Formulaic language in English for Academic Purposes

Philip Durrant

NB. This is the author's pre-print version of a chapter which appears in:

Siyanova-Chanturia, A and A. Pelicer-Sanchez (Eds). (2019) *Understanding formulaic language: A second language acquisition perspective*. London: Routledge

Abstract

Formulaic language has been a central concern in recent work on English for Academic Purposes. This chapter reviews the main motivations for EAP research in this area before critically discussing four issues which have been prominent in the literature: how analysis of formulaic language can help us understand the nature of academic language; how formulaic language relates to originality and criticality in academic work; how use of formulaic language influences the grades students receive; and how appropriate formulas for teaching can be identified.

Introduction

English for Academic Purposes (EAP) is concerned with understanding and teaching the use of English in university settings. Its primary motivation is the practical one of helping students who do not speak English as their first language succeed in English-medium university study. A number of reasons have been put forward for a focus on formulaic language (FL) in EAP. Some of these echo reasons given for a focus on FL in general; others are more specific to the EAP context.

The most fundamental reason for focusing on FL comes from the theoretical position that formulas are basic linguistic units. This view is influenced by Sinclair's (1991) *idiom principle* and by *pattern grammar* (Hunston & Francis, 2000) and *construction grammar* (Goldberg, 2006). It is captured well by O'Donnell et al.'s observation that "the phrase is the basic level of language representation where form and meaning meet with greatest reliability" (O'Donnell, Roemer, & Ellis, 2013, pp. 83-84). From the learner's perspective, this translates to the point that formulas are often irreducible syntactic/semantic units, which cannot be properly understood if they have not been specifically learned. It is this motivation which drives, for example, Hsu's (2014) list of semantically opaque formulas commonly found in academic readings and DeCarrico and Nattinger's (1988) similar list of items found in lectures.

A second motivation is found in the ideas that formulas have a special psycholinguistic status and that they play an important role in language acquisition. The former point suggests that phrases support fluent processing, making the learner's job correspondingly easier and freeing up resources for focusing on the content of the language which they are producing or trying to understand (e.g., Cortes, 2006; Simpson-Vlach & Ellis, 2010). The latter, based on models of first language learning suggested by, amongst others, Tomasello (2003) and Lieven (2008), and adapted for second language learning most prominently by Ellis (2008), sees the learning of formulas as a crucial stage in language learning, with command of the creative language system (grammar) being constructed through the gradual analysis of formulas (e.g., Lewis, 2000; Nattinger & DeCarrico, 1992).

A further rationale for incorporating formulas in EAP is the claim that there is a link between learners' use of FL and their perceived proficiency in the language (Chen & Baker, 2016;

Cortes, 2004; Staples, Egbert, Biber, & McClair, 2013). Accordingly, it has been suggested that appropriate use of formulas enables learners to achieve higher grades in assessments (AlHassan & Wood, 2015; Jones & Haywood, 2004). At its most simple, this has sometimes been equated to the claim that learners who use more formulas will be perceived as more proficient (Haswell, 1991), but recent work has shown that the relationship is considerably more complex than this (see discussion below).

The link between FL and proficiency is closely related to Pawley and Syder's (1983) widely-cited suggestion that appropriate use of formulas is important to achieving 'nativelike' production. The term 'nativelike' is probably misleading here as it suggests the issue is about sounding like someone born into a community, which is not quite right. The key point is rather that particular discourse communities (including those, like the various EAP communities, which people enter into later in life, and hence into which no one is really born) develop conventional ways of expressing frequently-occurring meanings, which often take the form of formulaic utterances. Because such formulas are familiar to members of the community, they are considered 'natural' (Hoey, 2005), and because formulas can be specific to particular communities, they can act as powerful signals of group membership (Wray, 2002). Since EAP is centrally concerned with helping learners meet the expectations of particular academic communities, these aspects of FL become central: by far the most commonly-cited reasons for exploiting FL in EAP are couched in terms of their importance in *meeting expectations* (e.g., Li & Schmitt, 2009; McKenny, 2006), achieving *appropriateness* (e.g., Byrd & Coxhead, 2010; Chon & Shin, 2013), *naturalness* (e.g., Ackermann & Chen, 2013; AlHassan & Wood, 2015) or *idiomaticity* (e.g., Ädel & Ermann, 2012), and in signalling *membership* of a group (e.g., Ädel & Ermann, 2012; Davis & Morley, 2015; Hyland, 2012). It has been argued, further, that meeting expectations and

signalling membership is a mark of "competent participation" (Hyland, 2008b, p. 5) in the community, implying that evaluations of the quality of learners' academic work are likely to be influenced by their effective use of formulas (AlHassan & Wood, 2015).

