Chapter 5

Formulaic sequences with ideational functions in L1 student and expert academic writing in English

Ying Wang
Karlstad University

Corpus studies have revealed that formulaic sequences are prevalent in academic discourse in English. The predominant trend in this research area is to take a frequency-based approach (e.g., lexical bundles, *n*-grams), relying on the computer to retrieve continuous word sequences that occur frequently in a given corpus. Such an approach has helped bring to light a rich repertoire of FSs with textual or interpersonal functions (e.g., *on the other hand*, *it is possible to*) that characterises successful academic writing. However, the use of formulaic language that is central to the construction of disciplinary knowledge has received relatively little attention partly due to the limitations of the identification method. Through manual identification and annotation of FSs in context, the present study examines successful L1 student and expert writing. The results reveal that both are highly formulaic in quantitative terms, and ideational FSs account for approximately 70% of all FSs identified. However, each has its own distinct features in terms of the variety of FSs used. In general, the student corpus employs more everyday FSs which are often highly idiomatic, whereas the expert counterpart yields more FSs associated with research and reasoning processes. It is also argued that knowledge of conventional usage patterns for what seem to semantically transparent and syntactically flexible FSs in academic discourse is not necessarily an inherent part of native speakers' linguistic competence, but needs to be acquired incrementally through formal instruction and training by non-native and native students alike.

Ying Wang. 2021. Formulaic sequences with ideational functions in L1 student and expert academic writing in English. In Aleksandar Trklja & Łukasz Grabowski (eds.), *Formulaic language: Theories and methods*, 113–137. Berlin: Language Science Press. DOI: 10.5281/zenodo.4727669

1 Introduction

Formulaic sequence (FS) is defined by Wray (2002: 9) as "a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar". FS is used in the literature as an umbrella term to mean anything from idioms, phrases, collocations, to clusters or multi-word units/expressions. Generally speaking, what makes a word sequence appear to be prefabricated can be either its high frequency of occurrence in a given situation, or the internal fixedness of the form, or sometimes both (Siyanova-Chanturia 2013). Depending on the type of FS under investigation, different methodologies have been used in previous studies to identify target sequences. The predominant trend in formulaic language research so far is to take a frequency-based approach (e.g., lexical bundle, *n-gram*), relying on computational tools to automatically identify frequently occurring word sequences in large text corpora. While this approach has the advantage of being methodologically straightforward and efficient, its inherent limitations have also been increasingly recognised (see Ädel & Erman 2012; Wang 2018). Among other things, some highly salient FSs tied up with a particular communicative context are difficult capture due to their relatively low frequency of occurrence and/or internal variability. More importantly, from a pedagogical perspective, such an approach often results in a large number of incomplete structural or semantic units (e.g., *although it is, that can be*) that are of limited use to language learners and novice writers, for whom the key information about FSs is rarely which sequences are the most frequent per se, but what functions they fulfil and what forms they tend to employ as well as the degree of variation allowed in a given context (Durrant & Mathews-Aydınlı 2010). In short, as Biber (2009) suggests, there is still a need to embrace new and complementary methodological approaches. The present study is a step forward in that direction by incorporating a primarily manual approach in identifying word sequences, continuous or discontinuous, in an attempt to provide empirical evidence on what may have been missed in frequency-based studies and what those overlooked FSs can tell us about formulaicity in language use.

Over the past decade, corpus studies, often utilizing a comparative approach, have revealed that FSs are prevalent in academic discourse[1] and offer an important means of differentiating disciplinary practices and groups of writers – the

[1]In this paper, terms such as *academic discourse* and *academic writing* are used to mean *academic discourse/writing in English*, and the claims made about them may not apply to other languages.

appropriate choice of a FS among a range of alternative expressions marks the writer as a member of the discourse community (e.g., Biber et al. 2004; Cortes 2004; Hyland 2008b; 2012; Durrant 2017). To date, differences between non-native (L2) and native (L1) or expert production have received most attention, often with the aim of outlining the difficulties experienced by L2 writers (either students or novice academics) (e.g., Hyland 2008a; Chen & Baker 2010; Ädel & Erman 2012). L1 novice writers, as Hyland (2016) points out, have been largely marginalized in studies of academic writing. Indeed, in studies focusing on FSs, L1 student writing, if involved, often serves as the benchmark against which non-native data are evaluated, with the assumption that the use of FSs is part of native speakers' inheritance (Wray & Perkins 2000). While this is true for everyday language use, it has been increasingly realised that academic English is no one's first language and formulaicity in academic writing may not be an inherent skill but require prolonged formal education and training (Ferguson et al. 2011; Pérez-Llantada 2014). The present study addresses this somewhat neglected line of research by putting L1 students under the spotlight. Through comparing successful L1 student disciplinary writing with published expert writing, the study aims to shed some light on the development of formulaicity specific to academic discourse among native speakers.

