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

Measurement of formulaic sequences

John Read and Paul Nation
Victoria University of Wellington

Introduction

Most of the research on formulaic sequences until now — particularly that done before the advent of computers and the field of corpus linguistics — has primarily involved descriptive work to exemplify and classify multiword units which scholars have considered to function lexically rather than grammatically in the language. However, if work in this area is to advance and to move into the mainstream of applied linguistic research, it is necessary to address some important methodological issues that arise in the investigation of these lexical units. This chapter draws on insights from research methodology and language testing to identify particular problems of measurement in dealing with formulaic language and propose how they might be solved. We will illustrate some of our points by reference to the work reported in other chapters of this volume.

One of the exciting developments in recent years is the realisation that formulaic sequences have been of long-standing interest to scholars in a whole variety of disciplines both inside and outside applied linguistics. Thus, in a sense, we are currently in a phase of surveying and attempting to integrate the insights that have been gained by researchers working in different fields all around the world without necessarily being aware of what others were doing. This is well illustrated by Wray's (2002) excellent book, which draws together work in general linguistics, phraseology, lexicography, corpus linguistics, first and second language acquisition, language teaching, neurolinguistics and other disciplines. It is important to note that scholars in these various fields not only bring their own theoretical perspectives to bear on the study of formulaic language but also have distinctive methodological approaches to their work. This of course is a familiar situation in an interdisciplinary field like applied linguistics, but what it means is that it would be unrealistic for us to attempt to impose a single research paradigm on the study of formulaic sequences. Thus, in this chapter we will attempt to focus on general principles and issues of measurement that need

to be taken into account regardless of the particular research paradigm that the investigator is working within.

Use of the term measurement may suggest that we favour quantitative or statistically based methods of investigation rather than qualitative ones. However, we are adopting a broad definition of measurement which includes criteria for the identification of multiword units as formulaic sequences and for classifying them into categories, even if no further counting of relative frequencies or any other form of statistical analysis is then applied. In addition, we argue that an adequate account of formulaic units as they function in language acquisition and language use can come only from a combination of quantitative and qualitative analyses. The same already applies, of course, in word-based vocabulary studies. Although it may seem quite straightforward to the naïve observer to identify and count words, linguists and vocabulary researchers are well aware of the problematic nature of the word as a linguistic concept. A purely formal definition of a word as word form is of limited value in itself, as illustrated by one of the early computer-based word frequency counts (Carroll, Davies and Richman, 1971), where *people*, *People*, *people's*, *People's*, *peopled*, *peoples* and *Peoples* are all listed as separate items. Thus, vocabulary scholars have developed more meaningful conceptual units, such as the lemma, homonym, word family, lexeme or lexical unit, and the raw output of a frequency count needs to be classified at least partially by means of human judgement into one or more of these categories in order to be usable for further analysis. Some of these categories already involve units consisting of more than one word form, such as compound nouns, phrasal verbs and idiomatic expressions. Once we shift the attention to the whole range of multiword units, the basic elements are rather more difficult to identify than individual word forms are and so both quantitatively and qualitatively, more sophisticated procedures are required to locate and classify them.

In this chapter we intend to do the following: We will consider a definition of formulaic sequences and then look at reliability and validity issues in their identification, eventually focusing on the importance of triangulation. Finally, we consider the procedures used in several of the studies included in this volume.

Definition of the construct

In modern validity theory in educational measurement, a crucial step initially is to define the construct at a conceptual level. This then provides a basis for

judging the adequacy of operational measures of the construct. In the case of formulaic sequences, Wray (2002: 9) has proposed a definition which is likely to be very influential but it also needs to be subject to critical scrutiny. If her definition is adopted, then the ultimate goal of an analysis will be to identify sequences that are "stored and retrieved whole from memory at the time of use". This is a challenging goal because the means of storage and retrieval of the same sequence can differ from one individual to another, and can differ from one time to another for the same individual depending on a wide range of factors such as changes in proficiency, changes in processing demands, and changes in communicative purpose.