The association between formulas and discourse communities is related to the more general point that formulas tend to be highly context-dependent. Formulas are associated not only with particular communities, but with particular topics, particular registers, and particular genres. This feature has made FL an important analytical tool for researchers interested in language variation. EAP researchers have used formulas to study, for example, variation across academic disciplines (e.g., Durrant, 2017; Groom, 2005; Hyland, 2008b) and genres (e.g., Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Groom, 2005), and across writers at different stages of their academic careers (e.g., Ädel & Römer, 2012; Cortes, 2004; Hyland, 2008a) or from different linguistic and cultural backgrounds (e.g., Chen & Baker, 2010; McKenny & Bennett, 2011; Pérez-Llantada, 2014).

One reason that analysis of this sort can be enlightening is that formulas are strongly associated, not only with particular contexts, but with particular meanings. Thus, as well as showing *where* texts vary, they also give clues as to *how* they vary in terms of the types of meanings that are created. In Hyland's (2012, p. 153) words, they "reveal lexico-grammatical community-authorized ways of making-meanings", so offering insights into the ways those communities, and the texts they produce, work. From the language learners' perspective, formulas offer ready-made ways of expressing the most common meanings which are required of them, and may even serve as "triggers to help thinking" (Davis & Morley, 2015, p. 28). Formulas may, in other words, act as a kind of scaffolding which supports learners in constructing academic texts, offering a set of ready-made meanings and ways of expressing

those meanings. Like all forms of scaffolding, they provide support by restricting movement and biasing their users towards particular positions. While they serve an important pragmatic function, therefore, we also need to ask to what extent formulas constrain and limit originality in academic work and student learning.

Of the points just outlined, those related to the status of formulas in linguistic theory, in psycholinguistics, and in language acquisition are discussed in greater detail elsewhere in this volume (see Chapters 1, 2, 5, and 7). This chapter will rather focus on four issues which are more distinctive of EAP and which capture much of the important research done in this area. Specifically, it will address the following questions:

- What can corpus analysis of FL, and the ways in which it varies across texts, tell us about the language which EAP students need to learn?
- How does FL relate to originality and critical thought in academic work?
- Does using FL improve students' grades?
- How can we identify appropriate formulas to teach for academic purposes?

What can analysis of formulaic language in academic texts tell us about the language which EAP students need to learn?

Corpus analysis has shown that academic texts make extensive use of recurrent expressions. The majority of these are used to 'frame' novel stretches of language, showing the reader how a particular piece of information is to be interpreted, how it fits into the surrounding text, or what aspect of a phenomenon the writer or speaker wants to focus on (Biber et al., 2004). Biber et al. (2004) describe such sequences as *anchors* for new information. They note that

recurrent four-grams usually function as the openings of clauses or phrases, providing "structural 'frames', followed by a 'slot'" (2004, p. 399). This is seen in cases such as:

1. ***It is important to*** note that Derrida does not assert…[1]
2. I want to talk ***a little bit about*** process control from that point of view.

(Biber et al., 2004, pp. 391, 395)

Similarly, Durrant and Mathews-Aydınlı (2011) show how the common rhetorical function of signalling the organisation of an upcoming text comprises a subject+verb combination which is highly formulaic, followed by a much more creative object/complement, e.g.:

3. ***Section 3 evaluates*** the consequences of strategic assortment reductions…
4. ***Part III outlines*** the significance of intimate discrimination at a structural level

Cortes (2013) adds to this picture the important distinction between formulas acting as *triggers*, which signal the start of a rhetorical move (as in 5), and those acting as *complements*, which appear later within a move (as in 6). In both cases, however, the formula provides a routine frame for the content which follows:

5. However, ***little is known about the*** narrative skills of…
6. The objectives here are to determine if hedge funds exhibit persistence ***in the sense that*** some funds consistently have higher returns than others…

In examples 1-6, the balance between formulaic frame and novel content shows in microcosm the balance between the formulaic and novel which academic texts need to

achieve. On the one hand, to be of value, a text needs to present new information or a new perspective. On the other, to count as a valid contribution to an academic discipline, it needs to locate this information or perspective in an accepted framework of understanding. It is these shared frameworks which distinguish academic disciplines from pre-theoretical ways of understanding the world, and learning these frameworks is at the heart of learning how to do academic work. Much of the interest of FL in EAP rests on the fact that formulas reflect these frameworks.

