*Applied Linguistics* 2011: 32/2: 129–148 
doi:10.1093/applin/amq035 Advance Access published on 9 November 2010

# The Processing of Malformed Formulaic Language

NEIL MILLAR

University of Birmingham
E-mail: n.j.millar@bham.ac.uk

The prevalence of formulaicity in naturally occurring language use points to an important role in the way language is acquired, processed, and used. It is widely recommended that second-language instruction should ensure that learners develop a rich repertoire of formulaic sequences. If this is justified, it follows that learner failure to use formulaic sequences should present some barrier to communication. However, it seems that few researchers have sought to objectively evaluate how learner deviations from the target-language (formulaic or otherwise) impact on online processing. Operationalizing formulaic sequence through collocation, this article reports the combination of corpus-based approaches and psycholinguistic experimentation to investigate the processing by native speakers of learner collocations that deviate from target-language norms. Results show that such deviations are associated with an increased and sustained processing burden. These findings support the widely asserted claim that formulaic sequences offer processing advantages and provide empirical support for the importance of formulaic sequences in language learning. Usage-based models form the basis for some hypotheses concerning cognitive processes that underlie the increase in processing demands.



## INTRODUCTION

It has long been recognized that natural language use often takes the form of recurrent clusters of words (Kjellmer 1994). Corpus linguistics has demonstrated that the simple combination of grammar and lexis alone cannot account for the lack of variation shown in naturally occurring language (Kjellmer 1984; Altenberg 1993). As Pawley and Syder (1983: 193) famously put it, 'native speakers do *not* exercise the creative potential of syntactic rules to anything like their full extent'. Formulaicity in natural language use appears, therefore, not to be peripheral, as was once thought, but rather to play a significant role in the way in which we acquire, process and use language. The literature on the subject suggests two main factors as motivating the prevalence of formulaic language: the sociofunctional and the psycholinguistic. Sociofunctional explanations posit that formulaic language fulfils human beings' 'desire to sound like others in the speech community' (Wray 2002: 74). From the psycholinguistic perspective, it is widely proposed that formulaic

language offers processing advantages (Kuiper and Haggo 1984; Wray 2000, 2002; Wray and Perkins 2000; Conklin and Schmitt 2008). That is, prefabricated units support comprehension because the recipient does not need to analyse every word bottom up. An utterance is recognized as formulaic by the hearer and the corresponding phrase is accessed within the mental lexicon *as a unit*, thus avoiding computationally demanding analyses of the input (Skehan 1998: 38). Formulaic sequences, in this way, appear to 'make it possible for us to protect our own interests by producing language that is fluent and easily understood' (Wray 2002: 281).

The functions of formulaic language proposed above have implications for second or foreign language learning. With regard to the sociofunctional claims, Skehan (1998: 39) comments that learners who restrict themselves to production based on the simple combination of grammar and lexis 'will be for ever marked as non-members of the speech community they aspire to'. The present study is however concerned with the psycholinguistic claims. Based on these, it can be hypothesized that a close alignment of the formulaic sequences used by the speaker to those internalized in the mental lexicon of the hearer will result in lesser processing demands and lead to gains in communication (Skehan 1998: 38). Although the evidence in support of the psycholinguistic claims is strong, only a few studies have investigated the processing of formulaic sequences by second language learners (e.g. Underwood *et al.* 2004; Jiang and Nekrasova 2007; Conklin and Schmitt 2008), and none have looked at the processing of formulaic sequences *produced by language learners*. The self-paced reading experiment reported in this article investigates whether, and how, learner deviations from formulaic norms impact on the addressee. It thus seeks evidential support for the role of formulaic sequences in language learning and teaching. The following section reviews research that has explored processing of formulaic sequences, discusses the role of formulaic language in second or foreign language learning, and outlines the hypothesis that is addressed in this study. The subsequent sections report the methodology and results of the experiment. The results are then discussed and some speculative hypotheses regarding the processing of malformed formulaic language are proposed. Finally, in the conclusion the findings are summarized and some methodological implications are explored.

## PSYCHOLINGUISTIC PERSPECTIVES ON FORMULAIC LANGUAGE

Wray (2002: 9) defines a formualic sequence as word combinations that are, or appear to be, prefabricated—that is stored holistically and retrieved as a whole 'from memory at time of use, rather than being subject to generation or analysis by the language grammar'. This definition of formulaicity encompasses phenomena ranging from strings which are idiomatic, immutable and syntactically irregular (e.g. *by and large*) to sequences which although transparent

and syntactically decomposable (e.g. constructions with open slots and lexical collocations) appear to be accessed as single items (Wray and Perkins 2000: 1–2). As the present research is concerned with *collocation*, this 'inclusive' definition of a formulaic sequence provides a useful starting point. Prefabricated language may represent a mechanism by which we are able to minimize the effects of limitations of working memory—a limited-capacity system (Baddeley 2000). If required to encode/decode meanings on a word-by-word basis in sequential fashion, the demands on working memory would be great—particularly in circumstances where additional constraints are in place. It has been widely argued that frequently occurring strings of language can be stored as whole units, 'chunks', in the long-term memory (Pawley and Syder 1983; Bybee and Scheibman 1999; Wray 2002; Schmitt and Carter 2004). Access to these pre-fabricated chunks by working memory requires substantially less processing capacity than access to, and subsequent manipulation of, their individual components. This reduces the processing burden. Chunking of linguistic information, that is 'the development of permanent sets of associative connections' (Ellis 1996: 107), may therefore represent a mechanism by which we are able to minimize the effects of limitations of working memory, and which enables the maintenance of fluency. The term here is used to mean the speed, fluidity and accuracy of *both* productive and receptive skills—characteristics that 'correspond to psychologically measurable aspects of complex cognitive performance' (Segalowitz 2000: 200). In the field of language pedagogy/second-language acquisition (SLA), however, *accuracy* is often differentiated from *fluency*. Brumfit (2000: 63), for example, sees them as distinct and polarized—he notes, however, that this is 'essentially a methodological distinction, rather than one based in psychology or psycholinguistics'.

The identification of what, according to the above definition, is essentially a psychological phenomenon has often relied on frequency of occurrence in corpora—the more frequently a sequence is required, the more likely it is to be holistically represented and processed. From a psycholinguistic perspective, language users' sensitivity to frequency of occurrence and co-occurrence is of course well established (see Ellis 2002a, 2002b; Bybee 2007). However, while acknowledging the link between frequency and formulaicity, Wray and Perkins (2000: 6–7) suggest that frequency may not necessarily be a defining feature, citing the relative rarity in real texts of formulaic sequences with a high culturally-based familiarity.