There is some evidence for this variability from the study of idioms. Grant (2003) did an exhaustive study of what she called *core idioms*, which are non-compositional (the meaning of the parts does not give the meaning of the whole) and non-figurative (the image created by the unit does not relate to the meaning of the unit). They must also consist of words that can occur in other places. Grant found that English has about 104 core idioms. About 25% are frozen, and only 10 had a literal equivalent in the British National Corpus. Even among such a narrowly defined group of items, where we would expect to find extreme formulaicity, the norm seems to be that there is considerable variation. Here some of the variants of the core idiom *pull someone's leg*:

> *pull my blue leg, somebody's leg was being pulled, having his leg pulled, leg pulling, a leg pull, a leg puller, tugged my leg, yank somebody's leg, leg tugged/yanked.*

There is a similar set of variants for *put your foot in your mouth*:

> *put your foot in it, putting his foot in his mouth to the kneecap, put his foot well and truly in his mouth, with her foot in his mouth, foot and mouth, foot-in-mouth moments, foot-and-mouth soldiers, put your feet in your mouth.*

Most of these are low in frequency but there is a lot of variation, even without considering the numerous versions of the object or verb form. This variability however does not prove that all uses of the idiom are not formulaic. It is clear that some of the variations are deliberate attempts to add humour by playing with something that is typically fixed. The evidence from the study of core idioms suggests that there are probably very few sequences, if any, that are always formulaic, and thus the most valid criteria for deciding formulaicity will be those that take account of features that are present in each particular use of a possible sequence.

Wray's (2002: 9) definition of formulaic sequences is deliberately inclusive. It

goes only a short way towards specifying the form in which a sequence is stored and it states explicitly that the sequence need not be continuous. That is, there may be insertions in it, such as when *right bloody* is inserted into *came a cropper*: *came a right bloody cropper*. The definition also seems to exclude substitution of items within a sequence, such as the following variations within the 'pull' and 'person' components of *pulling my leg*:

pull	his	
pulled	her	
pulls	my	
pulling	your	leg
yank etc.	our	
tug etc.	someone's	
	his sister's	

Similarly, transformations of a sequence would not be included: *chew the fat, fat-chewing, fat-chewers*. These substitutions and transformations would be excluded because they would involve "generation or analysis of the language grammar" (Wray 2002: 9).

The definition does not specify the form of the items in storage. If it is verbatim storage, where the actual words of the sequence are stored without the possibility of substitution or transformation, then Grant's (2003) research suggests we are dealing with only a small number of sequences that are rather infrequent. This definition of a formulaic sequence is one that Kuiper (this volume) seems to follow. It is relatively easy to identify such sequences because of their fixed form, and most researchers would readily consider them formulaic. However, much further along a possible scale of formulaicity are the numerous examples of collocational prosody such as *bordering on*, where the formula is at a rather abstract level. These sequences allow insertion, inflection, substitution, deletion, and transformation which all involve "generation or analysis by the language grammar". The term formulaic sequence could not be sensibly applied to such patterns.

Thus, Grant's (2003) findings challenge the adequacy of Wray's definition of the construct. The interest in formulaic sequences is partly a reaction to the lack of description of semantic patterning in previous descriptions of language. However, semantic patterning and formulaic sequences are not the same thing and so the definition needs to take account of this distinction if it is to be comprehensive enough to cover the phenomena to be investigated. Given the variability in formulaic language that we noted above, the definition of these sequences may need to be tailored to some degree to the specific objectives of each research study.

Sources of evidence

Once conceptual issues have been addressed, an essential requirement for the identification of formulaic sequences is to have a source of examples of multiword units for analysis. From a measurement perspective, the key issue in choosing a suitable source is one of sampling: how to ensure that there are sufficient examples to allow reliable generalisations to be made and, where applicable, that the sample is representative enough to provide the basis for a valid classification system.

There is a long-standing practice among grammarians and linguists of building up a collection of examples of idioms or other formulaic sequences, based on their own introspective knowledge of the language plus instances that they encounter through their reading, conversational interaction and other communicative activities in the language. Some scholars such as Pawley and Syder (1983) and Nattinger and DeCarrico (1992) adopted a more structured approach, drawing on transcriptions of spoken discourse and/or written texts of various kinds but without giving specific details of the scope of the source material. Their work has proved to be very important in applied linguistics in drawing attention to the pervasiveness of formulaic sequences and highlighting the variety in both the forms they take and the functions they perform. However, in sampling terms, this general approach will typically create a "convenience" sample, which is subject to uncontrolled bias. For work in this area to advance, it is necessary to complement such informal collections of examples with more systematic data-gathering procedures that can challenge the perceptions of individual investigators.

The obvious source of more systematic evidence is some kind of text database. These now commonly take the form of computer corpora, providing very large samples of language, which can then be searched in an efficient manner. Corpus software generates frequency counts and a whole variety of other quantitative measures. In addition, it can supply lists of words and word strings that meet particular specifications as the basis for qualitative analyses of idiomaticity, semantic transparency, semantic vs. pragmatic meaning, and so on.

There are a number of options when it comes to the choice of a corpus for the analysis of formulaic sequences.