This insight has been put to good effect in helping analysts understand the nature of academic texts and how they vary both from non-academic texts and from each other. This can be seen most clearly, perhaps, in comparisons of FL across academic disciplines. Durrant (2017), for example, shows how a quantitative analysis of overlaps in the use of 4-grams between texts can map similarities and differences in language use across students working in different subject areas. Analysis of the recurring word sequences (*lexical bundles*) distinctive of texts in humanities/social sciences on the one hand and science/technology on the other reflects the distinct approaches to knowledge of each, with the former focusing on, e.g.:

- abstract constructs (***the idea of a***; ***the nature of the***)
- autonomous agents (***the role of the; the power of the***)
- evaluations (***at the heart of; one of the main***)
- multiple contingent viewpoints (***it can be argued that***; ***can be seen as***)
- setting things in interpretive viewpoints (***in relation to the; in the context of***)

and the latter focusing on, e.g.:

- the physical world (***the presence of a; the shape of the***)
- passive instruments (***it can be used; will be used to***)
- quantification (***a large number of; the difference between the***)
- received knowledge (***it is thought that; was found to be***)
- cause and effect (***is due to the; the reason for this***)

Analysis of repeated forms has also revealed characteristic differences between different genres of academic discourse. Biber et al. (2004), for example, show how most recurrent sequences found in university textbooks are based around noun and prepositional phrases and have referential functions. Characteristic formulas of this type include ***the size of the***; ***the nature of the***; ***in the case of***; ***in the absence of***. This contrasts with non-academic conversation, where most recurrent forms are based around verb phrases or dependent clauses and have a more interpersonal function (which Biber et al. call *stance*). Characteristic examples include ***I don't want to***; ***you don't have to***; ***I was going to***. Interestingly, classroom teaching turns out to combine these styles, drawing on both the NP/PP-based referential formulas of academic writing and the VP/dependent clause-based stance formulas of conversation, resulting in a hybrid, and highly-formulaic, form of discourse.

How does formulaic language relate to originality and critical thought?

While contemporary writers have focused mostly on the benefits of FL, George Orwell's classic essay *Politics and the English Language* (1946) took a more critical view. The cognitive processing efficiencies which researchers have seen as so important to attaining fluency are cast by Orwell in a very different light: "The attraction of this way of writing is that it is easy…By using stale metaphors, similes and idioms, you save mental effort, at the cost of leaving your meaning vague, not only for yourself but for your reader" (p. 259).

Where applied linguists have seen formulas as providing a useful way of indicating 'in-group membership', Orwell sees numbing conformity:

> Orthodoxy, of whatever colour, seems to demand a lifeless, imitative style…When one watches some tired hack on the platform mechanically repeating the familiar phrases - *bestial atrocities, iron heel, bloodstained tyranny, free peoples of the world, stand shoulder to shoulder* - one often has a curious feeling that one is not watching a live human being but some kind of dummy […] A speaker who uses that kind of phraseology has gone some distance towards turning himself into a machine. The appropriate noises are coming out of his larynx, but his brain is not involved as it would be if he were choosing his words for himself. If the speech he is making is one that he is accustomed to make over and over again, he may be almost unconscious of what he is saying, as one is when he utters the responses in church. And this reduced state of consciousness, if not indispensable, is at any rate favourable to political conformity. (Orwell, 1946, p. 261)

Orwell's argument is an important reminder that both the cognitive and the social aspects of FL are not unambiguously beneficial. They provide a cognitive crutch by automating some aspects of language production, allowing us to focus more attention on other aspects of our message. However, the fact that the automated parts of our message go uninspected may lead to a lack of clarity or a failure to fully grasp our own assumptions. They provide a way of identifying with a group and access to ways of constructing knowledge which have evolved within our chosen disciplines, but identification with a group may also imply unthinking conformity and self-alienation.

This issue finds an echo in the concern sometimes expressed by teachers that students use formulaic phrases without properly understanding the ideas they express (Davis & Morley, 2015). It is also reflected more broadly in the debates between pragmatic approaches to EAP, which aim to teach students the conventions of their disciplines, more critical approaches which argue that EAP should help students actively challenge the practices of academic discourse, and academic literacies approaches, which address the issues of alienation that students may feel if forced to adopt a voice they feel is not their own (Hyland, 2006).

