Applied Linguistics 2012: 33/3: 299–320 © Oxford University Press 2012
doi:10.1093/applin/ams010 Advance Access published on 15 April 2012

A Phrasal Expressions List

[1,2,]*RON MARTINEZ and [1,]**NORBERT SCHMITT

[1]University of Nottingham and [2]San Francisco State University
*E-mail: ronmartinez@sfsu.edu
**E-mail: Norbert.Schmitt@nottingham.ac.uk

There is little dispute that formulaic sequences form an important part of the lexicon, but to date there has been no principled way to prioritize the inclusion of such items in pedagogic materials, such as ESL/EFL textbooks or tests of vocabulary knowledge. While wordlists have been used for decades, they have only provided information about individual word forms (e.g. the *General Service List* (West 1953) and the *Academic Word List* (Coxhead 2000)). This article addresses this deficiency by presenting the *PHRASal Expressions List* (*PHRASE List*), a list of the 505 most frequent non-transparent multiword expressions in English, intended especially for receptive use. The rationale and development of the list are discussed, as well as its compatibility with British National Corpus single-word frequency lists. It is hoped that the PHRASE List will provide a basis for the systematic integration of multiword lexical items into teaching materials, vocabulary tests, and learning syllabuses.

THE IMPORTANCE OF FORMULAIC LANGUAGE

One of the most important findings from corpus research is that language is made up of not only individual words, but also a great deal of formulaic language. Formulaic language has been defined in a number of ways (e.g. Wray 2002, 2008), but in essence, most definitions indicate that individual formulaic sequences behave much the same as individual words, matching a single meaning or function to a form, although that form consists of multiple orthographic or phonological words. For example, in the sentence *Increasingly, extreme weather events indicate that climate change is upon us,* the concept of 'a situation becoming noticeably prevalent' is realized by the single word *increasingly*, but it could be equally well realized by the formulaic sequence *more and more*. This article deals with a particularly semantically opaque subset of formulaic language (*phrasal expressions*), but it is useful to first discuss formulaic language and its importance in language as a whole.

Whereas formulaic language was once considered a peripheral phenomenon (Ellis *et al.* 2008), research has now established that it is fundamental to the way language is used, processed, and acquired in both the L1 and L2. Evidence for this strong statement is now widely available, and perhaps most fully outlined in books by Sinclair (1991), Nattinger and DeCarrico (1992), Moon (1998), Biber *et al.* (1999), Wray (2002, 2008), Schmitt (2004, 2010), Meunier and Granger (2008) and Corrigan *et al.* (2009a, 2009b). However, some of the key evidence for the essentialness of formulaic language includes the following.

Formulaic language is ubiquitous in language use. Nattinger and DeCarrico (1992: 66) were among the first to assert that formulaic language makes up a large proportion of any discourse, and subsequent research has borne this out. Although various studies have used different methodologies—and differ in what each considers a formulaic sequence—they consistently produce high figures. Erman and Warren (2000), for example, calculated that formulaic sequences of various types constituted 58.6 per cent of the spoken English discourse they analyzed and 52.3 per cent of the written discourse, whereas Biber *et al*. (1999) found that around 30 per cent of the words in their conversation corpus consisted of lexical bundles, and about 21 per cent of their academic prose corpus. Thus, it seems clear that while estimates may vary, formulaic language is a substantial constituent of language overall.

Meanings and functions are often realized by formulaic language. One reason that formulaic language items are so widespread is that they realize a wide number of referential, communicative, and textual functions in discourse. They can be used to express a concept (*take into account [You must also take into account the rush hour]* = the necessity of considering something in one's calculations), transact routinized meanings (*Tell me about it!* = a statement of strong agreement), state a commonly believed truth or advice (*Money talks* = money is persuasive), signpost discourse organization (*on the other hand* signals a contrasting point), and even provide technical phraseology that can transmit information in a precise and efficient manner (e.g. *figure of speech* is a word/phrase used in a non-literal way) (Schmitt and Carter 2004). In fact, it has been suggested that for every recurrent communicative need, there is typically conventionalized language (i.e. formulaic sequences) available to realize this need (Nattinger and DeCarrico 1992: 62–63) in all genres, including scientific and academic discourse (e.g. Biber *et al.* 2004; Dorgeloh and Wanner 2009; Hyland 2008; Wulff *et al.* 2009).

Formulaic language has processing advantages. Pawley and Syder (1983) and Kuiper and Haggo (1984) were some of the first to assert that formulaic sequences offer processing efficiency, and there is now considerable converging support for this. Research into the processing of idioms (e.g. Gibbs *et al.* 1997) provides evidence that L1 readers quickly understand formulaic sequences in context and that they are not more difficult to understand than literal speech. Formulaic sequences are consistently read more quickly than non-formulaic equivalents by L1 readers (and sometimes by L2 readers), as shown by eye-movement studies (Siyanova-Chanturia *et al.* 2011; Underwood *et al.* 2004) and self-paced reading tasks (Conklin and Schmitt 2008). Grammaticality judgments of formulaic items were both faster and more accurate than the judgments of matched non-formulaic control strings (Jiang and Nekrasova 2007). Similarly, Millar (2011) found that when L2 learners' collocations in written production did not match conventionalized formulaic forms, L1 readers required more time to process them. Moreover, there is little doubt that the automatic use of acquired formulaic sequences allows chunking, freeing up memory, and processing resources (Ellis 1996; Kuiper 1996). In short, formulaic language promotes efficient and effective communication.

Formulaic language can improve the overall impression of L2 learners' language production. As noted by Ellis and Sinclair (1996), '[t]he attainment of fluency, in both native and foreign languages, involves the acquisition of memorized sequences of language' (p. 234). Boers *et al.* (2006), for example, showed that L2 speakers were judged as more proficient when they used formulaic sequences. The same applies for written discourse. Ohlrogge (2009) examined 170 written compositions from an EFL proficiency test and concluded that those with higher scores also tended to use more formulaic expressions than the lower scoring group. Likewise, Lewis (2008) found in her analysis of EFL university compositions in Sweden that 'as the use of formulaic language increases, so do the grades' (p. 104).

THE NEED FOR A LIST OF FORMULAIC SEQUENCES

Given the importance of formulaic language, it can be argued that it needs to be part of language syllabuses. Moreover, it would naturally have a prominent place in language teaching textbooks and materials, as well as tests of language achievement and proficiency. Unfortunately, this is generally not the case. A perusal of almost any EFL/ESL textbook or test yields a paucity of formulaic sequences targeted for explicit attention/noticing, and even for those that do occur, there does not seem to be much principled basis for selection (Koprowski 2005; Gouverneur 2008; Hsu 2008).

This is not particularly surprising given that formulaic sequences are often difficult to intuit (Fox 1987). While some formulaic sequences are quite obvious (e.g. idioms like *raining cats and dogs*), others like *take place* (i.e. 'occur') are not. An easy illustration of this is attempting to determine the most frequent formulaic sequences in English by intuition alone. While it is probably possible to think of a number of these sequences, it is unlikely that the list would be very comprehensive, or that the relative frequency of occurrence could be stated with any confidence (cf. Alderson 2007).

