Annual Review of Applied Linguistics (2012), **32**, 45–61.
© Cambridge University Press, 2012, 0267-1905/12 $16.00
doi: 10.1017/S0267190512000074

The Processing of Formulaic Language

Kathy Conklin and Norbert Schmitt

It is generally accepted that we store representations of individual words in our mental lexicon. There is growing agreement that the lexicon also contains formulaic language (*How are you? kick the bucket*). In fact, there are compelling reasons to think that the brain represents formulaic sequences in long-term memory, bypassing the need to compose them online through word selection and grammatical sequencing in capacity-limited working memory. The research surveyed in this chapter strongly supports the position that there is an advantage in the way that native speakers process formulaic language compared to nonformulaic language. This advantage extends to the access and use of different types of formulaic language, including idioms, binomials, collocations, and lexical bundles. However, the evidence is mixed for nonnative speakers. While very proficient nonnatives sometimes exhibit processing advantages similar to natives, less proficient learners often have been shown to process formulaic language in a word-by-word manner similar to nonformulaic language. Furthermore, if the formulaic language is idiomatic (where the meaning cannot be understood from the component words), the figurative meanings can be much more difficult to process for nonnatives than nonidiomatic, nonformulaic language.

What is stored in our mental dictionary, the lexicon, is an open question. For example, do we store the word *dog* as well as its plural form *dogs*? Or do we only store *dog* and have a rule (NOUN + s = plural) to compute the plural form? If we assume that under normal circumstances we simply store the word *dog* and the plural form is computed, will this be different for someone who has three dogs and is frequently using the plural form? Thus the important question is whether to facilitate processing, forms and formulaic sequences that occur frequently in language (*fish and chips* or *Watch out!*), are stored. It makes sense that our brains would make use of a relatively abundant resource (long-term memory) to compensate for a relative lack in another (working memory) by storing frequently occurring formulaic sequences. These could then be easily retrieved and used without the need to compose them online through word selection and grammatical sequencing (Pawley & Syder, 1983). If native speakers are able to decrease demands on cognitive capacity because formulaic sequences are, in a sense, ready to go, are nonnative speakers able to do the same? This is an

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

important issue, as some evidence seems to show that second language (L2) learners neglect phrases, focusing instead on individual words (e.g., Foster, 2001).

THE IMPORTANCE OF FORMULAIC LANGUAGE

Before we turn to how our brains store and represent formulaic language,[1] it is useful to ask why we study formulaic sequence processing at all. It is becoming increasingly clear that formulaic language is an important element of language learning and use. Perhaps the best evidence for this is simply its ubiquity. Normal discourse, both written and spoken, contains large (but not yet fully determined) percentages of formulaic language. Oppenheim (2000) counted the multiword stretches of talk that occurred identically in practice and final renderings of a short speech on the same topic and found between 48 percent and 80 percent (overall mean of 66 percent) consisted of identical strings. Sorhus (1977) calculated that speakers in her corpus of spontaneous English Canadian speech used an item of formulaic language once every five words. Erman and Warren (2000) calculated that 52–58 percent of the language they analyzed was formulaic, and Foster (2001) came up with a figure of 32 percent using different procedures and criteria. Biber, Johansson, Leech, Conrad, and Finegan (1999) found that around 30 percent of the words in their conversation corpus consisted of lexical bundles, and about 21 percent of their academic prose corpus. Howarth (1998) looked at frequent verbs in a social science/academic corpus and found that they occurred in either restricted collocations or in idioms in 31–40 percent of the cases. Rayson (2008) found that 15 percent of text is formulaic according to a Wmatrix analysis. Overall, these studies suggest that formulaic language makes up between one third and one half of discourse.

For formulaic language to be so widespread, it must provide some useful purpose in communication. Schmitt and Carter (2004) listed a number of these purposes:

- Expressing a message or idea (*The early bird gets the worm* = do not procrastinate)
- Realizing functions ([*I'm*] *just looking* [*thanks*] = declining an offer of assistance from a shopkeeper)
- Expressing social solidarity (*Yeah it is* = expressing agreement)
- Transacting specific information in a precise and understandable way (*Cleared for takeoff* = permission to enter a runway and commence takeoff)
- Signaling discourse organization (*on the other hand* = conversely)

In addition to this list, it has been suggested that formulaic language has another benefit: It helps language users be more fluent. This idea was first outlined by Pawley and Syder (1983; followed closely by Kuiper & Haggo, 1984). They argued that native speakers have cognitive limitations in how quickly they can process language, presenting evidence that the largest unit of novel discourse

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

that native speakers can process is a single clause of 8–10 words. Nevertheless, these speakers are also able to produce language seemingly beyond this limitation, for example:

You can lead a horse to water, but you can't make him drink.