Large general corpora

Mega-corpora such as the Bank of English and the British National Corpus lend themselves well to certain kinds of research on formulaic sequences, for similar reasons to the enormous contributions they have made to lexicography, word-based vocabulary studies, and descriptive grammars, among others. However, depending on the particular focus of the research, they also have some limitations.

- There is bias in the sample of texts they include. The most obvious one is that spoken language is underrepresented, but there is also bias in style (overrepresentation of formal, informative prose) and genre (journalistic texts in the Bank of English).
- Even in such large corpora, particular kinds of formulaic sequence may have quite low frequency, as Moon (1998) found in her research on idioms, proverbs and similes.
- Although corpus software is getting more sophisticated all the time, there are still limits on what it can find in a large corpus.
- The particular kinds of text that are of interest (eg learner language; storytelling to schoolchildren) may not be in the corpus at all.

Specialized corpora

There are a fast growing number of more specialized corpora which offer opportunities to investigate formulaic sequences in more particular varieties of language. These include corpora of spoken language (the London Lund Corpus, the Cambridge and Nottingham Corpus of Discourse in English — CANCODE), learner language (the International Corpus of Learner English — ICLE), child language (The Child Language Data Exchange System — CHILDES), regional varieties (the International Corpus of English — ICE — corpora, the Brown corpus of American English and the various parallel corpora of other national varieties), and discipline-specific corpora.

The issues involved in selecting a particular corpus include considering whether the corpus fits the particular requirements of a proposed formulaic sequence study, whether it is accessible by other researchers (than the original compilers), whether the corpus is large enough to satisfy reliability requirements, and whether certain crucial kinds of information about the texts are available in the corpus, for example, the specific sources of written texts or particular phonological notation for oral texts. Given the pragmatic dimension to the meaning of many formulaic sequences, especially in oral language use, the researcher may require richer contextual information than the corpus provides.

A further category includes collections of written or oral texts that may not be thought of as constituting a corpus, such as the reanalysis by Foster (2001) of the transcripts from the Skehan and Foster research on task-based language learning.

Purpose-built databases

If existing corpora do not meet the research requirements, it will be necessary to build a set of data from scratch. This does not necessarily involve compiling a "whole" corpus (whatever the minimum dimensions of that might be). It may simply be the kind of data-gathering that sociolinguists, discourse analysts and others routinely engage in to collect samples of language use, either by unobtrusive recording of "natural" speech events or by elicitation procedures. Kuiper's studies of race callers, auctioneers and checkout operators are good examples of these (see Chapter 3).

Procedures for identification and classification

As previously indicated, in its present stage of development the study of formulaic sequences still faces fundamental problems in identifying the units of analysis within a database or corpus. Wray (2002: Chap 2) gives a comprehensive discussion of the criteria that have been proposed or applied in previous research. We will summarize the criteria here and explore the measurement issues.

Intuition

The status of the intuition of an individual investigator is dubious from a modern "scientific" perspective. The exercise of this kind of subjective judgement is likely to be more acceptable if one or more of the following conditions apply:

- a definition of what is meant by a formulaic sequence is carefully formulated in advance, as previously discussed.
- the investigator communicates the definition to a second person, who then attempts to replicate the investigator's identification of the formulaic units.

- instead of relying on the researcher's judgement, a panel of judges is formed to analyse the database and a multiword unit is accepted as formulaic only when most, if not all, the judges identify it as such.

In other words, what is required is intersubjectivity or, in measurement terms, a high degree of inter-rater reliability.

Nevertheless, as Wray (2002:20–25) points out, even meeting these basic conditions is not straightforward in the case of formulaic language. Corpus linguists such as Sinclair (1991) argue that their research reveals intuition to be a very fallible means of investigating the facts of language use, with regard to the relative frequency of linguistic features, typical meanings of lexical items, characteristic patterns of collocation, and so on. Secondly, in the context of second language acquisition research, the native speaker intuitions of the researcher are often brought to bear to account for the language production of learners, who may or may not have an intuitive basis for what they say or write in the second language. This means that the formulaic status of sequences in learner language is even more difficult to establish by means of intuition than in the case of native speaker production. A third difficulty identified by Wray is that recognition of formulaic language may depend on the shared knowledge which comes from membership of a particular speech community rather than being universal among users of the language concerned. This represents just one more limitation on the value of intuition as an investigative procedure.