The frequency-based approach has helped uncover a rich repertoire of lexical-grammatical resources available for writers to organise their texts (e.g., *on the other hand, in addition*), take a stance towards its content (e.g., *it is possible to*), and to engage with the readers (e.g., *note that*). While such FSs with textual or interpersonal functions have received considerable coverage in previous studies, those that are associated with the propositional content typical of a given discipline, including core disciplinary concepts (e.g., *positive rights, position vectors*), methodologies and research procedure (e.g., *scale up to, at low/high stresses*), norms for reasoning (e.g., *rule out, a plausible explanation for*), have been largely neglected. In the few studies which do involve what they call "research-oriented" expressions, only a handful of roughly defined sub-categories have emerged, e.g., location (e.g., *at the beginning of*), quantification (e.g., *a wide range of*), attribute (e.g., *the structure of the*), and procedure (e.g., *the use of the*) (Cortes 2004; Biber et al. 2004; Hyland 2008b). This imbalance in coverage may be partly due to the limitations of the identification approach. Textual and interpersonal FSs tend to be longer word combinations – textual FSs in particular are likely to be invariable word sequences (Wang 2019), which means they are more easily captured by automatic retrieval methods than FSs with ideational meanings which often involve two or more core lexical items with a great deal of formal variability.

Using a partly manual approach in the identification of FSs and a more comprehensive classification framework derived from Systemic Functional Linguistics (SFL) (cf. §2.3), the current study is part of an on-going project that sets out to investigate the use of FSs that distinguishes successful L1 student and expert academic writing, while at the same time exploring the potential and feasibility of the proposed methodology. The results of textual and interpersonal FSs can be found in Wang (2018) and Wang (2019), respectively. This paper focuses on ideational FSs, and by comparing the results with those of textual and interpersonal FSs, it will also provide an overall picture of the distribution of the three categories of FSs in L1 student and expert academic writing.

2 Data and procedure

2.1 Data

The present study used the same data as used in Wang (2018, 2019), involving two small corpora of approximately 100,000 words, representing successful L1 student and expert writing, respectively (see Table 5.1).

Table 5.1: Data used in the study

	No. of texts	No. of words
Student corpus	15	46,722
Expert corpus	11	52,626
Total	26	99,348

The student texts were randomly drawn from one subset of the BAWE corpus (Nesi & Gardner 2012), containing ”essays” with a ”distinction” grade, written by L1-English students in their final year of undergraduate studies. The texts are also evenly distributed across a number of disciplines so that they should provide a broadly representative sample of successful L1 student writing at the chosen level.

It is extremely difficult, if not impossible, to find a control corpus containing texts that are exactly equivalent to student writing (Callies 2015). In the present study, the keywords that occur in the titles of the student texts were used to search for published research articles in order to minimise the effect of topic on lexical features (Caines & Buttery 2017). In addition, all the articles were drawn from SCI indexed journals to ensure the quality of writing is reasonably high. In

terms of genre, while the published articles may be considered as representing a homogenous text type to a great extent, the "essay" genre in the BAWE corpus is by definition quite broad, where the students "are expected to develop ideas, make connections between arguments and evidence, and develop an individualized thesis" (Nesi & Gardner 2012: 38). An examination of the selected student essays revealed that indeed there can be variations across and within disciplines, but most of the essays seem to bear a great deal of resemblance to the expert counterparts in terms of the structure of the text and the type of arguments and evidence involved (e.g., empirical or theoretical). That said, student assignments are by nature different from published articles with regard to communicative purposes; therefore, the comparison between the two must be treated with caution.