This phrase is clearly beyond the limit of 8–10 words, yet native speakers can say it without hesitation. This kind of evidence led to Pawley and Syder's argument that the storage of formulaic language in abundant or unconstrained long-term memory can compensate for limited working memory.

Some of the earliest empirical evidence supporting Pawley and Syder's (1983) assertions came from Dechert (1983), who found that the spoken output of a German learner of English was smoother and more fluent when she used formulaic language. These formulaic sequences were so useful in providing a platform for more fluent and accurate output that Dechert called them "islands of reliability," suggesting that they may anchor the processes necessary for planning and executing speech in real time.

There is now converging evidence that formulaic language is processed both more quickly and potentially differently from nonformulaic language, which makes the processing of formulaic language an increasingly interesting topic. This article will survey recent research (largely drawn from the field of psychology) into how both native and nonnative speakers process various types of formulaic language. In particular, we will focus on literal and figurative meanings of formulaic language, processing of various types of formulaic language, the relationship between corpus-extracted formulaic language and its psycholinguistic processing, and the processing/storage of individual words versus formulaic language.

FORMULAIC LANGUAGE: A FOCUS ON IDIOMS

In the formulaic sequence literature, idioms have received a fair amount of attention. One of the reasons for this is that many idioms allow for two distinct interpretations: figurative and literal.[2] Much of the research on idiom comprehension in native speakers has addressed the following issues: (a) activation of idioms' figurative versus literal meanings and (b) processing of idiomatic expressions versus novel (nonformulaic) phrases. With regard to the former, a number of models have been proposed (e.g., Bobrow & Bell, 1973; Cacciari & Tabossi, 1988; Swinney & Cutler, 1979). One of the influential models of idiom comprehension is the lexical representation hypothesis by Swinney and Cutler (1979). They propose that idioms are represented in the mental lexicon much like morphologically complex words are. They argue that the computation of the literal meaning and the retrieval of the figurative one are initiated simultaneously, as soon as the first word of the expression is encountered. However, because computation of the literal meaning is more time-consuming than the retrieval of the figurative one, the latter meaning should become activated first.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Another prominent theory of idiom processing puts forward the idea of an *idiomatic key*, which refers to the place where the expression becomes recognizable as idiomatic (Cacciari & Tabossi, 1988). According to this theory, dubbed the configuration hypothesis, the individual words and their literal meanings are activated until the key has been reached. Once the idiomatic key is reached, the idiomatic configuration emerges and the figurative meaning is accessed, while the literal meaning is rejected as no longer viable. Cacciari and Tabossi (1988), as well as Tabossi and Zardon (1993), pointed out that this is only true in the absence of a language context, which would prepare the reader for either figurative or literal rendering. With the aid of context, an idiomatic meaning may become available earlier.

In addition to exploring the activation of literal versus figurative interpretations of idioms, researchers have looked at the processing of idioms versus novel phrases. For example, Swinney and Cutler (1979) studied native speakers and found that idioms (*break the ice*) were processed more quickly than nonidiomatic phrases (*break the cup*). The findings of Gibbs (1980), Gibbs and Gonzales (1985), and Van Lancker, Canter, and Terbeek (1981) also suggested that idioms enjoy faster processing (in comprehension, as well as production) than matched novel strings for native speakers. The idiom decomposition hypothesis (Gibbs, Nayak, & Cutting, 1989) addresses when the figurative meaning of conventional language is activated in relation to novel language. This theory holds that idiom processing is highly dependent on whether an idiom is decomposable (the meanings of component idiom words are related to the overall figurative meaning) or nondecomposable (there is no obvious link between the meanings of words and the overall figurative meaning). Gibbs et al. argued that only in the case of decomposable idioms (*pop the question*) should idioms be faster to process than their novel control phrases (*ask the question*), because their individual components contribute to the idiom's figurative meaning. For nondecomposable idioms (*kick the bucket*) where no such link exists, no processing advantage should be observed for idioms over their novel matches (*fill the bucket*). However, this proposition is not supported by Tabossi, Fanari, and Wolf's (2009) results. In a semantic judgment task, native speakers responded more quickly to decomposable and nondecomposable idioms (as well as clichés) than to matched literal expressions. This suggests that idiomatic expressions are processed more quickly than compositional language; however, whether an idiom's constituents are, or are not, related to the idiom's overall figurative meaning does not seem to affect its processing.

Despite the differences that exist among idiom theories, the evidence seems to support the view that idioms are processed more quickly than nonformulaic language by native speakers. What remains an open question is whether for native speakers the literal or figurative meaning of an idiom is activated more quickly. Another important question is how idioms are processed by nonnative speakers. Similar to native speakers, one of the issues in the literature on nonnative speakers is whether there is difference in the processing of literal versus figurative meanings.

