A function-first approach to identifying formulaic language in academic writing

Abstract
There is currently much interest in creating pedagogically-oriented descriptions of formulaic language. Research in this area has typically taken what we call a 'form-first' approach, in which formulas are identified as the most frequent recurrent forms in a relevant corpus. While this research continues to yield valuable results, the present paper argues that much can also be gained by taking a 'function-first' approach, in which a corpus is first annotated for communicative functions and formulas are then identified as the recurrent patterns associated with each function. We demonstrate this approach through a comparative analysis of introductions to student essays and research articles. Focusing on one particularly common communicative function, the analysis demonstrates that (1) this function is more common in student essays than in articles; (2) both the choice to use the function and the choice of linguistic forms that realize the function vary across subject areas in research articles, but not in student essays; (3) research articles tend to be more formulaic in expressing the function than student essays; (4) some parts of the forms used are highly formulaic, while others are more open. The key formulas are described and suggestions made regarding their pedagogical presentation.

Keywords
formulaic language, corpus, genre analysis, academic writing, student essays, English for specific academic purposes

Introduction

Formulaic language and academic writing

Recent years have seen much interest in the phenomenon of formulaic language (e.g., Schmitt, 2004; Wray, 2002). For researchers and teachers of English for specific purposes, perhaps the most pertinent theme in this research is the claim that formulas can facilitate idiomatic production and so mark a speaker or writer as an 'insider' in a given discourse community (Wray, 2002, pp. 88-90). Formulas develop within communities, it is argued, as recurrent responses to recurrent communicative situations. For a qualified community member, such formulas feel like the 'natural' thing to say in a given situation (Pawley & Syder, 1983). Since a community's choice of one particular expression as the standard is inevitably arbitrary, individuals who do not have sufficient 'insider' experience, may not hit on the most appropriate expression, and their production may therefore seem 'not quite right' to insiders (Kjellmer, 1990). This has led a number of writers to suggest that mastering the appropriate use of formulas is an essential part of achieving idiomatic production (see Prodromou, 2008: Chapter 3 for a recent review).

This poses researchers with the question of what formulas learners need to learn to take the most effective 'shortcut' into the discourse community. A number of recent studies have addressed this issue with reference to academic language (Biber, 2009; Hyland, 2008; Simpson-Vlach & Ellis, 2010) and thanks to this work we are starting to build up a clearer picture of academic formulaicity. It is widely acknowledged, however, that the methods currently used for identifying formulaic language are not entirely satisfactory (see Wray, 2008: Chapter 8 for a recent review). With this in mind, Biber (2009) has recently stressed that researchers must embrace a range of different methodological approaches.

In the spirit of Biber's suggestion, the current paper describes an approach to the study of formulas which we believe is capable of providing important information which current methods do not provide. After outlining our conception of formulaic language, we explain why we believe this approach is necessary, and give an example of the approach in use to describe formulaicity in academic writing. We hope to demonstrate that this approach can provide pedagogically and linguistically useful information and therefore constitute a useful addition to our range of methods for studying formulaic language.

A pedagogical definition of formulaic language

Formulaic language has – notoriously – been defined in many different ways (Wray, 2002, p. 8). Three main orientations can be identified in the literature:

- 'phraseological' approaches (e.g., Cowie, 1998) focus on the non-compositionality of certain expressions, defining formulaicity in terms of either the degree to which the meaning of a word combination is predictable from the meaning of its parts or the degree to which words with similar meanings can be substituted into the phrase. Non-compositional phrases include idioms (e.g., *kick the bucket*, *spill the beans*) and certain collocations (e.g., *curry favour*, *French window*). The 'formal idioms' (Fillmore, Kay, & O'Connor, 1988) of construction grammar (e.g. *what's* NP *doing* Y; *the* ADJ-*er* the ADJ-er) can also be included in this category as items which cannot be easily understood and/or produced without specific learning.
- 'frequency-based' approaches (e.g., Biber, 2009; Hoey, 2005; Sinclair, 2004; Stubbs, 1995) focus on the tendency for certain linguistic combinations to appear with high frequency in text, defining formulas as strings of linguistic items (including words, parts of speech, and semantic fields), which have a statistical tendency to co-occur in

corpora. Examples include high frequency collocations (e.g., *hard work; as shown in Table* N); colligations (e.g., preposition + *the naked eye;* complement function + *consequence*); semantic preferences (e.g., 'words related to *express*' + *true feelings*; 'words related to *logic*' + *consequences*); and semantic prosody (e.g., negative concept + *happen;* positive concept + *provide*);

- 'psychological' approaches (e.g., Hoey, 2005; Wray, 2002) focus on the efficient mental processing and storage of language, defining formulas as strings of linguistic items which speakers remember and process as wholes, rather than constructing them 'online' with each use.

The differences between these orientations should not be overstated. Non-compositionality and high-frequency of occurrence can both be cited as evidence for holistic mental storage, and non-substitutability of parts can be evidenced in terms of co-occurrence frequencies in a corpus. The three approaches therefore overlap. Moreover, common to all is the idea that formulas are linguistic strings which, though they have the potential to be analysed into multiple components, are – for one reason or another – better left unanalysed. In pedagogical terms, this idea is perhaps best translated into the point that some combinations of linguistic items are best *learned* as wholes. This recalls Palmer's original definition of collocations as:

> "successions of words…that (for various, different and overlapping reasons) …must or should be learnt, or is best, or most conveniently learnt as an integral whole or independent entity, rather than by the process of placing together their component parts" (Palmer, 1933, p. 4).