The limitations of intuition mean that language teachers, textbook writers, and test developers require a more principled manner of identifying and ranking formulaic sequences. The obvious solution is a list of frequent or useful formulaic sequences to which they can refer. Wordlists have a long history as useful pedagogic tools. Back in the 1930s, vocabulary management made possible by wordlists facilitated the creation of the graded readers in Michael West's *Reading Method*, which helped second-language readers to more easily access texts. The *General Service List* (GSL) (West 1953) has been influential in helping to grade the vocabulary inserted into both first- and second-language teaching materials. The *Academic Word List* (AWL) (Coxhead 2000) has helped to raise awareness of academic support vocabulary, and has also led to a plethora of pedagogic materials to teach and test AWL words. More recently, wordlist-based tools on the Internet have put frequency analysis of texts within the reach of the average practitioner (e.g. *Lextutor,* www.lextutor.ca). However, for all the benefits of wordlists, they possess a key deficiency: they have hitherto been restricted to individual words. Those individual words, in turn, are often only the tips of phraseological icebergs.

This article reports on the construction of a list of the most frequent formulaic sequences in English, a list that necessarily involved both automated and manual selection of items. The following sections discuss how the list was compiled, including a brief discussion of its qualities. The list itself is provided in the Appendix in online supplementary material in full.

CONCEPTUALIZING A LIST OF FORMULAIC LANGUAGE

The starting point for any lexical list is a determination of its purpose(s). We mainly wished to create a list that would have pedagogic utility, mirroring purposes similar to the GSL, and AWL lists, but for formulaic sequences. These purposes include, but are not limited to, the following:

- a guide for language learners and educators to include formulaic sequences in their learning and teaching, particularly for receptive purposes.
- a means of including formulaic sequences in tests that assess receptive L2 knowledge and receptive skills.
- an aid in monitoring vocabulary acquisition progress.

Pedagogic purposes like these dictate that the list needed to focus on the most frequent formulaic sequences in English. It is widely accepted that frequency of occurrence is one of the best indicators of usefulness of individual words in general English (e.g. Leech *et al.* 2001; Nation 2001). For example, the GSL, the model of a pedagogically based wordlist, used a number of selection criteria, but the essential one was frequency. This is true to the extent that it was often used as an indicator of the most frequent 2,000 word families in English before more modern word counts came along. There is no reason to believe that this frequency–usefulness relationship does not also apply to formulaic language (Nation and Waring 1997: 18).

Another frequency-based issue to consider was the extent of the list. While frequency is a valid indicator of usefulness, the list must stop at some point. The GSL contains about 2,000 entries, and this is one possible answer to the question of extent. Based on early research, this may have seemed adequate (Schonell *et al.* 1956). However, more recent research by Nation (2006) indicates that it actually takes 6,000–7,000 word families to comprehend a range of spoken discourse and 8,000–9,000 families for written discourse, based on 98 per cent coverage. Hence a list of formulaic sequences stopping at the same frequency as the 2,000 word family frequency level is obviously too small, but a list extending to the 9,000 level would become too unwieldy for practical use. We decided to compromise at including phrases that matched the frequency of words up to the 5,000 level as it 'represents the upper limit of general high-frequency vocabulary' (Read 2000: 119).[1]

There are two general approaches to identifiying formulaic sequences: one which uses frequency as the main criterion, the other which primarily considers semantics/grammar, or what Nesselhauf (2005) has called the 'frequency-based approach' and 'phraseological approach', respectively. We did not want to be completely driven by frequency in compilation of the list as we could end up including sequences such as *is the* or *is of a*, which encode very little meaning in themselves (cf. De Cock 2000). We felt a pedagogical list should include only formulaic sequences that realize meanings or functions, in order to be of the maximum utility. We therefore decided to highlight the meaning or function aspect as a selection criterion when going through our initial n-gram (i.e. frequency-based) corpus extraction. This meant we would only accept those sequences that conveyed a discrete, identifiable meaning or function. However, with a view to the usefulness of the list, we also considered the transparency of the formulaic sequences' meaning. Consider, for example, the following three expressions:

- *at all*
- *at all costs*
- *at all times*

Although all three expressions tend to occur as phrases according to the British National Corpus (BNC), they differ in *compositionality*, as shown in Figure 1.

Lewis (1993) observed that expressions vary in terms of the degree to which 'the meaning of the whole is not immediately apparent from the meanings of the constituent parts' (p. 98), and called this varying compositionality a 'spectrum of idiomaticity' (ibid.). In Figure 1, it could be argued that, while precise divisions are impossible to pinpoint, the expression *at all times* can be understood relatively easily from the meanings of its three component words. Grant and Bauer (2004) would deem *at all times* as compositional, since its meaning is still retained when each lexical word is replaced with its own definition (p. 52). The phrase *at all costs* has a more figurative, and likely less transparent,

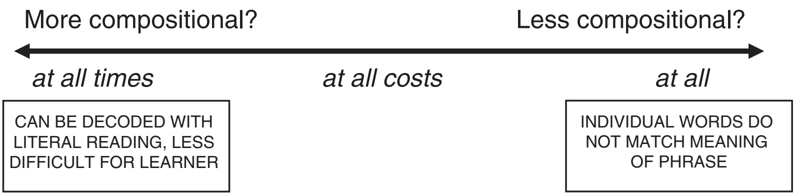

More compositional?
Less compositional?
at all times
at all costs
at all
CAN BE DECODED WITH LITERAL READING, LESS DIFFICULT FOR LEARNER
INDIVIDUAL WORDS DO NOT MATCH MEANING OF PHRASE

Figure 1: Degrees of compositionality for formulaic item selection

meaning—potentially making that item more difficult for a learner unfamiliar with that expression (Cooper 1999; Spöttl and McCarthy 2004). On the far right of the spectrum lies *at all* (e.g. *Do you exercise* ***at all****?*), whose individual components offer no more help to someone meeting that expression for the first time than do the individual letters that spell a word. In essence, *at all* behaves much like a word in terms of form-meaning link. In terms of pedagogic value—especially with respect to so-called 'receptive skills'—it does not seem efficient to include compositional phrases which can be easily analyzed for meaning. Therefore, in our list, we will not include items judged as lying to the far left of the spectrum of idiomaticity (*at all times*), preferring instead ones that learners may find difficulty in interpreting.

We therefore ended up with selection criteria that revolved around high frequency, meaningfulness, and relative non-compositionality. (See the next section for how these were operationalized in detail.) These criteria would select formulaic sequences that would in many ways be comparable with the individual words in a typical frequency-based wordlist. This was done on purpose, as we also wanted our list of selected formulaic sequences to be meaningfully comparable with these wordlists. In this way, it would be sensible to combine our formulaic sequences into these wordlists in a way that would create a much more inclusive overall description of the most frequent (and therefore useful) lexical items of English, both individual- and multi-word. It would also make it possible to insert these formulaic sequences into frequency-based vocabulary tests (e.g. the *Vocabulary Size Test* [VST], Nation and Beglar 2007) in order to gain a more valid measure of overall receptive vocabulary knowledge.

