Applied Linguistics 29/1: 72–89
doi:10.1093/applin/amm022

Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?

KATHY CONKLIN and NORBERT SCHMITT

University of Nottingham

It is generally accepted that formulaic sequences like *take the bull by the horns* serve an important function in discourse and are widespread in language. It is also generally believed that these sequences are processed more efficiently because single memorized units, even though they are composed of a sequence of individual words, can be processed more quickly and easily than the same sequences of words which are generated creatively (Pawley and Syder 1983). We investigated the hypothesized processing advantage for formulaic sequences by comparing reading times for formulaic sequences versus matched nonformulaic phrases for native and nonnative speakers. It was found that the formulaic sequences were read more quickly than the nonformulaic phrases by both groups of participants. This result supports the assertion that formulaic sequences have a processing advantage over creatively generated language. Interestingly, this processing advantage was in place regardless of whether the formulaic sequences were used idiomatically or literally (e.g. *take the bull by the horns* = 'attack a problem' vs. 'wrestle an animal'). The fact that the results also held for nonnatives indicates that it is possible for learners to enjoy the same type of processing advantage as natives.

INTRODUCTION

Formulaic sequences[1] are a major component of almost all types of discourse. This is true in terms of both degree and scope of usage. Research suggests that at least one-third to one-half of language is composed of formulaic elements (Howarth 1998a; Erman and Warren 2000; Foster 2001), although the percentage is affected by both register and mode (Biber *et al.* 1999, 2004). Moreover, formulaic sequences are used in a wide variety of ways. They can be used to express a concept (*put someone out to pasture* = retire someone because they are getting old), state a commonly believed truth or advice (*a stitch in time saves nine* = it is best not to put off necessary repairs), provide phatic expressions which facilitate social interaction (*Nice weather today* is a non-intrusive way to open a conversation), signpost discourse organization (*on the other hand* signals an alternative viewpoint), and provide technical phraseology which can transact information in a precise and efficient manner (*blood pressure is 140 over 60*) (Schmitt and Carter 2004).

Much of the research done on formulaic language has been done on written discourse, but it is equally, if not more, important in spoken discourse (Altenberg 1998; McCarthy and Carter 2002; O'Keeffe *et al.* 2007). For example, Sorhus (1977) calculated that speakers in her 131,536-word corpus of spontaneous Canadian speech used an item of formulaic language once every five words. She included one-word fillers, such as *eh, well, OK,* and *please* in her counts, but even without these, her analysis still shows a very high frequency of formulaic sequences such as *for example, at times,* and *a lot of.* Some of her 'fixed expressions' (her term) were especially frequent (e.g. *I think, you know, of course*); Sorhus calculated that the nineteen most frequent ones accounted for 41 per cent of the total number of fixed expression occurrences she found. Moreover, formulaic sequences are particularly essential in spoken discourse that occurs under heavy time constraints, such as auctioneering and sports announcing (Kuiper 2004). Biber *et al.* (1999), as part of their descriptive grammar, show the wide extent of contiguous formulaic strings (*lexical bundles* in their terminology) in speech, where they made up 28 per cent of the conversation corpus. Interestingly, there is evidence that individual lexical bundles are generally preferred in *either* spoken or written discourse, but seldom in *both*, at least in academic discourse (Biber *et al.* 2004).

Formulaic sequences are much more than strings of words linked together with collocational ties. Indeed, it is becoming increasingly obvious that much of the communicative content of language is tied to these phrasal expressions. They are often linked to a single meaning/pragmatic function, which gives them considerable semantic/pragmatic utility. Furthermore, building on Wittgenstein's (1958) observation that words are more than simple, discrete descriptions of phenomena in the world, corpus analysis has confirmed Firth's (1935) proposal that some of a word's meaning is derived from the sequences in which it resides. For example, while *border* usually means a physical 'edge' or 'boundary', when used as part of the phrase *bordering on X*, the meaning often shifts to 'nearing an undesirable state of mind', as in *bordering on arrogance* (Schmitt 2005). Moreover, the different meaning senses of a word will often be realized with quite different phraseological configurations (Sinclair 1966).

Formulaic sequences become particularly important in language use when we consider their pragmatic value. For instance, they are very often used to accomplish recurrent communication needs. These recurrent communicative needs typically have conventionalized language attached to them, such as *I'm (very) sorry to hear about* ___ to express sympathy and *I'd be happy/glad to* ___ to comply with a request (Nattinger and DeCarrico 1992: 62–3). Because members of a speech community know these expressions, they serve as a quick and reliable way to achieve the desired communicative effect. Formulaic sequences also realize a variety of conversational routines and gambits and discourse (Coulmas 1979, 1981). They are typically used for particular purposes and are inserted in particular places in discourse.

For instance, formulaic sequences regularly occur at places of topic-transition and as summaries of gist (Drew and Holt 1998). Overall, understanding the pragmatic role of formulaic language can tell us much about the nature of interaction (McCarthy and Carter 2002).

Moreover, formulaic sequences do more than just carry denotative meaning and realize pragmatic function. They can often have a type of register marking called *semantic/collocational prosody* (Stubbs 2002; Sinclair 2004). For example, *bordered on X* (*bordered on the pathological, bordered on apathy*) often has a negative evaluation, but semantic prosody can also be positive, as in collocations that form around the word *provide* (*provide information, provide services*). This semantic prosody is one means of showing a speaker/writer's attitude or evaluation. For example, their stance can be indicated concerning the knowledge status of the proposition following the formulaic item (*I don't know if X* indicates uncertainty about X), their attitude towards an action or event (*I want you to X* shows a positive attitude towards this action), and their desire to avoid personal attribution (*it is possible to* avoids a directly attributable suggestion) (Biber *et al.* 2004). Likewise, the choice of formulaic sequences can reflect an author's style and voice (Gläser 1998). Formulaic sequences can also be used to encode cultural ideas, as Teliya *et al.* (1998) have demonstrated for Russian.