While we accept the spirit of Palmer's definition, some elaboration is required if it is to be of practical use. First, we need to specify the "various, different and overlapping reasons" why certain strings are best learned as wholes (and so qualify as formulas). Three main reasons can be cited[1], roughly corresponding to the three orientations to formulaic language described above:

- *comprehending/producing non-compositional sequences:* idioms (e.g., *kick the bucket*), non-compositional collocations (e.g., *curry favour*), and idiosyncratic grammatical forms (e.g., *the* ADJ-*er* the ADJ-er) cannot be easily understood (and so, *a fortiori*, produced) without specific learning;
- *producing arbitrarily preferred sequences*: some semantically transparent sequences require special learning because they are arbitrarily preferred to possible synonyms. These include situational formulas, in which particular phrases are linked to particular contexts (e.g. *long live the king; as shown in Table* X), and collocations, in which particular collocates are linked to particular nodes (e.g. *commit a crime*; *answer the phone*);
- *increasing fluency:* it has long been hypothesised, and evidence is now starting to accumulate (e.g., Arnon & Snider, 2010; Ellis, Simpson-Vlach, & Maynard, 2008) that high-frequency combinations are stored as units in the mind, and that drawing on such units can facilitate fluent language processing.

A second elaboration required to Palmer's definition is that – as the examples given above make clear – formulas are not only successions of words, but may also include other

[1] A fourth reason sometimes proposed (e.g., Ellis, 2003) suggests that effective second language acquisition could be based on the stockpiling, and gradual 'decomposition' of memorized word sequences. However, little reliable evidence currently exists for such an acquisitional route in adult learners.

linguistic entities, such as parts of speech or words from a particular semantic set. Such combinations include what Sinclair (2004) has called *colligations* and *semantic preferences*, the *patterns* of pattern grammar (Hunston & Francis, 2000), and the *formal idioms* of construction grammar (Fillmore, et al., 1988).

Taking these points onboard, we propose revising Palmer's formulation to define formulas pedagogically as:

> successions of linguistic entities that are best learned as integral wholes or independent entities, rather than by the process of placing together their component parts, either because (a) they may not be understood or appropriately produced without specific knowledge, or (b) because they occur with sufficient frequency that their independent learning will facilitate fluency.

Using corpora to identify formulaic language

If formulas are defined in these terms, what role can corpora play in their identification? As we noted above, the frequency distributions and internal-fixedness of forms found in a corpus can provide indirect evidence for both non-compositionality and holistic storage. However, such evidence is indirect and best used in conjunction with other types of data (such as semantic analyses and psycholinguistic experiments). Where corpora truly come into their own is in determining which items are most often preferred in particular circumstances.

Such preferences can be divided into two types. In the first, a linguistic entity is conventionally associated with a particular *cotext*. This is seen in collocations and colligations, in which one item is commonly linked with another. Such associations have

been extensively studied. Since simple frequency is not always a good guide to how strongly forms are associated (pairings such as *arbiters-taste* and *bated-breath* are relatively rare, but strongly associated, whereas *and-in* and *of-a* are very common but not strongly associated) a number of statistical measures have been developed to quantify the strength of associations between words (Manning & Schütze, 1999).

In the second type, a formula may be associated with a particular communicative *context* – or, in the terms we will use in this paper, with a particular *function*. Corpus researchers typically identify such formulas by extracting linguistic strings which occur in a corpus above a certain frequency threshold and then identifying the functions of those high-frequency forms (e.g., Biber, 2009; Hyland, 2008; Simpson-Vlach & Ellis, 2010). This approach has much to commend it. It allows large corpora to be processed quickly and reliably and thus can uncover regularities not evident to the unaided observer. However, it also has important limitations. One is that the presumed correlation between frequency and formulaicity often breaks down. Wray (2002, p. 30) has pointed out that many apparent formulas have low frequencies of occurrence, even in very large corpora. Examples include situationally-specific phrases such as *long live the king*, idioms like *kick the bucket*, and even simple collocations, such as *criminal gang* or *personal apology*. Such phrases tend to be infrequent simply because the messages they express are relatively rare. This, Wray argues, suggests that frequency is not by itself an adequate guide to formulaicity. Rather, we need to know how regularly speakers make use of a particular form *when they need to express a particular message*. Since "[s]ome messages are much more common than others", Wray suggests, it is this "ratio of message-expression that will best help us to understand how some expressions of a given message are favoured over others" (2002, p. 31).

This is, of course, in direct parallel to the case mentioned above for collocations. Just as frequency is not by itself a sufficient guide to how strongly a word is associated with its cotext, it is also not a satisfactory guide to how well a phrase is associated with its communicative function. The corpus-based techniques typically used in this area are not able to provide this information because they take what we will call a 'form-first' approach to identification. That is, patterns are identified entirely on the basis of the recurrence of linguistic forms, with information about how the forms are used only being integrated at a later stage of analysis (e.g., Biber, 2009; Hyland, 2008; Simpson-Vlach & Ellis, 2010). To access the type of information Wray describes, however, the communicative context needs to be integrated into the analysis from the start. Specifically, the corpus needs to be tagged for communicative functions, and formula identification grounded in these functions. This is the approach which we adopt in the present paper.