The last issue of conceptualization concerned nomenclature. Terminology in the area of phraseology has always been messy, with Wray (2002: 9) finding over 50 terms to describe the phenomenon of formulaic language. In an attempt to lend some consistency to the field, Schmitt (2010) has suggested *formulaic language* as the umbrella term for the range of phrasal units that occur in language, and *formulaic sequence* as the term for each individual case of this phenomenon. (This article has followed these conventions.) However, we cannot claim to have produced an exhaustive list of all 'formulaic sequences': the formulaic sequences to be identified by our selection criteria will clearly be a limited subset of formulaic language, and need a discrete descriptive name. We therefore decided to opt for what we felt was the most transparent term, and named our particular category of formulaic language *phrasal expressions*. A phrasal expression is hence defined as a fixed or semi-fixed sequence of two or more co-occurring but not necessarily contiguous words with a cohesive meaning or function that is not easily discernible by decoding the individual words alone. Thus, hereafter we will refer to the list as the *PHRASal Expressions List*, or *PHRASE List*.

COMPILING THE PHRASE LIST

The first step in the compilation process was to operationalize the general criteria reached in the conceptualization stage. As discussed in the previous section, it was determined that the multiword items in the PHRASE List should generally require an understanding of the phrase as a whole, remaining consistent with the underlying constructs of wordlists like the GSL and the pedagogical instruments derived from them, such as the VST. Automated identification can reliably identify only some of the potential PHRASE List candidates that would meet this condition (cf. Blackwell 1987; Leech *et al.* 2001; Ellis *et al.* 2008). Although inroads have been made in recent years towards their automated extraction using a combination of semantic and grammatical tagging (e.g. Katz and Giesbrecht 2006; Korkontzelos and Manandhar 2009), no computer application yet designed can replicate key qualitative judgments regarding individual multiword items that most native speakers apparently make unconsciously, intuitively, consistently, and instantaneously (e.g. Deignan 2009; Wulff 2009).

Therefore, in order to arrive at a list of multiword items that went beyond mere probabilistic and indiscriminate word combinations, it was necessary to use a mixed-methods, two-step methodology: an exhaustive computer-assisted search for co-occurring words (n-grams) with respective frequency, statistical, and distributional data, followed by a manual vetting of those items with the guidance of pre-determined selection criteria. Such criteria to guide the identification of formulaic sequences have been used fruitfully in previous studies, most notably Simpson-Vlach and Ellis (2010), Shin and Nation (2008), and Wray and Namba (2003). The criteria used in those studies bore features that both resembled and differed from the criteria in our research.

Simpson-Vlach and Ellis (2010) sought to compile a list of the most useful formulaic sequences used in Academic English. As in our research, Simpson-Vlach and Ellis first extracted formulaic candidates using quantitative corpus search methods (e.g. n-grams, multual information, log likelihood), and then incorporated a winnowing phase for those candidates. According to the authors, the purpose of this more qualitative phase was to ensure that the list was to be pedagogically relevant, and therefore judges (with language testing and teaching experience) were asked to rate a stratified random sample of the formulas on the basis of the following criteria (p. 10):

> A. whether or not they thought the phrase constituted 'a formulaic expression, or fixed phrase, or chunk' [...];
>
> B. whether or not they thought the phrase has 'a cohesive meaning or functions, as a phrase' [...];
>
> C. whether or not they thought the phrase was 'worth teaching, as a bona fide phrase or expression' [...].

Simpson-Vlach and Ellis were then able to correlate the qualitative judgment data with the quantitative statistics and, through multiple regression, arrive at a metric that could be applied to all quantitatively derived formulas and predict which ones would be worth teaching (or 'formula teaching worth'—FTW). Therefore, the items in the Academic Formulas List (AFL) are in theory prioritized by this FTW metric (Table 1), with formulaic sequences most likely to be deemed useful listed first.

Although the methodology involved in the development of the AFL in many ways is similar to our own, one key difference is Simpson-Vlach's and Ellis's rejection of subjective judgments as a determinant of item inclusion. The judges, who only examined a subset of the formulaic sequences, were used to help inform the multiple regression alone—their judgments did not directly influence the selection of items. As the authors point out, such strict adherence to statistically derived phrase selection virtually eliminates possible 'claims of subjectivity' (p. 4); however, as the criteria (A, B, and C above) did not actually guide the selection, many items in the AFL—particularly those with lower FTW ratings, might be seen as only marginally having 'cohesive meaning' as a 'bona fide phrase' (see sample in Table 1). Moreover, it is important to note that the items are not ranked by how commonly they occur in discourse, which is also a departure from most current wordlists, including our own PHRASE List.

In another formulaic-list-related study, Shin and Nation (2008) sought to identify the most frequent collocations in spoken English, and established six criteria involving such aspects as frequency and grammatical well-formedness.

Table 1: Spoken AFL Top 10 (Simpson-Vlach and Ellis 2010)

		Speech		Writing		
		Raw frequency	Frequency per million	Raw frequency	Frequency per million	FTW
1	be able to	551	256	209	99	2.96
2	blah blah blah	62	29	0	0	2.92
3	this is the	732	340	127	60	2.77
4	you know what I mean	137	64	4	2	2.27
5	you can see	449	209	2	1	2.12
6	trying to figure out	41	19	2	1	2.05
7	a little bit about	101	47	0	0	2.00
8	does that make sense	63	29	0	0	1.99
9	you know what	491	228	4	2	1.99
10	the University of Michigan	76	35	1	0	1.98

Shin and Nation also considered semantics, particularly individual senses of collocations with the same form (e.g. ‘looking up’ meaning ‘to improve’ and ‘looking up’ as in to find a word in a dictionary). The researchers report identifying 4,698 collocations using the criteria, with each criterion always met (p. 343). Our study resembles the Shin and Nation methodology in some key ways—particularly regarding frequency criteria and the analysis of individual senses for different expressions—but differs in that semantic transparency was not considered, nor degree of potential ‘usefulness’ to learners, teachers, and testers. Therefore, while Shin and Nation do identify many items which we also include in our PHRASE List (e.g. *a bit*, *as well*, *in fact*), they also include a number of sequences that for their transparency would not be included (e.g. *this year*, *very good*, *in the morning*). (See comparison in Table 2.) In addition, unlike the criteria used for the PHRASE List, the criteria in the Shin and Nation study were cumulative—all criteria had to be met in their study in order to include a collocation.