Corpus analysis

Computer corpus analysis has added a powerful new tool to the range of procedures available for the study of formulaic sequences. Moving beyond the concept of locating and counting individual word forms, corpus software can search for specified headwords, combinations of words and even discontinuous sequences of words. Thus, if the investigator can specify particular words or word strings that are potentially formulaic (or known to be so on the basis of other evidence), the software can instantly assemble all of the examples in the corpus for inspection and further analysis. An alternative approach is a purely statistical procedure that identifies sequences of two, three or more words that regularly co-occur throughout the corpus beyond a threshold level of probability. This second approach has produced a great deal of data that turns out not to be formulaic, depending on the definition of formulaic language adopted, but on the other hand it has shown its potential to give new insights into multiword units that are not available through intuition. In both cases, the quantitative evidence supplied by the software needs to be evaluated by the application of human judgement to determine which of the word sequences are formulaic — and if a classification system is involved, which ones fit in which categories.

Concordance software such as that included in Wordsmith Tools and SARA can be used to find collocational clusters in corpus data. The most flexible software allows the researcher to specify a search word or words and to gather and count the occurrences of collocates for several positions on either side of the search node. Such software is an extremely valuable tool for research on formulaic language. However, it is essential for the researcher to examine each instance of the data to make sure that it is relevant. One way to demonstrate this point is by means of a training exercise employing the SARA software on the British National Corpus. The task is to use corpus data to answer the question, "Are men beautiful?". That is, do *men* and *beautiful* collocate? A corpus search with *men* as the node and *beautiful* as the collocate, using a 6 to the left 6 to the right span, found 38 instances. In only five of these were they really collocates. A more limited search of the same corpus using 3 to the left and right produced ten instances of which only four were collocates. Excluding right hand occurrences of *beautiful* would not change the result substantially. Here are the ten instances.

to see if she were as **beautiful** as men told
who felt the need to dress up and be **beautiful** for their men
made love to the most brilliant and **beautiful** men of your generation
Next to him were two brothers, tall **beautiful** men with liquid eyes
There are some **beautiful** men's clothes around
You are so **beautiful** that men would die for you
stunningly **beautiful** to boot. Men would
Men and **beautiful** women also join in.
If you were in Prague, two **beautiful** men like you,
There are some very **beautiful** young men there.

Clearly, valid cluster analysis requires manual checking of the data.

Another limitation of concordance software is that it can automatically locate only contiguous sequences. In order to locate non-contiguous ones, it is necessary for the researcher to enter in the search request either a contiguous subpart of the whole sequence or at least one key lexical component of it. This of course assumes the whole sequence is already known to be formulaic. It is very likely

that a substantial proportion of the formulaic language in English remains to be discovered; the non-contiguous nature of the sequences involved means that they fall below the threshold of recognition, whether it be by human intuition or automated computer search.

In addition to the limitations of corpus analysis we have already noted, Wray (2002:28–30) discusses two others. One is the big discrepancy in the estimates by different researchers of the proportion of the corpus they analysed which could be considered to consist of formulaic sequences. Leaving aside any problems with the reliability of the individual analyses, there are clearly validity issues here related to differing theoretical and operational definitions of formulaicity. Secondly, Moon (1998) among others has found that numerous formulaic expressions that are very familiar to native speakers do not occur at all even in the mega-corpora.

Structural analysis

A variety of formal criteria have been proposed to assist in the identification of formulaic sequences. The two mostly widely recognised ones are non-compositionality and fixedness, which are characteristics of some idioms and other formulaic expressions to a lesser degree. Noncompositionality means that the sequence is not interpretable as a literal statement. It may contain individual words that never occur except as part of that expression. Fixedness refers to the degree to which either the order of the words in the sequence can be changed, individual words can be replaced by others, items can be inserted, or items can be inflected.

The fact that these criteria turn out to be continua contributes to the difficulty in drawing the line between formulaic and non-formulaic expressions.

Phonological analysis

In the case of spoken language, certain phonological features have been investigated as possible indicators of formulaic sequences. These include speech rate, pausing, stress patterns and clarity of articulation. The investigation of phonological criteria is likely to involve elicitation of data by means of a structured research design rather than analysis of an existing corpus. Apart from the relatively limited size of spoken corpora, the transcription of the oral texts in a general corpus may not meet the specific requirements of a phonological analysis. In addition, there are certain variables that need to be controlled in the interests of internal validity, such as whether the talk is spontaneous or prepared, what the topic is and the nature of the speaking task to be performed.

As with other kinds of research involving the elicitation of spoken language data, there is tension between the control and manipulation of key variables needed to obtain interpretable results and the desirability, in the interests of external validity, of recording speech which is as natural and unmonitored as possible.

