both in total duration measures on the word, as well as in measures of the region containing the phrase as a whole. In cases where a context is built up, the bilinguals seem to be attending to this information, using it to build a representation, and only later, in integration, do they realize that the input they saw was anomalous. Conversely, without the help of a surrounding phrasal environment (i.e. Experiment 2b), bilinguals seem to encounter greater difficulty in early measures, particularly on the target word, in cases where the orthography is anomalous. In this case, without a context to facilitate processing, the degraded bottom-up cues incur a processing penalty earlier, where in the absence of a local context, bottom-up information may be more heavily relied on. These data show a clear interaction between bottom-up and top-down information, namely how more semantic transparency and compositionality lead to a better ability to overcome slight changes in misspelling.

Despite the merits of the findings from Experiments 1 and 2, there are limitations to the present set of studies. First, the sample size in Experiments 1b and 2b are less than those in Experiments 1a and 2a. This diminishes the ability to generalize across works, and more importantly, the lack of power limits the ability to see true effects. Due to multiple comparisons, effect sizes were smaller in the data from Experiments 1b and 2b; by including comparable participant groups in both parts of each experiment, this would better elucidate whether the comparisons made were truly insignificant or more a result of insufficient power. By collecting more data in Experiments 1b and 2b, this issue would be alleviated.

Further, in all of the sentences, the target expressions were all preceded by a preamble. While this demonstrates natural reading of idioms and literals when the context supports the interpretation, this design ultimately masks the time course of when a configuration is recognized. Namely, by incorporating sentences where the MWE occurred earlier in the sentence and the preamble postposed to after the expression, we may be able to more clearly see how configuration recognition manifests in natural reading. Even though the current study was designed to examine the interaction between bottom-up and top-down cues in processing, and context is a major top-down cue, the results of this study are ultimately not entirely insightful in this regard.

Further, the design of Experiments 1 and 2 necessitated that the idioms and literals be read in sentence contexts to be able to answer questions about literal and nonliteral language in natural sentence processing. Due to the study design, the results, thus, are not informative with respect to how idioms and literals are processed along another important domain—in isolation. Language is often experienced in context, and while the results from the eyetracking experiments show how differences in semantic opacity manifest in natural reading, isolated presentation is a better method for answering questions of how perceptions of meaning are affected by semantic opacity. Jolsvai and colleagues (2013) endeavored to answer this question in their phrase decision task. By asking participants to make a decision about whether a string was a possible string of the language, the authors inferred from this task that participants found no difference in the meaningfulness of frequency-matched idioms and literals. While this investigation certainly gets at the possibility or grammaticality of a phrase in a language, it is unclear whether perceptions of possibility are conflated with meaningfulness. Ultimately, a phrase decision task is akin to a grammaticality judgment, which does not necessitate activation of the meaning of the string. Rather, a phrase decision requires syntactic knowledge of the language to know which combinations of words are possible. It is unclear from the results of Jolsvai et al. (2013) whether their results were impacted more by syntactic, grammatical knowledge or perceptions of meaning. To test whether semantic opacity affects perceptions of plausibility and meaningfulness in a language, I conducted Experiment 3.

2.5 Experiment 3a – Norming

Participants

Prior to the primary task, 32 participants were recruited via Amazon Mechanical Turk (range: 20-59 years, mean age=33.94; 19 female) for norming. Participants had a range of highest level of education completed; 6 reported having completed high school; 13 completed part of college; 10 completed an undergraduate program, and 3 either were in or completed graduate school. All participants reported English as their native language.

Materials

Norming items were three-word strings from one of three conditions: idioms, literal phrases, or fragment strings. There were 40 items per condition, and each participant saw all of the items in a randomized order. Idioms were selected from Bulkes & Tanner (2017) and were matched for frequency (COCA; Davies, 2008) of the trigram, both bigrams, and each of the three unigrams with literal expressions taken from the same database. Fragment strings were also extracted, which were matched for trigram and bigram frequency with both the idiomatic and literal expressions (see Table 38 for idiom properties; see Table 38 for idiom properties, see Table 39 for descriptive statistics of stimuli frequency, and see Table 40 for norming statistics; see Appendix B for stimuli used in Experiment 3).

Table 38. Properties of the idioms used in Experiment 3

	Frequency	Meaningfulness	Global Decomposability	Literal Plausibility	Predictability
Mean	3.59	4.66	0.58	3.43	0.36
SD	0.52	0.25	0.21	1.18	0.28

N.B. Values obtained from ratings in Bulkes & Tanner (2017), where frequency, meaningfulness, and literal plausibility were Likert scale ratings (1=low, 5=high); global decomposability is expressed as the proportion of people who rated the expression decomposable (No=0, Yes=1); and predictability is the proportion of participants who provided the idiom-final word in a cloze task.

Table 39. Descriptive statistics for trigrams included in Experiment 3

	Trigram	Bigram 1	Bigram 2	Unigram 1	Unigram 2	Unigram 3
Fragments						
Average	5.83	37.90	230.65	841410.23	89613.90	1699594.28
St. Dev	14.88	153.02	978.47	1597776.97	128244.86	2618279.99
Idioms						
Average	9.48	54.08	240.95	136255.40	13411970.65	49016.25
St. Dev	16.76	200.27	1054.02	363201.51	11607282.76	71805.10
Literals						
Average	8.90	63.25	241.88	146828.80	14310849.58	49731.60
St. Dev	16.28	205.82	1056.19	358903.80	10948650.05	73853.92

N.B. Frequency information was obtained from the COCA's N-gram database

Table 40. Paired t-tests, comparing trigram, bigram, and unigram frequencies of stimuli in Experiment 3

	t	df	p-value
Trigram frequency			
Literal v. Idiom	-0.16	77	.88
Idiom v. Fragment	1.03	77	.31
Fragment v. Literal	-0.88	77	.38
Bigram 1 frequency			
Literal v. Idiom	0.20	77	.84
Idiom v. Fragment	0.41	77	.69
Fragment v. Literal	0.63	77	.53
Bigram 2 frequency			
Literal v. Idiom	0.00	77	1.00
Idiom v. Fragment	0.05	77	0.96
Fragment v. Literal	-0.05	77	0.96
Unigram 1 frequency			
Literal v. Idiom	0.13	77	.90
Idiom v. Fragment	-2.72	77	.01
Fragment v. Literal	2.68	77	.01
Unigram 2 frequency			
Literal v. Idiom	0.36	77	.72
Idiom v. Fragment	7.26	77	.01
Fragment v. Literal	-8.21	77	.00
Unigram 3 frequency			
Literal v. Idiom	0.04	77	.97
Idiom v. Fragment	-3.99	77	.01
Fragment v. Literal	3.98	77	.00

Procedure

Participants were asked to rate on a Likert scale how plausible the strings of words were as strings of English (1=highly implausible; 7=very plausible). Participants were told to make their judgments based on their knowledge of English, not on whether they would actually see the word string ever presented in isolation in real life. An example sentence was provided, such that participants were told they should provide high plausibility ratings for expressions like "He read a" and "read a book" because both strings

could appear in "He read a book", but that "He a book" would be rated as implausible as a possible string of English.

Data Analysis

Ten catch trials were included in the norming to ascertain attention to the task, and an a priori threshold of 50% was set as an exclusionary criteria. No participants were excluded due to this. Table 41 shows the descriptive statistics from the norming study.

Table 41. Descriptive statistics from the plausibility norming task

	Mean	SD
Idiomatic expressions	6.42	0.26
Literal expressions	6.34	4.18
Fragment strings	4.18	0.75

Paired t-tests showed the difference between idiom and literal expressions was not significant (t=1.37, p=.17). These items were those used in the judgment tasks described below in Experiments 3a and 3b.

2.6 Experiment 3a

2.6.1 Method

Participants

For Experiment 3a, 150 participants were recruited via Amazon Mechanical Turk (range: 18-66 years old; mean age=35.51 years; 76 female). Participants had a range of highest level of education completed: 21 had completed high school; 56 had completed part of college; 63 had completed an undergraduate program; and 17 were either in or had completed graduate school. All participants reported English as their native language.

Materials

Experimental items were those included in the norming (40 idioms, 40 literal phrases, 40 fragment strings; see Table 42 for example stimuli).

Table 42. Example stimuli from Experiment 3

	Fragments	Idioms	Literals	Idioms, scrambled (filler)	Literals, scrambled (filler)
Examples	its loans as	test the waters	walk a block	The waters test	A block walk
	its soul if	scratch the surface	led the nation	The surface scratch	The nation led
	memory fact that	cover your tracks	set the meeting	Your tracks cover	The meeting set
	my eyelids so	join the club	allow the user	The club join	The user allow
	my mother if	throw a fit	raise the taxes	A fit throw	The taxes raise

There were also 80 fillers distributed throughout the experiment that were the same strings from the idiom and literal conditions but scrambled, such that the second bigram was presented first, and the remaining word was presented in string-final position. To ensure participants were paying attention to the task, 10 catch trials were included, where the sentence "Are you still paying attention?" appeared instead of a word string, and participants were asked to press the opposite button that they had been pressing for "Yes". An a priori threshold of 50% incorrect responses on catch trials was used to exclude participants on the basis of not paying attention; nobody was excluded based on this criteria.