A very concrete example of the relationship between formulaicity and originality in student writing is seen in discussions about plagiarism. Plagiarism has been described as a form of *transgressive intertextuality* (Borg, 2009) - an illicit re-use of prior texts. One reason for students' problems with plagiarism is that most academic work actively requires the use of licit forms of intertextuality, and distinguishing the licit from the illicit is not always an easy task. Indeed, judgments about which individual cases constitute plagiarism can differ dramatically between professional academics (Borg, 2009; Davis & Morley, 2015; Pecorari & Shaw, 2012). Pecorari and Shaw (2012), for example, find disagreement as to what constitutes 'common knowledge' and thus not in need of citation. Importantly for our topic, they also disagreed as to what language is 'common property', and so not requiring a reference.

This raises the point that FL is itself a licit form of intertextuality. A key challenge for students is to learn to re-use prior language in ways that are expected of them while avoiding the types of re-use which will be seen by their teachers as transgressive. Moreover, they need to do this in a context where, as with plagiarism in general, what is considered legitimate re-

use can vary widely from discipline to discipline, from genre to genre, and even from section to section of a single paper (Flowerdew & Li, 2007; Pecorari & Shaw, 2012).

Flowerdew and Li (2007) and Davis and Morley (2015) report surveys of faculty attitudes towards language re-use which offer some guidance on what distinguishes re-usable formulas from plagiarism. From both studies, the central message is that faculty perceive a sharp distinction between the *form* and the *content* of texts, which drive their perceptions of plagiarism. Davis and Morley report that re-use was seen as particularly unacceptable where language expresses a particular opinion or judgment (e.g. *deliberately and decisively debunks this myth*) or is stylistically distinctive (e.g. *Dawkins is deaf to theology*).

This view of legitimate FL tallies well with the descriptive work of Biber et al. (2004), described in the second section in this chapter, which portrayed formulas as structural frames which opened up slots for novel content. To describe these as entirely lacking in content is perhaps inaccurate. As I have argued above, formulaic frames encode the ways of thinking and writing which construct academic ways of making sense of the world. The academics surveyed by Flowerdew and Li and Davis and Morley, however, clearly perceive a difference between this type of underlying disciplinary thought (what one of Davis and Morley's respondents called the disciplinary *lingua franca* (2015, p. 27)) and more specific content.

Does using formulaic language improve students' grades?

A number of researchers have looked into the relationship between use of FL and perceptions of proficiency[2]. Studies have been of two main types: those which look at learners' use of sequences which are thought to be formulaic in native-speaker English, and those which look at sequences which are frequent within a learner corpus. Within the first type, a number of

different methods have been used. AlHassan and Wood (2015) focus on 65 formulas which were taught to students as part of an intervention. In students' responses to an IELTS-style writing prompt at the end of the intervention, they found strong correlations (r=.71 and .60) between the number of target formulas used and the grades awarded to texts for two of the three raters who independently marked each text. Paquot (2017) takes a more wide-ranging approach, looking at the extent to which texts use adjective-noun, adverb-adjective and verb-direct object combinations from Ackermann and Chen's (2013) *Academic Collocations List* (see following section). In a corpus of research papers written by French EFL students for a university linguistics course, and which had been rated at one of three CEFR levels (B2, C1, and C2), she found no systematic differences across levels.

Whereas these studies looked at lists of selected sequences which had been identified in advance as formulaic, others have been more comprehensive, quantifying the formulaicity of a text by checking the frequency of all of its constituent word sequences against a reference corpus. Granger and Bestgen (2014), for example, automatically extracted all bigrams (two-word sequences) in a corpus of L2 English essays which had been graded against the Common European Framework of Reference (CEFR). The status of the extracted bigrams as formulaic was checked based on their frequencies in the British National Corpus (BNC). Formulaicity was quantified using two statistical measures of association:

- *T-score*: a measure of the degree of certainty with which we can claim that a word pair occurs more frequently than chance would predict. This measure emphasizes high-frequency collocations, e.g.: ***other hand; long time; little bit***
- *Mutual information (MI)*: a measure of the strength of the association between two word pairs. This measure emphasizes the exclusivity of the two words to each other.

> Words which are not often found without each other tend to have high mutual information scores, e.g.: ***pop music; juvenile delinquency; vicious circle***

Essays scoring in the 'advanced' C1/C2 range of the CEFR were found to have higher proportions of bigrams with high mutual information scores and lower proportions of bigrams with high t-scores than essays scoring in the 'intermediate' B1/B2 range. In a separate study which used a similar methodology with a different corpus of L2 timed writing, and quantifying frequencies using the Corpus of Contemporary American, the same authors (Bestgen & Granger, 2014) found that the mean mutual information score of bigrams in a text positively correlated with quality ratings, while the mean t-score showed no correlation. Similarly, Paquot (2017) looked at the mean mutual information score of adjective-noun, adverb-adjective and verb-direct object combinations in the corpus of CEFR-rated EFL research papers mentioned above, with mutual information scores for each item being based on their frequencies in a reference corpus of published L2 research. She also found significant increases in mean scores across proficiency levels.