Finally, Wray (2008) outlines and describes a set of 11 criteria, first used in Wray and Namba (2003), designed to help researchers justify intuitions regarding what may or may not be formulaic. As the criteria were designed to guide researchers in assessing any potential formulaic candidate, the diagnostics are broader in scope than those used for the PHRASE List, which has a more specific intended application. The Wray and Namba criteria, however, are similar to the ones used for the PHRASE List because they are not cumulative (i.e. not all criteria are meant to necessarily be met), they are to be used post-hoc (i.e. to help justify strings first identified by computer), and are in support of qualitative judgments (i.e. not intended to ‘micro-analyze’ the sequences). Those features match the intended use of the six criteria we developed for selecting items in the PHRASE List, outlined below, divided into ‘core criteria’ and ‘auxiliary criteria’. The core criteria are those that

Table 2: Top 10 items— Shin and Nation (2008) compared with the PHRASE List

Shin and Nation (2008)		PHRASE List	
1	you know	1	have to
2	I think (that)	2	there is/are
3	a bit	3	such as
4	used to {INF}	4	going to {future}
5	as well	5	of course
6	a lot of {N}	6	a few
7	{No.} pounds	7	at least
8	thank you	8	such a(n)
9	{No.} years	9	I mean
10	in fact	10	a lot

were used to determine the candidacy of a given n-gram to inclusion in the list, while the auxiliary criteria were occasionally consulted to add support to decisions.

PHRASAL EXPRESSIONS: CORE CRITERIA

1 Is the expression a Morpheme Equivalent Unit (MEU)? Wray (2008) has suggested that one definition of a phraseological lexical item is that it is processed as if it were one morpheme 'without recourse to any form-meaning matching of any sub-parts it may have' (Wray 2008: 12), and especially among high-frequency expressions, there is psycholinguistic evidence for this assertion (Sosa and MacFarlane 2002; Kapatsinski and Radicke 2009). Such a criterion is consistent with the construct of 'word'. After all, a person reading the word *might* does not break it down into any subparts: it is clearly one morpheme, processed as such. An example of an MEU, then, would be *might as well*, as one who knows the expression is unlikely to resort to form-meaning matching of its subparts. As noted by Wray, however, 'morpheme equivalence' is more of a 'theoretical position' that certain wordstrings 'contain semantically viable parts that are not taken into account' when we read them or write them, for instance (Wray 2009: 31). Therefore, deeming a formulaic sequence to be an MEU is not a hard and fast science, but we start with this 'theoretical position' first, and use indicators (below) to justify our judgments (Wray 2008: 113).

2 Is the expression semantically transparent? To reiterate, the general idea regarding the items to be included in the PHRASE List is that they should be ones that are identified as potentially causing difficulty for learners of English, particularly on a receptive level. The expression *at this time*, for example, may qualify as an MEU because it means essentially the same thing as 'now', but even a learner who has never met this expression before and who encounters it in a text for the first time would stand a very good chance of unpacking its meaning simply by virtue of understanding *at* + *this* + *time* (i.e. the meaning remains even if each component word is replaced with its own definition). However, like all the criteria used, this one was applied with careful subjective evaluation for each potential item. As has been suggested by Taylor (2006), '[f]ull compositionality is rarely the case' (p. 61) in multiword exressions, and '[t]he distinction between the idiomatic and the non-idiomatic may not be so clear-cut...' (p. 62) – hence 'even the simplest of collocations may contain difficulty for learners' (Lewis 2000: 136). We bore this in mind for every expression considered.

3 Is the expression potentially 'deceptively transparent'? This question is also related to the issue of compositionality. Laufer (1989) has shown

that some lexical items in English can be 'deceptively transparent'—words learners 'think they know but they do not' (Laufer 1989: 11). Examples include *every so often* (which can be misread as 'often') and *for some time* (potentially misunderstood as 'a short amount of time'). When selecting phrasal expressions, an item was also often judged to potentially fit into this category when the most common and familiar meaning of at least one of the words in the expression was likely to pose confusion, especially if even a dictionary would not offer clear-cut help. For example, a survey of three advanced learner dictionaries for the word *further* shows that the first and highlighted senses of the word are to do with distance and extent. However, in the multiword item *a further* ('another') that meaning does not hold, and in the dictionaries surveyed the definition of 'additional' is not found until the third or fourth senses in the entry (senses are usually listed in order of frequency in corpus-informed dictionaries).

PHRASAL EXPRESSIONS: AUXILIARY CRITERIA

1 Does the expression have a one-word equivalent? For example, *put up with* is synonymous with *tolerate*. This is evidence of that expression being an MEU. Indeed, even if there is no one-word equivalent in English, but there is one in another language, it may also be evidence that the expression represents a single morpheme (Zgusta 1967). For example, there is no other English equivalent for *used to*, but there is evidence in Spanish (*solía*) and Portuguese (*costumava*) that it represents an MEU—not a series of separate words. Although not all items in the list necessarily must meet this criterion, the ability to roughly equate a multiword lexeme to a single one also facilitates its ability to be included in vocabulary tests with item formats that require form-meaning matching, like the VST.

2 Could the learner's L1 negatively influence accurate interpretation? Take, for instance, the expression *out there*, which on the surface may seem marginally semantically transparent. Although there is a metaphorical mapping at work (THE WORLD IS OUTSIDE), which ostensibly could make it easy to understand by a learner, the source domain does not necessarily operate the same way in languages other than English. The sentence *She wants a job but there's simply nothing* ***out there*** *right now* would be translated thusly in Portuguese: *Ela quer um emprego mas por enquanto não tem nada* ***por aí*** ('. . . there's nothing around . . .'). This criterion is also related to cognate words, of course. The word 'addition' in the phrases *in addition to* and *in addition* might seem at first glance to be easily decodable by a speaker of a Romance language, for example, but when one considers the formulaic equivalents in such languages (in Spanish: *aparte de, ademas*; Portuguese: *alem de, mais*; French: *en outre, en plus de*; Italian: *in piu,*

oltre che, etc.), it seems plausible that a focus on the cognate may actually render a spurious interpretation.

3 Does the meaning and/or opacity of a word change due to the grammar of the expression? The expression *no doubt* may violate the precepts outlined in Wray's MEU definition (since recourse to sub-parts may occur), but consider the discoursal difference between *I have no doubt she'll arrive* and *The president has no doubt taken his share of criticism*. While the first sentence is likely readily interpretable, the grammar of the expression has changed in the second: it is still preceded by a subject, but as an adverbial rather than a direct object. As such, it also potentially qualifies as being 'deceptively transparent'. This criterion was often partcularly relevant to passive constructions. For example, the fact that a beginner recognizes the meaning of the word 'know' does not mean that the same learner will understand a sentence like 'He's *been known to* do that before'. The verb 'expect', according to most learner dictionaries, is related to what one 'thinks will happen'. However, that meaning does not really remain in examples such as 'bathers *are expected to* shower before entering the pool' (= 'bathers *must* shower before entering the pool) and 'as a host *I'm expected to* be courteous' (= 'as a host *I'm supposed to* be courteous').