Given their utility, it is not surprising that the use of formulaic sequences is widespread in language. Erman and Warren (2000) calculated that formulaic sequences of various classes constituted 58.6 per cent of the spoken English discourse they analysed and 52.3 per cent of the written discourse. Foster (2001) had raters look for formulaic language in transcripts of unplanned native speech, and these raters judged that 32.3 per cent of the speech was made up of formulaic sequences. Biber *et al.* (1999) found that 3- and 4-word lexical bundles made up 28 per cent of the conversation and 20 per cent of the academic prose they studied. Howarth (1998a) looked at 238,000 words of academic writing and found that 31–40 per cent was made up of collocations and idioms. As all of the above studies used different criteria and procedures, it is not surprising that their results varied to some degree. However, the basic conclusion that formulaic sequences make up a large part of any discourse seems inescapable (Nattinger and DeCarrico 1992). This has led Sinclair (1991) to argue that the structure of language is dominated by the *idiom principle* rather than the *open-choice principle*.

Although most of the research reported above has focused on English, formulaic sequences have been studied in a number of languages. Cowie (1998a, 1998b) discusses the widespread work done on phraseology in Russian (mainly concerning categorization systems and lexicography), which clearly indicates the existence of formulaicity in that language. Some of the other languages where formulaic sequences have been noted include French (Arnaud and Savignon 1997; Cardey and Greenfield 2002), Spanish (Butler 1997), Italian (Tognini-Bonelli 2002), German (Bahns 1993; Gläser 1998), Swedish (Bolander 1989), Polish (Zabor 1998), Arabic (Farghal and

Obiedat 1995), Hebrew (Laufer 2000), Turkish and Greek (Tannen and Öztek 1981), and Chinese (Xiao and McEnery 2006). Not only do formulaic sequences exist in many languages, but Spöttl and McCarthy (2004) found that their multilingual participants were largely able to transfer the meaning of formulaic items across L1, L2, L3, and L4. Although it is much too early to confidently declare formulaic sequences as a universal trait of all languages, the widespread existence of formulaicity in the above languages strongly suggests that such an assumption is not unreasonable, and is probably worth allowing until proven otherwise.

If formulaic sequences are so important to language use and are so widespread in discourse, it follows that proficient speakers must have knowledge and mastery of them at some level. Some scholars claim that this phrasal knowledge is extensive. Jackendoff (1995) analysed a small corpus of TV quiz show language and concluded that formulaic sequences may be of equal if not greater significance than the lexicon of single words. Pawley and Syder (1983: 213) suggest that the number of 'sentence-length expressions familiar to the ordinary, mature English speaker probably amounts, at least, to several hundreds of thousands.' Their estimate presumably does not include phrase-length expressions, nor collocations, nor the number of greater than sentence-length expressions we know (such as poems, nursery rhymes, song lyrics, speeches from plays). The addition of these types of formulaic sequences would suggest that the phrasal component of the mental lexicon is likely to be truly vast. However, we know of no study that has directly attempted to quantify the size of a proficient user's phrasal lexicon, and so we are currently left with the assumption that because the formulaic component of discourse is large, the phrasal lexicon must be too. Still, the undisputed existence of formulaic sequences in discourse means that, at a minimum, they must have some consequences in terms of how language is acquired, processed, and used.

THE PROCESSING OF FORMULAIC SEQUENCES

There must be a reason why formulaic sequences are so widespread. The above discussion gives a sociofunctional explanation for the pervasiveness of formulaic sequences, but there is also a psycholinguistic explanation, which has perhaps been articulated most clearly by Pawley and Syder (1983): formulaic sequences offer processing efficiency because single memorized units, even if made up of a sequence of words, are processed more quickly and easily than the same sequences of words which are generated creatively. In effect, the mind uses an abundant resource (long-term memory) to store a number of prefabricated chunks of language that can be used 'ready made' in language production. This compensates for a limited resource (working memory), which can potentially be overloaded when generating language on-line from individual lexical items and syntactic/discourse rules.

It must be said that this explanation, even though intuitive, is still more assertion than demonstrated fact. However, there is considerable indirect evidence to support it, much of it coming from studies of speech production. Speech production is cognitively challenging, as de Bot (1992) illustrates:

> When we consider that the average rate of speech is 150 words per minute, with peak rates of about 300 words per minute, this means that we have about 200 to 400 milliseconds to choose a word when we are speaking. In other words: 2 to 5 times a second we have to make the right choice from those 30,000 words[2] [in the productive lexicon]. And usually we are successful; it is estimated that the probability of making the wrong choice is one in a thousand. (de Bot 1992: 11)

The holistic processing explanation suggests that one way the mind meets this daunting cognitive challenge is by using formulaic sequences. A number of speech production studies seem to support this argument. For example, Dechert (1983) studied the spoken output of a German learner of English as she narrated a story from six cartoons. He found that some of her output was marked with hesitations, fillers, and corrections, while other output was smooth and fluent. The fluent output was characterized by what Dechert labelled 'islands of reliability', which essentially describe formulaic language. Dechert suggests that islands of reliability may anchor the processes necessary for planning and executing speech in real time. Oppenheim (2000) found that the speech of her six nonnative participants contained between 48 and 80 per cent recurrent sequences with an overall mean of 66 per cent. This is further evidence that nonnatives rely on formulaic language a great deal in their efforts to produce fluent speech. Similarly, Bolander (1989) found that her learners of Swedish commonly relied on formulaic sequences in their speech. In addition, there are a number of studies on 'smooth talkers', such as auctioneers and sports announcers, who use formulaic sequences a great deal in order to fluently convey large amounts of information under severe time constraints (see Kuiper 2004).