What corpus studies do show is that fixed or semi-fixed word combinations make up a substantial proportion of natural language use. This is perhaps best exemplified by the phenomenon of collocation as observed in corpus linguistics—the co-occurrence of lexical items 'with greater than random probability' (Hoey 1991: 6–7). Sinclair (1991) and Stubbs (1996) suggest that *all* lexical items have collocations. The phenomenon extends beyond the two-word level—larger chunks of language bound together by collocational ties are common. In their corpus-based grammar of written and spoken English, Biber *et al.* (1999) found that three- to four-word *lexical bundles*,

contiguous recurrent word combinations, account for as much 28 per cent of conversational text and 20 per cent of academic texts. Lexical bundles can vary in frequency, function and form according to register (Biber *et al.* 1999, 2004; Hyland 2008). The prevalence of seemingly prefabricated language led Sinclair (1991: 110) to propose that we deal with language using two systems: the *open choice principle* in which we select individual words to fill slots, constrained only by grammaticalness; and the *idiom principle*, in which we have at our disposal a large number of formulaic sequences. The latter forms the default mode, and only when 'there is good reason, the interpretative process switches to the open-choice principle, and quickly back again' (Sinclair 1991: 14)—for example, when lexical choices are unexpected. Biber *et al.* (2004: 400) also reach a psychological conclusion, stating that reliance on certain lexical bundles in classroom talk is 'evidence that lexical bundles are stored as unanalysed multi-word chunks'.

Evidence of formulaic language as a processing short-cut is supplied by studies of speech production in situations requiring fluent transfer of large quantities of information under severe constrains. Kuiper (1996, cited in Kuiper 2004) examines the language of auctioneers and sports commentators, so-called 'smooth talkers'. Fast-paced commentaries on horse races contain more formulaicity than commentaries on slower paced cricket. He attributes the highly formulaic nature of language in these contexts to the need to lighten the high processing load and the pressure to react instantly. Wray (2002: 78) suggests that this points to a 'close correlation between processing pressure and information to be computed and expressed within a given time-frame, and the proportion of formulaicity in language'. Phonological coherence may provide further evidence of formulaicity in spoken language. Van Lancker *et al.* (1981) analyses differences between literal and non-literal readings of idiomatic formulaic sequences (e.g. *skating on thin ice*). They find idiomatic readings are articulated faster. Literal readings were characterized by more, and longer, pauses, more changes in pitch, slower articulation of key lexical items and an overall less precise pronunciation (Van Lancker *et al.* 1981: 331–4).

Several recent studies have used experimentation to investigate the processing of formulaic sequences. Also focusing on idioms, Underwood *et al.* (2004) use an eye-tracking paradigm to measure the number and duration of fixations on the final word of twenty, mainly idiomatic, formulaic sequences (e.g. *hit the nail on the head*). It was found that for native speakers the fixations on the terminal words of the formulaic sequences were shorter and fewer than when the same words were read in a non-formulaic context. They interpret this as evidence that for native speakers the idioms were holistically represented and processed (Underwood *et al.* 2004: 167) (for non-native speakers the results were inconclusive). In a related study, Conklin and Schmitt (2008) use similar texts in a self-paced reading experiment. This time both native and non-native participants read the formulaic sequences more quickly, which they conclude 'supports the assertion that formulaic sequences are involved in more efficient language processing' (Conklin and Schmitt 2008: 86). Idioms may, however,

represent a special case of formulaicity—by definition, the meaning cannot be determined on the basis of the individual components alone. It is also worth noting that, according to corpus data, 'core idioms' are, by and large, a relatively infrequent phenomenon (Grant 2005).

Schmitt *et al*. (2004) use oral dictation to investigate whether frequency is a good indicator of holistic representation by speakers of English. A total of 25 corpus-derived clusters were selected to reflect a number of attributes (including length, frequency and transparency of meaning or function), some of which the researchers thought 'seemed likely to be stored holistically by proficient speakers (*as a matter of fact*) and some which were quite questionable in this regard (*in the number of*)' (Schmitt *et al.* 2004: 130). Their results suggest that for native speakers, although some of the clusters appear to be stored holistically, a large number may not be; and for non-native speaker participants, clusters did not appear to be holistically stored. Jiang and Nekrasova (2007) use online grammaticality judgement experiments to investigate reaction times of native and non-native speakers to corpus-derived formulaic and non-formulaic sequences. Both groups reacted significantly faster, and with fewer errors, to the formulaic sequences. The researchers take this as evidence in support of the claims of holistic representation and processing.

While Schmitt *et al.* (2004: 147) advise caution in assigning too much psycholinguistic importance to frequency data alone, on the basis of recent experiments Ellis and Simpson-Vlach (2009: 73) argue that 'formulaic sequences, *statistically defined and extracted* from large corpora of usage, have clear educational and psycholinguistic validity' (emphasis added). Using a range of language processing tasks, they investigate how several corpus-derived measures (length, frequency and mutual information) impact on processing. Their results show that although all these metrics affect processing, for native speakers it is predominantly the information (MI) of formulae, which predicts processability (see also Siyanova and Schmitt 2008). MI (Church and Hanks 1990) is a statistical formula from the field of information theory which is commonly used in corpus linguistic study of collocation (McEnery *et al.* 2006: 56–57). Based on distributional information, it provides a measure of the strength of the association between two words in a corpus.

Usage-based models of language acquisition offer an explanation for the relationship between distributional information and processability. Bybee and Scheibman (1999: 582), for example, argue that frequent exposure to multiword sequences can transform them into fused storage and processing units. Hoey's (2005) theory of *Lexical Priming*, which specifically addresses the phenomenon of collocation, argues that a consequence of the frequent co-occurrence of words, for example the pattern *AB*, is that when language users encounter *A*, they are *primed* to anticipate *B*. As with repetition priming (Scarborough *et al.* 1977), and semantic priming (Meyer 1971), prior activation is likely to be advantageous for processing. He argues that each word which we acquire is 'cumulatively loaded with the contexts and co-texts in which it is encountered' (Hoey 2005: 8).

Psycholinguistic research by McDonald and Shillcock (2003a; 2003b) provides empirical evidence in support of Hoey's theory of Lexical Priming. In their experiments verb–noun bigrams are differentiated on the basis of *transitional probabilities*: the simple probability that word *B* will follow word *A* in a bigram *AB* (i.e. $F_{AB}/F_A$). Using eye tracking, they demonstrate that in stimuli such as *The dictionary editor tried to resolve disputes/meanings in a relatively impartial way*, transitional probabilities calculated on the basis of frequency data in the BNC (*resolve disputes* = high probability; *resolve meanings* = low probability) have a significant effect on the duration of fixations on the target noun, longer fixations being shown for low probability transitions. They suggest that a probabilistic model of language processing can account for their results: 'the brain is able to draw upon statistical information in order to rapidly estimate the lexical probabilities of upcoming words' (2003a: 648) (however, see also Frisson *et al.* 2005). Connectionist theories (e.g. Rumelhart *et al.* 1986) provide a model for how probability of this sort might be realized at the lower levels of 'implementation' (Jurafsky 2003). Such theories propose that knowledge in the brain is represented by numerous inter-connected units and that it is through experience (i.e. repeated exposure) that these connections are strengthened. Probability may thus be encoded as weights on connections or resting activations.