Procedure

The reaction time experiment was built using IbexFarm, a javascript suite for conducting online reaction time experiments. The experiment was hosted on the spellout.net server, and participants were provided a link to the survey via Amazon Mechanical Turk. Participants were asked to turn off all other sources of distraction wherever they were and to place their fingers on the "1" and "2" keys on their keyboard. They could use either the number pad or the numbers above the letters. Upon starting, an instruction screen told participants that they would be asked to read strings of words and decide whether the string could be a possible string of English with a Yes/No button press. Response fingers were counterbalanced across participants, such that half of the participants pressed 1 for "Yes" and 2 for "No", and the other half of

participants used the opposite key assignments. Participants were told to judge all legal strings of English as possible. They were given the examples “went to the” and “to the store” and were told that both should get a “Yes” response, because both could appear in a sentence like “Bob went to the store,” but a string like “Bob to store” would not be possible and that this string should receive a “No” response. All expressions appeared in the center of the screen. The experiment was built so that expressions could be presented in 16-point font, but if participants had alternate viewing (i.e. zooming in/out) settings activated on their personal computer, this would change the size of presentation.

2.6.2 Results

A two-step cleaning procedure was implemented: In the first step, I eliminated trials with RTs equal to or below 150ms, as absolute outliers. Next, I calculated the mean plus or minus 2.5 standard deviations for each person to eliminate relative outliers. Trials outside of this window per participant were excluded from further analysis (634 trials, or 3.04% of the data). Next, I re-computed the mean for each person across all of the conditions, and calculated a new mean and standard deviation. Individuals whose mean RT was beyond three standard deviations of new mean were also excluded from analysis; 3 participants were excluded based on this criterion. The data were then fit to a linear mixed-effects (LME) model. Frequency and norming data were all mean-centered. First, I fit a full model with Phrase Type as a fixed effect, and whole string, both bigram frequencies, and all unigram frequencies as covariates; participant and item were included as random intercepts with no random slopes (due to computational resources). None of bigram 1, bigram 2, unigram 1, unigram 2 and unigram 3 contributed to the model (all ps>.2). A reduced model was then run with Phrase Type as a fixed effect and whole-string frequency as a covariate.

Mean RTs and standard errors are provided in Table 43, which show that idioms, overall, had fastest RTs, literals had longer RTs and fragments had the longest RTs. Mean accuracy percentages are provided in Table 44; only trials where participants were accurate in indicating the phrase was a possible string of the language were included in further analyses. The model output is available in Table 45. Results showed a main effect of Phrase Type, such that people were fastest to say idioms were possible

phrases of English, slower to say literals were possible phrases of English, and slowest to respond to fragments. Whole-string frequency contributed significantly to the model as a covariate (see Figure 9 for visualization of the data). Pairwise comparisons showed that RTs in the idiom condition were shorter, overall, than RTs in the literal condition and in the fragment condition, and that RTs in the literal condition were shorter than RTs in the fragment condition, and that these differences were significant.

Table 43. Mean reaction times per condition for the phrase decision task in Experiment 3a

	Mean	SD
Idiomatic expressions	1160.62 (11.05)	856.79
Literal expressions	1380.40 (15.78)	1178.23
Fragment strings	1790.86 (34.18)	1681.98

Values represented in milliseconds. Values in parentheses indicate standard error of the mean.

Table 44. Accuracy percentages per condition for Experiment 3a

	Mean	SD
Idiomatic expressions	0.98 (0.002)	0.14
Literal expressions	0.92 (0.004)	0.27
Fragment strings	0.48 (0.007)	0.50

Values represented in milliseconds. Values in parentheses indicate standard error of the mean.

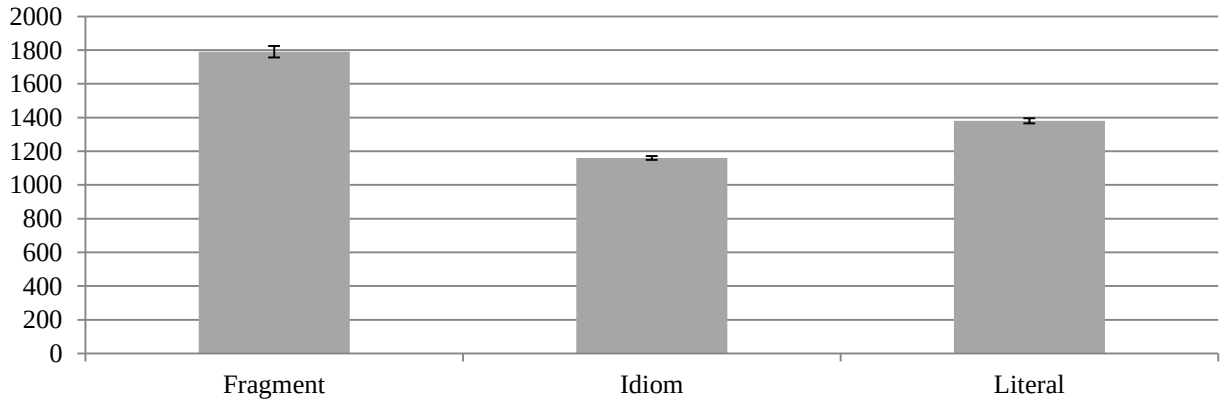

Figure 9. Average reaction times (ms) per condition. Error bars illustrate the standard error of the mean.

Table 45. Model for reaction time data from the phrase decision task in Experiment 3a

Effect	df	X^2	p-value
Phrase Type	2	164.35	<.001
Trigram frequency	1	12.65	<.01

Output constructed using mixed function with "afex" package in R. Phrase Type is a main effect; Trigram frequency is a covariate.

Contrasts	β	St. Error	df	t	p-value
Fragment - Idiom	664.3384	41.74	163	15.92	<.001
Fragment - Literal	440.70	41.77	167	10.55	<.001
Idiom - Literal	-223.64	36.72	119	-6.09	<.001

Output constructed using "lsmeans" package in R. All comparisons are pairwise within the level Phrase Type.

Experiment 3a summary

This task was designed to replicate the phrase decision task done in Jolsvai, McCauley, & Christiansen (2013), to determine whether people's perceptions of a string's legality were related to their perceptions of the a string's meaningfulness. In the aforementioned paper, while the authors describe a meaningfulness judgment task, their instructions indicate the distinction between these two constructs may not have been so clear. Further, the results of the current task—which instructed participants to make a possibility judgment—did not replicate those of Jolsvai et al. (2013), namely that, when controlling for whole-string and substring frequencies, there were no significant differences in the RTs of idioms and literals. Instead, the results of the current study demonstrate that, when these factors are controlled for, participants were faster to say that idioms were possible strings of English than they were to say the same of literal strings, and that they were faster to say both idioms and literals were strings than illegal, fragmented combinations of words. While fragment strings were not matched with idioms and literals with respect to unigram frequency, fragment strings had higher unigram frequencies than both idioms and literals, which suggests that this effect is not driven by the frequency of the lexical components of the strings, alone.

However, despite not replicating the results of Jolsvai and colleagues so far, it is still unclear from the results of Experiment 3a, alone, how or whether perceptions of plausibility in a language are related to or qualitatively different from perceptions of meaning. To test this, I followed up Experiment 3a with

Experiment 3b, using the same stimuli, to understand how directing people to these two notions separately affects the perception of semantically opaque and transparent strings.

2.7 Experiment 3b – Norming

Participants

Prior to the primary task, 30 participants were recruited via Amazon Mechanical Turk (range: 20-65 years, mean age=33.93; 20 female) for norming. Participants had a range of highest level of education completed; 8 reported having completed high school; 9 completed part of college; 12 completed an undergraduate program, and 3 either were in or completed graduate school. All participants reported English as their native language.

Materials

Norming items were the same three-word strings from Experiment 3a, namely those from one of three conditions: idioms, literal phrases, or fragment strings (see Table 40 for a review of the whole-string and substring frequencies for items from each of the three conditions). The same stimuli were used to compare perceptions of plausibility as a string in a language and how meaningful the string is in the language. To do this, the same stimuli used in Experiment 3a were used in Experiment 3b.

Procedure

Participants were asked to rate on a Likert scale how meaningful the strings of words were as strings of English (1=No clear meaning; 7=Very clear meaning). Participants were asked to rate how meaningful each string was as a string of English. They were asked to rate expressions lowly, with a 1, if they would never find the string to be meaningful, and to rate the string highly, with a 7, if they found the string to be a meaningful, easily interpretable string of English. Participants were asked to make their judgment on whether the expression, as it occurred in isolation, was meaningful, not on whether they could conceive of a context in English where it would be meaningful, despite not being meaningful on its own. They were given an example, where in judging "visit the store" and "the store which", they should rate the first with

a 7, as it could occur in isolation, whereas the latter should be rated with a 1, as it requires a context to be meaningful.