However, just because we find formulaic sequences in speech production (and in corpora), it does not necessarily mean that they are stored as wholes in the mind. There is some preliminary production evidence that this is not always the case. Schmitt *et al.* (2004) embedded recurrent sequences derived from a corpus analysis into a passage that was then used in a dictation task wherein the individual dictation bursts were long enough to overload working memory. This meant that the dictated language needed to be reconstructed rather then being repeated rote from memory. Since the task was to repeat the dictated bursts exactly, it was assumed that the native participants would draw upon any of the target formulaic sequences they had stored in memory. Since they would be stored as wholes in memory, it was also assumed that they would be repeated fully intact, without hesitation, and with a normal stress profile. The results showed that many of the recurrent sequences were not repeated in such a manner, or even produced at all by the speakers. This suggests that these recurrent sequences may not in fact be stored as formulaic sequences in the minds of these participants.

While most speech production research and corpora investigations have shown the importance of formulaic sequences, research into the mental processing of formulaic sequences has not kept up. Studies on the processing of formulaic language have focused on the processing of idiomatic expressions (e.g. the straw that broke the camel's back), metaphorical language (e.g. Bill is a bulldozer), irony/sarcasm, proverbs, and indirect speech. Research on the processing of formulaic language has focused on the question of whether people first analyse the complete literal interpretation of such sequences (pragmatic view) or whether they understand the nonliteral meaning without first analysing the complete literal meaning of an expression (direct access view). Under the pragmatic view formulaic sequences should take longer to process and be harder to understand than literal speech. However, most studies have found that many kinds of nonliteral language are comprehended as quickly as literal speech when they are presented in context (Gibbs 1994; Glucksberg 1998). Such studies show that comprehenders quickly understand nonliteral speech in context and that it is not more difficult to understand than literal speech.

Gibbs *et al.* (1997) showed that idiomatic expressions were processed just as quickly as their literal paraphrase. Gibbs and colleagues presented participants with formulaic sequences (e.g. *He blew his stack*), a literal paraphrase (e.g. *He got very angry*), or a control phrase (e.g. *He saw many dents*) in story contexts. Following a target phrase participants performed a lexical decision on a word related to the meaning of the formulaic sequence, or a control word that was unrelated to either the formulaic sequence, the literal phrase, or the control phrase. Gibbs *et al.* found that reading times for formulaic sequences were faster than those for control phrases, but that there was no reading time difference for formulaic sequences and their literal paraphrases. Response times to the words related to the meaning of the formulaic sequences were faster following the formulaic sequences than literal paraphrases or control phrases. This was taken as an indication that participants accessed the conceptual metaphors when understanding formulaic sequences, but significantly less so when processing their literal paraphrases. Overall findings from studies like those of Gibbs *et al.* provide evidence that readers quickly understand formulaic sequences in context and that they are not more difficult to understand than literal speech. Crucially such results indicate that it is unlikely that readers first access the complete literal meaning of an expression before the nonliteral one. If this were the case, longer reading times for formulaic sequences than their literal paraphrases should have been found.

While the literature examining native speaker processing of nonliteral speech has focused on the processing of formulaic sequences like metaphors, idioms, irony/sarcasm, and proverbs, second language research has focused on whether second language learners can interpret indirect speech (e.g. Takahashi and Roitblat 1994; Taguchi 2005, in press). For example, Takahashi and Roitblat presented native and nonnative speakers with short stories that induced either a direct or indirect interpretation of a request like '*Can you park the car over there?*' Contexts and requests were followed by a paraphrase of either the direct or indirect interpretation of the request. There was no difference in reading speed for direct and indirect requests by either the native or nonnative speakers. Paraphrases of indirect requests were read more quickly than direct ones. However, paraphrases were read more quickly when they corresponded to the interpretation appropriate to the context. These results suggest that both L1 and L2 speakers access both the direct and indirect interpretations of a request. Similar to the native speaker literature there is NO processing cost for the nonliteral/indirect speech.

While the above studies inform about the processing of formulaic sequences from a 'literal versus idiomatic meaning' perspective, they do not substantiate the hypothesis put forward by Pawley and Syder (1983): if formulaic language is hypothesized to be used by the mind because it is somehow easier or quicker than the generation of language through other means (e.g. syntactic rules with lexical fillers), then that processing ease or advantage needs to be demonstrated. Two studies attempted to directly address this issue. Underwood *et al.* (2004) used the eye-movement 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*). The results were then compared to measurements of the same words in non-formulaic 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 requires 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 non-formulaic contexts, and for a shorter duration. This is evidence for a processing advantage for formulaic sequences vs. creative language. Moreover, this advantage was partially shared by proficient nonnatives as well. In a related study, Schmitt and Underwood (2004) used a participant-paced word-by-word reading task to examine at which point in a formulaic sequence recognition occurs (e.g. after the 1st word, 2nd word, etc.). Unfortunately, the native participants processed all words so quickly that it was impossible to determine the recognition point.