In sum, for native speakers at least, there is evidence in support of claims that formulaicity is associated with processing advantages. That the picture for second-language learners may be less clear (e.g. Schmitt *et al.* 2004; Underwood *et al.* 2004) is, given the inherent variability of learner language (Ellis 1994: 22), not surprising. Non-native speakers' total target-language exposure is small compared with that of native speakers. It seems probable that the formulaic sequences contained in their mental lexicon will differ greatly from those of native speakers (Wray 2002). For example, based on the results of cloze tests of native and EFL adolescents and adults, Kuiper *et al.* (2009: 233) suggest that, for the participants in their study, 'EFL speakers' acquistion of vernacular phrasal vocabulary is an order of magnitude lower than that of native speakers'. How *learner* formulaic sequences are processed, however, is a question that has not been investigated and is one with implications for the field of SLA.

In recent years, it has been widely argued that formulaic sequences (variously referred to as *lexical phrases*, *multi-word units*, *formulas,* etc.) should be assigned prominence in second- or foreign-language learning (Nattinger and DeCarrico 1992; Lewis 1993, 2000; Wray 2000; Foster 2001). In a recent review of SLA research aimed at language teachers, Ellis (2005: 210) suggests that language teaching needs to ensure that, as well as developing a rule-based competence, learners develop 'a rich repertoire of formulaic expressions'. If such an approach is justified, it follows that formulaicity should contribute to the learner's communicative competence, which includes knowledge of what constitutes *appropriate* and *effective* language behaviour (Ellis 1994: 7). The present study sets out to investigate whether this is the case.

Might the psycholinguistic claims outlined above provide support for the foregrounding of formulaic sequences in language teaching? Obviously, classroom conditions cannot fully replicate the sociolinguistic and psycholinguistic learning conditions that ultimately lead to reduced processing for native speakers. Nor is it reasonable to expect that learners could acquire a repertoire of formulaic sequences anywhere near comparable in size with that of native speakers—Pawley and Syder (1983: 215) hypothesize that 'complex lexical items' form by far the largest component of a native speaker's lexicon, numbering in the range of several hundred thousand items. However, a better understanding of the relationship between second-language learners' use of formulaic sequences and communicative competence may contribute to decisions about the role of such sequences in language teaching (see also Nesselhauf 2007). We may hypothesize that close alignment of the formulaic sequences in the learner's interlanguage with those in the target-language will lead to gains for both the learner and the interlocutor, which in turn can result in more efficient communication. The two types of gain are as follows:

1 Gains in fluency for learner production and comprehension.
2 Gains in comprehensibility for the addressee.

With regards to the first gain, the automatization of appropriate target-language formulaic sequences allows learners to devote more attention to other tasks—in this sense, they act as 'islands of reliability' (Dechert 1983), allowing learners to plan the next part of their utterance. Furthermore, correctly memorized formulaic sequences are accurate utterances which can be articulated without hesitation, enabling the learner to come across as a more fluent speaker. A study by Boers *et al.* (2006) showed that the number of formulaic sequences used correlates well with oral proficiency ratings—a finding which lends support to this hypothesis. However, a recent study by Siyanova and Schmitt (2008) indicates that even though learners may be able to produce target-like formulaic sequences accurately (adjective–noun collocations in this case), compared with native speakers, advanced learners require longer to process the sequences and have poor intuitions regarding their frequency.

Close alignment of formulaic sequences might lead to a second gain. If a learner uses the appropriate formulaic sequence, the language produced will be more easily comprehensible, because it places fewer demands on processing (Skehan 1998: 38). A corollary of this is that learner deviations from formulaic norms will create *greater* processing demands for the addressee. There is a general consensus that, even for advanced learners, target-like mastery of these norms (collocation in particular) is especially difficult (Wray 2000; Nesselhauf 2005). Learners use fewer types of collocations than native speakers; however, a small number of collocations are overused (Hasselgren 1994); and they tend to overuse non-restricted collocates which might be considered 'safe options' (Hasselgren 1994). While the restriction (i.e. idiomaticity) of combinations appears to influence the difficulty (e.g. Bahns and Eldaw 1993; Lorenz 1999), it seems that L1 transfer plays a stronger role in

learner use of non-target-like L2 collocations (Farghal and Obiedat 1995; Granger 1998; Nesselhauf 2005).

This study set out to investigate how learner errors (specifically collocation errors) impact on native speakers—a topic that in recent years has received little attention. In the 1970s and 1980s error evaluation studies abounded. However, these studies, in which native speakers made offline judgement concerning the gravity of learner errors, have been criticized for being subjective and producing inconclusive results (Ellis 1994: 67; Ellis and Barkhuizen 2005: 67). Even Nesselhauf (2007: 310–1), who recently argues for the importance of evaluating the impact of learner collocation errors on the addressee, is sceptical about the practicality of her own suggestion, stating that 'it is almost impossible to measure disruption objectively'. The experiment reported below combined corpus-based approaches and online psycholinguistic methods to investigate if and how learner deviations from formulaic norms (specifically collocation) impact on processing by the addressee.



## METHOD

The experiment set out to test the hypothesis that *learner failure to use the appropriate formulaic sequence will create greater processing demands for the addressee*. The notion of formulaic sequence was operationalized through the corpus linguistic phenomenon of collocation (namely, statistically defined bigrams—that is collocations within a span of ±1). Collocations are, thus, regarded as instances of formulaic sequences. Collocations were extracted from learner and native speaker corpora. The processing burden on native speakers posed by learner collocation errors (relative to the native speaker equivalent) was then measured using a self-paced reading task. The *self-paced reading task* is based on the premise that eye movement data can reflect moment-to-moment cognitive processes in reading (Rayner 1998: 372). This technique essentially involves holding the gaze still and moving the text into the centre of vision—a task which, compared with eye-tracking, is technically less demanding. Typically, a sentence is presented one word at a time on a computer screen with the reader controlling the pace by pushing a button. Divergences in mean reading times (the time between pushes of the button) between conditions provide quantitative evidence for differences in attention deployed in reading particular words. It was therefore hypothesized that for a collocation pair *AB* (e.g. *best/ideal partner*), irrespective of whether the deviation is located on the first or the second word of the bigram, a difference in reading times between conditions would be observed on the second word *B* (i.e. *partner*).