Data Analysis

Ten catch trials were included in the norming to ascertain attention to the task, and an a priori threshold of 50% was set as an exclusionary criteria. No participants were excluded due to this; Table 46 shows the descriptive statistics from the meaningfulness norming study. Paired t-tests showed the difference between idiom and literal expressions was significant ($t=3.00$, $p<.01$).

Table 46. Descriptive statistics from the meaningfulness norming task

	Mean	SD
Idiomatic expressions	6.21	0.38
Literal expressions	5.86	0.63
Fragment strings	1.80	0.44

These stimuli are those that were used in Experiment 3b.

2.8 Experiment 3b

2.8.1 Method

Participants

For Experiment 3b, 150 participants were recruited via Amazon Mechanical Turk (range: 20-60 years old; mean age=34.49 years; 65 female). Participants had a range of highest level of education completed: 25 had completed high school; 60 had completed part of college; 60 had completed an undergraduate program; and 7 were either in or had completed graduate school. All participants reported English as their native language.

Materials

Experimental items were those included in the norming (40 idioms, 40 literal phrases, 40 fragment strings), and were the same items included in Experiment 3a. The same 80 fillers from Experiment 3a were also distributed throughout the experiment that were the same strings from the idiom and literal

conditions but scrambled, such that the second bigram was presented first, and the remaining word was presented in string-final position. To ensure participants were paying attention to the task, 10 catch trials were included, where the sentence "Are you still paying attention?" appeared instead of a word string, and participants were asked to press the opposite button that they had been pressing for "Yes". An a priori threshold of 50% incorrect responses on catch trials was used to exclude participants on the basis of not paying attention; nobody was excluded based on this criterion.

Procedure

The reaction time experiment was built and run the same as Experiment 3a. In this task, participants were told that they would be asked to read strings of words and decide whether the string was a meaningful, interpretable string of English with a Yes/No button press. Response fingers were counterbalanced across participants, such that half of the participants pressed 1 for "Yes" and 2 for "No", and the other half of participants used the opposite order. Participants were told to judge strings that were meaningful in English as such; any string that would not be meaningful in isolation—in the absence of a disambiguating context—should be responded to with a "No" response. They were given the examples "read a book" and "a book that" and were told that while the first was meaningful in isolation and should get a "Yes" response, the second would not; even though both could appear in a sentence like "I read a book that I liked," the second string was not meaningful on its own. All expressions appeared in the center of the screen. The experiment was built so that expressions could be presented in 16-point font, but if participants had alternate viewing (i.e. zooming in/out) settings activated on their personal computer, this would change the size of presentation.

2.8.2 Results

The same data cleaning method employed in Experiment 3a was used for Experiment 3b. A two-step cleaning procedure was implemented: In the first step, I calculated the mean plus or minus 2.5 standard deviations for each person. Trials outside of this window per participant were excluded from further analysis (881 trials, or 4.23% of the data). Next, I re-computed the mean for each person across all of the

conditions, and calculated a new mean and standard deviation. Individuals whose mean RT was beyond three standard deviations of new mean were also excluded from analysis; 3 participants were excluded based on this criterion. Frequency and norming data were all mean-centered. First, I fit a full model with Phrase Type as a fixed effect, and whole string, both bigram frequencies, all unigram frequencies, and the meaningfulness norming ratings as covariates; participant and item were included as random intercepts with no random slopes (due to computational resources). None of bigram 1, bigram 2, unigram 1, unigram 2 and unigram 3 contributed to the model (all ps>.2). A reduced model was then run with Phrase Type as a fixed effect and whole-string frequency and the norming ratings as covariates.

Mean RTs per condition are available in Table 47. The data show the shortest RTs for idioms, longer RTs for literal expressions, and the longest RTs for fragments. Accuracy percentages are shown in Table 48; only accurate trials were included in analysis. The model output is provided in Table 49. Results showed a main effect of Phrase Type, such that reaction times were fastest for idioms, slower for literals, and the slowest for fragments (see Figure 10Figure 10. Average reaction times (ms) per condition. Error bars illustrate the standard error of the mean. for visualization of the data). Pairwise comparisons within the factor Phrase Type showed that RTs in the idiom condition were shorter, overall, than RTs in the literal condition and in the fragment condition, but that RTs in the literal condition were not significantly shorter than the RTs in the fragment condition.

Table 47. Mean reaction times per condition for the meaningfulness judgment task in Experiment 3b

	Mean	SD
Idiomatic expressions	1082.76 (9.81)	737.35
Literal expressions	1302.79 (11.51)	815.69
Fragment strings	1339.33 (15.69)	982.01

Values represented in milliseconds. Values in parentheses indicate standard error of the mean.

Table 48. Accuracy percentages per condition from Experiment 3b

	Mean	SD
Idiomatic expressions	0.97 (0.002)	0.18
Literal expressions	0.90 (0.002)	0.34
Fragment strings	0.80 (0.003)	0.40

Values represented in milliseconds. Values in parentheses indicate standard error of the mean.

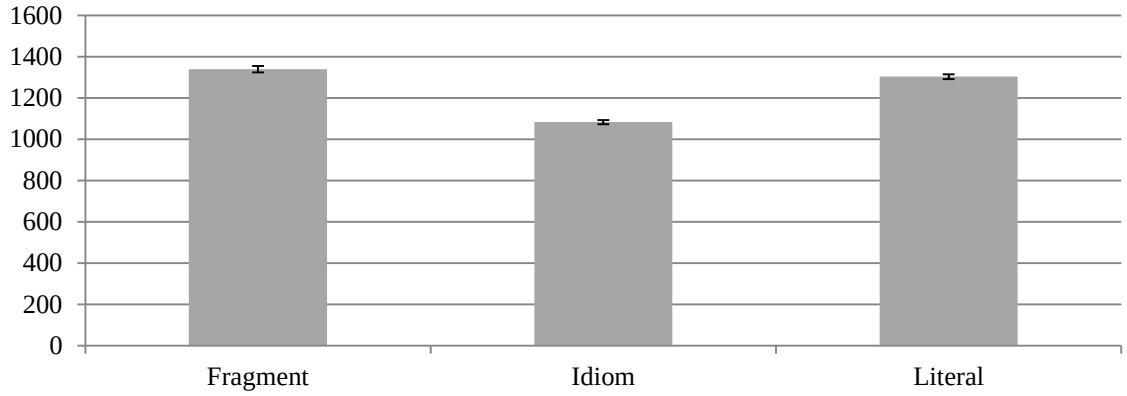

Figure 10. Average reaction times (ms) per condition. Error bars illustrate the standard error of the mean.

Table 49. Model for reaction time data from the meaningfulness judgment task in Experiment 3b

Effect	df	X^2	p-value
Phrase Type	2	54.00	<.001
Trigram frequency	1	7.47	<.01
Norming ratings	1	23.91	<.001

Output constructed using mixed function with "afex" package in R.

Contrasts	β	St. Error	df	t	p-value
Fragment - Idiom	278.57	29.19	119	9.54	<.001
Fragment - Literal	53.09	29.47	124	1.80	.17
Idiom - Literal	-225.48	27.70	114	-8.14	<.001

Output constructed using "lsmeans" package in R.

Experiment 3 was designed to examine how perceptions of a phrase's plausibility in a language differ from the perceptions of its meaningfulness, specifically focusing on expressions that varied with respect to semantic opacity. To accomplish this, wording was used in the instructions prior to each norming and each reaction time study to direct participants' attention to, first, how plausible the string was as a possible string of English, and second, to how meaningful the string was.

In both experiments, reaction time data show a clear advantage for idioms compared to literals, and this comparison is of theoretical interest. Ultimately, both tasks failed to replicate the results of Jolsvai et al. (2013), which showed no difference between idioms and literals. There are a few possible explanations for this. It is possible there was more variability in the current set of stimuli—although, paired t-tests showed that the frequency differences between idioms and literals were not significantly different. Second, the instructions for the tasks in Experiment 3 were purposefully articulated to

distinguish plausibility from meaningfulness, something Jolsvai and colleagues did not do. Further, their participants were asked to make a possibility judgment about the expressions they saw, yet the authors deduced from this data that people were equally able to activate the meanings of idioms and literals. Further, in each of the reaction time tasks in Experiment 3, there were 150 people with a wide range of ages and educational backgrounds, compared to the 40 undergraduates who participated in the 2013 study. First, the current analysis has greater statistical power, and with more power comes better test sensitivity. Second, it is possible that the participants from the current dataset had greater overall experience with the input. With simply more time amassed speaking the language, these speakers may have been more familiar with configurational expressions, like idioms, and this is what led to the idiom advantage here. Further, Turkers offer an advantage compared to college undergraduates, such that more variability is achieved within the sample, something that is harder to do when recruiting on a college campus.