A further limitation of 'form-first' approaches concerns the pedagogical information they provide. For language learners, the key information about formulas is rarely which word sequences are the most frequent *per se*. Rather, learners need to know what functions they are likely to need to express, what forms most appropriately fulfil those functions, and what variation those forms permit when they are fitted into specific contexts. Similar to Wray's proposal above, this suggests the need for an analysis which, rather than starting from linguistic forms, starts from semantic functions and works towards deriving the range of recurrent forms associated with those functions.

The idea of working from function to form is not an entirely new one, and can be seen as a special case of the broader debate between 'top-down' and 'bottom-up' approaches to corpus analysis (Biber, Connor, & Upton, 2007; Swales, 2002). Particularly relevant in this context

is Flowerdew's (1998) suggestion that applied corpus research could be made more pedagogically useful if integrated with textlinguistic analyses grounded in systemics, genre, or discourse analysis. Envisaging an approach similar to that used in the present paper, she proposes that corpora could be tagged according to the generic 'moves' they fulfil (e.g. background, scope, purpose section) and corpus searches then sorted according to the functional/discoursal roles of stretches of text. A small number of studies have subsequently adopted this approach (see Flowerdew, 2009 for a recent review). However, we are not aware of any systematic attempt to study formulaic language in this way.

Methodology

Corpora used

The main corpus used in this study was a subset of the British Academic Written English Corpus[2], a collection of assignments written by students at British universities. All assignments in the corpus had received at least an 'upper-second class' grade, and so can be deemed examples of 'successful' student writing. We restricted our investigation to the genre of 'essay', the most common text type in the corpus. In particular, we looked at essays produced by students on social science MA courses. Moreover, analysis was restricted to relatively substantial pieces of writing – essays shorter than 2,000 words were not included. 96 texts were found meeting these criteria. However, two texts (from the 'Drama and Theatre' discipline group) were eliminated from our investigation because we were not satisfied with their classification as 'social science' assignments.

[2] BAWE was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800). More details can be found at the corpus website: http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/.

The article corpus comprised 94 papers from recent issues of prominent journals. The spread of subject areas matched that in the student corpus and 3 different journals were used for each area. The spread of texts across disciplines for both text types and the titles of the journals used are shown in Table 1.

*****Table 1 about here*****

Identifying functions

The first stage in our analysis was to annotate the student essays for communicative functions. This annotation was based on Swales' notion of 'generic moves' (1990). A 'move' is defined as "a discoursal or rhetorical unit that performs a coherent communicative function". A move "is a functional, not a formal unit" and so is "flexible in terms of linguistic realization". It may be realized, at "one extreme, by a clause, and, at the other, by several sentences". (Swales, 2004, pp. 228-229).

As producing a complete move analysis of 94 essays was beyond the means of the present project, we decided to focus only on the introduction sections. A number of previous studies have analysed the generic moves found in the introductions to student texts (Bunton, 2002; Dudley-Evans, 1986; Henry & Roseberry, 1997; Hyland, 1990; Kusel, 1992). However, as none of these focuses on the specific text type studied here (MA social science essays), we decided to develop our own analysis of the moves found in our corpus, drawing where we could on concepts from existing frameworks.

Our analysis followed a multi-step iterative process. First, the first author read through the texts and attempted to apply move types from the literature. As, following Swales (2004, p.

22), we have defined moves as functional, rather than formal, units, moves were not taken necessarily to correspond to syntactic units or to respect paragraph boundaries. Single stretches of text were taken to be capable of performing more than one move, and so could receive multiple codings. Move types were adapted and new types added as required, with definitions being written for each type. This initial coding process involved several 'sweeps' of the texts, with the coding process starting again from the beginning each time the inventory of types had been substantially altered.

In the second stage, the second author was provided with the inventory of move types and their written definitions. She then read a random subset of texts (N=10) and attempted to code them for moves. The first and second authors then met to discuss their codings. Agreement was reached on any discrepancies and move definitions were adjusted where required. Both authors then read a second random subset of texts (N=20) and attempted to apply the adjusted definitions. They then met again to discuss their codings. Agreement was reached on any discrepancies and category definitions again adjusted where required. Finally, the first author re-read all texts using the revised definitions and adjusting previous codings where required. Any ambiguous cases were set aside and coded later in discussion with the second author.

Identifying forms

To identify the formulas associated with particular communicative functions, we first grouped together all instances of a particular move. Within moves, more specific functions were then identified and grouped together. For example, the 'justification' move, in which authors provide a rationale for their paper, may involve one or more of: *negative justification* (in which some problem with the current state of affairs is noted); *positive justification* (in

which current interest in the subject or positive outcomes from the study are emphasised); *justifying limitations* (in which shortcomings of the present work are discussed); *justifying approaches* (in which the particular approaches/frameworks used in the study are justified). We will adopt Swales's term of 'steps' (1990, p. 140ff) to label these more specific functions.