Van Lancker-Sidtis (2003) looked at whether prosodic cues helped native and proficient nonnative speakers distinguish between the two idiom

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

interpretations. Participants listened to sentences that contained idioms used either figuratively or literally and then had to identify the intended meaning. Results suggested that prosodic cues enabled native participants to successfully differentiate between idioms used figuratively and literally, whereas even highly proficient nonnatives were unable to do so.

In a study by Cieslicka (2006), nonnative participants listened to nondefining sentences containing familiar idioms (*George wanted to bury the hatchet soon after Susan left*) and performed a lexical decision on one of four targets: a word related to the idiom's figurative meaning (*forgive*), its control (*gesture*), a word related to its literal meaning (*axe*), or its control (*ace*). Faster response times to targets related to the literal meaning than to ones related to the figurative one suggest that literal idiom interpretations are activated more strongly than figurative ones. Thus, according to Cieslicka, in nonnative idiom comprehension, the literal meaning enjoys a processing advantage over the figurative meaning. However, perhaps, it is not surprising that upon hearing the word *hatchet*, there is a strong facilitation for the word *axe*, since the two words are strongly semantically related.

In a direct exploration of the reading speed of formulaic language, Conklin and Schmitt (2008) used a self-paced line-by-line reading paradigm with native and nonnative speakers. Lines with formulaic sequences (*everything but the kitchen sink*) and their matched nonformulaic control strings (*everything but the kitchen sink*) were embedded in story passages. Participants read the passages, one line at a time, by pushing a button to bring up each new line. By comparing the timing of the button-pushes for the two types of strings, Conklin and Schmitt found that their participants read formulaic sequence more quickly than the matching nonformulaic control strings. They interpreted this as evidence that formulaic language is more easily processed than nonformulaic language, and crucially, this effect applied to proficient nonnatives as well as natives.

The use of eye-movement apparatus is an even better methodology for studying reading, because it eliminates the behavioral interface (i.e., the physical button-pushing), which can add variability to the results. Another advantage is that the text can be displayed all at once rather than word-by-word or line-by-line, which is a closer approximation of the natural reading process. In eye-movement studies, participants merely look at a computer monitor, and the apparatus tracks the movement of the eyes and determines which part of the screen (e.g., which words) are being focused upon and for how long. Underwood, Schmitt, and Galpin (2004) were the first to use this methodology to explore the recognition of formulaic sequences in texts. They embedded idioms into reading passages and then measured how often and for what duration the final words in those idioms were fixated upon (e.g., *met the deadline by the skin of his teeth*) by native and nonnative speakers. The results were then compared to measurements of the same words in nonformulaic contexts (e.g., *the dentist looked at his teeth*). This tested the hypothesis that once an idiom is recognized from the first few words, the final word will require less attention, because it is already known from familiarity with the idiom. Underwood et al. found that native speakers indeed fixated less on the terminal words in formulaic than nonformulaic contexts, and for a shorter duration. This is evidence for a processing

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

speed advantage for formulaic sequences versus creative language. Moreover, this advantage was partially shared by proficient nonnatives as well.

Finally, Siyanova-Chanturia, Conklin, and Schmitt (2011) carried out an eye-movement version of the Conklin and Schmitt (2008) study with native and nonnative participants. Stimuli similar to the Conklin and Schmitt passages were used, containing idioms (*left a bad taste in my mouth*) and matched control phrases (*the bad taste left in his mouth*). Whole passages were presented on the monitor screen, and the participants' eye movements were tracked while they read these passages. The eye-movement analyses of first-pass reading time, total reading time, and number of fixations showed that the native-speaking participants processed the idioms significantly faster than the nonformulaic controls. For the nonnative participants, there was no evidence that the idioms were processed any faster than the matched controls; on the contrary, the figurative readings seemed to be read more slowly than literal readings.

Taken together, the research is mixed on whether nonnatives process idioms faster than matched novel strings. However, the findings reported here suggest that even for highly proficient nonnative speakers, processing may be slowed when idioms are used figuratively. Much of the research on formulaic sequence processing has been focused on idioms, and this is problematic for a number of reasons. First, idioms, while salient, tend not to be very frequent, and therefore nonnative speakers and first language (L1) children may not have much exposure to them. Thus they may not be the best test case for whether learners are sensitive to regularly occurring patterns in language. Second, the meaning of idioms can be transparent (or decomposable) to varying degrees. Sometimes the individual component words contribute overtly to the overall meaning of the expression (*add fuel to the fire* = make bigger) and sometimes not (*kick the bucket* ≠anything to do with kicking or buckets). Thus other kinds of consistently transparent formulaic sequences might provide a better test case for whether or not frequently occurring expressions that can be computed online actually are. Third, because many idioms have two possible senses—figurative and literal—they are ambiguous, and the processing system has to select from competing semantic representations that could slow processing. Thus, while idioms have been widely studied, other formulaic sequences may provide a more accurate picture about formulaic processing.