A rather different approach to understanding the relationship between formulaicity and perceptions of text quality is taken by studies which have looked at word sequences which are repeated within the learner corpus itself (Appel & Wood, 2016; Biber & Gray, 2013; Chen & Baker, 2016; Staples et al., 2013; Vidakovic & Barker, 2010). The key conceptual difference between these and the studies reviewed above is that they focus on sequences which are common in student writing, rather than those which are common in the target academic language community. Though I have not been able to locate any studies of this sort looking directly at authentic academic language (i.e. language written for university-based

teaching or research), three have looked at writing from academic language proficiency exams.

Staples et al. (2013) studied lexical bundles in low, medium and high-rated TOEFL writing scripts. They found that the median frequency of bundles decreased as proficiency increased, implying that there was greater repetition of bundles in the lower-level texts. They also found that lower-rated texts tended to use more bundles lifted directly from the task prompts. Similarly, Appel and Wood (2016), comparing high- versus low-scoring texts written for the Canadian Academic English Language (CAEL) assessment, found that the low-scoring texts used a greater number of bundles and made greater use of longer bundles and bundles taken from the task prompts. While Staples et al. did not find any functional differences between the bundles used in texts at different levels, Appel and Wood found that low-scoring texts made greater use of stance bundles (especially opinion statements) and discourse-organizing bundles (used to reference materials from the readings which formed part of the text prompt), while high-scoring texts made greater use of referential bundles.

Biber and Gray (2013) looked at the use of both lexical bundles and collocations in written and spoken responses to the TOEFL iBT. As with the studies reviewed above, they found that many bundles were recycled from task prompts. Whereas the other studies found a decrease in quantity of bundles as grades increased, Biber and Gray found a slightly more complex pattern, with texts achieving mid-level grades having more bundles than both those achieving the lowest and those achieving the highest grades. They argue that this implies a developmental pattern whereby learners at the lowest levels have not yet acquired fixed expressions, intermediate levels overuse the expressions they have learnt, and learners at the highest levels move back towards greater creativity. A similar pattern is also suggested by

their analysis of the collocations associated with five high-frequency verbs (*get, give, have, make* and *take*), identifying which collocates are used within a three-word window of these nodes at least five times per 100,000 words. As with the analysis of lexical bundles, the mid-level texts were found to have more collocations than either the lowest or the highest-scoring texts. For lexical bundles, this regular pattern is slightly modulated by the discourse function of bundles, with discourse organizing bundles in speech and epistemic bundles in writing being most common at the highest levels. These final points stand in contrast to Appel and Wood (2016), who, found extensive stance and discourse bundles to be associated with lower-level texts.

These complex results suggest that there is much still to be understood about the relationship between use of formulas and the grades students are likely to receive. Moreover, it is noteworthy that, with the exception of Paquot (2017), none of the studies reviewed analysed authentic academic texts. The tasks from academic language proficiency tests studied by various researchers are clearly of some relevance to understanding academic writing, but given what we know about the contextual-specificity of FL use, it is also plausible that the way such language is evaluated in genuine academic contexts will differ sharply from these. Given its importance to understanding the roles of FL use in student writing (and, accordingly, their potential roles in teaching and testing), the relationship between use of formulas and the perceived quality of academic texts is one which requires substantial further research.

How can we identify appropriate formulas to teach?

Originating in the broader *English for Specific Purposes* (Hutchinson & Waters, 1987) movement, a founding principle of EAP is that language teaching can be made more effective

if we first identify what learners will be using the language for. This, the thinking goes, allows us to narrow the scope of teaching from the unattainable target of 'English' as a whole, to a more manageable sub-section of the language. By identifying the language which is most important for the learner's target areas of use, we can make their learning both more efficient and more motivating.

In line with this philosophy, several studies have asked how we can select academic formulas for teachers and learners to focus on. The most common approach has been to produce a list, modelled on the lists of academic words (especially Coxhead's (2000) influential *Academic Word List*) which are already widely-used in EAP programmes (Burkett, 2015). As with word lists, the rationale for formula lists is that, given the huge number of formulas in English, teachers need to prioritise particular items for their learners to focus on. Again as with word lists, the primary criterion for prioritising items is that of frequency, which, it is assumed, gives a good indication of the importance of particular formulas for learners.