### Corpus data

The source of learner data was a corpus of English essays (c. 180,000 tokens) written by Japanese University students ($N = 960$) studying English as a foreign language (author). First, all bigrams (adjacent word pairs) were extracted

(c. 43,000 bigrams). These were then filtered to leave only pairs that met the statistical criteria for collocation. The criteria for collocation were as follows: (i) frequency of the bigram is $\geq 2$; (ii) the bigram is produced by two or more learners; and (iii) the bigram is statistically significant ($P < 0.05$, Fisher's exact test).[1] Of the remaining learner collocations, only those which did *not* occur in the British National Corpus (BNC) were retained (c. 150 bigrams). This procedure was based on the assumption that *absence of a collocational pattern from a large native-speaker corpus of English can function as a proxy for non-native-like selection.* The 100 million of the words of the BNC represent, in terms of quantity, a very large portion of a British native speaker's total language exposure (Aston and Burnard 1998: 28). From this list of candidates, intuition-based judgements made by a native speaker were used to remove the native-like bigrams (e.g. *Japanese teachers*, *learns English*). The resulting thinned list consisted of solely non-native-like learner collocations (c. 100).

Each word in the learner collocations was submitted to collocation analysis in the BNC. The aim of this procedure was to identify a semantically equivalent native speaker bigram that differed from the learner bigram in terms of only a single word (*marriage life* vs. *married life*, where the former is the learner bigram), but not both (*freed times* vs. *free time*). Where no native speaker equivalent was readily identifiable, the learner collocation was discarded. Native speaker collocation equivalents were required to be statistically significant ($P < 0.05$, log likelihood) and have a high associative strength (MI $\geq 3$).[1] Hunston (2002: 71) suggests that an arbitrary mutual information score of three or greater can be taken as evidence of strong collocation. The resulting 70 collocation pairs were then categorized according to how the surface structure of the learner collocation deviated from the corresponding native speaker collocation. The collocations were categorized as either (i) *lexical misselection* or (ii) *misformation*. Lexical misselection refers to the choice of an inappropriate lexeme—for example, *best (ideal) partner*, *cheap (low) cost*, *volunteer people (workers)*, where the word in parentheses constitutes the target-like native speaker equivalent. Misformation refers to the choice of an inappropriate form of the lemma—for example, *marriage (married) life, culture (cultural) background, child abusing (abuse)*.

## Materials

Thirty-two collocation pairs were selected as experimental items (native speaker collocations: mean MI = 6.54, SD = 2.55). A plausible context in which the native speaker bigram could occur was identified through examination of naturally occurring corpus examples and, based on these, stimuli were written (see supplementary material). Thus, items were embedded in contrived single sentences, so that each sentence had two conditions—a learner collocation error and a native speaker target-like equivalent. True/false statements on which participants were questioned after reading were also written to ensure

participants focus on meaning. Aaronson and Scarborough (1976) show that without this, participants fall into a pattern of pressing the button at a steady rate with little variability. In order to avoid this, in self-paced reading participants are often required to recall part of the text in order to answer a simple comprehension question. An example is provided below—experimental item highlighted in bold.

> *Condition 1 (learner collocation): Thanks to the Internet, Kevin was able to find his* ***best partner*** *through online dating.*
>
> *Condition 2 (native speaker collocation): Thanks to the Internet, Kevin was able to find his* ***ideal partner*** *through online dating.*

Statement
*Online dating was successful for Kevin* (*true*)

Each sentence thus differed only in terms of a single word—the deviation and its 'corrected' version. To ensure comparability, and to enable measurement of reading times relative to surrounding context, all stimuli were approximately the same length (mean=15.03 words; SD = 1.31), and all items were embedded in the latter half of the sentence, at least three words from the final word of the sentence. Low-frequency vocabulary was minimized, so that word recognition would be unlikely to pose problems to a native speaker. However, due to the relatively small pool of items, it was not possible to control for the part of speech and syntactic role of the item. Sixteen non-experimental filler items were also written, the purpose of which was to distract participant attention from the focus of the experiment. The filler sentences met the similar design criteria outlined above, and in place of experimental items were native speaker bigrams which were statistically significant (BNC, $P < 0.05$) with a high association strength (MI >3).

To ensure the linguistic plausibility of the contrived contexts and assess the reliability of introspective judgements made by the researcher, 20 native speaker informants (who did not take part in the experiment) rated each sentence as one of the following: (i) *perfectly acceptable*; (ii) *a little odd*; (iii) *strange*; (iv) *very strange*; or (v) *unacceptable*—where *perfectly acceptable* was not chosen, informants identified the feature of language that they considered unacceptable. As with the self-paced reading experiments, informants saw only one condition for each item and conditions were counterbalanced. *Condition 1* items (learner collocations) were judged, to varying degrees, as unacceptable (mean acceptability=2.97, SD = 1.15); *Condition 2* items and the contexts as acceptable (mean acceptability = 1.04, SD = 0.32), and the differences between conditions were shown to be statistically significant for all but one item. Where *Condition 1* stimuli were judged as unacceptable (137 out of a total of possible 160), in all cases it was the experimental items, as opposed to the context, which were cited as the unacceptable feature of the language. In no cases did the majority of informants rank a *Condition 2* stimulus as unacceptable, and where

*Condition 2* stimuli were identified as unacceptable, there was little or no consensus as to which feature of the language was unacceptable.

## Design

A repeated-measures design where each participant reads both conditions would have allowed for direct comparison of reading times enabling control for individual differences. Individual differences in reading speed are considerable—Jackson and McClelland (1975), for example, show mean reading speeds to vary from 260 words per minute for average readers to 586 words per minute for fast readers. Such a design would, however, be likely to lead to repetition priming—'a change in the processing of a stimulus... due to prior exposure to the same or a related stimulus' (Gabrieli 1998: 100). A between-groups design with counter-balanced conditions was, therefore, used. Two lists were created, each comprising 32 experimental items, with the conditions counter-balanced so that each list contained only one condition for each of the experimental items. In addition, each list contained 16 filler sentences, the purpose of which was to ensure that Condition 1 items were not too obvious. Participants were randomly assigned to one of the two groups. Thus, each participant read all sentences but saw only one condition for each experimental item. The numbers of participants in each group were balanced. In all experiments, the order of stimuli was pseudo-randomized across participants to minimize any ordering effects, and no more than two Condition 1 items (learner collocations) were presented back-to-back.