PROCESSING OF NONIDIOMATIC FORMULAIC LANGUAGE

The position that formulaic language has a processing speed advantage is now supported by a number of research approaches. For example, Kuiper (1996, 2004) looked at actual language use in the real world and found that the speech of so-called smooth talkers (people who need to produce fluent speech under severe time pressure, such as auctioneers and sports announcers) was largely formulaic in nature. In another study, grammaticality judgments by L2 speakers for formulaic items were not only more accurate but also faster than judgments for matched nonformulaic control strings (Jiang & Nekrasova, 2007).

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Sosa and MacFarlane (2002) used an auditory word-monitoring task for the function word *of* in two-word combinations varying in frequency (*sort of* and *kind of*) with native speakers. They found that reaction times of *of* in high-frequency combinations were significantly slower than those in low-frequency ones, indicating that very frequent combinations are stored as wholes. The slowed responses observed for *kind of* (high-frequency) over *piece of* (low-frequency) could not simply be due to phonological reduction (i.e., *kinda* for *kind of*), as phonological reduction was equally prevalent across the frequency groups for their stimuli. Further, the number of correct responses produced was very low for the high-frequency collocations (37 percent compared to 60 percent of correct responses produced for the low-frequency group), which supports the view of holistic storage of frequently occurring multiword units. Sosa and MacFarlane argued that their results indicated that when such multiword units were used frequently, they became chunked and were subsequently stored as a unit.

Bod (2000, 2001) investigated the processing of frequently occurring sentences by native speakers. Bod's participants read frequent three-word SVO (subject-verb-object) sentences (e.g., *I like it*) and low-frequency control sentences (e.g., *I keep it*) that were matched for lexical frequency, complexity, and surface structure. Participants responded faster to high-frequency sentences than low-frequency ones, suggesting that frequently occurring sentences are stored holistically in long-term memory. However, the processing cost found for the less frequent sentences may be a consequence of these phrases being less natural. Because of the tense and aspect of some of the low-frequency experimental items, they may have sounded less natural to the participants than high-frequency ones (e.g., *I test it* sounds less natural than *I love it* because it's encountered more often in the past and future tense and in present progressive; e.g., *I tested it* / *I will test it* / *I'm testing it*). Thus it is difficult to say whether the processing cost observed for the low-frequency phrases over the high-frequency ones is due to properties of the stimuli or to the holistic storage of the high-frequency sentences.

In a similar study, Arnon and Snider (2010) investigated the role of frequency in the comprehension of compositional four-word phrases (i.e., the phrase is comprehensible from the meanings of the individual words, e.g., *don't have to worry*) with native speakers. They compared reading times for phrases, which differed in phrasal frequency but whose individual components were controlled for length and frequency. They found that the more frequent phrases were processed reliably faster than the less frequent ones. The authors concluded that language users appear to notice, learn, and subsequently store frequency information not only about words but also with regard to multiword phrases, even when they are entirely compositional. Although informative with respect to the role of phrasal frequency, Arnon and Snider's study is limited to highly compositional phrases that are rather different from highly familiar fixed or semifixed formulaic sequences, such as binomial expressions or collocations.

Bannard and Matthews (2008) compared monolingual children's production of phrases that differed in the frequency with which they appeared in child-directed speech (e.g., *a drink of milk* vs. *a drink of tea*). Young children

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

(two–three years old) were found to be reliably faster and more accurate at repeating higher frequency phrases than lower frequency ones. This shows that children as young as two years are sensitive to the frequency with which multiword strings occur in their input, and that frequent phrases like *a drink of milk* are represented in young children's lexicon.

Tremblay and Baayen (2010) used phrase recall and electrophysiological (ERP) measures to investigate the processing of four-word sequences (*in the middle of*) by native speakers. They found that the frequency of occurrence of the four-word sequences improved participants' recall and that whole-string probability modulated P100 and N100 amplitudes,[3] which are usually associated with the perception of input and typically occur before the onset of semantic/syntactic processing. Although these results were taken to suggest that multiword forms are stored both as parts and wholes, it is unclear whether they support such a view or instead are indicative of more general attentional processing, as these early ERP components have been implicated in attention. Similarly, in a self-paced reading study, Tremblay, Derwing, Libben, and Westbury (2011) compared the processing of sentences containing lexical bundles (*don't worry about it*) and matched control phrases. They found that sentences containing lexical bundles were read faster than control sentences, and were more likely to be remembered and recalled correctly than sentences with novel phrases. Taken together, these findings may suggest that the more frequent lexical bundle is, the more likely it is to be represented as a chunk in memory, which eases the burden on working memory during initial processing and subsequent recall.