Beyond simple counts of items, three more sophisticated types of frequency information are also commonly included:

- *Dispersion*: the extent to which items are used across a wide range of texts, rather than being concentrated in a few sources. In some studies (Ackermann & Chen, 2013; Eriksson, 2012), items are required to appear in more than a set number of distinct texts. In others, they are required to reach a particular frequency threshold across a range of distinct disciplinary areas (Ackermann & Chen, 2013; Durrant, 2009; Liu, 2012; Simpson-Vlach & Ellis, 2010). The latter is an attempt to ensure that formulas are relevant to a wide range of students.

- *Keyness*: the extent to which items are distinctive of the target text type. It is likely that many formulaic items which are frequent in academic texts are also frequent in the language as a whole. To identify items which are distinctively academic, some studies (Durrant, 2009; Simpson-Vlach & Ellis, 2010) incorporate information on the relative frequencies of items in an academic corpus versus a corpus of general English, and focus on those items where the gap between these is largest. A different use of keyness is made by Eriksson (2012), who focuses on items which are more common in published academic texts than in texts written by students. The logic here is that these are likely to be the items which students need but do not yet use.
- *Association*: measures of association usually focus either on how confident we can be that a combination of words occurs more frequently than chance would predict (see the description of *t-score* in the previous section) or on how strongly words in a formula are associated with each other (see the description of *mutual information* in the previous section). Many researchers have claimed that these provide a more valid way of identifying important collocations than frequency alone (e.g., Simpson-Vlach & Ellis, 2010)

While frequency data are key to formula list construction, they are not by themselves sufficient to generate a pedagogically-useful list. One issue is that items which are identified on frequency grounds alone are frequently not meaningful units (e.g. *as well but*). Relatedly, frequency-based lists return a number of items which are clearly overlapping variants of a central formula (e.g. *were statistically significant; statistically significant differences between; no statistically significant*). Further, some teachers and researchers wish to focus on particular subsets of items (e.g. items which are semantically opaque or items which do not translate directly in the L1) which cannot be identified on frequency grounds alone. For these

reasons, studies often supplement their frequency data with manual filtering of items. Ackermann and Chen (2013), for example, manually exclude from their list linguistically incomplete items, fixed combinations, adverbs of time/frequency, common transparent adjectives, concrete geographical references and combinations which are often hyphenated. Hsu (2014) uses a checklist to identify formulaic sequences (FSs) which are both meaningful and require holistic learning.

With extensive lists such manual analysis is obviously arduous. Simpson-Vlach and Ellis (2010) take the ambitious step of estimating qualitative judgements about the teach-worthiness of items using frequency data alone. After asking a panel of teachers to rate a subset of 108 candidate formulas for teaching value, they use a regression analysis to determine a 'formula teaching worth' formula, which they use to estimate how teachers would have rated the remainder of their items. While this is an intriguing approach, it is striking that Simpson-Vlach and Ellis do not report the goodness of fit between their model and teacher ratings. They also don't report how well the model generalized beyond the initial set of 108 items. The real effectiveness of their approach, and whether a frequency-based formula can effectively substitute for manual judgment, therefore remain unclear.

I noted above that the motivation for academic phrase lists is parallel to that for academic word lists. However, one important difference needs to be acknowledged. A key part of the justification for academic word lists is that the majority of the language we meet is made up of a relatively small number of distinct words (Durrant (2016) provides a discussion of the principles behind word lists). Coxhead (2000), for example, reports that the 570 words in her *Academic Word* List achieves 10% coverage of an academic corpus, while Durrant (2016) shows that the most frequent 587 lemmas in the Academic Vocabulary List (Gardner &

Davies, 2014) account for 25% of lexical words in a corpus of student writing. There is little research into whether similar patterns exist for FL, however. Intuitively, it seems likely that the pay-off will be substantially less: FSs are, by definition, more specific, and hence rarer, than words. As we saw above, they also vary strikingly between even closely-related contexts. We therefore cannot expect the same high levels of coverage as have been achieved by word lists.

The existing work on phrase lists shows that coverage can vary based on the type of formula studied and the type of academic writing studied. Durrant's (2009) list of 1,000 collocations appeared with a total frequency of between 17,677 (in Arts and Humanities disciplines) and 35,306 (in Social Science disciplines) occurrences per million words. Since each item comprises two words, and assuming each token in the corpus only appears in only one collocation (a mostly plausible assumption for two-word combinations), this translates into a coverage of between 3.5 and 7.1% of words. Hsu (2014) also reports a reasonable coverage of 2.08% for a list of 475 2-5-word sequences. It is important to note, however, that Durrant's list comprises largely combinations of lexical + function words (e.g. *associated with; as shown*), while Hsu's list includes combinations of function words (e.g. *along with; as to*) and a small number of colligations (e.g. *[auxiliary verb] + hardly*). In contrast, Ackerman and Chen's (2013) list, which includes only combinations of lexical words and which (as described above) excluded a number of particular types of combinations, achieves a coverage of only 1.4% with 2,468 lemmatised collocations. While there is great variation in these figures, none comes close to the high levels of coverage achieved by word lists, suggesting that listing may not be a particular effective pedagogical approach for formulas.