Finally, the common forms associated with each function were identified. This was achieved by identifying, in the first instance, relatively abstract forms which were shared across a large number of instances. These forms were then grouped together and more concrete (i.e., lexically fixed) repeated forms of each were identified.

Results

Move analysis

Three main generic functions were identified in the essay introductions: *Background information, justifying research*, and *essay focus*. Within each of these, several steps were identified. Figure 1 shows our final inventory of moves and steps.

*****Figure 1 about here*****

Formulaicity in Indicating Structure *steps*

Since space limitations do not allow discussion of all moves and steps, we will focus here only on *indicating structure* (IS) steps. These serve to describe the structure of the essay. For example:

> This essay will first analyse the general causes of the increasing practice of DEI. In the second part, emphasis will be laid on two of the most typical and popular forms of

> direct employee involvement: communication and teamworking. What are the motives of applying, and how far have they met the objectives are the key issues to be discussed. The last part will be a brief conclusion with some implications and suggestions.

IS steps were chosen for this analysis both because they were numerically prominent and because they appeared to show signs of formulaicity. Our analysis proceeds from the abstract to the concrete. We start by considering the extent to which writers use IS steps. We then look at a small number of relatively abstract constructions which feature in a large percentage of these steps. Finally, we look at how these constructions are instantiated more concretely at the lexical level.

Numerical spread of IS steps

We have argued that, both for the sake of quantifying how formulaic a given construction is and for the sake of determining whether a formula is pedagogically useful, we need first to know the frequency of the function it expresses. The first stage in our analysis is therefore to quantify the use of the IS step itself.

Overall, 71% of students essays and 53% of research articles used an IS step. The difference between text types was statistically significant (χ^2=6.54, p<.05). However, this overall difference disguises strong disciplinary preferences amongst articles. Not all disciplines have enough texts for useful generalisations to be drawn, but, as Table 2 shows, of those represented by eight texts or more, articles evidence a very broad range: from Sociology and Anthropology, where IS steps appear to be avoided, through Business and Politics, where they are optional, to Law, where they appear to be obligatory. Student essays do not show

such specialisation, none of these disciplines having fewer than 50% texts with IS steps. This suggests that, for these student writers, the decision to use an IS step is a relatively common one, regardless of subject area, whereas for article writers, the choice to use an IS step is strongly dependent on academic discipline.

***** Table 2 about here*****

Abstract constructions used in IS steps

We have defined formulaic language as combinations of linguistic items which learners ought to learn as wholes. While the most commonly studied types of formula have been combinations of words, it should not be forgotten that more abstract constructions can also fall within the scope of formulaic language, if they are employed with sufficient regularity. This section will consider such relatively abstract constructions, while the following section will consider concrete lexical formulas.

IS steps always include at least two pieces of information:

1. What will happen in the text
2. Where in the text this will happen

In some cases, these are represented by separate forms ('what' forms are shown in italics; 'where' forms are shown in bold):

In the final section *we discuss implications of this research for …*

In other cases, they overlap:

> ***Section I*** *addresses some stylized facts on the causes and resolution of…*

In our corpus, each function was realised by three main constructions:

'What' constructions:

1. *text + verb:* the article/essay or some section of it is construed as an agent which will carry out an action such as describing, analysing, discussing, etc.:

    - **Section III lays out** the model and basic assumptions.
    - **The paper begins with** a discussion of…

2. *passive:* the action to be carried out is expressed through a passive verb:

    - In the first three sections, the definition of value and labor and their interrelation **will be discussed** in detail.
    - Another important factor, that is, domestic reasons for making state actors prefer economic integration, **will be addressed** in the next section.

3. *pron + verb:* the author(s) (referred to as 'I' or 'we') are described as carrying out an action, e.g.:

- **We conclude with** a discussion (§5) of …
- In the second part of this essay **I analyze** recent literature on…

'Where' constructions:

1. *text*: the 'text' part of a 'text + verb' construction (see 1, above) can also mark position:

    - **Section III** lays out the model and basic assumptions.
    - **The next section** will focus on…

2. *adverbial*: position is marked with an adverbial phrase:

    - **In section two** Jafee and Russell (1976) and Stiglitz and Weiss (1981) are developed and discussed.
    - **Secondly**, a more detailed discussion on…

3. *verb*: the reporting verb used in the 'what' construction inherently signals position:

    - This article **begins** with the review of…
    - …**will be followed** by an analysis of…

It should be clear that these two sets of constructions are strongly interrelated. 'Where' can only be indicated with *text* if 'what' is indicated with *text* + *verb*. While *adverb* markers of place can occur with any of the 'what' structures, they are far less likely to occur in *text* +

verb constructions than in the other two (averaged across the two text types, 18% of *text + verb*, 86% of *passive*, and 80% of *pron + verb* constructions are paired with such an adverbial). *Verb* markers, meanwhile, tend to avoid *passive* constructions (13% of *text + verb,* 16% *of pron + verb*, and 2% of *passive*).

Table 3 shows the numbers and percentages of IS steps in each text type featuring each of these constructions. Individual IS steps are counted as any stretch of language which indicates what will happen in a particular part of the text. Thus, the example quoted above is divided into three steps:

1. This essay will first analyse the general causes of the increasing practice of DEI.
2. In the second part, emphasis will be laid on two of the most typical and popular forms of direct employee involvement: communication and teamworking. What are the motives of applying, and how far have they met the objectives are the key issues to be discussed.
3. The last part will be a brief conclusion with some implications and suggestions.