A recent eye-tracking study by Siyanova-Chanturia, Conklin, and van Heuven (2011) investigated processing by native and nonnative English speakers of formulaic sequences imbedded in sentences that differ in phrasal frequency. Participants read sentences containing three-word binomial phrases (*bride and groom*) and their reversed forms (*groom and bride*), which are identical in syntax and meaning but that differ in phrasal frequency. Mixed-effects modeling revealed that native speakers and nonnative speakers, across a range of proficiencies, read more frequent formulaic sequences more quickly than less frequent ones. Furthermore, regardless of frequency, the typical binomial configurations were processed more quickly than the reversed forms, which indicates that something more than a pure frequency effect is influencing the processing. Crucially, this provides strong evidence that binomial phrases are entrenched in memory in some way.

Although they are not typically considered formulaic language, research on compounds (*chalkboard*) can inform the ongoing debate regarding the trade-off between the storage and computation of multimorphemic words. Badecker (2001), Badecker and Allen (2002), Juhasz (2007), and Libben (1998) took a compositional approach to the processing of compounds. They argued that compounds are decomposed during their recognition. In contrast, Pollatsek, Hyona, and Bertram (2000) proposed that individual words (e.g., *blue* and *berry*), as well as compounds (e.g., *blueberry*), are stored in the lexicon, and that access to a compound can occur via the individual words or via the holistic representation.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Likewise, Mondini, Jarema, Luzzatti, Burani, and Semenza (2002) and Mondini, Luzzatti, Saletta, Allamano, and Semenza (2005) maintained that retrieval of a compound entails the activation of its individual components, as well as the whole form of the compound. Mondini et al. (2002) investigated processing of compounds of the type Adj + N and N + Adj (e.g., *natura morta* – "still life") and matched novel combinations (e.g., *natura bella*—"beautiful nature") by two nonfluent aphasic patients. In Italian, adjectives agree with the grammatical gender of the noun in both compounds and novel combinations. Mondini et al. (2002) hypothesized that if compounds are stored holistically, then participants should have difficulty making noun-adjective agreement for novel combinations (because of their language impairment), but not compounds (which can be retrieved whole from memory). They found that both participants performed significantly better on compounds than on novel noun-adjective combinations. This suggests that for novel combinations the participants retrieved the adjective and noun separately and then applied agreement rules. Compounds, on the other hand, were retrieved as wholes, and therefore, no morphosyntactic operations were necessary. Interestingly, one of the participants was also able to repeat compounds significantly more accurately than noncompounds. According to Mondini et al. (2002), this implies that compounds require less working memory, which provides further evidence that they are stored holistically. Such results suggest that compounds are stored and processed as wholes, rather than computed online. This may also be true of other multimorphemic combinations that are hypothesized to be stored and retrieved as lexical units such as collocations, binomials, and idioms. Later, we turn to studies that address the issue of multiword storage and representation in the mental lexicon.

Other studies with impaired participants have also contributed to our knowledge of formulaic sequence processing. Van Lancker and Kempler (1987) investigated the processing of familiar phrases (e.g., idioms, speech formulas, etc.) and novel phrases by left- and right-brain damaged participants using a picture-matching auditory comprehension task. The authors predicted a larger role of the right hemisphere in familiar phrase processing, as it has been implicated in the processing of idiomatic, nonliteral language. Their results revealed that despite impaired syntactic processing, the left-brain damaged group performed significantly better on familiar phrase recognition than the right-brain damaged group. The latter group, on the other hand, performed better in the novel phrase recognition task. The authors concluded that familiar phrases such as idioms, collocations, routines, and clichés are represented in the brain differently from the novel language. Following up on this finding, Van Lancker-Sidtis and Postman (2006) examined occurrences of nonnovel expressions in the spontaneous speech of normal, right- and left-hemisphere damaged participants, respectively. They found that left-hemisphere damaged participants used significantly more fixed expressions than the normal control group, whereas right-hemisphere damaged participants produced fewer fixed expressions than the control group. This, the authors argued, provides support for the view that novel language is left-hemisphere lateralized, while fixed expressions are right-hemisphere lateralized.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

THE RELATIONSHIP BETWEEN CORPUS-EXTRACTED FORMULAIC LANGUAGE AND ITS PSYCHOLINGUISTIC PROCESSING

Identification of the formulaic sequences used in many of the studies reviewed in this chapter have relied on corpus evidence. This makes the relationship between the two (formulaic sequences extracted from corpora and their psycholinguistic bases in the mind) a very interesting issue. Some scholars believe that collocation is an entirely textual phenomenon and is not indicative of how language is represented in the mind (e.g., Bley-Vroman, 2002). They believe that collocations arise spontaneously in text as an epiphenomenon of the meaningful use of language in context, rather than being linguistic patterns that can be learned and used. For example, the words *dark night* occur together simply because nights are dark, and so people will naturally use these words together when speaking about the nighttime. However, given all the evidence for the processing advantages of formulaic language in this chapter, it is difficult to believe that it does not somehow exist in the mind (for an alternative view, see Ellis 2002a, 2002b).