The processing of formulaic sequences may seem like a specialist issue, but it is fast becoming essential for the wider field of applied linguistics, as some current views of language acquisition and use conclude that formulaic sequences are an integral part of the process. For example, connectionism suggests that repeated input can lead to memory traces between particular linguistic elements becoming strengthened. This seems a plausible explanation for how the words within a formulaic sequence become associated (although other views are possible), and would imply that formulaic sequences form a major part of any developing awareness of the structure of language. In essence, connectionist/emergentist accounts of language acquisition suggest that the key to learning is the ability to extract patterns from language input. Because formulaic sequences are frequent and relatively salient (because they are typically linked to specific meanings or functions), they are likely to be included in the extracted linguistic patterns. (See Hopper 1998; MacWhinney 1999; and Ellis 2003, 2006a, 2006b for more on this approach to language acquisition.) Thus, a better understanding of formulaic sequences and how they are processed can only help to develop and refine these developing theories.

The most elementary place to develop this understanding is to empirically confirm (or not) the assertion by Pawley and Syder and others that formulaic sequences indeed facilitate language processing. The above studies point in this direction, but more research is required before such an important assumption can be accepted as a given. For example, Underwood *et al.* (2004) showed processing advantages for terminal words in a sequence, but studies looking at the processing behavior of complete sequences are also needed.

In order to determine whether formulaic sequences indeed hold a processing advantage, one must consider the context in which they reside. After all, formulaic sequences do not exist in isolation, but rather in discourse, and many factors have been shown to affect processing speed. Research shows that word frequency affects word recognition (e.g. Rayner and Balota 1989). In general, findings indicate that high frequency words are processed more quickly than low frequency words. Gernsbacher (1984) demonstrated that familiarity ratings for experimental stimuli may be a better predictor of response times than written frequency norms, especially for low frequency words. She showed that overall the relation between familiarity and frequency is highly linear. However, for low frequency words the relationship is less linear. Based on Gernsbacher's results it appears that for low frequency words, familiarity may be a better predictor of response time than frequency. Word length also appears to influence the speed word recognition. However, a length effect independent of word frequency has been hard to find. In naming tasks, naming time increases as a function of the number of syllables in a word (e.g. Eriksen *et al.* 1970; Klapp 1974). Priming is another well-known effect, whereby word recognition is speeded by presentation of a previous related word (e.g. Meyer and Schvaneveldt 1971). It is interesting to note that words in a collocation also prime each other (Schooler 1993, cited in Ellis 2006a). In order to explore whether it is actually the formulaicity of formulaic sequences which leads to any processing speed advantage, we will need to consider and control for as many of these extraneous factors as possible.

Controlling for these other factors, we will focus directly on the issue of processing advantage using the mode of reading. We ask the following research question:

1 Are formulaic sequences read more quickly than equivalent non-formulaic sequences?

The study will also provide an opportunity to explore the processing of idiomatic versus literal interpretation of formulaic sequences:

2 Are figurative renderings of formulaic sequences (*a breath of fresh air* = a new approach) read more quickly or slowly than the literal renderings of the same sequences (*a breath of fresh air* = breathing clean air outside)?

METHODOLOGY

If formulaic sequences such as *everything but the kitchen sink* are processed as chunks rather than word by word, they should be read more quickly in context than control phrases like *everything in the kitchen sink*. The hypothesized processing advantage for formulaic sequences was tested by comparing the reading time of formulaic sequences versus control phrases.

Instruments

This study is concerned with the processing, rather than identification, of formulaic sequences, and so we wished to use unambiguous cases as our targets. We chose mostly idioms, because idioms are clearly formulaic in nature since they represent idiosyncratic meanings which cannot generally be derived from the sum of the individual words in the string. We extracted some formulaic sequences from the materials used in Underwood *et al.* (2004), and added a number more from the *Oxford Learner's Dictionary of English Idioms* (Warren 1994). The candidate formulaic sequences were subjected to a frequency analysis based on the British National Corpus (BNC). Candidates with relatively low frequencies were deleted from the list. We also wished the formulaic sequences to be well-known to the native participants. This had already been established for the sequences from the Underwood *et al.* study, but we needed to confirm this for the nine sequences taken from the idiom dictionary. Thus all the formulaic sequences were embedded in modified cloze tests with short contexts, such as the following example:

> I had a bunch of problems the last few days of the semester. My house was burgled, my bicycle was stolen, and my computer crashed so I had to hand in my assignment late. I could go on and on about my disasters, but to make a lo__________ st__________ sh__________, this was one of my worst semesters ever.

These cloze contexts were given to ten native first-year undergraduates, and the nine sequences were all well-known, being produced by 8–10 students. Finally, the most frequent and best-known twenty formulaic sequences were chosen for the study.

Twenty control phrases were created to match the formulaic sequences by rearranging the words in the formulaic sequences, taking particular care that the controls could not be interpreted with either the idiomatic or literal meanings of the formulaic sequence, for example *hit the nail on the head* → *hit his head on the nail*. Usually this rearrangement required the substitution of one or two words for the control to make sense in the passage. Care was taken to ensure that factors known to influence processing speed were controlled for. Therefore, the main content words were always kept, with only the function words being substituted. In the above example, *the* (the most frequent word in English) was substituted by *his* (the 26th most frequent word). In the majority of cases, function words were substituted with ones of a higher frequency, and in all cases the differences in terms of rank frequency order were relatively small, as virtually all function words are among the most frequent words in English.[3] The number of words in the formulaic and control sequences were the same in all cases, as were the number of syllables in all but two cases. By using essentially the same words in the controls as formulaic sequences, we were able to control for any individual characteristics of the words that might have contributed to variability between the three conditions, such as word length, word frequency or part of speech.

The target phrases (i.e. formulaic sequence or control phrase) were embedded into twenty passages. The context of the passages was written to force either an idiomatic or literal interpretation of the formulaic sequences (e.g. *take the bull by the horns* = 'attack a problem' vs. 'wrestle an animal'). In all, there were 60 target phrases: 20 formulaic sequences presented in contexts supporting their idiomatic interpretation, the same 20 sequences presented contexts supporting their literal interpretation, and the 20 control phrases. Thus each formulaic sequence had three conditions (idiomatic, literal, control). Each of the three conditions appeared once for each formulaic sequence, although not in the same passage. Each passage had at least two target phrases and no more than four.