As (2) illustrates, a single step can include more than one 'what' structure type. Similarly, two 'where' types are sometimes used in combination within a single step. For these reasons, in Table 3, the total number of steps in which at least one of the constructions appears is lower than the sum of steps featuring each construction type, and the total number of 'what' types is lower than the total number of 'where' types.

****Table 3 about here ****

Both text types show a strong preference for these constructions: in student essays, the three 'what' constructions appear in 85% of IS steps and the three 'where' structures in 88%; for journal articles, the coverage of these forms is higher still, at 93% and 98% respectively. The higher level of use in journal articles is statistically significant, ('what': χ^2=8.27, p<.005; 'where': χ^2=16.35, p<.001).

At this level of relatively abstract constructions, therefore, both text types are highly formulaic in their choice of language, and articles appear to be more formulaic than student essays. There are also differences between text types in the specific forms chosen: while both types show an overall preference for *text* + *verb* forms, articles make significantly greater use of *pron* + *verb* structures than do essays (χ^2=38.31, p<.001), but significantly less use of passives (χ^2=20.26, p<.001). Students thus seem to pay far more heed to the traditional admonition against the use of personal pronouns than do expert writers.

As Table 4 shows, journal articles demonstrate clear disciplinary preferences for particular forms (only disciplines with more than 10 examples of IS steps are included in this analysis), i.e.:

- Economics and Law papers prefer *text* + *verb* forms for 'what' combined with *text* forms for 'where';
- Business and Politics papers prefer *pron* + *verb* forms for 'what' combined with *adverb* for 'where'.

Law and Business in particular are highly formulaic, with 85% and 74% respectively of IS steps employing a single construction for 'what'. Student essays (Table 5) show less variation between disciplines: all except Sociology favour the *text* + *verb* form for 'what'; all except

Economics and Law prefer *adverb* for where. Again, therefore, it seems that student essays are more uniform across subject areas than are articles.

****Table 4 about here ****

****Table 5 about here about here****

Lexically specified IS *forms*

The constructions looked at so far are highly formulaic in that they are strongly associated with the IS step. However, the primary focus of research in formulaic language has been on more lexically specific forms. We will now move our attention to such forms. Space limitations do not allow a full discussion of the lexis used in all of the forms described above. We will therefore focus in detail only on the most common form: 'text + verb'. Summary information will then be provided for the other forms.

Text + verb:

As Figure 2 illustrates, this construction has three main parts: the 'text' section, which acts as subject; a verb; and an object/verb complement.

***** Figure 2 about here*****

While this is the paradigm case, it is not invariable. Most prominently, adverbials are often included in the form:

Parts II and II **also** argue, **however**, that neither the validity of the idea…

Further variation occurs where a single sentence includes more than one statement of what will happen, as in Figure 3.

***** Figure 3 about here*****

Verbs can also be doubled up, as in Figure 4.

*****Figure 4 about here*****

A final notable variation on the form is that in some cases the verb is preceded by the auxiliary 'will':

The next section **will** focus on the effectiveness of secondary action.

This form is far more common in student essays than in journal articles. While precisely 50% of *text + verb* forms in the essay corpus included *will*, it was found in only one article.

The three main parts of the form show varying degrees of formulaicity. The object/complement slot is, as the above examples suggest, highly variable. This slot carries the main informational load of the statements and their content is dependent on the content of the paper in which they appear. The 'text' and 'verb' slots, on the other hand, do exhibit formulaicity. We will look at each in turn.

The most frequent instantiations of the 'text' slot for the two text types are shown in Tables 6 and 7. It should be noted that the formulas identified here do allow some internal variation. To represent this, words in italics indicate the exact words used in the phrase; words not in italics stand for a small set of possible words with a similar meaning; 'N' represents a numerical; 'X' represent a wide range of possible words; words in brackets are optional extensions; "/" indicates two alternative instantiations. For example,

'*the* X (section (*of* text))'

could be instantiated as: *the next section of the article*; *the first part*; *the second,* etc.

Formulaicity is quantified here in two different ways. First, it is shown as the percentage of total IS steps in the corpus which use this particular form. This figure indicates how regularly this meaning is expressed using this form. Second, it is shown as the percentage of cases of the *text + verb* construction which take this particular form. This figure indicates the degree of variability evidenced within the form itself.

****Table 6 about here ****

****Table 7 about here ****

In both text types, a relatively small number of forms accounts for a large percentage of IS steps. This is especially the case for journal articles: 43% of all IS steps use a text + verb form starting '*part/section* N', while 55% use a text + verb form starting with one of the four forms listed. In the student essays, 4 forms account for 38% of IS steps. Again, the greater formulaicity of journal articles is statistically significant ($\chi^2 = 12.70$, $p < .001$) In terms of the

variation allowed within the text slot itself, in both text types, this appears to be highly restricted. Of the two types, the journal articles in this sample are slightly more formulaic (though the difference is not statistically significant): a single form accounts for 74% of occurrences of the *text + verb* form, and the four forms together account for 96% of occurrences of this slot . The essays also demonstrate strong formulaicity within this slot, with 4 forms accounting for 92% of occurrences.

