A function-first approach to identifying formulaic language

There has been much recent interest in creating pedagogically-oriented descriptions of formulaic language (e.g., Biber, Conrad, & Cortes, 2004; Ellis, Simpson-Vlach, & Maynard, 2008; Hyland, 2008; Shin & Nation, 2008). Corpus-based research in this area has typically taken what might be called a ‘form-first’ approach, in which candidate formulas are identified as the most frequent collocations or n-grams in a relevant corpus. The functional correlates of these high frequency forms are then determined at a later stage of analysis. This approach has much to commend it. It enables fast, reliable analysis of large bodies of data, and has produced some important insights, retrieving patterns which are often not otherwise evident.

While this research continues to yield valuable results, the present paper argues that much can also be gained by taking what I will call a ‘function-first’ approach to identifying formulaic language. On this approach, a corpus is first annotated for communicative functions; formulas are then identified as the recurrent patterns used to express each function. This method has two key advantages. First, it enables the analyst to identify not only how frequent a given form is, but also what proportion of attempts to express a particular message use that form. As Wray (2002, p. 30) has pointed out, information of this sort is often vital, alongside overall frequency information, in determining whether an item is formulaic.

The second advantage of a function-first approach is that it enables the analyst to take account of the range of variation associated with a particular expression. Form-first approaches are required to specify in advance the formal features that two strings of language must have in common to count as instances of a single formula. This has usually meant restricting analysis to simple forms such as two-word collocations or invariant ‘lexical bundles’. Since many formulas are to a certain degree internally variable, this restriction inevitably involves the loss of potentially important information. The function-first approach aims to account for the range of variation typical in the expression of a given function and so to include a fuller description of such variable formulas.

While these features of the function-first approach have the potential to be of much theoretical interest, the primary motivation for the approach is pedagogical. For language learners, the key information is often not which strings of language are the most frequent, but rather which functions they are most likely to need, what formulas will most appropriately meet these needs, and how those formulas can be manipulated. This is precisely the information which a function-first approach aims to provide.

This paper will demonstrate a function-first approach to identifying formulaic language through an investigation of a collection of introductions to academic essays written by MA students in the social sciences, taken from the corpus of British Academic Written English. The methodological issues involved will be discussed and sample results presented.