## Participants

Thirty native speakers of British English at Lancaster University and the University of Bath participated in the experiment. Participants were recruited by the researcher from undergraduate students (24) and staff (6) in accordance with the ethics procedures at Lancaster University. Participants were given a café voucher in return for participation.

## Equipment and procedure

The experiment involved word-by-word self-paced reading with each word centred on the computer display. As Rayner (1998: 391) notes: 'one virtue when only one word is presented is that a processing-time measure is associated with each word in the text (i.e. words cannot be skipped)'. The experiment was scripted using *PsyScript*[2] and administered on a 2.4 GHz Macintosh computer. Participants controlled the pace of presentation by clicking the button on an external mouse. With each click the word on the screen was replaced with the next word in the sentence. The times between mouse clicks were recorded, as were the participants' true/false judgements. Participants were asked to read at a pace that would facilitate answering of a comprehension question. In order to familiarize participants with the task, six practice

items which did not contain non-native-like language were provided prior to starting the experiment.

## RESULTS

The experiment lasted on average between 10 and 15 min per participant. The mean reading time per word was 494 ms (SD = 167 ms), excluding sentence initial and sentence final words, where the participants showed a tendency for prolonged pauses. Comprehension accuracy for Condition 1 and Condition 2 stimuli was 94.4 and 93.5 per cent, respectively—the difference between these means is not significant. For the two conditions, mean reading times per word across a region ±2 words relative to all experimental items (AB) are plotted in Figure 1. In the sample sentence, Condition 1 items are exemplified by the bigram *best partner* and Condition 2 by *ideal partner*.

This graphical representation of the data appears to confirm the initial hypothesis that for learner collocations the reading times of the target words would be slower. A substantial divergence in reading time between conditions (77 ms) is observed on the target word (*B*) (631 ms, SD = 391 in Condition 1; and 554 ms, SD = 326 in Condition 2). However, what is particularly notable is that these differences appear to extend to the right of the target word. Within the two-word region (*B*+1 and *B*+2) to the right of the target word, the difference in reading times between conditions is sustained (88 ms on *B*+1; 71 ms on *B*+2). In light of this, the analyses were extended (*post hoc*) to cover this region.

To correct for skewed distribution values, reading times were converted to natural logarithms and submitted to a multi-level mixed model analysis of variance (ANOVA) using the SPSS module 'linear mixed model'. The purpose

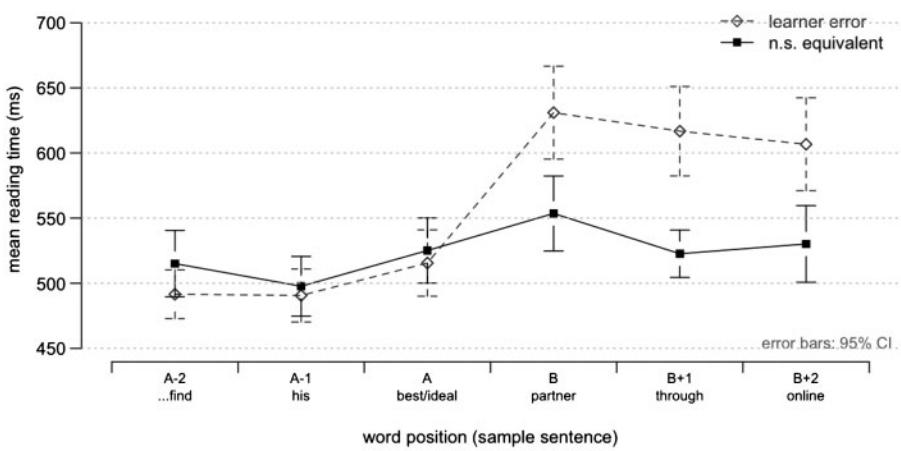


*Figure 1: Mean reading times across region ±2 words relative to all experimental items*

of the analysis was to take account of any random effects as a source of variability, and to determine which factors distinguish the sentences. The following factors were submitted to the model as main effects and interactions: *condition* (learner bigram vs. native speaker bigram) and *position* (whether the deviation occurred on the first or the second word of the bigram). Participants and items were submitted to the model as random effects. The results of the ANOVA showed a significant condition effect on the target word, $F_B(1,928) = 8.93$, $P = 0.003$ ($\beta = 0.00598$–standardized regression coefficient); one word to the right of the target word, $F_{B+1}(1,928)=22.52$, $P < 0.001$ ($\beta = 0.00769$); and two words to the right of the target word, $F_{B+2}(1,928)=17.66$, $P < 0.001$ ($\beta = 0.00796$). No other main effects or interactions were observed.

In the above analysis, for 23 of the 32 items the deviation is located on the first word of the bigram (e.g. *'best/ideal' partner*). For the remaining nine items, the deviation was located on the second word of the bigram (e.g. *volunteer 'people/workers'*). It is important to note that for this second set of items comparison of reading times of the target words is not based on direct equivalence (i.e. comparison of *people* and *workers*). It is well-established that words take longer to identify depending on factors such as word length, word frequency, and morphological complexity (Rayner and Pollatsek 1994). In light of this, the analysis was repeated with only the 23 items where target words were directly comparable (i.e. those items with the deviation on the first word of the bigram). The results of the ANOVA showed significant condition effect on the target word, $F_B(1,688)=11.42$, $P < 0.001$ ($\beta = 0.00857$); one word to the right of the target word, $F_{B+1}(1,688)=21.58$, $P < 0.001$ ($\beta = 0.01023$); and two words to the right of the target word, $F_{B+2}(1,688)=14.58$, $P < 0.001$ ($\beta = 0.01053$). This finding lends support to the initial analysis.

Finally, a *post hoc* analysis was carried out to explore possible difference in the relative effects of the two different types of learner collocations on which the experimental items were based: *lexical misselection* and *misformation*. Target words in native speaker bigrams based on lexical misselection were read an average of 83 ms faster than target words in the equivalent learner bigrams (629 ms, SD = 393 in Condition 1; and 546 ms, SD = 304 in Condition 2). Target words in native speaker bigrams based on misformation were also read faster, but only by an average margin of 40 ms (622 ms, SD = 369 in Condition 1; and 582 ms, SD = 405 in Condition 2). Using the same model, natural logarithms of mean reading times for the target words were submitted to two separate ANOVAs. A significant condition effect was observed for items based on lexical misselection, $F(1,748)=11.45$, $P < 0.001$ ($\beta = 0.00782$), however items based on misformation did not reach significance, $F(1,208)=0.287$, $P = 0.593$. In interpreting these results, two important points must be noted. First, as there were only seven items based on misformation, the experiment was heavily biased towards items based on lexical misselection (25 out of 32)—the lack of a significant condition effect for misformation items is, therefore, arguably due to a type II error (i.e. a result of there not being enough data). Furthermore, for

half of the misformation items the deviation was located on the target word, and, as a result, the comparison is not based upon direct equivalence. These analyses, therefore, do not provide a solid basis for drawing inferences about the relative effects of collocation error type.