Ultimately, the results of Experiment 3 support an idiom advantage for expressions presented in isolation. As participants were faster to indicate idioms were both possible as well as meaningful compared to literal expressions, this set of results supports prior work attesting that idioms can be more quickly recognized than literals (i.e. Swinney & Cutler, 1979). Further, these results support the notion that frequency is not the only determining factor in the ease or difficulty with which a phrasal meaning is accessed, contrary to studies supporting frequency as crucial in this (e.g. Arnon & Snider, 2010; Jolsvai, et al. 2013). As other studies have suggested, it is possible that, when presented in isolation, the configuration of the idiom is what makes it so recognizable. Literal collocations also have a configuration, but as I argued in the discussion of Experiments 1 and 2, literal expressions are more open-ended, such that synonyms of expected completions will also form felicitous combinations; in idioms, this does not work. In order to activate the idiomatic meaning, idioms have to occur in a particular order with specific component parts. When participants saw a recognizable configuration onscreen, this may be what made idioms faster to respond to as both plausible—as people may have experienced them before—and meaningful—as idioms carry a specific meaning as a phrase.

3. Discussion

In Experiment 1, I investigated how bottom-up and top-down cues in processing work together to facilitate prediction, namely with respect to orthographic, visual cues and semantic opacity. In Experiment 2, I asked how reading in a second language would be affected by the relative semantic opacity or transparency of expressions, and how the quality of the visual input would impact this. In Experiment 3, I studied how semantic opacity affects perceptions of plausibility and meaningfulness judgments in a language, and how semantic opacity affects meaning retrieval.

Starting with prediction, the data from Experiments 1 and 2 show little evidence that either idioms or literals are more predictive of reading environments compared to the other. Namely, the skipping data from Experiment 1a show that while native English speakers are sensitive to the level of degradation of the input in the parafovea, they do not plan eye movements any differently when reading an idiom compared to a literal collocation. Despite the notion that idioms may be represented as chunks in the lexicon, these results do not support idioms as a more predictive environment compared to literals. Namely, the difference in results from Experiment 1a to 1b suggest that skipping behavior may be more loosely supported by the knowledge that a person knows she is reading a meaningful expression. Even without knowledge of what a particular expression means, syntactic knowledge alone would more highly constrain what completions are possible in an expression, and this difference is demonstrated by the skipping data in Experiment 1. Further, for some speakers, this may have been compounded by co-occurrence information, where in the case of a familiar configuration, a person might have even more specific expectations for a kind of lexical item. In the absence of a local context—an MWE—to guide expectations for a phrasal completion, it is reasonable to suggest that people were not predicting at all, and were more loosely comprehending the incoming input, which would explain the similarity among the three levels of Letter in first fixation duration in Experiment 1b. In Experiment 1a, even though the constraint of the sentences, overall, was not considered highly constraining up to the target word, the phrase-final word would have made sense as a phrasal completion. In Experiment 1b, with more possible

ways to continue the sentence, this would make processing more coarse-grained, ultimately leading to comparable rates of skipping across conditions.

Interestingly, the results of Experiment 3 present a different picture than what was seen in Experiments 1 and 2. In Experiment 1, processing of idioms and literal collocations was comparable in cases where the input was intact. Overall, eye movement measures demonstrated that processing of idioms and literals diverged only in late processing and only in cases where the bottom-up input was more degraded (i.e. SUB targets). In Experiment 2, L1 Mandarin-L2 English bilinguals demonstrated an idiom penalty, which manifested most clearly in downstream, integrative processing. However, bilinguals showed the biggest benefit when processing literal expressions in cases where the visual input was either intact or slightly misspelled. In Experiment 3, there was an idiom benefit compared to literal expressions during phrase and meaningfulness judgment tasks, and a possible explanation for this difference may be how the method impacted results. While the meaningfulness judgment task, in particular, provided a window into the relative ease or difficulty of activating the meanings of the phrases presented, judgment tasks have much less temporal resolution than eyetracking. Whereas a judgment task yields an end result from participants (i.e. reaction time, accuracy), eyetracking allows the study of multiple stages of processing, where researchers are afforded a more nuanced view into how a linguistic representation unfolds over time.

Another reason for this difference might be the presentation method. In both Experiments 1 and 2, participants read sentences containing these expressions for as long as they wanted to, with the ability to go back and revisit anomalous parts of the sentence if needed. Each stimulus contained a sentence preamble with context that fit with the meaning of the expression. In Experiment 3, however, expressions were presented in isolation, and participants were asked to respond as quickly as possible in order to investigate the processing load that comes with identifying different types of strings as possible or meaningful in the language. In the absence of a surrounding context, all that the participants had access to when making decisions were the individual lexical items and the configuration of those items. This explanation would support idioms as unitarily represented in lexicon; when presented in isolation, idioms

are arguably more word-like than literals and their meanings can be retrieved as such. Compared to idioms, literal expressions have more flexibility; in literal language synonyms of an expected word can usually convey the same message. This may be another reason why reaction times diverged between the phrase types.

However, the data in this dissertation suggest additional factors—i.e. context—play a larger role in determining how idioms are processed, not just the idiom-specific properties. For example, as Experiment 3 shows, idioms can be identified and processed in isolation when a person acknowledges the configuration of the lexical items as significant. However, when in the context of a longer sentence, compositional analysis may play a larger role, as that is the mode of natural reading, and this is demonstrated by both Experiments 1 and 2. Findings from the three studies ultimately support theories of idiom comprehension that say idioms are compositionally analyzed (e.g. Siyanova-Chanturia et al. 2011) as well as theories that posit more chunk-like retrieval (e.g. Cacciari & Tabossi, 1988). The data demonstrate that both routes to comprehension are possible, and that it depends on the context which processing strategy is most appropriate and ultimately used. For example, in L2 reading, bilinguals demonstrated an idiom penalty and overcame letter position manipulations more easily when reading literals compared to idioms. In Experiment 1, native speakers also showed greater difficulty with idioms in cases where the visual input was severely impoverished (i.e. targets containing substitutions).

Going back to prediction, it does not seem to be the case that idioms facilitate prediction any differently from literals, as is evidenced by the skipping data in Experiment 1a. Participants planned eye movements comparably across phrase types with respect to the information available in the parafovea, where differences in skipping were driven by visual information, not phrase-level semantic opacity or transparency. Further, the quality of the information in the parafovea also affected the ease of identifying and accessing the meaning of a target, as was demonstrated in Experiment 1a's first fixation duration measures. When reading in a supportive context, higher quality information in the parafovea supported native readers in early processing, and the more degraded the information, the less this facilitated initial processing of the target.

While idioms in the current project did not prove to be more predictive environments than literals, results did demonstrate the influence of recognizing a configuration as meaningful in idiom processing. This is supported by the data from Experiment 1, namely the difference in total duration measures between idioms and literals when the target contained a substitution. When the visual input was only slightly degraded (i.e. targets with transpositions), monolinguals showed no differences between idioms and literals. However, when the target contained a substitution, idioms incurred a greater penalty, suggesting there was something impactful about significantly impoverishing the visual input in an idiom compared to a literal. This pattern of results concurs with the notion that recognizing an idiom's configuration is important in processing. It may be that while semantic opacity does not differentiate idioms from literals in natural reading, when things like length and frequency are controlled for, it might be that the configuration of an idiom is more recognizable than the configuration of a literal, and that this recognizability can facilitate processing. In a natural reading environment, like in Experiments 1 and 2, idioms with Identity targets did not show a benefit over literals, as in both conditions, the expressions were supported by their contexts. In the TL condition, it is possible that a certain degree of degradation in the visual input is permissible without incurring a processing burden, but once this degradation surpasses a certain threshold, the more fixed nature of idioms is what drives the penalty in the SUB condition. The results from Experiment 3 support this, too; in the absence of context or anomalous visual features, the configuration seems to be what sets idioms apart from literals in the phrase decision and meaningfulness judgment tasks. In isolated presentation, the configuration may be more salient for idioms compared to when presented within a sentence.

While the idioms in Experiment 3 yielded faster reaction times than frequency-matched literals, it would be interesting to compare the same idioms to highly frequent literal strings. This would help to better understanding whether a literal expression can be represented as a configuration, like an idiom. To do this, expressions would be needed that co-occur much more frequently than the current set. Ultimately, while idioms are known expressions to some, they are comparatively infrequent in the language. Biber and colleagues (1999) suggest a phrase with a frequency of 10 per million is a good candidate for

representation in the lexicon; analysis requiring highly frequent expressions would do well to consider this as a potential threshold when considering items for inclusion in a stimulus list. In Experiment 3, it is possible that the idiom benefit is driven by the recognizability of the idiomatic configurations. As the literal expressions included in the current study were well below Biber's threshold, more frequent literal expressions that exhibit more of a chunk-like representation due to whole-string frequency should demonstrate the recognizability of a literal configuration. While Arnon & Snider (2010) would suggest that frequency is the dominating factor in determining lexical processing ease, Experiment 3 suggests there may be something, too, to be said for recognizability. It is possible that, with frequency, comes recognizability; however, Experiment 3 results seem to suggest there is something more recognizable about idioms, when whole-string and substring frequencies across idioms and literals are matched. Further work would help to disambiguate this relationship.