Turning to the 'verb' part of the construction, the use of specific verbs in this slot in essays and articles is summarised in tables 8 and 9 respectively (which list all verbs accounting for at least 3% of IS steps).

****Table 8 about here****

****Table 9 about here****

The very high frequency of the 'text + verb' form means that a relatively small number of verbs accounts for a reasonably high proportion of IS steps: in essays, 9, and in articles, 7 different verbs account for 25% of steps. However, unlike the 'text' part of the form, there is also a wide range of other instantiations used outside of these more frequent forms. In total, the 126 student uses of this form employed 52 different (lemmatised) verbs; the 174 forms in journal articles used 66 verbs.

<u>Other constructions:</u>

Turning now briefly to look at the other main constructions used in the IS step, 'pron + verb' forms contain two potentially formulaic elements. The 'pronoun' part is obviously highly fixed – always being realised by *we* or *I*. The 'verb' part shows a distribution similar to that

seen for the verb part of 'text + verb' constructions: though a small number of verbs are rather more frequent than others (in articles, *discuss* is used in 8% of cases; in essays *analyse* is used in 17%), there is also extensive variation (in articles, 55 different verb are used, 34 of them once only; in essays, 26 different verbs are used, 16 once only). In 'passive' forms, articles show little evidence of formulaicity in their choice of verb (unsurprisingly, given the relative infrequency of this form), with 8 different verbs found in the 10 different uses of the form; while essays show some preference for *discussed* (used in 15% of cases), again there is much variation (32 different verbs are used in total, 24 of them only once).

Instantiations of the 'text' form have been dealt with above in the context of the *text + verb* form, where we saw that these forms are highly predictable, though more so in articles than in essays. The vocabulary used in 'adverb' constructions is also rather formulaic. In articles, just seven forms accounted for 94% of uses: *in section/part* X (21%); *then* (21%)*; in the* X *section/part* (16%)*; first* (11%)*; next* (9%)*; after* (7%)*; finally* (7%). Essays are again a little less formulaic, with the top seven forms accounting for 83% of uses: *in the* X *section/part* (28%); *finally* (18%); *firstly/secondly/thirdly* (14%); *then* (11%); *next* (4%); *first/second/third* (4%); *after* (4%). Finally, the majority of verbs used to express place are also chosen from a very small pool. In articles, 3 verbs account for 80% of cases: *conclude* (37%); *begin* (31%); *follow* (11%); in essays, four verbs are prominent, together accounting for 85% of uses: *begin* (42%); *conclude* (19%); *start* (12%); *follow* (12%).

Summary and Conclusions

This paper has introduced a 'function-first' approach to studying formulaic language. We have attempted to show that this approach can both provide information which is of practical

use to teachers and give insights into the nature of academic discourse communities and student writers' place within them.

Our first main finding concerns the extent to which the IS step is used. This step was extremely common in successful student writing, being found in 71% of the essays. Interestingly, while the articles appear to show strong disciplinary preferences for the use or non-use of IS steps, student writing does not exhibit such specialisation. This suggests that students are being taught to use the IS step, and that this instruction does not distinguish between disciplines. This may be because, in writing classes full of students from various disciplines, it is easier to teach broadly generic introduction steps and structures. Alternatively, there may truly be different expectations for student writers than for expert ones. Apprentice writers of all disciplines might be expected to include more explicit, up-front signposting of the structure of their essays, while professional writers are expected to be capable of making their texts clear and understandable without spelling out the structure beforehand.

Three main forms were found to be used in expressing each of the 'what' and the 'where' aspects of IS steps. While students' use of these forms was not as predictable as that of article writers, they nevertheless accounted for the overwhelming majority of IS steps in these essays. As with the decision to use or not use an IS step, the preference for one or other form appears to be discipline specific amongst article writers, but not amongst students.

There is a long-running debate in the field of EAP regarding whether teaching should focus on general or discipline specific academic language (see Hyland, 2006 for a review), and these findings provide information which can contribute to this discussion. However, we

would caution that no direct line can be drawn between our descriptive findings and any normative prescriptions for teaching. The step from *is* to *ought* requires us, as always, to supply our own value judgements. If we believe that the existing practices of student writers (and, by implication, their teachers) are an adequate model for future generations, then these results might reinforce the case for generic writing. If, on the other hand, we believe that students should be encouraged to emulate more closely the practices of journal authors, then our findings can serve as a critique of existing practices. In short, while empirical findings can help to illuminate the question of English for general or specific academic purposes, it is important to remember that the dispute cannot be solved on empirical grounds alone.

Regarding lexical formulas, two main findings emerged. First, the various constructions studied differed in the extent to which they were lexically formulaic. While the *text* part of *text* + *verb* constructions and the *adverb* and *verb* forms used to indicate 'where' something will happen are each associated with a very limited range of lexical instantiations, the choice of verbs used to describe 'what' will happen was far more diverse. Formulaicity at the lexical level, therefore, appears to be a highly specialised phenomenon, with different aspects of meaning within a single generic step demonstrating different degree of fixedness. This suggests that a formulaic approach to teaching this step should not be primarily focused on lexical formulas. We would suggest instead an approach which takes as its basis the distinction between 'what' and 'where' and the three forms used to express each. Specific lexical forms could be usefully introduced for the 'where' parts of constructions, but tying the 'what' aspects too strongly to specific verbs may give learners an overly restrictive impression of how the constructions are used.