Content words in the target phrases were not used previously in the passage so as to avoid priming effects. The passages were subjected to frequency analysis through *The Compleat Lexical Tutor (v.2)* (Cobb) to ensure that low frequency vocabulary was kept to a minimum. Less than 6 per cent of the words were not within the first 2,000 words of English or on the Academic Word List (Coxhead 2000), and most of these were proper names (e.g. *John, Africa*), so the texts would pose no problems for native speakers and cause relatively few problems for nonnatives. Finally, a multiple-choice comprehension question for each passage was devised to ensure participants read the contexts conscientiously.

Three of the passages with their comprehension questions are illustrated in the Appendix (available online). For the reader's convenience, the idiomatic condition is marked in italics, the literal condition in bold, and the control condition is underlined. In the experiment, the target phrases were given in normal font.

Participants

Nineteen native English speakers from the University of Nottingham, mostly undergraduates, formed the native group. The nonnative group consisted of twenty L2 English speakers from the University of Nottingham. They were mostly postgraduates studying on the MA-ELT program and were chosen to include a mixture of L1s. All participants were paid £5 for their participation, with an additional £2 incentive payment if they answered all comprehension questions correctly under a time limit.

Procedure

Passages like those in the Appendix (available online) were presented on a computer using a participant-paced line-by-line reading procedure. Each trial began by asking a participant to press the 'R' button to indicate they were ready to begin reading a passage. Once 'R' was pressed, the first line in a passage appeared. Participants pressed the spacebar when they finished reading the line, causing the phrase they were reading to disappear and the subsequent line to be displayed. Participants were asked to read each passage for comprehension as quickly as possible. Reading times were collected for each line. When participants finished reading a passage they were asked a comprehension question about it. Before beginning the experiment, participants read instructions that described the task. Following the instructions, participants were given two practice passages to become familiar with the task. The practice trials did not contain any formulaic sequences. Participants were not made aware that their knowledge of formulaic sequences played any role in the experiment.

RESULTS

In this self-paced reading task, we are mainly interested in whether the reading times for the formulaic sequences are quicker than the reading times for the control phrases. Table 1 illustrates these times in milliseconds (ms). The mean reading times for both the native and nonnative speakers were submitted to separate one-way analyses of variance (ANOVA) with either participants or items as the random variable. The results of the native speakers will be discussed first. For native speakers, the main effect of phrase type was significant by both participants, $F_1(2,18) = 27.1$, $p < .001$, and by items $F_2(2,19) = 9.5$, $p < .001$. Therefore, the prediction that formulaic sequences are processed as a unit and should be read more quickly than

Table 1: Mean reading times (ms) for native and nonnative English speakers (standard error in parentheses)

Condition	Native speakers	Nonnative speakers
Idiomatic formulaic sequence	1,111.7 (90.1)	2,035.1 (130.5)
Literal formulaic sequence	1,137.6 (83.6)	2,133.8 (155.1)
Control phrase	1,376.2 (91.0)	2,291.9 (151.2)

control phrases was examined by way of paired t-tests. The t-tests indicated significantly shorter reading times for formulaic sequences used idiomatically than for control phrases by both analyses of participants t_1 (18) = 5.7, $p < .001$, and items t_2 (19) = 3.5, $p < .05$. There were also significantly shorter reading times for formulaic sequences used literally than for control phrases by analyses of participants t_1 (18) = 5.7, $p < .001$, and items t_2 (19) = 3.4, $p < .05$. However, there was no significant difference in reading times for formulaic sequences used idiomatically and literally by either participants t_1 (18) = 0.9, $p > .05$, or items t_2 (19) = 0.6, $p > .05$.

The mean reading times for each nonnative participant and item were also submitted to separate one-way analyses of variance (ANOVA) with either participants or items as the random variable. The main effect of phrase type was significant by participants, $F_1(2,19) = 7.4$, $p < .05$, but not by items $F_2(2,19) = 1.7$, $p > .05$. The lack of significance by items is probably due to the large variability in reading times for L2 participants. As with the native speaker results, the L2 English speakers' results were submitted to paired *t*-tests to see whether formulaic sequences are processed more quickly than control phrases. Significantly shorter reading times were found for formulaic sequences used idiomatically than for control phrases by analysis of participants t_1 (19) = 3.8, $p < .001$, and items t_2 (19) = 3.5, $p < .05$. As with native speakers, there were also significantly shorter reading times for formulaic sequences used literally than for control phrases by analysis of participants t_1 (19) = 2.3, $p < .05$, and items t_2 (19) = 3.5, $p < .05$. There was no significant difference in reading times for formulaic sequences used idiomatically and literally by either participants t_1 (19) = 1.5, $p > .05$, or items t_2 (19) = 0.5, $p > .05$.

As predicted, formulaic sequences in contexts supporting their idiomatic interpretation were read more quickly than control phrases for both natives and nonnatives. Interestingly, formulaic sequences in contexts supporting their literal interpretation were also read more quickly than control phrases by both groups. Furthermore, there was no reading time difference for formulaic sequences in contexts supporting their idiomatic interpretation and those in contexts supporting their literal interpretation. This pattern of results indicates that even when a formulaic sequence is used literally, it has a processing advantage. Moreover, this pattern held for both groups, indicating that even L2 speakers have a processing advantage for formulaic sequences no matter whether they are used idiomatically or literally.