2.2 Identifying formulaic expressions

The present study aims to be as inclusive as possible in the identification of FSs. Therefore, mixed criteria were adopted, given the rationale that "most examples will be captured one way or another" (Wray 2008: 110). If a multi-word sequence satisfies one of the following criteria, it was regarded as formulaic. [2]

2.2.1 Grammatical irregularity and/or semantic opacity

This means that as long as some aspect of the form or meaning of a word sequence is not strictly predictable from its component parts or from regular grammar, the expression is a FS, e.g., *take place*, *account for*, *run through* (Wray 2008; Schneider et al. 2014; Herbst 2015). Note that there is a continuum of fixedness, ranging from those resulting from a grammaticalisation or lexicalisation process (e.g., *as opposed to*, *with respect to*) to those that allow a certain degree of compositional freedom and semantic transparency (e.g., ***in*** *a similar way*, ***in*** *this* ***way***, *the* ***way*** **in** *which*). In the present study, dictionaries (primarily the *Oxford Learner's Dictionaries*)[3] and the list of phrasal expressions provided by Martinez & Schmitt (2012) were regularly consulted to avoid subjective judgement. If a word sequence is highlighted in the dictionaries (either as a separate entry or emphasised in bold type) or occurs on the list, it was considered to contain some kind of irregularity and therefore a FS.

[2]Some sequences may satisfy more than one of the criteria.

[3]This is an online source (https://www.oxfordlearnersdictionaries.com), which is home to the following dictionary and grammar reference titles:*Oxford Advanced Learner's Dictionary (9th edition),Oxford Advanced American Dictionary*, *Practical English Usage*, *Oxford Learner's Dictionary of Academic English*, and *Oxford Collocations Dictionary*.

2.2.2 Underlying frame

This refers to a formulaic frame that involves open slots to be filled, often by items of similar characteristics, e.g., *in the YEARs*, *in the Nth century*, *from YEAR to YEAR* (Wray 2008).

2.2.3 Situation/register/genre-specific formula

Expressions of this type are considered formulaic not because of their internal semantics or syntax, but rather the fact that they are the normal ways (judged by frequency of occurrence) of saying things in a particular situation (Wray 2008; Buerki 2016). In the case of ideational expressions in academic discourse, some examples are *the nature of, the structure of, research methods, public opinion polls*. To identify such FSs, the present study relied on an online tool, IdiomSearch (Colson 2016b; see also Colson 2016a). This program uses a built-in list of frequently occurring multi-word phrases (ranging from bigrams to sevengrams), derived from a multimillion-word reference database, to identify FSs in any given stretch of text. It has an advantage over the more commonly used tools such as AntConc particularly when dealing with small corpora where some FSs simply cannot reach the frequency threshold to be extracted. Clearly, one limitation of IdiomSearch is the difficulty in identifying FSs that are highly specific to a particular social practice or academic discipline (e.g., *Kant's critical philosophy, fluent aphasia*). However, such FSs are normally salient enough to be spotted manually and can be easily checked using either AntConc (whether they occur repeatedly in the given corpus) or Google Scholar (whether the same terms are used by other scholars).

The sequences identified by IdiomSearch were then manually sifted through to remove structural fragments without a clear meaning or function, such as *to be the*, *will give*, *is not a*, *we have a*. In some cases, an automatically identified sequence may contain more elements than needed for a complete semantic unit (e.g., *involves* in *involves the development of*) or only part of a semantic unit (*sequence of* in *an exact sequence of*, *a better way* in *in a better way*) (see Martinez & Schmitt 2012 and Buerki 2016 for the idea of semantic units). Human intervention means that the FSs identified will be self-contained semantic units (e.g., *the development of*, *an exact sequence of*, *in a better way*) that can be of utility for language teaching and learning purposes. There are also some cases that were not identified by IdiomSearch but were nevertheless included in the analysis because they contain the same core elements as in those that have been identified by the program, albeit with some formal variations. Take the combination of *ask* and

question for instance; while *asked questions about* was identified by the program, those involving changes of word order or form, or intervening elements as in *the questions asked*, *asked 10 blocks of questions*, *asking questions*, *asking knowledge questions about*, *the question being asked* were all missed by the computer as the exact sequences may not be frequent enough in the reference database. However, the exclusion of such variations would risk overlooking potentially important features of a given discourse community, and a manual approach was applied exactly to identify those non-contiguous FSs.

2.3 Classification of ideational functions

The classification of functions in the current study was based on Systemic Functional Linguistics (SFL), developed by Halliday (see Halliday 2014). SFL focuses on the underlying communicative functions of language and the systemic choices that are made available by the language system (Gledhill 2011). Central to the theory is the notion of three kinds of metafunctions – ideational, interpersonal, and textual – which underlie the organisation of language. In previous studies of lexical bundles such as Hyland (2008b) and Biber et al. (2004), the functional framework used was all based loosely on SFL. As discussed in the introduction, while textual and interpersonal functions have been extensively investigated in previous research, the ideational – also called "research-oriented" – functions are less well defined, often containing only a fairly small number of options. For a more comprehensive study of FSs with ideational functions, the present study turned to the original SFL framework for the purpose of deriving a workable annotation taxonomy.