## DISCUSSION

That target words in learner collocations are read significantly slower confirms the initial hypothesis, and is persuasive evidence that learner deviation from target language formulaicity places an increased processing burden on native speaker addressees. Importantly, words *to the right* of the target word were also read significantly faster when they followed a native speaker bigram. This would indicate that condition effect observed on the target word is sustained even *after* the reader has moved away from the target word. The term *spillover effect* is used to describe the sustained processing of a target word after eye fixations have progressed to the right (see, e.g. Balota *et al.* 1985). It may be possible that the strong condition effect observed on words to the right of the target indicates a spilling over of increased processing burden of the learner bigram.

Although it has been widely argued that approaches to second-language instruction should ensure that learners develop a rich repertoire of formulaic sequences (Ellis 2005: 210–11), there has been little empirical evidence to show that the formulaicity of learner language directly contributes to communicative competence. On the basis of this single experiment it is difficult to quantify the degree to which collocational deviations might actually impede communication. The conditions in this experiment had no measurable impact on comprehension accuracy—however, the instrument was not designed with that purpose in mind. Whether or not collocational deviations could lead to a breakdown in communication would conceivably depend on a complicated interplay of factors—for example, the context and gravity of the deviation and even the affective state of the addressee. Indeed, native speaker interaction with non-native speakers is likely to be marked by a multiplicity of accommodations—both linguistic and non-linguistic. Although definitive statements on how collocation errors impact on communication would be premature, on the basis of these results it is argued that they have the *potential* to hinder communication. This study therefore provides support for the widely asserted claim that formulaic sequences offer processing advantages, and empirical support for the importance of collocation (and formulaic sequences in general) in language learning.

Hoey's (2005) theory of Lexical Priming (see above) provides a basis for speculation about the processes that underlie the increased cognitive load associated with learner collocations. Participant perception of the first word of the bigram activation may lead to a spreading of activation across words which are, based on the constraints of the sentence up until that point and the individual's frequency of language experience, likely to follow in the input.

As learner collocations (*AB*) were absent from the BNC, the likelihood that on exposure to *A* activation would be spread to word *B* is low. For example, when participants encountered the word *tight*, the word *relationship* is not likely to be activated for use; hence the learner bigram *tight relationship* is likely to require longer to process. In contrast, for the native speaker collocations, distributional information (in particular, high MI scores) points to word *A* (*close*) predicting word *B* (*relationship*). This link may be reflected in the ready processability of these collocations (cf. Ellis and Simpson-Vlach 2009).

Observations made during the experiment provide some evidence (albeit anecdotal) pointing towards anticipatory processes similar to those proposed by Hoey (2005) and McDonald and Shillcock (2003a, 2003b). At times some participants read the text, or parts of the text, *aloud* (this was neither requested nor prohibited). On three occasions, when a learner collocation was encountered, the word verbalized was not the word on the screen, but rather the target-like word, which the participant appeared to have anticipated. Thus, the Condition 1 items *free times*, *child abusing,* and *daily behaviours* were verbalized as *free time*, *child abuse,* and *daily routine*. In fact prior to developments in eye-tracking technology, oral reading served as an important source of data for theories of reading. Goodman (1967), for example, on the basis of the observation that mistakes children make in oral reading are often semantically consistent with the actual word in the text, argues that fluent reading is a 'psycholinguistic guessing game' based on a series of preemptive guesses about what word will come next.

The results also generate further interesting questions. First, how long does the processing burden last for? Figure 1 suggests that by two words to the right (*B*+2) the difference between reading times for the two conditions has started to decrease. Extension of the analysis further to the right (i.e. to *B*+3,4. . .etc.) to identify if and when reading times for the two conditions re-converge would provide valuable insights into processing of errors. In the present experiment this was not possible as the stimuli used extended only three words or more to the right of the target word (sentence terminal words were excluded from the analyses). A second question of interest concerns the relative effects of the *type* of collocation error (misformation or misselection)—it has already been noted that this experiment provides inconclusive results on this point. However, it is hypothesized that the processing of morphological errors (misformation) and lexical errors (misselection) may differ with regards to the strength and duration of the effect. Above it was suggested that comprehension may involve anticipatory processes (a spreading of activation) based on experienced frequency (see also Pickering and Garrod 2007 for a similar recent proposal). Predictions made regarding the grammatical class of an upcoming word will be associated with a markedly greater certainty than predications on a lexical level (there are far more items in the lexicon than grammatical categories). As misformation items most often differ from the native speaker equivalent on the part of speech level (e.g. *responsibility person* vs. *responsible person*), they will often result in violation of a strong predication. For example, a forward prediction

about the upcoming part of speech made on the basis of *responsibility* would be heavily weighted towards the category of preposition. Based on BNC data, the probability of *responsibility* being followed by a preposition approaches 1. In contrast, predications made on a lexical level are spread over multiple types and are associated with low weightings (*best* precedes over 15,000 noun types in the BNC). It is therefore hypothesized that misformation items have the greater potential to disrupt sentence processing.[3]

A second hypothesis is that learner *overuse* of certain collocates may lead to an increased processing burden even though the collocation may not constitute an *error* as such. It is well documented that second learners' have a tendency to overuse non-restricted collocates—Hasselgren (1994) describes such choices as 'lexical teddy bears'. Non-restricted lexical items featured heavily in the learner corpus data used in the experiments (e.g. *do*, *make*, *big*, *best*, *bad*). Learner over-use of such 'teddy bears', in contrast to the types of deviant collocations used here as experimental items, may not represent an error as such. However, non-restricted choices often lack the predicative power of a more precise collocate (e.g. *nice memories* vs. *fond memories*). Although the impact of 'lexical teddy bears', on the level of the individual item, is likely to be small, their *cumulative* effect may in fact be measurable. Both of these hypotheses are purely speculative.

In considering directions for future research, it is important to note that the instrument used here is a *substitute* for natural reading. In natural reading about 10–15 per cent of all eye movements are in fact backwards (Rayner and Pollatsek 1994: 114). These *regressive saccades*, or *regressions*, are often made when readers have difficulty processing the text (Rayner 1998: 375). However, self-paced reading constrains movement to a left-to-right forward motion, in which participants are not able to make regressions. As the spill-over region appears to represent processing *difficulties* (i.e. a disorientation of some sort), we might predict that in natural reading this region would be characterized by regressions. To explore the hypotheses outlined above, future research might use an eye tracking methodology to generate information regarding the point at which the regressions occur, the location to which the reader regresses, and the duration of any subsequent fixations. Such data would be useful in gaining a more robust understanding of the processing of collocation errors. Another area for investigation is the exploration of the extent to which the processing demands of the sort described in this study impact on qualitative judgements regarding a candidate's written production—findings from such a study could have implications for language testing.