A further issue with formula lists is that, as discussed above, formulas tend to be highly contextually-specific. For researchers, this has been a boon as it has made formulas an excellent tool for identifying and characterising differences between text types. It is also the chief reason why a learner's use of the most apposite formula can be so impressive. The flip-side, however, is that a generic list for EAP students in general can be of only limited use. The formulas which would help a physicist write a research report will vary sharply from those which would help an historian write an essay.

Given these limitations of formula lists, it is worth asking whether there are other ways of helping learners decide what to focus on, preferably more closely tailored to their individual needs. Vincent (2013) offers one way forward here, setting out a methodology through which teachers can use corpus resources to identify FSs within specific texts that their learners are working with. His method involves identifying in a text candidate phrases which include particular high-frequency, closed-class words and then checking a reference corpus to determine their formulaic status. While Vincent's method is rather labour intensive, it is not hard to imagine systems which could make the process much easier and help learners to highlight formulas (of various types) in the texts they are reading/writing.

Conclusions and future directions

This chapter has explored a number of issues around the role of FL in EAP. I have suggested that formulas act as conventional frames in which academics set the novel content of their texts, that these frames carry important characteristic meanings which academics use to create knowledge in approved ways, and that they vary in interesting ways across academic communities and contexts. These features of formulas make them important, I have claimed, both for researchers interested in how academic texts work and for students who are learning

to create their own texts. We have seen that the conventions inherent in formulas are both useful in helping to construct meanings and potentially dangerous in that they embody conformity to a norm and in that it may be difficult for students to distinguish legitimate from illegitimate re-use of forms (otherwise known as *plagiarism*). While many researchers have claimed that use of appropriate formulas can make learners sound more proficient, we have seen that the evidence for precise relationship between formulaicity and perceptions of proficiency is currently ambiguous. I have also reviewed a number of attempts to create pedagogically-oriented lists of FL. While these offer an important resource, I have pointed out that they may not prove as useful as academic word lists have, and that alternative approaches to identifying language on which learners can focus may be needed.

On the basis of this review, a number of issues suggest themselves as important focuses for future work:

- The relationship between use of FL and perceptions of linguistic proficiency remains unclear. This is a crucial issue as it is this putative link which provides the ultimate rationale for students' learning FL. If formulas don't improve perceptions of learners' language, there seems little reason to study them. Moreover, a better understanding of the formula-proficiency link would help us enrich our understanding of the construct of academic language proficiency. This is crucial both for theoretically-oriented tasks such as understanding academic language development and for practically-oriented tasks such as establishing the validity of high-stakes tests of academic language proficiency, such as IELTS and TOEFL.
- Orwell's claim that using FL prevents us from thinking critically and originally is a striking one. Traditionally, applied linguists have stayed away from research on the

language-thought interface, and I am not aware of any systematic research on the implications of formulaicity for critical or original thought, though formulaic thinking is an issue touched on by Wray (2008) and some work has looked at the role of formulas in linguistic creativity more broadly (see Bell, 2012 for an interesting review). The recent resurgence in research perspectives on the relationship between thought and language (e.g., Pavlenko, 2014) may open up possibilities for work in this area.

- We do not yet have a satisfying means of identifying formulas for teaching. A 'list' approach has significant shortcomings, as discussed above. More 'responsive' methods which help students to identify formulas they encounter may offer a useful way forward here.

It can be hoped that future research will expand our understandings in these directions.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247.

Ädel, A., & Ermann, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approash. *English for Specific Purposes, 31*(2), 81-92.

Ädel, A., & Römer, U. (2012). Research on advanced student writing across disciplines and levels: Introducing the *Michigan Corpus of Upper-level Student Papers*. *International Journal of Corpus Linguistics, 17*(1), 3-34.

AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes, 17*, 51-62.

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: differences between high- and low-proficiency levels. *Language Assessment Quarterly, 13*(1), 55-71.