The ideational metafunction in SFL is concerned with the construction of knowledge or human experience, represented as a configuration of a process (a type of action or event), participants in that process (an actor or object), and circumstantial elements such as time, place and manner. Each of these three components gives entry to a more specific system with a variety of options. Table 5.1 presents a slightly simplified version of the original system of ideational functions (see Halliday 2014), excluding those that either are not normally associated with FSs or rarely occur in the type of discourse under investigation, such as the category of behavioural processes. Some of the functions as well as their explanations have been tailored to the discourse at hand and its features. For instance, verbal FSs in the present study are often related to reference to previous research (i.e., what other scholars say about something), definition, explanation, and argumentation. The circumstantial elements in the original framework were merged into a few main sub-categories. Among them, manner encompasses a number of elements,

such as angle and role, which are treated as separate sub-categories parallel to manner in the original taxonomy. The remaining sub-categories (e.g., matter, accompaniment) were put under the "other" category due to their low frequencies. Terminology was added to the framework to address the large number of specialist terms occurring in the data under investigation.

The corpus data were manually gone through to identify FSs based on the criteria presented earlier. The UAM corpus tool (O'Donnell 2013) was used for the annotation of functions according to the functional taxonomy presented above.

3 Results and discussion

This section presents the overall frequencies of ideational FSs in the two corpora before offering a more detailed analysis of a number of major sub-categories of FSs found in the two corpora.

3.1 An overall picture

Altogether, 9,558 FSs with ideational functions were identified in the two corpora. Table 5.3 presents both raw and normalised frequencies (per 10,000 words) of ideational FSs in each corpus. To give an overview of the distribution of FSs associated with all the three metafunctions, the results from Wang (2018; 2019) regarding interpersonal and textual FSs are also presented in Table 5.3; see also Figure 5.1 for a graphical representation of the distribution. The log-likelihood test was conducted throughout the study to calculate whether a difference between two raw frequency counts is due to chance or to a statistically significant difference between the two corpora.

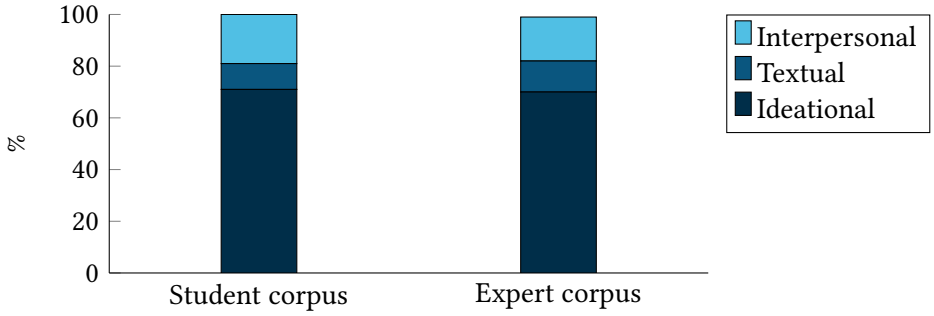

Figure 5.1: Distribution of the three types of metafunction in each corpus

Table 5.2: Sub-categories of the ideational metafunction

Process	
Material	Doing: action, movement, research procedure, e.g., *tidy X up to, turn away from, the operation of, search for, an examination of*
Mental	Perception, cognition, emotion, reasoning process, e.g., *make sense of, the understanding of, be expected to, take into account*
Verbal	Saying (normally associated with the reporting of previous research), explaining, defining, argumentation, e.g., *put forward, assert/proclaim that, as X put it, argue against, an explanation of*
Relational	Attributing, identifying, e.g., *consist of, be linked to, interaction with*
Existential	Existing, happening, e.g., *there be, there remain, the emergence of, take place*
Circumstance	
Location	Place, time, e.g., *in the world, in the Nth century, at the end of*
Manner	Means, comparison, degree, extent, angle, role e.g., *as a means of, quickly and easily, as opposed to, to the extent that, from the perspective of, in the form of*
Cause and contingency	Reason, purpose, condition, concession, e.g., *because of, as a result of, for the purpose of, in case of, in the absence of, in spite of*
Other	Matter, e.g., *with respect to*; accompaniment, e.g., *instead of, as well as*
Participant	
Attribute	Descriptive property, e.g., *the nature of, the character of*
Quantification	Quantity and category specification, e.g., *a small number of, a lot of, the majority of, a piece of, a type of*
Human or non-human entity	Normally non-specified, e.g., *human beings, ethnic minorities, a large audience*
Terminology	Specialist terms, e.g., *constant coefficients, international law, probability theory, amino acid, carbon dioxide, public health*