With respect to L2 processing of idioms, it is possible that cross-linguistic influence may have played a role in the results from Experiment 2. Namely, the simple notion that participants may have been unfamiliar with the expressions in the language could have influenced results, where the penalty for idioms may be due to a greater processing burden when reading idioms but it may also be due to simple unfamiliarity with the nonliteral phrases. Further, this is not only a possibility for the L2 learners but also for the native speakers, who without sufficient experience with these forms may also not have understood the meanings of some of the idioms. To gauge individual familiarity with the idioms used, all participants completed an exit survey at the end of the experimental session, where they were asked to rate on a scale of 1 to 5 (1=low, 5=high) how familiar they were with idioms used in the sentence processing task. However, while this Likert-scale rating system is analogous to that what is typically used in norming studies to gauge familiarity and subjective experience, this ultimately does not capture whether participants are actually familiar with an expression. For example, future research in this domain concerned with the familiarity of idiomatic forms may be better executed using a paraphrase task to investigate this. For example, participants could provide short paraphrases or definitions of what they think the expressions mean. While this would require additional work in data analysis—for example,

having to employ naïve raters to code responses manually—this would more clearly demonstrate knowledge of the phrase rather than a person's impression of how well they think they know the phrase in question. Such a task would not only provide insight into how familiar L2 speakers are with idioms but also the native speakers, as this is not just an issue for analysis of nonnative speaker data but also for data from native speakers.

Sidestepping the familiarity issue, the language processing question still remains: Is L2 processing of idioms qualitatively different from L1 processing of idioms? How these phrases are processed—namely the mode of processing—is at the heart of this question, and the results from Experiments 1 and 2 are insightful here. Specifically, the results support compositional analysis as the default processing mode employed by both L1 and L2 speakers in natural sentence reading. Greater exposure to the input over the lifespan led to no differences in reading between idioms and literals in native English speakers when the target either appeared as expected or contained a transposition. For L2 speakers, there was a penalty for idioms across the board. While compositional analysis is sufficient for literal language processing, it is insufficient for processing many idioms, yet the native speaker data show no differences between the idioms and literals, particularly in cases where the phrase contained an Identity target. This suggests that native speakers tapped into some other kind of knowledge in order to process the idioms with relatively the same amount of ease as the literals, something that the bilinguals showed less of—as evidenced by the idiom penalty. It is reasonable to attribute this to language experience, as this is the one systematic difference across the two groups in Experiments 1 and 2. Even if the L2 speakers were familiar with all of the idioms—something that, in actuality, varied across participants and items—reading an idiom in the L2, overall, seems to be harder. Where a native speaker may be able to deduce from compositional analysis that an expression is nonliteral, a nonnative speaker may be more likely to assume literality of a phrase and only when that interpretation is infelicitous with the context, would the person revise and entertain other possible interpretations. Here, only in cases of prior experience with the phrase would the person be able to activate the appropriate nonliteral meaning.

Ultimately, by acknowledging the colloquial, language-specific nature of idioms, it would be reasonable to suggest that idioms be harder to process in a second language compared to a first, native language.

It would also be interesting to conduct the phrase decision and meaningfulness judgment tasks with second-language learners to further study how the mode of presentation impacts processing. Results from prior studies (i.e. Conklin & Schmitt, 2008; Siyanova-Chanturia et al, 2011) and those from Experiment 2 suggest that compositional analysis is a reliable route used in L2 processing. However, as the results from Experiment 3 suggest, the mode of presentation might affect which processing strategies learners choose to implement. In a classroom setting, for example, learners are aptly able to recognize when an expression is a "colloquialism" or a phrasal expression with a meaning more than the sum of its parts. Although L2 users may primarily use compositional analysis in sentence reading, isolated presentation may elicit different results, which would also provide further nuance to the compositional-first strategy. If compositional analysis is always used, regardless of presentation rate, then L2 users should demonstrate comparable reaction times to frequency-matched idiomatic and literal strings. If, however, the presentation mode is key, and isolated presentation draws more attention to the configuration and relative recognizability of expressions, then L2 users, too, should show an advantage when making timed decisions about the meaningfulness of idioms. A phrase decision task, on the other hand, might elicit more grammaticality judgments, where idioms and literals may not show any difference. However, as frequency is a language-specific metric, such a task would require a subsequent task, asking participants to indicate familiarity with all of the experimental items—both idioms and literals—to account for individual variation in input exposure.

4. Conclusion

The current set of data is rich, and the analyses described here do not exhaust the ways in which this data could be used to better understand how idioms are processed. For example, by using a LME model to fit the data, the ratings from Bulkes & Tanner (2017) could be added to a model as fixed effects to determine whether and how dimensions like idiom familiarity, global decomposability, literal plausibility, and predictability affect reading behavior. Specifically, it is possible that idioms that were rated as more familiar demonstrate more of a penalty when read with degraded visual cues. Further, in Experiment 3, it is possible that these dimensions might also predict reaction times. Additionally, the stimuli from Experiment 1 and 2 also have constraint data available, namely the ratings obtained from the norming prior to the study. While whole-string and unigram frequencies were matched across idiom and literal lists, some of the sentences in either condition were more constraining than others. It would be insightful to see if constraint impacts things like skipping behavior, and whether semantic opacity differentiates this when comparing across phrases.

The findings from this dissertation illustrate a nuanced picture of idiom comprehension and the factors that impact their processing. Namely, the data show how bottom-up and top-down information can influence predictive mechanisms when reading literal and nonliteral collocations, and how the degree of degradation in the bottom-up visual stream affects processing when reading semantically transparent and semantically opaque strings. By showing how idioms can be both compositionally analyzed as well as identified and retrieved as chunks, this dissertation supports a dual-route processing model of idiom comprehension. Finally, by showing how both routes are possible—albeit, depending on the context—idiom scholars can have their cake and eat it too.

5. References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*, 278-289.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247-264.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth–Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development, 7*(2), 107-129.

Arnon, I., & Priva, U. C. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and Speech, 56*(3), 349-371.

Arnon, I., & Priva, U. C. (2014). Time and again: The changing effect of word and multiword frequency on phonetic duration for highly frequent sequences. *The Mental Lexicon, 9*(3), 377-400.

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

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology, 17*, 364-390.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning the effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*(3), 241-248.

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.

Beeman, M. (1998). Coarse semantic coding and discourse comprehension. In M. Beeman & C. Chiarello (Eds.), *Right hemisphere language comprehension: Perspectives from cognitive neuroscience* (pp.255-284). Mahwah, NJ: Erlbaum.

Beeman, M., & Chiarello, C. (Eds.). (1998). Right hemisphere language comprehension: Perspectives from cognitive science. Mahwah, NJ: Erlbaum.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). MIT Press.

Blanchard, H. E., Pollatsek, A., & Rayner, K. (1989). The acquisition of parafoveal word information in reading. *Perception & Psychophysics, 46*, 85-94.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from exemplars. *Linguistic Review, 23*, 291-320.

Boland, J. (2005). Cognitive mechanisms and syntactic theory. In A. Cutler (ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 23–42). Mahwah, NJ: Lawrence Erlbaum.

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

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135-149.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991-997.

Bulkes, N. Z. & Tanner, D. (2017). Going to town: Large scale norms and statistical analysis of 870 English idioms. *Behavior Research Methods, 49*(2). 772-783.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425-455.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change*, *14*(3), 261-290.

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

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J., & Scheibman, J. (1999). The effect of usage on degrees of constituency: the reduction of don't in English. *Linguistics*, *37*(4), 575-596.

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

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, *42*(4), 368-407.

Chomsky, N. (1965). Aspects of the theory of syntax Cambridge. *Multilingual Matters: MIT Press*.

Cieslicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research, 22*, 115-144.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*(1), 3-42.

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

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61.

Coulson, S., & Lovett, C. (2004). Handedness, hemispheric asymmetries, and joke comprehension. *Cognitive Brain Research, 19*, 275-288.

Coulson, S., & Williams, R. F. (2005). Hemispheric asymmetries and joke comprehension. *Neuropsychologia, 43*, 128-141.

Cowie, A. P. (1998). Phraseological dictionaries: some East-West comparisons. *Phraseology, Theory, Analysis, and Applications*, 209-228.