We found only one study that directly explored whether the formulaic sequences extracted from corpus data are psycholinguistically real. Schmitt, Grandage, and Adolphs (2004) identified a number of different types of formulaic sequence from corpus evidence and embedded them in a spoken dictation task with native and nonnative English speakers. Each burst of dictation was longer than short-term memory could hold (i.e., 20–24 words), so the respondents were not able to repeat a burst from rote memory. This meant they were forced to reconstruct the language. The researchers assumed that if the formulaic sequences in the bursts were stored holistically, then they would be reproduced intact, with no hesitation pauses or transformations. The results showed that many of the formulaic sequences in the dictation responses did meet this holistic criterion, but also that many did not. A sort of continuum of holisticness seemed to emerge. The authors concluded that many of the corpus formulaic sequences were not stored holistically, but that this varied from individual to individual. From this one study, it seems that any particular formulaic sequence extracted from a corpus may or may not be stored holistically by any particular person.

WHAT IS REPRESENTED IN THE LEXICON?

The body of research reviewed in this chapter points to the fact that adult native speakers, and most likely children and nonnative speakers who have had enough exposure to a language, appear to have representations not only for the words that make up formulaic sequences (*fish*, *and*, *chips*) but also for the sequence itself (*fish and chips*), which is in line with the view put forward by Wray (2008). Frequency seems to lead to a particular form being represented in the mental lexicon. However, if a form has not been encountered frequently enough, as in the case of lower proficiency nonnative speakers or very young children, it appears that it may not be well entrenched in memory.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

The finding that the frequency of a formulaic sequence affects the ease of processing is of importance for models of language use and processing. In the words-and-rules approach, a distinction is made between the lexicon (a collection of memorized and stored forms) and grammar (a collection of rules that are applied to these forms; Pinker, 1999; Pinker & Ullman, 2002). In line with this approach, frequency effects should only be observable in the processing of memorized forms (words). Researchers argue that frequency effects should not manifest themselves in the processing of compositional formulaic sequences. Thus, the words-and-rules approach is incompatible with the results reported in this chapter unless the lexicon is reconceived to include all formulaic sequences and not just idiomatic, nonliteral language.

However, usage-based (Bybee, 1998; Goldberg, 2006; Tomasello, 2003) and exemplar-based models (Abbot-Smith & Tomasello, 2006; Bod, 2006; Pierrehumbert, 2001) propose that the basic unit of language acquisition is a construction and that the task of a language learner is to acquire a set of constructions that vary in size, complexity, and level of abstractness (Goldberg, 2006; Tomasello, 2003). These theories propose that all linguistic information is represented and processed in the same way, and thus it should be similarly affected by frequency. New experiences with a linguistic unit, that is, a word or a phrase, are not decoded and then discarded; rather, they determine memory representations (Bybee, 2006). As Bod (2006) noted, what is represented is based solely on statistics. Thus, language should be viewed not as a set of grammar rules, but as a statistical accumulation of experiences that changes every time a particular utterance is encountered (e.g., Ellis, 2002a). This view predicts faster processing for all frequent sequences—words and phrases—over less frequent ones and is compatible with connectionist approaches to language acquisition and processing, which emphasize statistical properties of the input in language learning (Christiansen & Chater, 1999; Elman, 1990; Rumelhart & McClelland, 1986). In a connectionist approach, units do not exist in isolation; rather, they form and exist in relationships (networks) with each other. The frequency with which various linguistic exemplars occur together is a determining factor of the strength of the connections in the lexicon. It determines what and how speakers learn and eventually represent in their lexicon.

The results reviewed above indicate that formulaic sequences are represented in the lexicon and are processed faster than novel language. However, what it means to be represented in the lexicon and what underpins the processing advantage is unclear. One might argue that words that occur together frequently have strong connections. Thus when readers encounter *fish and*, activation quickly spreads to *chips*. Alternatively, faster processing of formulaic language could be explained probabilistically. In probabilistic models of language processing, information about word co-occurrences forms an integral part of speakers' knowledge of language (e.g., Jurafsky, 1996; McDonald & Shillcock, 2003). The probability of *chips* occurring after *fish and* is higher than the probability of *fish* appearing after *chips and*. Because *fish and chips* is a frequent expression, whereas *chips and fish* is not, one might therefore expect to see *chips* after reading *fish and*, which should facilitate reading; no such

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

expectations may exist for the reverse. Thus, the processing advantage for formulaic language may be due its predictability.

