Formulaicity across Academic Disciplines: A Function-to-Form Approach

Ying Wang (Uppsala University, Sweden)

1. Introduction

Research in corpus linguistics has shown that people belonging to a particular speech community have preferred ways of saying things, which are generally reflected in the use of formulaic sequences (FS), i.e. certain words (e.g. *at the same time*, *on the other hand*, *for example*) have "an especially strong relationship with each other in creating their meaning" (Wray 2008: 9) and play an important role in differentiating socially-situated practices (Biber et al. 2004; Hyland 2012). The predominant trend in this research area is to take what Durrant and Mathew-Aydınlı (2010) call a 'form-first' approach (e.g. lexical bundle, n-gram), relying on the computer to identify frequent recurrent forms in a given corpus, at the expense of disregarding their structural and semantic unity (e.g. *at the same, this paper we*) and multifunctionality, and overlooking discontinuous units (e.g. *not only….but also*), to mention a few of its limitations. The present project employs a function-to-form approach in the hope of providing a pedagogically-oriented understanding of formulaic language in academic discourse. Halliday's (2014) functional model of language, which views language as a set of systemic choices with underlying communication functions (Gledhhill 2011), is useful in understanding formulaic language, which is necessary for functional language use. Drawing on Halliday's framework of functions, this paper reports on a pilot study with the following main aims: 1) to develop an annotation scheme for a functional analysis of formulaic language in academic discourse, which can be modified and applied to other text types in future studies; 2) to shed light on possible disciplinary variation in the distribution and linguistic realisations of various functions; 3) to compare the results achieved by the traditional 'form-first' approach (i.e. the automatic extraction of FSs) and those by the function-first approach. The ultimate aim in this regard is to provide suggestions as to how to combine the advantages of both approaches in order to provide a full picture of formulaicity in language use.

2. Material and methods

In this pilot study, twelve texts (with a total of 19,592 words) were drawn from the British Academic Written English Corpus (BAWE), representing two broad disciplinary groupings: Art & Humanities (AH) and Physical Sciences (PS). All the texts are of the same genre (essay) and grade (D), and were written by English-speaking students in their last year of undergraduate studies or on master's courses (years 3 and 4). FS in this study is an umbrella term covering a number of sub-categories including set phrases that are semantic opaque or grammatically irregular in varying degrees (e.g. *in so far as, give way to, a great deal of*), underlying frames with one or more gaps

(e.g. *as X as Y, if X then Y*), word strings that can be replaced by a single word (e.g. *fail to – not, be able to – can, make a decision – decide*), and formulas that are not peculiar in terms of their internal semantics/syntax, but are genre specific in the sense that they are used to realise functions in a particular type of situation (e.g. *this means that, it is important to*). The selected texts were manually examined to identify word sequences that satisfy at least one of the criteria mentioned above. What is more, the sequences should form a complete semantic and structural unit, whether or not the main elements are contiguous.

The identified FSs fulfil a range of functions, which fall into three broad categories within Halliday's framework: i) ideational (or experiential) metafunction, including functions such as referring to previous research (e.g. *according to*), describing attributes (e.g. *the length of, be responsible for*), describing research procedures (e.g. *tidy this up, work our way through*), and manner (e.g. *as a means of, by doing so, in detail, in a straightforward manner*); ii) textual metafunction, including structuring signals (e.g. *see for example, we can show that, as follows, in conclusion*) and cohesive devices (e.g. *as a result, in contrast, on the grounds that*); iii) interpersonal metafunction, including evaluation (e.g. *it is interesting that*, *play a key role in*) and hedging devices (e.g. *seem to, a certain degree of*). In the present study, the UAM corpus tool (O'Donnell 2013) was employed for the annotation of the texts. In addition to a functional-analysis scheme, the annotation also contains information including the criterion each sequence satisfies in order to qualify as a FS and its structural make-up for the subsequent analysis.

Apart from the manual identification, the conventional form-first approach (i.e. automatic identification) was taken, using IDIOM Search (Colson 2016a), which is an online tool for the extraction of multi-word phrases, ranging from bigrams to sevengrams (see Colson 2016b for the algorithm of and improvements made by this tool in corpus-based computational phraseology). This tool is useful in capturing continuous word sequences that are not necessarily irregular or semantically and/or structurally complete, but nonetheless occur frequently enough and are often associated with a particular function in a given context.