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

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: top-down influences on the interface between audition and speech perception. *Hearing research*, *229* (1), 132-147.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369* (1634), 20120394.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*, 1117–1121.

DeLong, K. A., Quante, L., & Kutas, M. (2014). Predictability, plausibility, and two late ERP positivities during written sentence comprehension. *Neuropsychologia, 61*, 150-162.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Language, Cognition and Neuroscience*, 1-8.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New ideas in psychology*, *25*(2), 108-127.

Dikker, S., Rabagliati, H., & Pylkkänen, L. (2009). Sensitivity to syntax in visual cortex. *Cognition, 110*(3), 293-321.

Dikker, S., Rabagliati, H., Farmer, T. A., & Pylkkänen, L. (2010). Early occipital sensitivity to syntactic category is based on form typicality. *Psychological Science, 21*(5), 629-634.

Duñabeitia, J. A., Dimitropoulou, M., Grainger, J., Hernández, J. A., & Carreiras, M. (2012). Differential sensitivity of letters, numbers, and symbols to character transposition. *Journal of Cognitive Neuroscience, 24*, 1610-1624.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody picture vocabulary test: Manual*. Pearson.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641-655.

Ellis, N. C. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(1), 234-250.

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

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use? In S. Th. Gries & D. S. Divjak (Eds.) *Frequency effects in language learning and processing* (Vol. 1). (pp. 7-34). Berlin: Mouton de Gruyter.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014). Second Language Verb-Argument Constructions are Sensitive to Form, Function, Frequency, Contingency, and Prototypicality. *Linguistic Approaches to Bilingualism, 4*(4), 405-431.

Ellis, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179-211.

Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proc. National Academy of Sciences of the USA, 103*, 12203-12208.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*(4), 469-495.

Federmeier, K. D., McLennan, D. B., Ochoa, E., & Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology, 39*(2), 133-146.

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84.

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*(3), 149-161.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11-15.

Fine, A. B., Jaeger, T.F., Farmer, T.A., Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE 8*(10): e77661.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *The Quarterly Journal of Experimental Psychology, 39*(2), 211-251.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66*(1), 226-248.

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1461-1469.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *The Quarterly Journal of Experimental Psychology A, 50*(1), 119-148.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from Eye Movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1366-1383.

Frisson, S., & Pickering, M., J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. *Metaphor & Symbol, 16*(3&4), 149-171.

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, & Cognition, 31*, 862-877.

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

Gibbs, R. W., & Gonzales, G. P. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition, 20*(3), 243-259.

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

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: Chicago University Press.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences, 7*(5), 219-224.

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

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Process, 23*, 1-35.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In: Cowie, A. P. (Ed.), *Phraseology: Theory, analysis, and applications*. Oxford University Press, pp. 145-160.

Green, D., & Meara, P. (1987). The effects of script on visual search.*Second Language Research, 3*(2), 102-113.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Language Research, 28*(2), 191-215.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*(5669), 438-441.

Hasher, L., & Chromiak, W. (1977). The processing of frequency information: an automatic mechanism? *Journal of Verbal Learning and Verbal Behavior, 16*(2), 173-184.

Haynes, M., & Carr, T. H. (1990). Writing system background and second language reading: A component skills analysis of English reading by native speaker-readers of Chinese.

Holsinger, E. (2013). Representing idioms: Syntactic and contextual effects on idiom processing. *Language and Speech, 56*(3), 373-394.

Hoover, M. L., & Dwivedi, V. D. (1998). Syntactic processing by skilled bilinguals. *Language Learning, 48*(1), 1-29.

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research, 22*(3), 369-397.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain research, 1626*, 118-135.

Ito, A., Martin, A. E., & Nieuwland, M. S. (2016). How robust are prediction effects in language comprehension? Failure to replicate article-elicited N400 effects. *Language, Cognition and Neuroscience*, 1-12.

Jackson, C. (2008). Proficiency level and the interaction of lexical and morphosyntactic information during L2 sentence processing. *Language Learning, 58*(4), 875-909.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAS (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59*, 434-446.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*(3), 433-445.

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

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism, 4*(2), 257-282.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language*, *49*(1), 133-156.

Katz, A. N., & Ferretti, T. R. (2001). Moment-by-moment reading of proverbs in literal and nonliteral contexts. *Metaphor and Symbol, 16*(3&4), 193-221.

Katz, A. N., & Ferretti, T. R. (2003). Reading proverbs in context: The role of explicit markers. *Discourse Processes, 36*(1), 19-46.

Keating, G. D. (2009). Sensitivity to Violations of Gender Agreement in Native and Nonnative Spanish: An Eye-Movement Investigation. *Language Learning, 59*(3), 503-535.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kliegl, R., Grabner, E., Rolfs, M., & Engbert, R. (2004). Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology, 16*, 262-284.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, *80*(4), 450-460.

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

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, *11*(5), 539-550.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(12), 161-163.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual Review of Psychology*, *62*, 621.

Lenth, R. V. (2016). Least-Squares Means: The R Package lsmeans. *Journal of Statistical Software*, *69*(1), 1-33.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126-1177.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*, 193-198.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, *63*, 447-464.

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

Libben, M. R., & Titone, D. A. (2009). Bilingual lexical access in context: evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(2), 381.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. *Eye guidance in reading and scene perception*, 55-75.

Lowder, M. W., & Gordon, P. C. (2013). It's hard to offend the college: Effects of sentence structure on figurative-language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 993-1011.

Luke, S. G., & Christianson, K. (2012). Semantic predictability eliminates the transposed-letter effect. *Memory & Cognition*, *40*(4), 628-641.

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676.

MacWhinney, B. (1997). Second language acquisition and the competition model. *Tutorials in Bilingualism: Psycholinguistic Perspectives*, 113-142.

MacWhinney, B. (1998). Models of the emergence of language. *Annual Review of Psychology*, *49*(1), 199-227.

MacWhinney, B. (2001). The competition model: The input, the context, and the brain. *Cognition and Second Language Instruction*, 69-90.

Martin, C. D., Thierry, G., Kuipers, J., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language, 69*, 574-588.

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

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience, 7*(7), 703-704.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, *2*(3), 217-228.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A corpus based approach.* Oxford: Clarendon Press.

Morgan, E., & Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In *CogSci*, 1649–1654.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition, 157*, 384-402.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 92-103.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165.

Nelson, M. J., Brown, J. I., & Denny, M. J. (1960). *The Nelson-Denny Reading Test: Vocabulary, Comprehension, Rate*. Houghton Mifflin.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins.

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

Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*, 97-113.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication*. London, New York: Longman, 191-226.

Perea, M., & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita & S. J. Lupker (Eds.), Masked priming: State of the art (pp. 97-120). Hov, U. K.: Psychology Press.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language, 51*, 231-246.

Perea, M., Duñabeitia, J. A., & Carreiras, M. (2008). Transposed-letter priming effects for close versus distant transpositions. *Experimental Psychology, 55*, 397-406.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*(3), 105-110.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*, 329-392.

Pinker, S. (1998). Words and rules. *Lingua, 106*(1), 219-242.

Pinker, S. (1999). How the mind works. *Annals of the New York Academy of Sciences, 882*(1), 119-127.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*(11), 456-463.

Pollio, H., Barlow, J., Fine, H., & Pollio, M. (1977). *Metaphor and the poetics of growth: Figurative language in psychology, psychotherapy and education*. Hillsdale, NJ: Erlbaum.

Rayner, K. (1975). Parafoveal identification during a fixation in reading. *Acta Psychologica, 39*, 272-282.

Rayner, K., Well, A. D., Pollatsek, A., & Bertera, J. H. (1982). The availability of useful information to the right of fixation in reading. *Perception & Psychophysics, 31*, 537-550.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191-201.

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

Rayner, K., White, S. J., Johnson, R. L., & Liversedge, S. P. (2006). Raeding wrods with jubmled lettres there is a cost. *Psychological Science, 17*(3), 192-193.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr., C. (2012). *Psychology of Reading*. Psychology Press.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review, 16*, 1-21.

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.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6*(9), 382-386.

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid. *Formulaic sequences: Acquisition, processing and use*, 127-51.

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes*, *19*(3), 333-367.

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

Schweigert, W. A., & Moates, D. R. (1988). Familiar idiom comprehension. *Journal of Psycholinguistic Research, 17*, 281-296.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14*(03), 369-385.

Seidenberg, M. S. (1994). Language and connectionism: The developing interface. *Cognition*, *50*(1), 385-401.

Shapiro, B. J. (1969). The subjective estimation of relative word frequency. *Journal of verbal learning and verbal behavior*, *8*(2), 248-251.

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

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251-272.

Siyanova-Chanturia, A., Conklin, K., & Van Heuven, W. J. (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.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302-319.

Sosa, A. V., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of. *Brain and Language*, *83*(2), 227-236.

Stilp, C. E., & Kluender, K. R. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences*, *107*(27), 12387-12392.