To specifically address the question of predictability, Siyanova-Chanturia, Conklin, and van Heuven (2011) had participants perform a completion test for both binomials and their reversed forms (*fish and* . . . vs. *chips and* . . .). They then looked at whether scores on the completion test predicted reading times in their eye-tracking study. The completion test did not significantly add anything to their mixed-effects models of reading times. Importantly, the analyses revealed that predictability and phrase type (binomial vs. reversed form) were not entirely the same. Once test scores from phrase type were accounted for in the model, phrase type still had a significant effect. This shows rather convincingly that the processing advantage for familiar phrases extends beyond the "first word+*and*" (e.g., *fish and*) predicting the last one (e.g., *chips*). Rather, their results indicate the important contribution of phrasal frequency and entrenchment of a particular phrase in memory. Finally, it is worth pointing out that the predictability story per se does not go against a representational account: Each and every instance of a formulaic sequence (e.g., idiom, binomial, compound) is a highly predictable word combination in which subsequent words can be predicted from an initial one(s). Thus, being predictable is an intrinsic characteristic of a formulaic sequence. It appears that speakers are sensitive to the frequency with which formulaic sequences occur, and this leads to their entrenchment in memory.

CONCLUSION

Formulaic language is pervasive in language use, and the research reported in this article shows that it is easier to process. Virtually every study, using a variety of research methodologies, shows that formulaic language holds a processing advantage over nonformulaic language for native speakers. However, for nonnatives, this is often not the case, although higher proficiency levels increase the chances of also enjoying this advantage. The crucial role of frequency in processing clearly applies not only to individual words but also to formulaic sequences. It appears that frequency of exposure is a key aspect of learning formulaic sequences. Although native speakers will automatically obtain the required exposure by adulthood, in many cases, nonnatives will not. This explains why only relatively proficient nonnatives (who have acquired their L2 over a long period, allowing them the time to amass sufficient language exposure) begin processing formulaic language in the quick and automatic manner of native speakers. These acquisition themes are taken up in much more detail by Bannard and Lieven (this volume) for L1 acquisition and Ellis (this volume) for L2 acquisition.

NOTES

1 Terminology has always been fraught in the area of multiword units. We will follow Schmitt's (2010) convention of using *formulaic language* as the cover term for the phenomenon, and *formulaic sequence* for each individual instance of it. Various

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

categories of formulaic language will be referred to by their own terms, e.g., *idioms*, *binomials*, *lexical bundles*. In this article, we follow Wray's (2002, p. 9) broadly based definition of the phenomenon: "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, rather than being subject to generation or analysis by the language grammar."

2 It can be argued whether literally used idioms (e.g. *The ship broke the ice to clear the seaway*) are really idioms or not. The can just as easily be considered a form of creative language that happens to coincide lexically with a figurative idiom (*He broke the ice by cracking a joke*), just as homonyms *bank* (river side) and *bank* (financial institution) are orthographically and phonologically identical but semantically unrelated. Despite this, much psychological research conceptualizes idioms as having both figurative and literal renderings, and it is this research we report here.

3 In ERP studies, numbers such as P100 and N100 refer to electrical brain waves (either negative or positive) occurring 100 milliseconds after stimulus offset. Thus P100 refers to a positive brain wave 100 ms after the stimulus has been presented. The polarity and timing of these brainwaves are thought to reflect specific types of language processing.

ANNOTATED BIBLIOGRAPHY

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

Our article underscores the important role of phrasal frequency on the speed of processing and ultimately on entrenchment in memory. This article is a good complement because it highlights the role of frequency in the processing of formulaic sequences. However, the statistical analysis section is not for the fainthearted.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 1–22.

This article presents an accessible eye-tracking study, which is a useful methodology for studying reading of units larger than single words. It provides a comparison of processing by native and nonnative speakers of idioms used literally and figuratively, as well as novel control phrases.

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

Like Arnon and Snider's (2010) article, this focuses on the increasingly hot topic of phrasal frequency. Crucially, this article shows that something above and beyond the simple frequency of formulaic phrases is represented and strongly supports the idea of entrenchment. Another useful aspect of the article is the comparison of native and nonnative speakers.

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory and Cognition*, *37*, 529–540.

This recent article is very useful because it looks at the different theories of idiom processing. Basically, it shows that knowing an expression, rather than its idiomaticity or whether its meaning is transparent, is what leads to faster processing.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Van Lancker, D., & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain and Language*, *32*, 265–277.

Evidence from impaired populations can provide strong evidence for formulaic language being processed differently (or at least by different areas of the brain) from nonformulaic language. This article is a good example of such processing differences.

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

Wray's volume provides a comprehensive overview of the acquisition, use, and attrition of L1 and L2 formulaic language. It provides a useful complementary perspective to the mainly psychology-based studies reviewed in this article.

REFERENCES

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *Linguistic Review*, *23*, 275–290.

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

Badecker, W. (2001). Lexical composition and the production of compounds: Evidence from errors in naming. *Language and Cognitive Processes*, *16*, 337–366.

Badecker, W., & Allen, M. (2002). Morphological parsing and the perception of lexical identity: A masked priming study of stem homographs. *Journal of Memory and Language*, *47*, 125–144.

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*, 241–248.

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

Bley-Vroman, R. (2002). Frequency in production, comprehension, and acquisition. *Studies in Second Language Acquisition*, *24*, 209–13.