DISCUSSION

The main purpose of this study was to explore the commonly asserted and widely accepted notion that formulaic sequences are more easily processed than nonformulaic language. The results provide evidence that this is indeed the case. Natives read the formulaic sequences faster than the equivalent controls. This study, combined with the research of Gibbs and colleagues (1997), and the eye-movement results from Underwood *et al.* (2004), provide converging evidence to support the processing advantage of formulaic sequences, at least when reading. To date there is no direct evidence of this advantage for auditory processing, although speakers' preference for formulaic sequences when under heavy time constraints seems to provide convincing indirect evidence (e.g. Kuiper 2004).

Interestingly, the processing advantage for formulaic sequences seems to extend to proficient L2 speakers as well. Both Underwood *et al.* and this study show that nonnatives read formulaic sequences more quickly than equivalent non-formulaic language. Of course the reading speeds are slower than for natives, as one would expect, but even at this slower speed formulaic sequences show an advantage. This is good news because skilful use of formulaic sequences is generally considered to come late in the acquisition process, and it is not unreasonable to question whether mastery can ever truly be achieved. Indeed, control of formulaic sequences is one element that can normally be relied upon to distinguish between natives and even relatively advanced nonnatives. For example, many items on tests of advanced English focus on issues of collocation or phraseology. Nonnatives obviously benefit when they produce more pragmatically appropriate language through the skilful use of formulaic sequences, but this study suggests that they can also enjoy the same type of formulaic processing advantage as natives.

The second issue explored in this study is the processing differences between formulaic sequences when interpreted idiomatically versus literally. Similar to the work by Gibbs and colleagues, we found that both renderings were read more quickly than the control phrases. This shows that the formulaic sequences were processed more quickly than equivalent nonformulaic language, but it did not seem to matter much whether the sequences carried an idiomatic or a literal meaning. Any possible explanation for this involves several factors.

First, it appears that the type of idioms in this study are not normally used in discourse with literal meanings. The BNC was checked to compare idiomatic versus literal usages, and in most cases there were very few, if any, literal usages of the formulaic sequences in the study. The exceptions were *breath of fresh air* which was used idiomatically 60 times and literally 33 times and *his back against the wall* (idiomatic 12; literal 5).[4] Given the relative infrequency of literal renderings, formulaic sequences such as this may well be processed as wholes as a default. This would account for equally quick reading times for idiomatic and literal meanings in the study.

Second, the context before the formulaic sequences in the passages constrained the meaning to either an idiomatic or literal one before the sequence was read, yet there was no difference in reading speed. This suggests that even though formulaic sequences might be initially processed as wholes, the mind is quick to activate the literal meaning when necessary. If this was not the case, then one would expect the lines following a literal formulaic sequence to be processed relatively more slowly than other lines in the passage, as participants attempted to work through garden-path confusion when they discovered that the idiomatic meaning did not fit. We checked the latencies for lines following both literal and idiomatic formulaic sequences and could find no pattern of increased reading times after only the literal sequences.

Third, if one analyses formulaic sequences such as those used in this study, one discovers that the idiomatic renderings are often extensions of the literal meaning. For example, if one thinks about *scrape the bottom of the barrel*, it is not difficult to associate the literal digging into the bottom of a barrel for the last item of something with the idiomatic rendering of 'something of very low quality'. Howarth (1998b: 26) suggests that much of formulaic language is gradable in terms of idiomaticity, and gives the example of *to let off steam* (= 'to display anger'). He suggests that such sequences 'will be more or less analysed . . . [and that] speakers may vary in the degree to which the literal sense of *steam* (as 'water vapour') is activated.' Moreover, he believes that the preceding context can help to decompose a formulaic sequence into its constituent parts, making a literal interpretation more likely. For example, the literal meaning of *sailing* in the idiom *plain sailing* may become salient when preceded by the previous context: *He conceived the idea of crossing Antarctica. It was not all <u>plain sailing</u>*. The contexts used in this study's passages give just this kind of meaning prompt, and so it might be that literal interpretations can be processed as quickly as idiomatic interpretations simply because they are primed by the preceding context.

CONCLUSION

Similar to the findings of Gibbs *et al.* (1997), our study showed a significant processing advantage for formulaic sequences over nonformulaic language. Furthermore, this advantage appears to be in place for both idiomatic and literal renderings of the formulaic sequences. Crucial to our study this benefit was seen for both L1 and L2 English speakers. These results add to a slowly growing body of evidence supporting the view that readers quickly understand formulaic sequences in context and that they are NOT more difficult to understand than literal speech. Findings like ours and those of Gibbs and colleagues suggest that it is unlikely that readers first access the complete literal meaning of a formulaic sequence before the nonliteral one. If this were the case formulaic sequences used idiomatically should have taken longer to read than those used literally.

Because the formulaic sequences were read more quickly than the control phrases, our results support the assertion that formulaic sequences are involved in more efficient language processing. However, it must be noted that the absolute amount of evidence for this is still small, and more psycholinguistic research on this issue needs to be done. Using a methodology like eye-tracking, and in particular first pass reading times, would provide evidence about the fast and automatic processing of formulaic sequences from a different perspective (Frenck-Mestre 2005). If the current pattern of results were replicated using the eye-tracking methodology, this would provide even stronger evidence for the processing advantage of formulaic sequences. In addition, eye-tracking methodology might shed additional light on processing differences for formulaic sequences used idiomatically and literally. Given the importance of formulaic sequences in language use, anything we can do to understand how they are processed can only help to expand our understanding of language processing in general.

Final manuscript received December 2006

SUPPLEMENTARY DATA

Supplementary material mentioned in the text is available online to subscribers at http://applij.oxfordjournals.org.