Bell, N. (2012). Formulaic Language, Creativity, and Language Play in a Second Language. *Annual Review of Applied Linguistics, 32*, 189-205.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26*, 28-41.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics, 25*(3), 371-405. doi:10.1093/applin/25.3.371

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT test: a lexico-grammatical analysis. *TOEFL iBT Research Report, 19*.

Borg, E. (2009). Local plagiarisms. *Assessment and Evaluation in Higher Education, 34*(4), 415-426.

Burkett, T. (2015). An investigation into the use of frequency vocabulary lists in university intensive English programs. *International Journal of Bilingual and Multilingual Teachers of English, 3*(2), 71-83.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*, 31-64.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language learning and technology, 14*(2), 30-49.

Chen, Y.-H., & Baker, P. (2016). Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Chon, Y. V., & Shin, D. (2013). A corpus-driven analysis of spoken and written academic collocations. *Multimedia-Assisted Language Learning, 16*(3), 11-38.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397-423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education, 17*, 391-406.

Cortes, V. (2013). *The purpose of this study is to:* Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12*(1), 33-43.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly, 34*(2), 213-238.

Davis, M., & Morley, J. (2015). Phrasal intertextuality: The reponses of academics from different discipliens to students' re-use of phrases. *Journal of Second Language Writing, 28*, 20-35.

DeCarrico, J. S., & Nattinger, J. R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes, 7*(2), 91-102.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *Journal of English for Specific Purposes, 28*(3), 157-179.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes, 43*(1), 49-61.

Durrant, P. (2017). Lexical Bundles and Disciplinary Variation in University Students' Writing: Mapping the Territories. *Applied Linguistics, 38*(2), 165-193.

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *Journal of English for Specific Purposes, 30*(1), 58-72.

Ellis, N. C. (2008). Usage-based and form-focused language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372-405). London: Routledge.

Eriksson, A. (2012). Pedagogical perspectives on bundles: Teaching bundles to doctoral students in biochemistry. In J. Thomas & A. Boulton (Eds.), *Input, process and product: Developments in teaching and language corpora* (pp. 195-211). Brno: Masaryk University Press.

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics, 28*(3), 440-465.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics, 52*(3), 229-252.

Groom, N. (2005). Pattern and meaning across genres and disciplines: An exploratory study. *English for Academic Purposes, 4*(3), 257-277.

Haswell, R. (1991). *Gaining ground in college writing: Tales of development and interpretation*. Dallas: Southern Methodist University Press.

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

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System, 47*, 146-161.

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

Hutchinson, T., & Waters, A. (1987). *English for specific purposes*. Cambridge: Cambridge University Press.

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

Hyland, K. (2008a). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41-62.

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21.

Hyland, K. (2012). Bundles in Academic Discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 269-300). Amsterdam: John Benjamins.

Lewis, M. (Ed.) (2000). *Teaching collocations: Further developments in the lexical approach*. Boston: Thomson.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of Second Language Writing, 18*, 85-102.

Lieven, E., V.M., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 168-196). London: Routledge.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31*(1), 25-35.

McKenny, J. A. (2006). *A corpus-based investigation of the phraseology in various genres of written English with applications to the teaching of English for academic purposes.* (PhD), University of Leeds.

McKenny, J. A., & Bennett, K. (2011). Polishing papers for publication: palimpsests or procrustean beds? In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New Trends in corpora and language learning* (pp. 247-262). London: Continuum.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

O'Donnell, M. B., Roemer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics, 18*(1), 83-108.

Orwell, G. (1946). Politics and the English Language. *Horizon, 13*(76), 252-265.

Paquot, M. (2017). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 1-25.

Pavlenko, A. (2014). *The bilingual mind and what it tells us about language and thought*. Cambridge: Cambridge University Press.

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.

Pecorari, D., & Shaw, P. (2012). Types of student intertextuality and faculty attitudes. *Journal of Second Language Writing, 21*, 149-164.

Pérez-Llantada. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent use. *Journal of English for Academic Purposes, 14*, 84-94.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics, 31*(4), 487-512. doi:10.1093/applin/amp058

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*, 214-225.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life Writing examinations. *University of Cambridge ESOL Examinations Research Notes, 7-14*(41).

Vincent, B. (2013). Investigating academic phraseology through combinations of very frequent words: A methodological exploration. *Journal of English for Academic Purposes, 12*(1), 44-57.

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

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

[1] In this chapter, I will use bold italics to identify stretches of text which are assumed to be formulaic. This will not always reflect emphases made in the original texts from which I am quoting.

[2] Since most assessment in universities involves written, rather than spoken work, the former has been the main focus of attention and the current review will not deal with studies of spoken language.