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

Bod, R. (2000,). *The storage vs. computation of three-word sentences*. Paper presented at AMLaP2000, University of Leiden, Leiden, the Netherlands.

Bod, R. (2001,). *Sentence memory: Storage vs. computation of frequent sentences*. Paper presented at CUNY 2001, University of Pennsylvania, Philadelphia, PA.

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

Bybee, J. (1998). The emergent lexicon. *Chicago Linguistic Society*, *34*, 421–435.

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

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

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

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.

Dechert, H. (1983). How a story is done in a second language. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 175–195). London, UK: Longman.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Ellis, N. C. (2002a). 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. (2002b). Reflections on frequency effects in language acquisition: A response to commentaries. *Studies in Second Language Acquisition*, *24*, 297–339.
Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.
Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*, 29–62.
Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning. Teaching and testing* (pp. 75–93). Harlow, UK: Longman.
Gibbs, R. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory and Cognition*, *8*, 449–456.
Gibbs, R., & Gonzales, G. (1985). Syntactic frozenness in processing and remembering idioms. *Cognition*, *20*, 243–259.
Gibbs, R., Nayak, N., & Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of Memory and Language*, *28*, 576–593.
Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.
Howarth, P. (1998). The phraseology of learners' academic writing. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161–186). Oxford, UK: Oxford University Press.
Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, *91*, 433–445.
Juhasz, B. (2007). The influence of semantic transparency on eye movements during English compound word recognition. In R. van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 374–389). Amsterdam, Elsevier Science.
Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*, 137–194.
Kuiper, K. (1996). *Smooth talkers*. Hillsdale, NJ: Erlbaum.
Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 37–54). Amsterdam, the Netherlands: John Benjamins.
Kuiper, K., & Haggo, D. (1984). Livestock auctions, oral poetry, and ordinary language. *Language in Society*, *13*, 205–234.
Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, *61*, 30–44.
McDonald, S., & Shillcock, R. (2003b). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*, 648–652.
Mondini, S., Jarema, G., Luzzatti, C., Burani, C., & Semenza, C. (2002). Why is "red cross" different from "yellow cross"?: A neurophysiological study of non-adjective agreement within Italian compounds. *Brain and Language*, *81*, 621–634.
Mondini, S., Luzzatti, C., Saletta, P., Allamano, N., & Semenza, C. (2005). Mental representation of prepositional compounds: Evidence from Italian agrammatical patients. *Brain and Language*, *94*, 178–187.
Oppenheim, N. (2000). The importance of recurrent sequences for non-native speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 220–240). Ann Arbor: University of Michigan Press.
Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). London, UK: Longman.
Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam, the Netherlands: John Benjamins.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York, NY: Harper-Collins.
Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences*, *6*, 456–463.
Pollatsek, A., Hyona, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 820–833.
Rayson, P. (2008). *Software demonstration: Identification of multiword expressions with Wmatrix*. Paper presented at the Formulaic Language Research Network (FLaRN) conference, University of Nottingham, Nottingham, UK.
Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 216–271). Cambridge, MA: MIT Press.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual.* Basingstoke, UK: Palgrave Macmillan.
Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 1–22). Amsterdam, the Netherlands: John Benjamins.
Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences* (pp. 127–151). Amsterdam, the Netherlands: John Benjamins.
Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 1–22.
Siyanova-Chanturia, A., Conklin, K., & van Heuven, J.B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 776–784.
Sorhus, H. (1977). To hear ourselves—Implications for teaching English as a second language. *English Language Teaching Journal*, *31*, 211–221.
Sosa, A., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word *of*. *Brain and Language*, *83*, 227–236.
Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behaviour, 18*, 523–534.
Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory and Cognition*, *37*, 529–540.
Tabossi, P., & Zardon, F. (1993). The activation of idiomatic meaning in spoken language comprehension. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure, and interpretation* (pp. 145–161). Hillsdale, NJ: Erlbaum.
Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA & London, UK: Harvard University 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. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151–173). London, UK: Continuum International.
Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*, 569–613.
Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences*. Amsterdam, the Netherlands: John Benjamins.
Van Lancker, D., Canter, G., & Terbeek, D. (1981). Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech and Hearing Research*, *24*, 330–335.
Van Lancker, D., & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain and Language*, *32*, 265–277.
Van Lancker-Sidtis, D. (2003). Auditory recognition of idioms by first and second speakers of English. *Applied Psycholinguistics*, *24*, 45–57.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press

Van Lancker-Sidtis, D., & Postman, W. A. (2006). Formulaic expressions in spontaneous speech of left- and right-hemisphere damaged subjects. *Aphasiology*, *20*, 411–426.
Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge, UK: Cambridge University Press.
Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford, UK: Oxford University Press.

https://doi.org/10.1017/S0267190512000074 Published online by Cambridge University Press