## CONCLUSION

First and foremost, this study clearly shows that learner collocation errors place an increased and sustained cognitive burden on the addressee. How different types of collocation error impact on processing and the duration of the

processing burden (the hypotheses discussed in the previous section) will be the subject of future studies. It has been suggested that the findings appear to be consistent with usage-based theories of language acquisition and processing (Bybee and Scheibman 1999; Hoey 2005; Bybee 2007) and support Hoey's (2005: 7) claim that collocation is 'fundamentally a psychological concept'. Drawing on usage-based models, hypotheses for a speculative mechanism that readers work through when they encounter an error were also proposed. In sum, the findings add to the growing body of evidence showing that formulaic sequences are advantageous for processing. It has been widely argued that formulaic sequences should play a central role in language teaching. This study provides empirical psycholinguistic evidence that collocation (and by extension formulaic sequences in general) can contribute to a learner's communicative competence.

Online psycholinguistic techniques (e.g. reaction times) and corpus-based methodologies are used more and more being by SLA researchers (Juffs 2001; Granger 2004). However, they are rarely used to complement each other. In corpus linguistics the combination of corpus data with experimentation is increasingly being used to provide a more empirically robust approach (Gilquin and Gries 2009). This study has combined these approaches to investigate a topic that in recent years has received little attention—the impact of learner errors. Criticized for being subjective and producing inconclusive results, error evaluation studies have now been largely discontinued. Through a combination of corpus-data and experimentation, the present study has endeavoured to address concerns about subjectivity, and demonstrates that the 'disruption' caused by features of learner language can be objectively gauged. It is suggested that similar investigation of how features of learner language (not only collocation) impact on processing has potential to further our understanding of what it means to be competent in a second language.

## SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

## NOTES

1 All of the association measures used here are based on a 2 × 2 co-occurrence table of each bigram type and involve comparison of observed frequencies with the expected frequencies (if the expected co-occurrence were due to chance)—for details see Oakes (1998).

2 The 'PsyScript' application, published by the Psychology Department, Lancaster University, is available at http://www.psych.lancs.ac.uk/software/.

3 Differences in the processing of morphological and lexical deviations may also be hypothesized on the basis of classical models of sentence processing (e.g. so-called 'garden path' models, Frazier 1987), which assume

representational modularity (Fodor 1983). As syntactic information is thought to be locally integrated and resolved, while integration of semantic information continues for longer, one might therefore expect different patterns for the time course of parsing for misformation and misselection items.

## REFERENCES

**Aaronson, D.** and **H. S. Scarborough.** 1976. 'Performance theories for sentence coding: Some quantitative evidence,' *Journal of Experimental Psychology: Human Perception and Performance* 2/1: 56–70.

**Altenberg, B.** 1993. *Recurrent Verb Complement Constructions in the London-Lund Corpus*. Rodopi.

**Aston, G.** and **L. Burnard.** 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

**Baddeley, A.** 2000. 'The episodic buffer: A new component of working memory?,' *Trends in Cognitive Sciences* 4/11: 417–23

**Bahns, J.** and **M. Eldaw.** 1993. 'Should we teach EFL students collocations?,' *System* 21/1: 101

**Balota, D. A., A. Pollatsek,** and **K. Rayner.** 1985. 'The interaction of contextual constraints and parafoveal visual information in reading,' *Cognitive Psychology* 17/3: 364–90.

**Biber, D., S. Conrad,** and **V. Cortes.** 2004. 'If you look at . . . Lexical bundles in university teaching and textbooks,' *Applied Linguistics* 25/3: 371–405.

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

**Boers, F., J. Eyckmans, J. Kappel, H. Stengers,** and **M. Demecheleer.** 2006. 'Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test,' *Language Teaching Research* 10/3: 245–61.

**Brumfit, C.** 2000. 'Accuracy and fluency: The basic polarity' in H. Riggenbach (ed.): *Perspectives on Fluency*. University of Michigan Press, pp. 61–73.

**Bybee, J.** 2007. *Frequency of Use and the Organization of Language*. Oxford University Press.

**Bybee, J.** and **J. Scheibman.** 1999. 'The effect of usage on degrees of constituency: The reduction of don't in English,' *Linguistics* 37/4: 575–96.

**Church, K. W.** and **P. Hanks.** 1990. 'Word association norms, mutual information, and lexicography,' *Computational Linguistics* 16/1: 22–9.

**Conklin, K.** and **N. Schmitt.** 2008. 'Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers?,' *Applied Linguistics* 29/1: 72–89.

**Dechert, H.** 1983. 'How a story is done in a second language' in C. Færch and G. Kasper (eds): *Strategies in Interlanguage Communication*. Longman.

**Ellis, N.** 1996. 'Sequencing in SLA: Phonological memory, chunking, and points of order,' *Studies in Second Language Acquisition* 18/1: 91–126.

**Ellis, N.** 2002a. 'Frequency effects in language processing,' *Studies in Second Language Acquisition* 24/2: 143–88.

**Ellis, N.** 2002b. 'Reflections on frequency effects in language processing,' *Studies in Second Language Acquisition* 24/2: 297–339.

**Ellis, N.** and **R. Simpson-Vlach.** 2009. 'Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education,' *Corpus Linguistics and Linguistic Theory* 5/1: 61–78.

**Ellis, R.** 1994. *The Study of Second Language Acquisition*. Oxford University Press.

**Ellis, R.** 2005. 'Principles of instructed language learning,' *System* 33/2: 209–24.

**Ellis, R.** and **G. Barkhuizen.** 2005. *Analysing Learner Language*. Oxford University Press.

**Farghal, M.** and **H. Obiedat.** 1995. 'Collocations: A neglected variable in EFL,' *IRAL* 33/4: 315–31.

**Fodor, J. A.** 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

**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*. Longman, pp. 75–97.

**Frazier, L.** 1987. 'Sentence processing: A tutorial review' in M. Coltheart (ed.): *Attention and Performance, 12: The Psychology of Reading*. Erlbaum, pp. 559–86.

**Frisson, S., K. Rayner,** and **M. J. Pickering.** 2005. 'Effects of contextual predictability and transitional probability on eye movements during reading,' *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31/5: 862–77.

**Gabrieli, J. D. E.** 1998. 'Cognitive neuroscience of human memory,' *Annual Review of Psychology* 49/1: 87–115.

**Gilquin, G.** and **S. T. Gries.** 2009. 'Corpora and experimental methods: A state-of-the-art review,' *Corpus Linguistics and Linguistic Theory* 5/1: 1–26.

