As Hill seems to suggest: Variability in formulaic sequences with interpersonal functions in L1 novice and expert academic writing

Ying Wang

Uppsala University, Sweden

A R T I C L E   I N F O

A B S T R A C T

Formulaic sequences (FSs) are pervasive in natural language use and play an important role in differentiating socially-situated practices. The predominant trend in this research area is to take a frequency-based approach, relying on the computer to identify frequent recurrent forms in a given corpus, at the expense of disregarding their structural and semantic unity and multifunctionality, as well as overlooking discontinuous and infrequent sequences. Through careful manual identification and annotation of FSs in context, the present study provides additional insights into the use of interpersonal FSs that distinguish L1 novice and expert academic writing. The results show that the novice writers actually produced a wider range of FSs with interpersonal functions than did the expert writers. It is argued that less frequent FSs cannot be dismissed in FL research simply because of their low frequencies. Taken together, these seemingly idiosyncratic choices may reveal important functional and formulaic features that characterise a particular community. The main differences between the two groups of writers pinpoint areas (e.g., genre- and discpline-specific conventions, register awareness) that need further investigation and specific attention in the training of novice writers.

1. Introduction

Research into formulaic sequences (FSs), i.e., words that have 'an especially strong relationship with each other in creating their meaning' (Wray, 2008: 9), such as *by and large*, *of course*, *on the other hand*, has been one of the rapidly growing areas in applied linguistics over the past decade (Hyland, 2008, pp. 41—62). Corpus studies, in particular, have revealed that they are pervasive in natural language use (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Erman & Warren, 2000; Martinez & Schmitt, 2012), and play an important role in differentiating socially-situated practices (Biber, Conrad, & Cortes, 2004; Hyland, 2012). Research in English for academic purposes (EAP) demonstrates that professional academics and student writers alike draw on formulaic resources to 'develop their argument, establish their credibility and persuade their readers' (Hyland, 2008, p. 59).

The predominant trend in this research area is to take a frequency-based approach (e.g., lexical bundle, n-gram), relying on the computer to automatically identify frequently recurring word sequences in a given corpus, at the expense of disregarding their structural and semantic unity and multifunctionality, as well as overlooking discontinuous and infrequent multi-word

units. In the present study, FSs and their functions are identified and annotated manually in context and the subsequent analysis starts from the functions and moves towards outlining and understanding the range of forms that are associated with those functions, with a view to providing insights into formulaicity in academic discourse that may have been missed in the prevalent corpus-based studies. As part of an on-going project that examines formulaic language (FL) using the system of metafunction (ideational, textual, interpersonal) in Systemic Functional Linguistics (SFL), the present study focuses on the interpersonal metafunction (i.e., the use of language to evaluate and take a stance on the proposition projected, and/or to build up a relation between the text-producer and the text-receiver).

Much attention in previous research on formulaicity in academic writing has been given to novice writers, in particular non-native novice writers (e.g., Chen & Baker, 2010; Ebeling & Hasselgård, 2015; Hyland, 2008; Salazar, 2014; Ädel & Erman, 2012). One of the main findings in this regard is that non-native novice writers tend to restrict themselves to a small range of FSs that are overused (see also Wray, 1999; Wang, 2016, pp. 4—6 for an overview). Native (or L1) novice writers, if involved in these studies, appear to be sidelined, serving mainly as the comparative basis against which non-native data are evaluated. Part of the reason for this lack of focus may be that native speakers' use and processing of FL in general is considered to be fairly well understood. To put it in a nutshell, FSs are said to constitute the bulk of native speakers' mental lexicon — that is, they are stored and retrieved whole from memory rather than generated anew on each occasion when they are needed, and therefore function as processing short-cuts (Wray & Perkins, 2000). While this may well be true of everyday communication and general language use, Pérez-Llantada (2014) argues that formulaicity in academic writing may not be an inherent (i.e., language universal) skill; rather it is likely to be associated with expert (native or not) academic writing production through formal instruction as well as extensive academic reading and writing practices. Therefore, more needs to be known about the development of formulaicity in academic discourse from the perspective of native novice writers. This paper is an attempt to fill this gap. More specifically, it attempts to answer the following research questions:

  i) What interpersonal functions are more frequently used with FSs in L1 novice or expert writing?
 ii) What are the main similarities and differences between L1 novice and expert writing in terms of the range of forms associated with the functions?

The remainder of this paper is organised as follows. Section 2 introduces the concept of FS and justifies the need for more manual analyses in the field. Section 3 presents the data used for the study and an annotation scheme for the analysis of interpersonal functions based on SFL as well as the criteria used to identify FSs. Section 4 reports on and discusses the results, and finally the paper ends with a summary of the main findings and their implications in Section 5.

2. Formulaic sequences

The term 'formulaic sequence' is formally defined by Wray (2002, p. 9) as 'a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar'. Although not explicitly specified in the definition, there are a few points that are essential for the understanding of formulaicity. First of all, FL is a complex construct and there are many types of FSs (Biber, 2009). Generally speaking, what makes a word sequence 'appear to be prefabricated' can be either its high frequency of occurrence in a given situation, or the internal fixedness of the form, or sometimes both. This is why FS is used in the literature as an umbrella term to mean anything from idioms (e.g., *in a nutshell*), phrases (e.g., *by and large*), collocations (e.g., *deeply committed*, *highly recommended*), to clusters or multi-word units/ex-pressions (e.g., *at the end of*, *as can be seen*), which may vary enormously in their idiomaticity, invariability, and structural completeness.

Secondly, FSs are 'retrieved whole from memory at the time of use' to meet different needs, which in turn are closely related to communities of practice. In other words, FSs develop to serve important communicative needs of a given discourse community. The appropriate choice of an expression among a variety of alternatives marks the speaker/writer as a member of that community. The distribution of FSs can therefore help characterise different discourses. Many studies have attested to considerable variation across genres and registers according to the extent to which formulaicity is applied (e.g., Biber, 2006; Biber et al., 1999; Wang, 2017a, 2017b).

Most FL studies take a frequency-based approach; that is, FSs are identified entirely on the basis of the recurrence of uninterrupted linguistic forms, whether or not they make up a complete structural unit or have a cohesive meaning or function (e.g., *as a result of*, *due to the fact that*, *is always*, *that can be*, *in a slightly*, *although it is*). While the results of such an approach are valuable in revealing, for instance, variation across discourses and interactions, they are of limited use to language learners and novice writers, for whom the key information about FSs is rarely which sequences are the most frequent per se. According to Durrant and Mathews-Aydınlı (2011), what novice writers need to know are rather what functions they are likely to employ in a given situation and what forms most appropriately fulfil these functions, as well as what restrictions that are placed on the forms in specific contexts. This is an aspect to which the present study aims to contribute.

From a methodological point of view, while the frequency-based approach has the advantage of being straightforward and consistent, and can be scaled up to very large datasets, its inherent limitations have also been increasingly recognised. First of