ACKNOWLEDGEMENTS

We wish to thank Gabi Kasper and two anonymous *Applied Linguistics* reviewers for their very helpful comments on an earlier version of this paper.

NOTES

[1] One of the problems in the area of formulaic language is terminology. Wray (2002: 9) found over fifty terms to describe the phenomenon, but suggests using *formulaic sequences.* We use it in our paper, but wish it to be interpreted as a broad cover term, including all of the various types of multi-word units and collocations, including, but not limited to, idioms, lexical bundles, and lexical phrases.

[2] The estimates of how many words native speakers know vary. See Nation (2001: chapter 1) for an overview of vocabulary size.

[3] The frequency rankings were taken from: A. Kilgarriff, BNC database and word frequency lists. Internet site: <www.ibri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html>. Accessed 26 June 2005.

[4] Interestingly, extending the sequence by one word usually constrained the meaning (i.e. *for a breath of fresh air* was always used in the literal sense, while *was a breath of fresh air* led to an idiomatic interpretation). However, the control context in the passage did not use any of these 'extended sequence' words.

REFERENCES

Altenberg, B. 1998. 'On the phraseology of spoken English: The evidence of recurrent word-combinations' in A. P. Cowie (ed.): *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 101–22.

Arnaud, P. and **S. Savignon.** 1997. 'Rare words, complex lexical units and the advanced learner' in T. Huckin and J. Coady (eds): *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press, pp. 157–73.

Bahns, J. 1993. 'Lexical collocations: A contrastive view,' *English Language Teaching Journal* 47: 56–63.

Biber, D., S. Conrad, and **V. Cortes.** 2004. 'Lexical bundles in university teaching and textbooks,' *Applied Linguistics* 25: 377–405.

Biber, D., S. Johansson, G. Leech, S. Conrad, and **E. Finegan.** 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Bolander, M. 1989. 'Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish' in K. Hyltenstam and L. Obler (eds): *Bilingualism across the Lifespan*. Cambridge: Cambridge University Press, pp. 73–86.

Butler, C. S. 1997. 'Repeated word combinations in spoken and written text: Some implications for functional grammar' in C. Butler, J. Connolly, R. Gatward, and R. Vismans (eds): *A Fund of Ideas: Recent Developments in Functional Grammar.* Amsterdam: IFOTT, University of Amsterdam, pp. 60–78.

Cardey, S. and **P. Greenfield.** 2002. 'Computerized set expression dictionaries: Analysis and design' in B. Altenberg and S. Granger (eds): *Lexis in Contrast: Corpus-based Approaches*, pp. 231–8.

Cobb, T. *The Compleat Lexical Tutor (v.2).* Internet resources available at <http://www.lextutor.ca>.

Coulmas, F. 1979. 'On the sociolinguistic relevance of routine formulae,' *Journal of Pragmatics* 3: 239–66.

Coulmas, F. 1981. *Conversational Routine*. The Hague: Mouton.

Cowie, A. 1998a. 'Introduction' in A. P. Cowie (ed.): *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 1–20.

Cowie, A. 1998b. 'Phraseological dictionaries: Some East–West comparisons' in A. P. Cowie (ed.): *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 145–60.

Coxhead, A. 2000. 'A new academic word list,' *TESOL Quarterly* 34: 213–38.

de Bot, K. 1992. 'A bilingual production model: Levelt's ''speaking'' model adapted,' *Applied Linguistics* 13: 1–25.

Dechert, H. 1983. 'How a story is done in a second language' in C. Faerch and G. Kasper (eds): *Strategies in Interlanguage Communication*. London: Longman, pp. 175–95.

Drew, P. and **E. Holt.** 1998. 'Figures of speech: Figurative expressions and the management of topic transition in conversation,' *Language in Society* 27: 495–522.

Ellis, N. 2003. 'Constructions, chunking, and connectionism: The emergence of second language structure' in C. J. Doughty and M. H. Long (eds): *The Handbook of Second Language Acquisition.* Malden, MA: Blackwell, pp. 63–103.

Ellis, N. 2006a. 'Language acquisition as rational contingency learning,' *Applied Linguistics* 27: 1–24.

Ellis, N. 2006b. 'Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning,' *Applied Linguistics* 27: 164–94.

Eriksen, C., M. Pollack, and **W. Montague.** 1970. 'Implicit speech: Mechanisms in perceptual encoding?' *Journal of Experimental Psychology* 84: 502–7.

Erman, B. and **B. Warren.** 2000. 'The idiom principle and the open-choice principle,' *Text* 20: 29–62.

Farghal, M. and **H. Obiedat.** 1995. 'Collocations: A neglected variable in EFL,' *International Review of Applied Linguistics* 33: 315–31.

Firth, J. 1935. 'The technique of semantics,' *Transactions of the Philological Society*: 36–72.

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, and M. Swain (eds): *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Harlow: Longman, pp. 75–93.

Frenck-Mestre, C. 2005. 'Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings,' *Second Language Research* 21: 175–98.

Gernsbacher, M. 1984. 'Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy,' *Journal of Experimental Psychology: General* 113: 256–81.

Gibbs, R. 1994. *The Poetics of the Mind: Figurative Thought, Language, and Understanding*. New York: Cambridge University Press.

Gibbs, R., J. Bogadanovich, J. Sykes, and **D. Barr.** 1997. 'Metaphor in idiom comprehension,' *Journal of Memory and Language* 37: 141–54.

Gläser, R. 1998. 'The stylistic potential of phraseological units in the light of genre analysis' in A. Cowie (ed.): *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 125–43.

Glucksberg, S. 1998. 'Metaphor,' *Current Directions in Psychological Science* 7: 39–43.