Stites, M. C., Federmeier, K. D., & Christianson, K. (2016). Do morphemes matter when reading compound words with transposed letters? Evidence from eye-tracking and event-related potentials. *Language, Cognition, & Neuroscience, 31*(10), 1299-1319.

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.

Tanner, D., & Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia, 56,* 289-301.

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

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

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

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

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules* (Vol. 1950). Cambridge, MA: MIT Press.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. *Perspectives on formulaic language: Acquisition and communication,* 151-173.

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.

Tremblay, A., & Tucker, B. V. (2011). The effects of N-gram probabilistic measures on the recognition and production of four-word sequences. *The Mental Lexicon*, *6*(2), 302-324.

Tulving, E., & Gold, C. (1963). Stimulus information and contextual information as determinants of tachistoscopic recognition of words. *Journal of Experimental Psychology*, *66*(4), 319.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition, 4*(02), 105-122.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience, 2*(10), 717-726.

Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of Psycholinguistic Research, 30*(1), 37-69.

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

Van Assche, E., Drieghe, D., Duyck, W., Welvaert, M., & Hartsuiker, R. J. (2011). The influence of semantic constraints on bilingual word recognition during sentence reading. *Journal of Memory and Language, 64*(1), 88-107.

Van Berkum, J. J. A., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443-467.

Van Gompel, R. P., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language, 52*(2), 284-307.

Van Lancker, D., Canter, G. J., & Terbeek, D. (1981). Disambiguation of ditropic sentences: acoustic and phonetic cues. *Journal of Speech, Language, and Hearing Research, 24*(3), 330-335.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 394.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*(2), 176-190.

White, S. J., Johnson, R. L., Liversedge, S. P., & Rayner, K. (2008). Eye movements when reading transposed text: the importance of word-beginning letters. *Journal of Experimental Psychology: Human Perception and Performance, 34*(5), 1261.

Wicha, N. Y. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not Pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters,* 346, 165–168.

Wicha, N. Y. Y., Moreno, E. M., & Kutas, M. (2004). Anticipating words and their gender: an event-related brain potential study of semantic integration, gender expectancy, and gender agreement in Spanish sentence reading. *Journal of Cognitive Neuroscience, 16*(7), 1272-1288.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460.

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuroimage, 62*(1), 356-366.

Wlotko, E. W., & Federmeier, K. D. (2015). Time for prediction? The effect of presentation rate on predictive sentence comprehension during word-by-word reading. *Cortex, 68*, 20-32.

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.

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

APPENDIX A: Cloze probability for stimuli used in Experiments 1 and 2

Phrase type	Expression	Sentence	Cloze
Idiom	Broaden one's horizons	Jessica wanted to try new things and broaden her horizons by taking new classes.	0.90
Idiom	Bury the hatchet	After their argument, the sisters decided to make up and bury the hatchet that morning.	0.94
Idiom	Came on strong	Drunk and inappropriate, Jake's advances toward women came on strong that evening.	0.63
Idiom	Catch someone off guard	Despite of all his planning, Tim was caught off guard and surprised about the unexpected events.	0.90
Idiom	Coin the phrase	Loving the new slogan, the store wanted to coin the phrase for their campaign.	0.77
Idiom	Come down hard	Disappointed, the strict mother would come down hard with her discipline.	0.77
Idiom	Come to one's senses	Believing the earth was flat, the student had to come to her senses and see the truth.	0.90
Idiom	Count one's blessings	Making ends meet, the poor family would count their blessings before asking for more.	0.60
Idiom	Cover a lot of ground	The epic documentary would cover a lot of ground before rolling the credits.	0.55
Idiom	Crack a joke	To begin his routine, the comedian would crack a joke to get people laughing.	0.87
Idiom	Cramp someone's style	An unyielding woman, Betty's mom would cramp her style by grounding her.	0.87
Idiom	Cross her mind	Despite the mess, cleaning Barry's room didn't cross his mind as he went to bed.	0.94
Idiom	Cross one's fingers	Optimistic, the group of friends would cross their fingers to hope to win the lottery.	0.53
Idiom	Cross the line	An unjust man, the unethical boss would always cross the line before getting caught.	0.94
Idiom	Cross your path	If a Calico cat were to cross your path that wouldn't be considered bad luck.	0.90
Idiom	Draw the line	After working eighty hours and exhausted, William had to draw the line and go home.	0.84
Idiom	Drive a hard bargain	The saleswoman wouldn't negotiate and would drive a hard bargain to make shrewd deals.	0.90
Idiom	Drive someone crazy	Heather's kids' constant yelling and whining would drive her crazy for the rest of the day.	0.63
Idiom	Fall off the wagon	Sober for three years, Gary decided he wouldn't fall off the wagon for his health.	0.84
Idiom	Feeling under the weather	Suffering from a high fever, Flora was feeling under the weather that morning.	1.00
Idiom	Fighting a losing battle	Despite needing to pass physics, Andrew was fighting a losing battle and skipping lectures.	0.84
Idiom	Hate someone's guts	After Preston stole his candy, Alex decided to hate his guts before going to tell on him.	0.27
Idiom	Hear a pin drop	Quiet in the waiting room, the family could hear a pin drop as they waited for results.	1.00
Idiom	Hold down the fort	Alone in the office, Claire had to hold down the fort and make things run smoothly.	1.00
Idiom	Keep a low profile	The shy girl wasn't talkative and would keep a low profile on her first day.	0.94
Idiom	Keep a straight face	Enjoying the comedian's act, Joe couldn't keep a straight face as he laughed.	1.00

Idiom	Keep an open mind	Gabrielle led a sheltered life and would keep an open mind for new experiences.	0.77
Idiom	Lend a hand	The devout churchgoers always wanted to lend a hand when someone was in need.	0.97
Idiom	Living in a dream world	Fantasizing about being rich, Ralph was living in a dream world when he tried to buy a car.	0.81
Idiom	Make one's skin crawl	Allie hated spiders and they always made her skin crawl when she saw them.	0.94
Idiom	Make the first move	After the wonderful date, Felicity wanted to make the first move before her date left.	0.90
Idiom	Meets his match	Sandra hopes her difficult son will change when he meets his match in the near future.	0.26
Idiom	Move up the ladder	The new employee was eager to move up the ladder after the first meeting.	0.74
Idiom	Passed with flying colors	After many late nights studying, Susie passed with flying colors and graduated on time.	0.97
Idiom	Push your luck	After escaping the police, the criminals would really push their luck with their behavior.	0.71
Idiom	Rain on someone's parade	After Alyssa's success, no one could rain on her parade and make her upset.	1.00
Idiom	Read between the lines	After Sylvia scoffed at her outfit, Edith could read between the lines to know she didn't like it.	1.00
Idiom	Rock the boat	The school's plans to get new students would rock the boat to drastically improve enrollment.	0.58
Idiom	Roll up one's sleeves	Matt wanted to help out and would roll up his sleeves to help with the relief effort.	0.94
Idiom	Roll with the punches	Despite many surprises, Parker would roll with the punches when asked to take charge.	0.74
Idiom	Rolled out the red carpet	Preparing for her guests, Nora rolled out the red carpet to welcome them.	0.84
Idiom	Scratch her head	Hearing the strange argument, Bridget would scratch her head as she struggled to understand.	0.87
Idiom	Scratch the surface	Unfortunately, the robbery investigation couldn't scratch the surface with the minimal evidence.	0.77
Idiom	Show one's true colors	Accepting the bribe, the man would show his true colors as a dishonest person.	0.83
Idiom	Sign on the dotted line	The weary woman would sign on the dotted line as she finalized her divorce.	0.93
Idiom	Sink or swim	The family's small local business would sink or swim in their effort to be successful.	0.81
Idiom	Speak the same language	In politics, the brothers would never speak the same language and agree.	0.55
Idiom	Speak your mind	Upset about the poor conditions, the tenant would speak her mind to the landlord that night.	0.81
Idiom	Stand one's ground	Unwilling to do what wasn't right, Julie would stand her ground during the argument.	0.94
Idiom	Stand out from the crowd	Her hot pink hair would make Felicia stand out from the crowd to get noticed.	1.00
Idiom	Stand the test of time	A good black dress will always stand the test of time to be a great fashion choice.	0.94
Idiom	Steal the show	Debuting her new line, the designer's last look would steal the show to warrant applause.	0.87
Idiom	Take the cake	The exciting new magic show would really take the cake as the best of the carnival.	0.87
Idiom	Test the waters	Unsure of how people would react, Edie would test the waters to gauge responses.	0.83
Idiom	Threw caution to the wind	Riding a motorcycle without a helmet, Ian threw caution to the wind as he drove.	0.94