3. Initial results

Altogether, 5480 FSs were identified in the corpus; see Table 1 for the distribution of the FSs in the two sub-corpora representing the two broad disciplinary groupings. Overall, there was an overuse of FSs in the AH sub-corpus relative to the PS sub-corpus ($p<0.0001$).

Table 1. Corpus size and identified FSs

Sub-corpus	No. of words	No. of FSs
AH	7656	2288 (30%)
PS	11936	3192 (27%)
Total	19592	5480

Log-likelihood test: LL=16.34, $p<0.0001$

Out of the three main functional categories, the ideational metafunction was in clear dominance, accounting for 75% of all FSs, whereas textual functions made up 14% and interpersonal 11%. Within the ideational category, the function of manner stood out (11%), followed by the functions of describing attributes (9%) and argumentation (8%). The textual category was dominated by structuring signals (location, sequencing, introductory remarks, presenting results), which accounted for 41% of all textual functions, followed by cohesive devices (23%). Out of the interpersonal category, 33% were of the evaluation function, and the other two functions that may be interesting to look further into were obligation (12%, e.g. *it is necessary to, ought to, have to*) and personal opinion (7%, e.g. *in my view, I think*).

Table 2 presents the distribution of the three categories in the two sub-corpora. Again, significant difference was found between the two broad disciplinary groupings ($p<0.001$). Most noticeably, the text-oriented functions seemed to be employed to a particularly large extent by the students of PS. A closer look at the distribution of various functions of this category revealed that the main difference between the two sub-corpora rests with the functions related to cohesive devices.

Table 2. Distribution of main functional categories in the two sub-corpora

Sub-corpus	Ideational	Textual	Interpersonal	Total
AH	1807 (79%)	208 (9%)	273 (12%)	2288
PS	2313 (72%)	568 (18%)	311 (10%)	3192
Total	4120	776	584	5480

df=2, X^2=84.808, $p<0.001$

Comparing the figure resulted from manual identification with that of IDIOM Search, the latter retrieved fewer FSs (3888 units). Among those manually identified FSs, 30% are identical with those automatically retrieved, 29% are partly compatible and 41% failed to be captured by IDIOM Search, indicating a need to combine the two approaches in the study of formulaicity.

4. Concluding remarks

In conclusion, the initial results are significant enough to warrant further investigation. Among others, a qualitative analysis will be carried out to look at the few functions that stood out in the overall comparison above and to find out the particular discipline(s) where they tended to be employed. The results will be compared with those of previous studies (e.g. Hyland 2008) that employ the 'form-first' approach. Given the small size of the corpus, individual differences will also be brought up in the discussion. This pilot study will be expanded later on by including more texts from BAWE to provide more illuminating insights into disciplinary variation. In addition, I will include a selection of published research articles as well as essays written by L2 students, who are of the same academic level as those L1 students in the pilot study. Through a comparative analysis of student essays and published research articles, the study will shed light on the degree of formulaicity in

producing academic discourse across disciplines, and bring out similarities and/or differences between novice and expert writing, with pedagogical implications for the training of novice writers in scientific fields.

References

Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. *If you look at…*: lexical bundles in university teaching and textbooks. *Applied Linguistics* 25 (3): 371-405.

Bloor, Thomas and Bloor, Meriel. 2004. *The Functional Analysis of English* (2nd edn). London: Arnold.

Colson, Jean-Pierre. 2016a. IDIOM Search. http://idiomsearch.lsti.ucl.ac.be/index.html.

Colson, Jean-Pierre. 2016b. Set Phrases around GLOBALIZATION: An experiment in corpus-based computational phraseology. In F. Alonso Almeida, I. Ortega Barrera, E. Quintana Toledo & M.E. Sánchez Cuervo (eds.), *Input a Word, Analyze the World. Selected Approaches to Corpus Linguistics*, pp. 141–152. Newcastle: Cambridge Scholars Publishing.

Durrant, Philip and Mathew-Aydınlı, Julie. 2010. A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes* 30: 58–72.

Halliday, M.A.K. (revised by Christian M.I.M. Matthiessen). 2014. *Halliday's Introduction to Functional Grammar* (4th edn). Oxen: Routledge.

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

Hyland, Ken. 2012. *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: Cambridge University Press.

O'Donnell, Mick. 2013. UAM Corpus Tool, version 3.0.

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