As in the Wray and Namba (2003) research, the criteria used for selection of expressions in the PHRASE List were consulted to qualitatively 'reveal the basis of intuitions already made' (Wray 2008: 116) about items in the initial quantitatively derived list, and not as a cumulative list of prerequisites. However, all expressions had to meet at least one of the core criteria. What all the criteria had in common was that they were designed to help us justify why we think the items chosen might pose some difficulty for a learner on a receptive level. Nonetheless, in order to ensure that criteria arrived at could be applied consistently by other researchers and produce replicable results, a subset of the same data to which the authors applied the criteria was ultimately analyzed by a trained rater, and this inter-rater exercise achieved an agreement figure of 99.2 per cent.

Regardless, it is clear from the criteria that the intuition, subjectivity, and general heuristics involved in the decision-making process necessitated a qualitative approach that no computer can yet achieve. The criteria (and heuristics), in turn, were also guided by the authors' combined 40+ years of English language teaching experience in a broad diversity of educational contexts (e.g. monolingual, multilingual, ESL, EFL, test preparation, EAP, etc.). Although such a methodology is extremely time and labor intensive, the end result is a PHRASE List that we would argue is clearly enhanced pedagogically.

The next step was to decide on the corpus source of our language data. For our purposes the BNC was the best choice from among the publicly available large corpora. It is a balanced 100-million word corpus of written and spoken English. Although like all corpora, the BNC has limitations (e.g. mostly British, skewed towards written English), it was chosen because of its size, widespread and longstanding use as a research instrument, and especially because it is the corpus that has most recently been used in the construction of recent vocabulary lists and tests (e.g. Leech *et al.* 2001; Nation 2006; Nation and Beglar 2007; and the BNC-20 *Vocabulary Profiler* available on the *Lextutor* website)—instruments into which the multiword items on the PHRASE List could be usefully incorporated. It needs to be acknowledged that the BNC is primarily written (90 per cent), but it still contains 10 million words of spoken discourse (one of the largest spoken [sub]corpora currently available), and so should also provide some useful information about the frequent multiword items used in spoken discourse.

The lead author began the actual extraction process by using *WordSmith Tools* (Version 5.0) to interrogate the BNC. A complete index (i.e. identification and extraction, including information about surrounding text) of every word contained in the BNC was made, and then WordSmith was asked to search for and list any and all n-grams between two and four words long repeated in the corpus at least five times. This search rendered a list of over 4.2 million n-grams. The BNC index of individual words when lemmatized and organized into word families (and rank-listed into 1,000-word bands (i.e. 1st to 1,000th, 1,001st to 2,000th, etc.) indicated that any lexical item that occurred more than 787 times was frequent enough for the 5,000-word family cut-off. Therefore, all n-grams occurring at least 787 times were considered for inclusion in the PHRASE List. This lowered the n-gram candidate list to approximately 15,000 items (i.e. only around 15,000 n-grams occurred 787 times or more).

The time-consuming qualitative stage of analysis then began. The lead author meticulously went down the n-gram list item-by-item looking for 'plausibly formulaic' multiword items (Wray 2009: 41), guided by the selection (and exclusion) criteria listed above. Great care was taken to not overlook potential expressions that at first may not appear formulaic. The sequence *at that,* for instance, may on the surface appear incoherent, but when more carefully investigated reveals interesting idiomatic patterning as in the sentence *CEOs took a pay cut in 2009, and a big one* ***at that***. Corpus-informed dictionaries were also regularly consulted as external confirmation that the n-gram constituted a lexical item (including, especially, the *Macmillan English Dictionary for Advanced Learners* (2007, 2nd edition), the *Cambridge Advanced Learner's Dictionary* (2008, 3rd edition), and the *Collins COBUILD Advanced Dictionary* (2009).

An additional challenge in the selection of n-grams for the PHRASE List was phraseological polysemy. It was quickly discovered that the number of multiword items with unique form-meaning mappings was relatively limited, with the vast majority requiring further investigation in order to determine their true frequency in the corpus. An example is the expression *at first*. Superficially, it may seem obvious that *at first* is an adverbial ('initially'), but as with all potential PHRASE List candidates, a concordance was run. It then became clear that *at first* also has other formulaic manifestations, as in *love at* *first sight*. However, since an item like *at first* has a frequency of over 5,000 in the corpus, line-by-line searching was not a viable option. Therefore, a random sampling method was employed instead. WordSmith generated a random concordance sample of 100 lines, and each line was scrutinized and deleted if necessary, until the percentage of lines reflecting the desired use of the multiword item was arrived at. In order to validate this percentage, a second random sample was generated to check consistency. This method produced consistent results, and in cases of minor discrepancies the lower of the two percentages was used (e.g. the two random concordances for *at first* yielded 84 and 85 per cent, so the 84 per cent figure was used). In the rare cases in which the figures did not match so closely, additional random samples were generated until a reliable percentage figure could be derived. Finally, the frequency figure for each multiword item was calculated by multiplying the total frequency figure by the percentage figure as explained above. For *at first,* this calculation was 5,090 (raw frequency) × .84 (per cent of desired use) = 4,275 (adjusted final frequency).

Also, frequency figures sometimes increased from their original levels. Since the current BNC-derived wordlists are lemmatized and organized into word families, the same needed to occur in the multiword item list. The expression *take place*, for example, in its uninflected form had a frequency count of just 3,248. However, the form can also be lemmatized:

take place → *takes place, taking place, taken place, took place*

In the case of *take place*, after conflating all of the inflected forms, the count increased from 3,248 to 10,556.

On other occasions, a subtractive method could be employed in order to arrive at a more accurate frequency figure. For example, *opposed to* essentially has two manifestations: *(be) opposed to sth*, and *as opposed to*. The n-gram list is not much help on its own since the program was asked to identify all recurring two-to-four word strings, and therefore *opposed to* is subsumed in *as opposed to*. In order to focus on just *opposed to*, it was possible to simply subtract the number of occurrences of the string *as opposed to* (1,615) from the number of times the bigram *opposed to* appears in the corpus (2,674), which rendered a difference of 1,059. In other words, the true frequency of just *opposed to* is 1,059.

Finally, expressions were sometimes encountered that contained variable components. For example, in the BNC, the first exemplar of *shake one's head* is actually 'shook his head' (1,698 occurrences). When a phrase with a variable component such as this one was identified (in this case, mainly the pronoun), a careful follow-up search was conducted in order to indentify all variable forms of that expression and arrive at a more accurate frequency count of it. Therefore, after considering *shook his head* (1,698), *shook her head* (1,241), *shook my head* (114), *shake my head* (30), *shaking my head* (17) and so on, the final frequency tally was 3,250.

Irrespective of the method ultimately employed, it was absolutely essential to take time to carefully examine each and every potential item to be included in the PHRASE List in order to both ascertain whether it met the selection criteria and to establish its true frequency. The final list ended up consisting of 505 multiword items which met the combined frequency and qualitative criteria. (See appendix in online supplementary data for the complete list, frequency figures, and examples of usage.)