Hopper, P. 1998. 'Emergent grammar' in M. Tomasello (ed.): *The New Psychology of Languag*e. Hillsdale, NJ: Lawrence Erlbaum, pp. 155–175.

Howarth, P. 1998a. 'The phraseology of learners' academic writing' in A. Cowie (ed.): *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press, pp. 161–86.

Howarth, P. 1998b. 'Phraseology and second language proficiency,' *Applied Linguistics* 19: 24–44.

Jackendoff, R. 1995. 'The boundaries of the lexicon' in M. Everaert, E. van der Linden, A. Schenk, and R. Schreuder (eds): *Idioms: Structural and Psychological Perspectives*. Hillsdale, NJ: Lawrence Erlbaum, pp. 133–66.

Klapp, S. 1974. 'Syllable-dependent pronunciation latencies in number naming, a replication,' *Journal of Experimental Psychology* 102: 1138–40.

Kuiper, K. 2004. 'Formulaic performance in conventionalised varieties of speech' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 37–54.

Laufer, B. 2000. 'Avoidance of idioms in a second language: The effect of L1–L2 degree of similarity,' *Studia Linguistica* 54: 186–96.

McCarthy, M. and **R. Carter.** 2002. 'This that and the other: Multi-word clusters in spoken English as visible patterns of interaction,' *Teanga (Yearbook of the Irish Association for Applied Linguistics)* 21: 30–52.

MacWhinney, B. 1999. *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum.

Meyer, D. and **R. Schvaneveldt.** 1971. 'Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations,' *Journal of Experimental Psychology* 90: 227–35.

Nation, I. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nattinger, J. and **J. DeCarrico.** 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

O'Keeffe, A., M. McCarthy, and **R. Carter.** 2007. *From Corpus to Classroom: Language Use and Language Teaching.* Cambridge: Cambridge University Press.

Oppenheim, N. 2000. 'The importance of recurrent sequences for nonnative speaker fluency and cognition' in H. Riggenbach (ed.): *Perspectives on Fluency.* Ann Arbor: University of Michigan Press, pp. 220–40.

Pawley, A. and **F. Syder.** 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. Richards and R. Schmidt (eds): *Language and Communication*. London: Longman, pp. 191–225.

Rayner, K. and **D. Balota.** 1989. 'Parafoveal preview and lexical access during eye fixations in reading' in W. Marslen-Wilson (ed.): *Lexical Representation and Process*. Cambridge, MA: The MIT Press, pp. 261–90.

Schmitt, N. (ed.). 2004. *Formulaic Sequences*. Amsterdam: John Benjamins.

Schmitt, N. 2005. 'Grammar: Rules or patterning?' *Applied Linguistics Forum* 26/2. Internet resource available at <http://www.tesol.org/NewsletterSite/view.asp?nid=2857>.

Schmitt, N. and **N. Carter.** 2004. 'Formulaic sequences in action: An introduction' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 1–22.

Schmitt, N. and **G. Underwood.** 2004. 'Exploring the processing of formulaic sequences through a self-paced reading task' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 173–89.

Schmitt, N., S. Grandage, and **S. Adolphs.** 2004. 'Are corpus-derived recurrent clusters psycholinguistically valid?' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 127–51.

Sinclair, J. 1966. 'Beginning the study of lexis' in C. Bazell, J. Catford, M. Halliday, and R. Robins (eds): *In Memory of J. R. Firth*. London: Longman, pp. 410–30.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Lexis, Corpus, Discourse*. London: Routledge.

Sorhus, H. 1977. 'To hear ourselves—Implications for teaching English as a second language,' *English Language Teaching Journal* 31: 211–21.

Spöttl, C. and **M. McCarthy.** 2004. 'Comparing knowledge of formulaic sequences across L1, L2, L3, and L4' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 190–225.

Stubbs, M. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Taguchi, N. 2005. 'Comprehending implied meaning in English as a foreign language,' *The Modern Language Journal* 89: 543–562.

Taguchi, N. in press. 'Development of speed and accuracy in pragmatic comprehension in English as a foreign language', *TESOL Quarterly*.

Takahashi, S. and **H. Roitblat.** 1994. 'Comprehension process of second language indirect requests,' *Applied Psycholinguistics* 15: 475–506.

Tannen, D. and **P. Öztek.** 1981. 'Health to our mouths: Formulaic expressions in Turkish and Greek' in F. Coulmas (ed.): *Conversational Routine*. The Hague: Mouton, pp. 37–54.

Teliya, V., N. Bragina, E. Oparina, and **I. Sandomirskaya.** 1998. 'Phraseology as a language of culture: Its role in the representation of a collective mentality' in A. Cowie (ed.): *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, pp. 55–75.

Tognini-Bonelli, E. 2002. 'Functionally complete units of meaning across English and Italian' in B. Altenberg and S. Granger (eds): *Lexis in Contrast: Corpus-Based Approaches*. Amsterdam: Benjamins Press, pp. 73–95.

Underwood, G., N. Schmitt, and **A. Galpin.** 2004. 'The eyes have it: An eye-movement study into the processing of formulaic sequences' in N. Schmitt (ed.): *Formulaic Sequences*. Amsterdam: John Benjamins, pp. 155–72.

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

Wittgenstein, L. 1958. *Philosophical Investigations.* Oxford: Basil Blackwell.

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

Xiao, R. and **T. McEnery.** 2006. 'Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective,' *Applied Linguistics* 27: 103–29.

Zabor, L. 1998. 'Idioms as nontransferable structures'. Paper presented at the 11th International Conference on Foreign and Second Language Acquisition, Szczyrk, Poland.