Idiom	Throw in the towel	Tired of football practices, Preston would throw in the towel to play soccer instead.	0.97
Idiom	Turned over a new leaf	A reformed gambler, Stan had finally turned over a new leaf for his family.	0.97
Idiom	Weather the storm	After arguing with her parents, Trudy would weather the storm by hiding in her room.	0.93
Idiom	Went out on a limb	When no one volunteered, Tanya went out on a limb to guess at the answer.	0.87
Idiom	Worked like a charm	The woman was thrilled that the cleaning solution worked like a charm for stain removal.	0.97

Phrase type	Expression	Sentence	Cloze
Literal	Press a button	To preheat the oven, John would have to press a button to adjust the temperature.	1.00
Literal	Water the plants	Leaving town, Daphne asked her friend to water the plants and feed her pets.	0.87
Literal	Come in peace	A gentle species, the aliens would come in peace when arriving on Earth.	0.70
Literal	Drive up the price	When tickets are hard to get, this will drive up the price and make them more valuable.	0.93
Literal	Send an email	Not wanting to call in sick, the man would send an email to his boss.	0.71
Literal	Raise some money	The town held a fundraiser to raise some money as the bridge needed repairs.	0.84
Literal	Slam on the brakes	Distracted, the novice driver had to slam on the brakes that morning.	0.90
Literal	Take a shower	Each morning before work, Penny would take a shower before making breakfast.	0.47
Literal	Swear to tell the truth	Cameron needed to swear to tell the truth that morning during the trial.	0.97
Literal	Deserve a break	After cooking for hours, the chef would deserve a break to get off his feet.	0.74
Literal	Wash the windows	To clean the city's skyscrapers, they hired people to wash the windows every week.	0.94
Literal	Remain in power	The powerful dictator would remain in power until someone usurped him.	0.74
Literal	Draw a picture	With her charcoal pencils, the artist began to draw a picture to show her skills.	0.61
Literal	Leave a message	When no one answered, Max had to leave a message to tell his brother he'd be late.	0.71
Literal	Light the fire	At the campsite, Austin used a match to help light the fire that evening.	0.87
Literal	Wash your hands	After being outside, you should wash your hands by scrubbing thoroughly.	1.00
Literal	Serve a useful purpose	A helping hand will always serve a useful purpose to those in need.	0.63
Literal	Place an order	At the restaurant, the boys were eager to place an order for burgers and fries.	0.97
Literal	Break up a fight	The prison guard had to run to break up a fight as the prisoners got rowdy.	0.94
Literal	Singing in the shower	As he washed his hair, Thomas loved singing in the shower until he shattered the glass.	0.93
Literal	Want a balanced budget	The accounting firm would always want a balanced budget for their clients.	0.45
Literal	Shake his hand	Finalizing the deal with the manager, Nick would shake his hand before thanking him.	0.90
Literal	Visit the Web site	To learn about the company, Michelle went to visit the Web site	0.90

		and get information.	
Literal	Stand on the porch	While the kids played outside, the father would stand on the porch as he kept a close watch.	0.81
Literal	Want to help people	Juan went to medical school, knowing he would want to help people for a living.	0.67
Literal	Raise your right hand	To take an oath, you sometimes raise your right hand to make a pledge.	1.00
Literal	Want the same things	The couple broke up, deciding they would never want the same things and parted ways.	0.58
Literal	Break the rules	Cheating in the game, the man would break the rules to make sure he won.	0.93
Literal	Come in a wide range	The shopkeeper asked that the sweater come in a wide range of colors that season.	0.06
Literal	Know the whole story	To hear the truth, Christa had to know the whole story before making a decision.	0.87
Literal	Enter the work force	Excited to earn money, the boy would enter the work force and get his first job.	0.81
Literal	Open a bottle	Her friends wanted wine, so Martha went to open a bottle to share that evening.	0.97
Literal	Went down the drain	The garbage disposal came on and the scraps went down the drain after dinner.	0.97
Literal	Fasten your seat belt	To drive safely, you should fasten your seat belt before taking off.	1.00
Literal	Read a story	Ready for bed, the kids asked their father to read a story as they prepared for bed.	0.68
Literal	Leave for the airport	To catch his flight, Trent had to leave for the airport to avoid being late.	0.71
Literal	Listen to the music	Buying the new album, Eli sat down to listen to the music in the living room.	0.63
Literal	Leave the house	Snowed in, the woman wasn't able to leave the house and had to stay in on Saturday.	1.00
Literal	Talk on the phone	The teenage girl would constantly talk on the phone for hours.	1.00
Literal	Hang from the ceiling	For the greenhouse, Edward bought plants to hang from the ceiling to make it look nice.	0.68
Literal	Invest in the stock market	With his earnings, Vince would invest in the stock market when the price was right.	0.94
Literal	Wash the dishes	After dinner, the Johnsons would wash the dishes as a family.	0.81
Literal	Play the guitar	With his pick in hand, the acoustic artist would play the guitar with expertise.	0.90
Literal	Provide a good example	By using her manners, Trisha would provide a good example by acting properly.	0.50
Literal	Lived in the same house	Growing up together, the sisters had lived in the same house as children.	0.80
Literal	Clean the house	While the kids were at school, Anne would clean the house that afternoon.	0.87
Literal	Trying to lose weight	The chubby boy was always trying to lose weight to get in better shape.	1.00
Literal	Take a walk	Beautiful outside, Mark put on his shoes to take a walk that evening.	0.90
Literal	Answer the question	Intimidating her, the lawyer badgered the witness to answer the question during the trial.	1.00
Literal	Hang on just a second	With two minutes left, Oliver asked Michael to hang on just a second before they left.	0.30
Literal	Save a lot of money	The piggy bank would help Zach save a lot of money for the new train set.	0.90
Literal	Play the piano	Stephen loved soothing, classical music and would play the piano with great appreciation.	0.29
Literal	Smoke a cigar	To celebrate, the wealthy man sat back to smoke a cigar after the	0.83

		deal closed.	
Literal	Raise their hands	When participating in class, the kids had to raise their hands to be called on.	1.00
Literal	Spend most of their lives	From youth onward, domestic pets spend most of their lives as faithful companions.	0.30
Literal	Hang up the phone	On hold for hours, Brittany would finally hang up the phone to show her frustration.	1.00
Literal	Pick in the second round	The NBA draft announcers predicted the top pick in the second round with precision.	0.80
Literal	Cast your ballot	After indicating her vote, the voter goes to cast her ballot in the ballot box.	0.87
Literal	Awake most of the night	Unable to fall asleep, Ella was awake most of the night to her dismay.	0.93
Literal	Walk down the aisle	The bride was nervous to walk down the aisle as the ceremony began.	0.87

APPENDIX B: Stimuli used in Experiment 3

Fragment	Frequency	Idiom	Frequency	Literal	Frequency
a table or	89	cross the line	88	led the nation	88
barrier point that	1	draw the line	1	notice the door	5
costume boys as	1	reinvent the wheel	1	spraying the wheel	1
every life when	3	stand your ground	12	need a husband	14
every man as	1	pay the price	1	cut the price	1
every story as	1	test the waters	1	test the phrase	1
every stream or	1	scratch the surface	7	receives the majority	3
grail vessel as	1	cover your tracks	1	cover his advance	1
her head but	26	join the club	24	raise the taxes	20
her parents before	8	throw a fit	1	throw a veil	1
his staff because	2	have the munchies	2	have the linguine	2
hit arm or	1	set the stage	1	set the meeting	1
its body so	2	hit the mark	18	jump the fence	19
its colleges and	7	build a bridge	7	having a guest	7
its feet while	2	get a headstart	1	see a dartboard	1
its loans as	1	lose your edge	1	find your sister	3
its soul if	1	rack your brain	1	dull your brain	1
memory fact that	1	steal the show	1	forgive the show	1
my eyelids so	1	know the score	1	know the crew	1
my mother if	32	play the field	29	join the church	23
my style than	3	pass the torch	31	allow the user	30
needs pager or	1	cover your back	1	hurting your back	1
no ashes since	1	spread your wings	1	catch your limit	1
no boy or	2	lend an ear	7	walk a block	9
no corns or	1	burst his bubble	3	impress a dragon	3
no purpose except	8	speak your mind	7	read your book	3
no secret where	2	get the message	2	get the kitchen	2
no sun nor	2	leave your mark	2	open our gifts	2
no time since	8	break your heart	10	feel your heart	11
our banks so	1	drag his feet	1	push her car	1
our kids as	1	smell a rat	1	rotate a graph	1
sight teachers and	1	break the ice	2	forget the ice	2
their job as	6	catch your eye	14	felt her mouth	15
their players than	4	see the light	51	see the numbers	51
your basis if	1	earn your keep	2	earn your place	3
your bees or	1	crack the whip	1	enter the digits	2

your cell so	2	hate his guts	8	adds a dash	8
your guy because	3	waste your breath	9	hurt your feelings	7
your liver where	1	steal your thunder	1	ease their tensions	1
your perfume when	2	shoot the breeze	26	embrace the notion	9