DISCUSSION

The PHRASE List consists of a total of 505 multiword items. This is actually quite a substantial number, and indeed, if integrated into and calculated as part of the 5,000 most frequent word families, the 505 multiword items would constitute over 10 per cent of the total items. This figure also can be viewed in relation to the assertion sometimes made that the number of commonly occurring opaque multiword expressions in English is low (e.g. Moon 1998; Grant and Nation 2006; O'Keeffe *et al.* 2007), and thus 'should not be a major learning goal of a language learning programme' (Grant and Nation 2006: 11). While 10 per cent is arguably a low figure in relation to the other 90 per cent, it is certainly not ignorable, and is clearly enough to cause comprehension problems if not understood or misunderstood. According to the analysis conducted for the present study, there is a sharp increase in the number of phrasal expressions identified after around 12,000 occurrences, or the 1,000 (1K) word-family level, surging from 32 items in the first band to 85 items in the next (2K) level. This trend of increase appears to continue to the 4K level, and then levels off after 5K (Figure 2). This may be a reflection of a tendency for the most frequently recurring word combinations to sometimes become 'grammaticized' (e.g. Bybee 2003), often losing compositionality.

Lending strength to the assertion that the most common words in existing wordlists are merely the tips of phraseological icebergs, an analysis of the expressions in the PHRASE List shows that the 505 expressions are almost entirely comprised of the top 2,000 words in English, with the vast majority in the top 1,000. (Ninety-five per cent in the first 1,000 and 2.88 per cent in the second.) It is not unreasonable to guess that L2 learners processing those expressions might therefore actually believe they understand them (if they identify them) simply because the individual words are so well known, making them, in Laufer's (1989) terms, 'deceptively transparent' (cf. Martinez and Murphy 2011). Figure 3 and Tables 3 and 4 serve to exemplify how methods of text profiling that do not account for opaque phraseology risk underestimating the lexical complexity of a text.

The number of words 'off list' in the text in Figure 3 rises from a relatively manageable 7.46 per cent (Table 3) to a much more onerous 26.87 per cent when multiword expressions are accounted for, with their respective frequencies (Table 4). Therefore, assuming a learner knows only words within the 2,000-word family level, and none in the AWL, that coverage drops

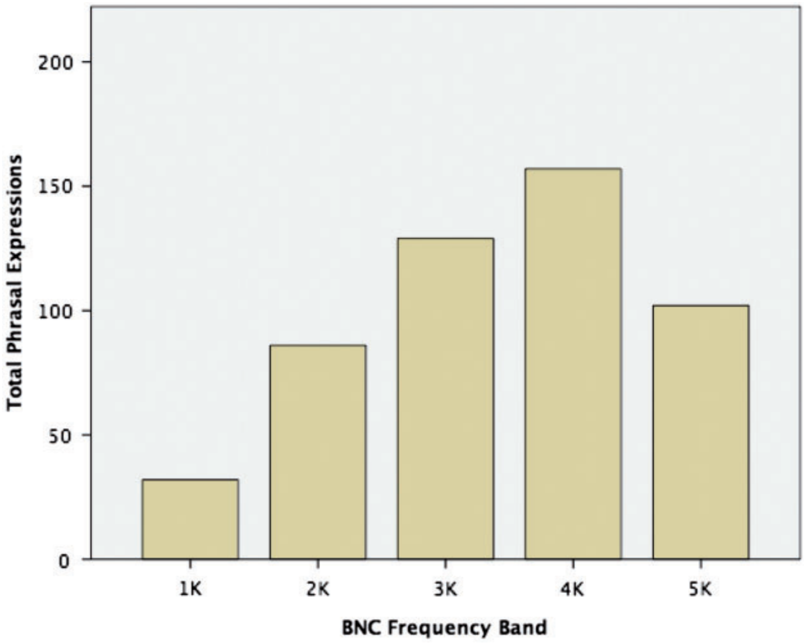
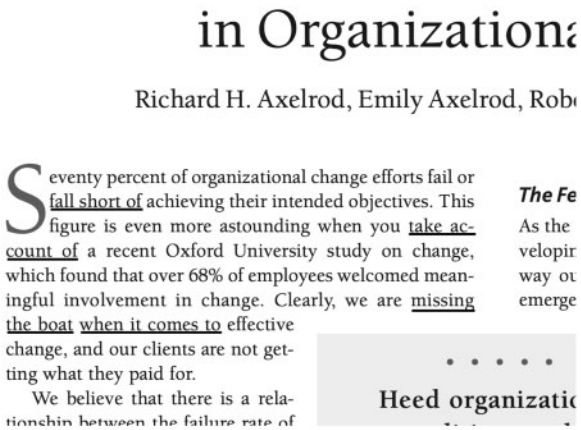


Figure 2: Phrasal expressions across frequency bands



in Organization

Richard H. Axelrod, Emily Axelrod, Rob

Seventy percent of organizational change efforts fail or fall short of achieving their intended objectives. This figure is even more astounding when you take account of a recent Oxford University study on change, which found that over 68% of employees welcomed meaningful involvement in change. Clearly, we are missing the boat when it comes to effective change, and our clients are not getting what they paid for.

We believe that there is a relationship between the failure rate of

The Fe

As the
velopir
way ou
emerge

• • • • •

Heed organizatio

Figure 3: Introduction from an authentic academic text (phrases underlined). Source: *Axelrod et al. (2006)*

from 88.06 per cent to a much more challenging 68.65 per cent (cf. Schmitt *et al.* 2011).

Above all, the PHRASE List was compiled with pedagogic purposes in mind, and it is hoped that it will be used to incorporate multiword items into existing

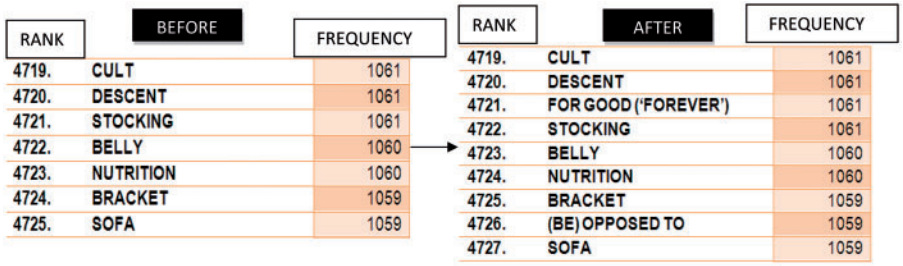
Table 3: Lexical profile of text in Figure 3 counting only single words

Frequency band	Words (types)	Text cover age (tokens)
0–1,000	*a account and are boat change clearly comes effective efforts employees even fail fall figure for found getting in is it meaningful missing more not number of on or our over paid recent seventy short study take that the their they this to university we welcomed what when which you*	86.57 per cent
1,001–2,000	*intended*	1.49 per cent
AWL	*achieving involvement per cent*	4.48 per cent
Off list	*astounding clients objectives organizational Oxford*	7.46 per cent
	Words in Top 2,000:	88.06 per cent
	+AWL words:	4.48 per cent
	Total text coverage:	92.54 per cent