Second, where formulaicity does exist, it appears to be stronger amongst researchers than amongst students. Some models of learning have suggested that formulaicity is a feature of early language use, with rote-learned phrases being gradually broken down and replaced with more creative usages as expertise develops (Ellis, 2003). Our data appear to suggest the opposite tendency; with usage being more fixed amongst more advanced writers.

We hope to have shown that a function-first approach to formulaic language has the potential to offer useful insights into written discourse. Space limitations have limited the analysis to one particular step, and our speculations concerning the relative formulaicity and disciplinary specificity of student and professional academic writing must be restricted accordingly. We suggest, however, that further research along these lines has the potential both to offer a rich pedagogical description of the language of academic discourse and to improve our understanding of the nature of academic discourse communities.

References

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of memory and language, 62*, 67-82.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*, 275-311.

Biber, D., Connor, U., & Upton, T. A. (2007). Discourse on the Move: Using corpus analysis to describe discourse structure. Amsterdam: John Benjamins.

Bunton, D. (2002). Generic moves in Ph.D. thesis introductions. In J. Flowerdew (Ed.), *Academic discourse* (pp. 57-75). Harlow: Pearson Education.

Cowie, A. P. (1998). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 1-20). Oxford: Oxford University Press.

Dudley-Evans, T. (1986). Genre analysis: an investigation of the introduction and discussion sections of MSc dissertations. In M. Coulthard (Ed.), *Talking about text: studies presented to David Brazil on his retirement* (pp. 128-145). Birmingham: English language research, University of Birmingham.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: the emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 63-103). Oxford: Blackwell.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 41*, 375-396.

Fillmore, C., J., Kay, P., & O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language, 64*, 500-538.

Flowerdew, L. (1998). Corpus linguistic techniques applied to textlinguistics. *System, 26*, 541-552.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: a critical evaluation. *International Journal of Corpus Linguistics, 14*, 393-417.

Henry, A., & Roseberry, R. L. (1997). An investigation of the functions, strategies and linguistic features of the introductions and conclusions of essays. *System, 25*, 479-495.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Hyland, K. (1990). A genre description of the argumentative essay. *RELC Journal, 21*, 66-78.

Hyland, K. (2006). *English for academic purposes: an advanced resource book*. London: Routledge.

Hyland, K. (2008). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*, 41-62.

Kjellmer, G. (1990). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111-127). London: Longman.

Kusel, P. A. (1992). Rhetorical approaches to the study and composition of academic essays. *System, 20*, 457-469.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). New York: Longman.

Prodromou, L. (2008). *English as a Lingua Franca*. London: Continuum.

Schmitt, N. (2004). Formulaic sequences: acquisition, processing and use. In. Amsterdam: John Benjamins.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, amp058.

Sinclair, J. M. (2004). The search for units of meaning. In *Trust the text: language, corpus and discourse* (pp. 24-48). London: Routledge.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative methods. *Functions of language, 2*, 1-33.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. (2002). Integrated and fragmented worlds: EAP materials and corpus linguistics. In J. Flowerdew (Ed.), *Academic discourse* (pp. 150-164). Harlow: Pearson Education.

Swales, J. (2004). *Research Genres: exploration and applications*. Cambridge: Cambridge University Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: pushing the boundaries*. Oxford: Oxford University Press.

Table 1: Contents of the corpora

Discipline	Texts	Journals
Anthropology	8	Evolutionary Anthropology Journal of Human Evolution American Journal of Physical Anthropology
Business	25	Academy of Management Journal Academy of Management Review Marketing Science
Economics	3	Journal of Political Economy Journal of Economic Literature Quarterly Journal of Economics
HLTM[3]	3	Tourism Management Annals of Tourism Research Leisure Sciences
Law	18	Harvard Law Review Columbia Law Review Texas Law Review
Politics	24	American Political Science Journal American Political Science Review Political Analysis
Publishing	3	Learned Publishing Logos Publishing Research Quarterly
Sociology	10	American Sociological Review American Journal of Sociology British Journal of Sociology

[1]Hotel, Leisure and Tourism Management

Table 2*:* IS step use by discipline

Discipline	Number of texts	% of texts with IS step	
		Essays	Articles
Anthropology	8	50	25
Business	25	68	56
Law	18	56	100
Politics	24	83	46
Sociology	10	100	10

Table 3: Abstract constructions used in IS steps

		Number of occurrences		% of total IS steps*	
		Essays	Articles	Essays	Articles
What	Text + verb	108	136	49	57
	Passive	47	9	21	4
	Pron + verb	39	93	18	39
	Total	**188**	**224**	**85**	**93**
Where	Text	59	124	27	52
	Adverb	113	98	51	41
	Verb	26	35	12	15
	Total	**194**	**234**	**88**	**98**

*total IS steps in essays = 221, in articles = 240

Table 4: Abstract constructions in articles by discipline

		'What'			'Where'		
Discipline	**Total IS steps**	**% Steps with *text + verb***	**% Steps with *passive***	**% Steps with *pron + verb***	**% Steps with *Text***	**% Steps with *Adverb***	**% Steps with *Verb***
Business	53	25	8	74	23	64	15
Economics	10	60	0	40	60	40	0
Law	118	85	0	14	78	21	14
Politics	46	30	7	59	28	59	15