**Goodman, K. S.** 1967. 'Reading: A psycholinguistic guessing game,' *Journal of the Reading Specialist* 6: 126–35.

**Granger, S.** 1998. 'Prefabricated patterns in advanced EFL writing: Collocations and formulae' in A. P. Cowie (ed.): *Phraseology: Theory, Analysis, and Applications*. Clarendon Press, pp. 145–60.

**Granger, S.** 2004. 'Discourse in the professions: Perspectives from corpus linguistics' in U. Connor and T. A. Upton (eds): *Language and Computers: Studies in Practical Linguistics*. Rodopi, pp. 123–45.

**Grant, L. E.** 2005. 'Frequency of core idioms in the British National Corpus (BNC),' *International Journal of Corpus Linguistics* 10/4: 429–51.

**Hasselgren, A.** 1994. 'Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary,' *International Journal of Applied Linguistics* 4/2: 237–58.

**Hoey, M.** 1991. *Patterns of Lexis in Text*. Oxford University Press.

**Hoey, M.** 2005. *Lexical priming: A New Theory of Words and Language*. Routledge.

**Hunston, S.** 2002. *Corpora in Applied Linguistics*. Cambridge University Press.

**Hyland, K.** 2008. 'As can be seen: Lexical bundles and disciplinary variation,' *English for Specific Purposes* 27/1: 4–21.

**Jackson, M. D.** and **J. L. McClelland.** 1975. 'Sensory and cognitive determinants of reading speed,' *Journal of Verbal Learning & Verbal Behavior* 14/6: 565–74.

**Jiang, N.** and **T. M. Nekrasova.** 2007. 'The processing of formulaic sequences by second language speakers,' *Modern Language Journal* 91/3: 433–45.

**Juffs, A.** 2001. 'Psycholinguistically oriented second language research,' *Annual Review of Applied Linguistics* 21: 207–20.

**Jurafsky, D.** 2003. 'Probabilistic modeling in psycholinguistics' in R. Bod, J. Hay, and S. Jannedy (eds): *Probabilistic Linguistics*. MIT Press, pp. 39–95.

**Kjellmer, G.** 1984. 'Some thoughts on collocational distinctiveness' in B. Aarts and W. Meijs (eds): *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research*. Rodopi.

**Kjellmer, G.** 1994. *A Dictionary of English Collocations: Based on the Brown Corpus*. Clarendon Press, pp. 163–71.

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

**Kuiper, K.** and **D. Haggo.** 1984. 'Livestock auctions, oral poetry, and ordinary language,' *Language in Society* 13/2: 205–34.

**Kuiper, K., G. Columbus,** and **N. Schmitt.** 2009. 'The acquisition of phrasal vocabulary' in S. H. Foster-Cohen (ed.): *Language Acquisition*. Palgrave Macmillan, pp. 216–40.

**Lewis, M.** 1993. *The Lexical Approach: The State of ELT and a Way Forward*. Language Teaching Publications.

**Lewis, M.** 2000. *Teaching Collocation: Further Developments in the Lexical Approach*. Language Teaching Publications.

**Lorenz, G. R.** 1999. *Adjective Intensification-Learners versus Native Speakers: A Corpus Study of Argumentative Writing*. Rodopi.

**McDonald, S. A.** and **R. C. Shillcock.** 2003a. 'Eye movements reveal the on-line computation of lexical probabilities during reading,' *Psychological Science* 14/6: 648–52.

**McDonald, S. A.** and **R. C. Shillcock.** 2003b. 'Low-level predictive inference in reading: the influence of transitional probabilities on eye movements,' *Vision Research* 43/16: 1735–51.

**McEnery, T., Z. Xiao,** and **Y. Tono.** 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. Routledge.

**Meyer, D. E.** 1971. 'Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations,' *Journal of Experimental Psychology* 90/2: 227–34.

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

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

**Nesselhauf, N.** 2007. 'The path from learner corpus analysis to language pedagogy: some neglected issues,' *Language and Computers* 62: 305–15.

**Oakes, M. P.** 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

**Pawley, A.** and **F. H. Syder.** 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. C. Richards and R. W. Schmidt (eds): *Language and Communication*. Longman, pp. 191–226.

**Pickering, M. J.** and **S. Garrod.** 2007. 'Do people use language production to make predictions during comprehension?,' *Trends in Cognitive Sciences* 11/3: 105–10

**Rayner, K.** 1998. 'Eye movements in reading and information processing: 20 years of research,' *Psychological Bulletin* 124/3: 372–422.

**Rayner, K.** and **A. Pollatsek.** 1994. *The Psychology of Reading*. Erlbaum.

**Rumelhart, D. E.** and **J. L. McClelland.** 1986. PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.

**Scarborough, D. L., C. Cortese,** and **H. S. Scarborough.** 1977. 'Frequency and repetition effects in lexical memory,' *Journal of Experimental Psychology: Human Perception and Performance* 3: 1–17.

**Schmitt, N.** and **R. Carter.** 2004. 'Formulaic sequences in action – an introduction' in N. Schmitt (ed.): *Formulaic Sequences Acquisition, Processing, and Use*. John Benjamins, pp. 1–22.

**Schmitt, N., S. Grandage,** and **S. Adolphs.** 2004. 'Are corpus-derived recurrent clusters psycholinguistically valid?' in N. Schmitt (ed.): *Formulaic Sequences Acquisition, Processing, and Use*. John Benjamins, pp. 127–49.

**Segalowitz, N.** 2000. 'Automaticity and attentional skill in fluent performance' in H. Riggenbach (ed.): *Perspectives on Fluency*. University of Michigan Press, pp. 200–19.

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

**Siyanova, A.** and **N. Schmitt.** 2008. 'L2 learner production and processing of collocation: A multi-study perspective,' *Canadian Modern Language Review/La Revue canadienne des langues vivantes* 64/3: 429–58.

**Skehan, P.** 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

**Stubbs, M.** 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Blackwell Publishers.

**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 Acquisition, Processing, and Use*. John Benjamins, pp. 153–72.

**Van Lancker, D., G. J. Canter,** and **D. Terbeek.** 1981. 'Disambiguation of ditropic sentences: Acoustic and phonetic cues,' *Journal of Speech and Hearing* 24: 330–5.

**Wray, A.** 2000. 'Formulaic sequences in second language teaching: Principle and practice,' *Applied Linguistics* 21/4: 463–89.

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

**Wray, A.** and **M. R. Perkins.** 2000. 'The functions of formulaic language: An integrated model,' *Language & Communication* 20/1: 1–28.