Pragmatic/functional analysis

Another analytical criterion recognises that formulaic sequences have important roles in the performance of speech acts and are commonly associated with particular speech events. This provides an alternative approach to identifying them when data-gathering focuses on the particular social setting in which they typically occur (see Kuiper, Chapter 3). It also gives another perspective on the lack of transparency that the more fixed formulaic sequences tend to exhibit. Idioms are said to lack *semantic* transparency because their meaning is not interpretable from knowledge of the individual lexical components. To this we can add *pragmatic* transparency, which refers to the need for knowledge of the social context in which particular formulaic expressions are used in order to be able to understand their role in the discourse.

The need for an eclectic approach

Overall none of the criteria outlined in the preceding section is adequate by itself for the identification of formulaic sequences. As Wray (2002) emphasises, researchers will generally need to apply more than one form of analysis in order to obtain valid results. The concept of triangulation, which has come to be an integral part of the qualitative research paradigm, is very relevant here.

Let us now look at some of the studies in this volume to see how this triangulation might be done.

Wray's (Chapter 12) fascinating study of a beginner's memorisation of sequences in Welsh uses evidence from pausing, errors, and changes to items in strings to examine the effect of the memorisation of sequences and analysis on the retention of immediately useable language items. This use of both quantitative and qualitative evidence provides interesting insights into the way language data is stored and changed.

In two innovative studies, Underwood, Schmitt and Galpin (Chapter 8), and Schmitt and Underwood (Chapter 9) used eye movement and self-paced reading methodologies to see if formulaic phrases embedded in a text were read any differently from other non-formulaic parts of the text. Considerable triangulation was used to ensure that the items being investigated were formulaic sequences. First a number of items were selected using intuition. Then their frequency was checked in a corpus (presumably the frequency of a fixed unchanging sequence), and then these were tested in a cloze text with initial letter cues to check that the items were indeed predictable.

The Schmitt, Dörnyei, Adolphs and Durow study (Chapter 4) uses a range of criteria including previous identification by other researchers, corpus frequency, and occurrence in language teaching texts to come up with a list of target sequences.

These examples illustrate the way forward in establishing a sound empirical basis from a measurement perspective for research in this rapidly developing area of vocabulary studies.

Reliability and Validity

As a summary of some of the main points of this chapter, let us consider the measurement of formulaic sequences in terms of the classic criteria of reliability and validity. To satisfy the internal reliability requirement, any measures need to be consistently applied. This means that the criteria for identification and classification should be clear and there should be a high level of agreement among at least two analysts (or raters) working independently through a substantial sample of the data, if not the whole data set. In some studies (Foster, 2001; Jones and Haywood, Chapter 13) several expert raters have been used and the identification of sequences as formulaic has relied on achieving consensus or near consensus among the raters. In other cases, where formulaic sequences are to be classified into a number of categories, the percentage of exact agreement in the classifications serves as the estimate of internal reliability.

External reliability requires the clear description of procedures so that the study could be replicated. For a corpus search, for instance, the necessary information includes a description of the corpus, the kind of search, search parameters (what span each side of the node was used), whether there was manual checking of the results of the search, and what criteria were applied when checking.

Validity issues are particularly problematic with formulaic strings, as the essential criterion — storage as a whole unit — is a difficult one to operationalise. For internal validity, there is a need for a clear definition of what a formulaic string is, both at the conceptual level and in operational terms. Research indicates that this may need to take account of the function of formulaic strings (Wray, 2002: Chaps 4 and 5). Where possible, there should be methodological triangulation: two or more methods should be employed to identify what is formulaic.

For external validity, the corpus — or whatever other data source is used — should represent target language use and be large enough to contain an adequate number of examples. This means that very large corpora are likely to be needed, which makes the problem of representativeness more difficult to solve.

References

Carroll, J.B., Davies, P., and Richman, B. 1971. *The American Heritage Word Frequency Book.* Boston MA: Houghton Mifflin.

Foster, P. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing,* M. Bygate, P. Skehan, and M. Swain (eds), 75–93. Harlow: Longman.

Grant, L. 2003. A Corpus-based Investigation of Idiomatic Multi-word Units. Unpublished PhD thesis. Victoria University of Wellington.

Moon, R. 1998. *Fixed Expressions and Idioms in English.* Oxford: Clarendon Press.

Nattinger, J.R. and DeCarrico, J.S. 1992. *Lexical Phrases and Language Teaching.* Oxford: OUP.

Pawley, A. and Syder, F.H. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication.* J.C. Richards and R.W. Schmidt (eds), 191–225. London: Longman.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford: OUP.

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