Table 5: Abstract constructions in essays by discipline

		'What'			'Where'		
Discipline	**Total IS steps**	**% Steps with *text + verb***	**% Steps with *passive***	**% Steps with *pron + verb***	**% Steps with *Text***	**% Steps with *Adverb***	**% Steps with *Verb***
Anthropology	11	55	30	20	0	27	18
Business	57	46	37	7	28	54	7
Economics	18	67	11	17	67	22	6
Law	30	70	7	7	60	33	0
Politics	64	50	14	25	16	50	23
Sociology	34	26	24	32	9	82	6

Table 6: Use of 'text' forms in articles

	Occurrences	**% of total IS steps***	**% of total text + verb forms****
part/section N (*of the article*)	102	43	74
the X *section/part* (*of the article*)	12	5	9
this part/section	12	5	9
the/this article	5	2	9
Total	**131**	**55**	**96**

(*total steps = 240)
(**total occurrences = 174, 37 of which are anaphoric, so excluded from these figures)

Table 7:Use of 'text' forms in essays

	Occurrences	**% of total IS steps***	**% of total text + verb forms****
the X (section (*of* text))	43	19	47
the/this essay	22	10	24
section N	14	6	15
the/this paper	5	2	5
Total	**84**	**38**	**92**

(*total IS steps = 221)
(**total occurrences = 126, 35 of which are anaphoric, so excluded from these figure

Table 8: Use of 'verb' forms in articles

	Occurrences	**% of total IS steps***	**% of total text + verb forms****
conclude	11	5	6
consider	9	4	5
describe	9	4	5
examine	9	4	5
show	8	3	5
explain	7	3	4

present	7	3	4
argue	6	3	3
begin	6	3	3
discuss	6	3	3
Total	**78**	**35**	**43**

(*total IS steps = 240)
(**total occurrences = 174)

Table 9: Use of 'verb' forms in essays

	Occurrences	**% of total IS steps***	**% of total text + verb forms****
examine	13	6	10
focus on	8	4	6
begin	7	3	6
conclude	6	3	5
look	6	3	5
Total	**40**	**19**	**32**

(*total IS steps = 221)
(**total occurrences = 126

Figure 1: Generic moves and steps with definitions

Move 1: *Background information*: provides information necessary to understand the paper;
- Step 1: *general topic background;*
- Step 2: *defining terms.*

Move 2: *Justifying research:* provides a rationale for the paper;
- Step 1: *negative justification*: indicates a current lack or undesirable state of affairs that has prompted the essay;
  - Step 1.1: *identifying a real-world problem*: notes a problem which needs to be addressed;
  - Step 1.2: *identifying a shortcoming of existing practices*: notes a limitation of current ways of doing something;
  - Step 1.3: *identifying a shortcoming of existing research*: notes a limitation of existing academic/theoretical work;
  - Step 1.4: *identifying controversy*: describes disagreement between authorities;
- Step 2: *positive justification*: indicates an intensity of current interest in or positive benefits to be gained from the discussion or the thing discussed;
  - Step 2.1: *claiming centrality of discussion*: emphasises the relevance/importance of the issues discussed, e.g. by citing large existing interest in the literature;
  - Step 2.2: *claiming centrality of object of discussion*: emphasises the relevance/importance of the object of the discussion, e.g. by citing widespread real-world interest in that object;
  - Step 2.3: *positive outcome from analysis*: argues that the paper itself will have a beneficial outcome;
  - Step 2.4: *positive outcome from object of discussion*: argues that the object under discussion has important benefits;
- Step 3: *justifying limitations*: explains why certain issues have not been addressed/data not included.
- Step 4: *justifying approaches:* explains why particular approaches/frameworks have been employed

Move 3: *Essay focus:* described the contents of the paper
- Step 1: *stating focus*: states in general what the essay will do; this may be 'stated' in the form of questions
- Step 2: *stating limitations*: notes and issues which have not been addressed/data not included
- Step 3: *indicating structure*: describes the structure of the essay; may incorporate statements of focus
- Step 4: *stating approaches*: notes any approaches/frameworks to be employed; may incorporate statement of focus
- Step 5: *stating the argument*: describes the conclusion or position that will be defended

Figure 2: text + verb forms

Text	**Verb**	**Object/Complement**
Section 3	evaluates	the consequences of strategic assortment reduction on consumer welfare.
The article	ends	with a discussion of the main findings and their implications for future research.

Figure 3: Multiple statements in text + verb forms

Text	Verb1	Object/Complement1	Conj	Verb2	Object/Complement2
Part III	outlines	the significance of intimate discrimination at a structural level	and	describes	how law and policy create hierarchies of subordination
The section that follows	defines	conditions of "risk" and "uncertainty"	and	derives	results about the evolving firm structure and allocation of ownership rights

Figure 4: Multiple verbs in text + verb forms

Text	Verb1	Conj	Verb2	Object/Complement
Section 4	summarizes	and	concludes	with managerial implications.
Section 4	presents	and	discusses	the results of service quality assessment for seven representative tourist farms.