Table 4: Lexical profile of text in Figure 3, phrases accounted for

Frequency band	Words (types)	Text coverage (tokens)
0–1,000	*a and are change clearly effective efforts employees even fail figure for found getting in is meaningful more not number of on or our over paid recent seventy study that their they this university we welcomed what when which you*	67.16 per cent
1,001–2,000	*intended*	1.49 per cent
AWL	*achieving involvement per cent*	4.48 per cent
Off list	*astounding clients fall short of missing the boat objectives organizational Oxford take account of when it comes to*	26.87 per cent
	Words in Top 2,000:	68.65 per cent
	+AWL words:	4.48 per cent
	Total text coverage:	73.13 per cent

wordlists, as exemplified in Figure 4. Not only would such integrated lists facilitate the systematic inclusion of both single-word and multiword expressions in tests and textbooks, they would also implicitly encourage their being perceived as a single construct in pedagogy. There is also the potential, for example, for the development of an automated lexical profiling tool such as *Range* (Heatley and Nation 1994) that instead of only analyzing a text for individual words (as in Table 3), also carries out a 'sweep' for phrases



BEFORE

RANK		FREQUENCY
4719.	CULT	1061
4720.	DESCENT	1061
4721.	STOCKING	1061
4722.	BELLY	1060
4723.	NUTRITION	1060
4724.	BRACKET	1059
4725.	SOFA	1059

AFTER

RANK		FREQUENCY
4719.	CULT	1061
4720.	DESCENT	1061
4721.	FOR GOOD ('FOREVER')	1061
4722.	STOCKING	1061
4723.	BELLY	1060
4724.	NUTRITION	1060
4725.	BRACKET	1059
4726.	(BE) OPPOSED TO	1059
4727.	SOFA	1059

Figure 4: Example of integrated list of phrasal expressions and single words

(Tom Cobb, personal communication) to more accurately reflect its lexical complexity, while simultaneously flagging up multiword items that may be worth including for explicit instruction or testing. In short, having a data- and practice-informed list of phrases, should take some of the guesswork out of what lexical items to teach and test, much like the GSL and AWL did for individual words. We would be pleased if the PHRASE List leads to future pedagogic materials including more multiword items, such as textbooks, graded readers, and language tests.

Finally, there will undoubtedly be many disagreements regarding certain individual items that were included in (or even excluded from) the list. Many expressions, such as *on the other hand* and *take for granted* are readily identifiable as formulaic, while others, such as *no one* (*'I can think of* ***no one*** *better'*) and *a good* (*'It takes* ***a good*** *three days'*) may not fit the stereotype of 'formulaic expression' in an obvious way. Users of the PHRASE List are advised to carefully consider such expressions in the light of the established criteria, and how what may at first seem easily understandable may in fact not be—even in context (Bensoussan and Laufer 1984; Haynes 1993)—especially for lower-proficiency learners. Nonetheless, critical evaluation of the list is welcome.

CONCLUSION

The importance and prevalence of formulaic language in the lexicon is now clear, as is the need for a principled way to more systematically include formulaic sequences in L2 pedagogy. The authors therefore sought to create a list of multiword lexical items that would serve a pedagogic purpose similar to that of well-established wordlists like the GSL and AWL—used, for example, in test and syllabus design—and through a mixed-method corpus analysis identified 505 phrasal expressions whose frequency and potential difficulty for learners make for a list that should address those needs. In the end, the PHRASE List is more than just a list of multiword items: it is also a list that is construct-matched with lists of individual words (Figure 4), and as such provides the means and justification for minimizing or even eliminating any 'special treatment' of multiword items. It is mostly to the advantage of all interested parties that formulaic vocabulary be eventually seen as simply being 'vocabulary'.

SUPPLEMENTARY DATA

Supplementary material is available at *Applied Linguistics* online.

NOTES

1 It is important to note that the 5,000-word threshold finds independent validation in the literature from various sources. Hindmarsh (1980) in his *Cambridge English Lexicon*, for example, found that 4,500 words would provide coverage to FCE (Cambridge First Certificate in English) level. Hindmarsh's lexicon, in turn, was used in conjunction with a number of other corpora by the English Profile Wordlists project in 2009 to compile a wordlist with levels aligned with the CEFR A1-B2—ultimately arriving at a list totaling 4,667 items (Capel 2010). This is also consistent with Milton (2009), who affirms that '[s]tudents who take advanced level examinations would probably be expected to recognize over 4500, or 90% or more, of this corpus (of 5000 words)' (p.180).

REFERENCES

Alderson, J. C. 2007. 'Judging the frequency of English words,' *Applied Linguistics* 28/3: 383–409.

Axelrod, R. H., E. Axelrod, R. W. Jacobs, and **J. Beedon.** 2006. 'Beat the odds and succeed in organizational change,' *Consulting to Management* 17/2: 1–4.

Bensoussan, M. and **B. Laufer.** 1984. 'Lexical guessing in context in EFL reading comprehension,' *Journal of Reasearch in Reading* 7: 15–32.

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. Johansson, 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.

Blackwell, S. 1987. 'Syntax versus orthography: problems in the automatic parsing of idioms' in R. Garside, G. Leech, and G. Sampson (eds): *The Computational Analysis of English*. Longman, pp. 110–19.

Bybee, J. 2003. 'Mechanisms of change in grammaticization: the role of frequency' in B. D. Joseph and R. D. Janda (eds): *The Handbook of Historical Linguistics*. Blackwell, pp. 602–23.

Capel, A. 2010. 'A1-B2 vocabulary: insights and issues arising from the English Profile Wordlists project,' *English Profile Journal* 1/e3: 1–11.

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.

Cooper, T. C. 1999. 'Processing of idioms by L2 learners of English,' *TESOL Quarterly* 33/2: 233–62.

Corrigan R., E. A. Moravcsik, H. Ouali, and **K. M. Wheatley** (eds). 2009a. *Formulaic Language Volume 1: Distribution and Historical Change*. John Benjamins Publishing Company.

Corrigan R., E. A. Moravcsik, H. Ouali, and **K. M. Wheatley** (eds). 2009b. *Formulaic Language Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations*. John Benjamins Publishing Company.

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

De Cock, S. 2000. 'Repetitive phrasal chunkiness and advanced EFL speech and writing' in C. Mair and M. Hundt (eds): *Corpus Linguistics and Linguistic Theory: Papers from ICAME 20 1999*. Rodopi, pp. 51–68.

Deignan, A. 2009. 'Searching for metaphorical patterns in corpora' in P. Baker (ed.): *Contemporary Corpus Linguistics*. Continuum, pp. 9–31.

Dorgeloh, H. and **A. Wanner.** 2009. 'Formulaic argumentation in scientific discourse' in R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (eds): *Formulaic Language Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations*. John Benjamins Publishing Company, pp. 523–44.

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

Ellis, N. C. and **S. G. Sinclair.** 1996. 'Working memory in the acquisition of vocabulary and syntax: putting language in good order,' *The Quarterly Journal of Experimental Psychology* 49A/1: 234–50.

Ellis, N. C., R. Simpson-Vlach, and **C. Maynard.** 2008. 'Formulaic language in native and second-language speakers: psycholinguistics, corpus linguistics, and TESOL,' *TESOL Quarterly* 42/3: 375–96.

Erman, B. and **B. Warren.** 2000. 'The idiom principle and the open choice principle,' *Text* 20/1: 29–62.

Fox, G. 1987. 'The case for examples' in J.M. Sinclair (ed.): *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Harper Collins, pp. 137–49.

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

Gouverneur, C. 2008. 'The phraseological patterns of high-frequency verbs in advanced English for general purposes: a corpus-drived approach to EFL textbook analysis' in F. Meunier and S. Granger (eds): *Phraseology in Foreign Language Learning and Teaching*. John Benjamins Publishing Company, pp. 223–43.

Grant, L. and **L. Bauer.** 2004. 'Criteria for re-defining idioms: are we barking up the wrong tree?,' *Applied Linguistics* 25/1: 38–61.

Grant, L. and **P. Nation.** 2006. 'How many idioms are there in English?,' *ITL – International Journal of Applied Linguistics* 151: 1–14.

Haynes, M. 1993. 'Patterns and perils of guessing in second language reading' in T. Huckin, M. Haynes, and J. Coady (eds): *Second Language Reading and Vocabulary Learning*. Ablex Publishing Corporation, pp. 46–66.

Heatley, A. and **P. Nation.** 1994. 'Range,' Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/. Accessed 2 April 2012].

Hindmarsh, R. 1980. *Cambridge English Lexicon*. Cambridge University Press.

Hsu, J. T. 2008. 'Role of the multiword lexical units in currect EFL/ESL textbooks,' *US-China Foreign Language* 6: 27–39.

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

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

Kapatsinski, V. and **J. Radicke.** 2009. 'Frequency and the emergence of prefabs: evidence from monitoring' in R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (eds): *Formulaic language Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations*. John Benjamins Publishing Company, pp. 499–520.

Katz, G. and **E. Giesbrecht.** 2006. 'Automatic identification of non-compositional multi-word expressions using latent semantic analysis,' *ACL 2006: Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, pp. 12–19.

Koprowski, M. 2005. 'Investigating the usefulness of lexical phrases in contemporary coursebooks,' *ELT Journal* 59/4: 322–32.

Korkontzelos, I. and **S. Manandhar.** 2009. 'Detecting compositionality in multi-word expressions,' *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pp. 65–8.

Kuiper, K. 1996. *Smooth Talkers*. Lawrence Erlbaum.

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

Laufer, B. 1989. 'A factor of difficulty in vocabulary learning: deceptive transparency,' *AILA Review* 6: 10–20.

Leech, G., P. Rayson, and **A. Wilson.** 2001. *Word Frequencies in Written and Spoken English Based on the British National Corpus*. Longman.

Lewis, M. 1993. *The Lexical Approach*. Language Teaching Publications.

Lewis, M. 2000. *Teaching Collocation*. Language Teaching Publications.

Lewis, M. 2008. 'The idiom principle in L2 English: assessing elusive formulaic sequences as indicators of idiomaticity, fluency and proficiency,' unpublished doctoral thesis, Stockholm University.

Martinez, R. and **V. Murphy.** 2011. 'Effect of frequency and idiomaticity in second language reading comprehension,' *TESOL Quarterly* 45/2: 267–90.

Meunier F. and **S. Granger** (eds). 2008. *Phraseology in Foreign Language Learning and Teaching*. John Benjamins Publishing Company.

Millar, N. 2011. 'The processing of malformed formulaic language,' *Applied Linguistics* 32/2: 129–48.

Milton, J. 2009. *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters.

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

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

Nation, I. S. P. 2006. 'How large a vocabulary is needed for reading and listening?,' *The Canadian Modern Language Review* 63/1: 59–82.

Nation, I. S. P. and **D. Beglar.** 2007. 'A vocabulary size test,' *The Language Teacher* 31/7: 9–13.

Nation, I. S. P. and **R. Waring.** 1997. 'Vocabulary size, text coverage and word lists' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge University Press, pp. 6–19.

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 Publishing Company.

Ohlrogge, A. 2009. 'Formulaic expressions in intermediate EFL writing assessment' in R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (eds): *Formulaic Language Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations*. John Benjamins Publishing Company, pp. 375–86.

O'Keeffe, A., M. McCarthy, and **R. Carter.** 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge 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. 191–225.

Read, J. 2000. *Assessing Vocabulary*. Cambridge University Press.

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

Schmitt, N. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan.

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

Schmitt, N., X. Jiang, and **W. Grabe.** 2011. 'The percentage of words known in a text and reading comprehension,' *Modern Language Journal* 95: 26–43.

Schonell, F. J., I. G. Meddleton, B. A. Shaw, and **M. Routh.** 1956. *A Study of the Oral Vocabulary of Adults*. University of Queensland Press.

Shin, D. and **P. Nation.** 2008. 'Beyond single words: the most frequent collocations in spoken English,' *ELT Journal* 62/4: 339–48.

Simpson-Vlach, R. and **N. C. Ellis.** 2010. 'An academic formulas list: new methods in phraseology research,' *Applied Linguistics* 31: 487–512.

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

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

Sosa, A. V. and **J. MacFarlane.** 2002. 'Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*,' *Brain and Language* 83: 227–36.

Spöttl, C. and **M. McCarthy.** 2003. 'Formulaic utterances in the multi-lingual context' in J. Cenoz, B. Hufeisen, and U. Jessner (eds): *The Multilingual Lexicon*. Kluwer, pp. 133–51.

Taylor, J. R. 2006. 'Polysemy and the lexicon' in G. Kristiansen, M. Achard, R. Dirven, and F. J. Ruiz de Mendoza Ibanez (eds): *Cognitive Linguistics: Current Applications and Future Perspectives*. Mouton de Gruyter, pp. 51–80.

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*. John Benjamins, pp. 153–73.

West, M. 1953. *A General Service List of English Words*. Longman, Green and Co.

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

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

Wray, A. 2009. 'Identifying formulaic language: Persistent challenges and new opportunities' in R. Corrigan, E. A. Moravcsik, H. Ouali, and K. M. Wheatley (eds): *Formulaic Language Volume 1: Distribution and Historical Change*. John Benjamins Publishing Company, pp. 27–51.

Wray, A. and **K. Namba.** 2003. 'Formulaic language in a Japanese-English bilingual child: a practical approach to data analysis,' *Japan Journal for Multilingualism and Multiculturalism* 9/1: 24–51.

Wulff, S. 2009. *Rethinking Idiomaticity*. Continuum.

Wulff, S., J. M. Swales, and **K. Keller.** 2009. ''We have about seven minutes for questions': the discussion sessions from a specialized conference,' *English for Academic Purposes* 28: 79–92.

Zgusta, L. 1967. 'Multiword lexical units,' *Word* 23: 578–87.