Table 5.3: Raw and normalised frequencies of FSs associated with the three metafunctions. SC: Student corpus; EC: Expert corpus.

	Ideational				Textual			
	SC	EC	G2	p	SC	EC	G2	p
No. of FSs	4484	5074	0.05	> 0.05	631	890	18.88	< 0.0001
per 10k words	960	964			135	169		
	Interpersonal							
	SC	EC	G2	p				
No. of FSs	1189	1243	3.38	> 0.05				
per 10k words	254	236						

As mentioned in the introduction, it is the textual and interpersonal metafunctions that have attracted most attention in previous studies of lexical bundles, *n*-grams and other types of FSs. However, as shown in Figure 5.1, the two categories of FSs together only account for approximately 30% of all the FSs identified in each corpus, whereas ideational FSs make up the remaining 70%, which, for some reason, have not been investigated systematically. Out of the 9,558 ideational FSs retrieved from the corpora, only 3,103 (32%) were captured by the frequency-based program, while 3,618 (38%) were completely missed; the remaining 2,829 (30%) instances were partially identified by the program in the sense that some of them may be part of a complete formulaic unit and some may contain elements outside a complete unit (cf. §2).

In terms of overall frequencies, Table 5.3 reveals a great similarity between the two corpora regarding ideational FSs; in fact, a statistically significant difference was only found in textual FSs between the two corpora. In other words, both student and expert texts are highly formulaic. Previous studies such as Chen & Baker (2010) and Ädel & Erman (2012) have observed a lack of formulaicity in L2 undergraduate students' academic writing, in comparison to either L1 student writing of the same academic level or expert writing. The results presented above suggest that successful L1 student writing is fairly close to expert writing in terms of formulaicity, at least quantitatively. This in turn may lend support to the advantage that native speakers have over non-native students in the use of ready-made multi-word expressions, which are considered part of native speakers' linguistic competence that non-native speakers have limited access to (Wray 2002; 2008; Kecskes 2016).

The following sub-section looks more closely at the ideational FSs based on the distribution of the three main categories of the ideational metafunction as well as the sub-categories as presented in Table 5.2.

3.2 FSs of different ideational functions

Table 5.4 presents the frequency counts and proportions of FSs with different ideational functions in each corpus.

Table 5.4: Distribution of FSs with different ideational functions in each corpus. Only significant p values are shown in the table. SC: Student corpus; EC: Expert corpus.

	SC		EC		Log-likelihood	
Sub-category	#	%	#	%	G2	p
Process						
Material	589	30	747	34	4.65	< 0.05
Mental	443	23	466	18	1.06	
Verbal	392	20	491	22	2.47	
Relational	330	17	350	16	0.61	
Existential	187	10	211	10	0.00	
Total	1941	100	2201	100	0.05	
Circumstance						
Location	451	31	497	31	0.11	
Manner	670	46	674	43	4.29	< 0.05
Cause and contingency	270	18	268	17	2.15	
Other	79	5	139	9	10.36	< 0.01
Total	1470	100	1578	100	1.76	
Participant						
Quantification	287	27	267	21	5.06	< 0.05
Attribute	479	45	535	41	0.02	
Terminology	268	25	432	33	21.74	< 0.0001
Human and non-human entity	39	4	61	5	2.62	
Total	1073	101	1295	101	2.80	

The two corpora resemble each other again in the distribution of the three broad functional categories. The identified FSs are most likely to be involved in processes (43%), followed by circumstances (31–33%) and participants (24–26%). However, within each category, significant differences between the two corpora were found in some sub-categories: material processes, manner and "other" circumstantial elements, as well as quantification and terminology. In what follows, some of these sub-categories will be examined further with examples drawn from the dataset.

Starting with material processes, as shown in Table 5.4, the expert writers used significantly more FSs than did the students. Table 5.5 gives some examples of such FSs, divided according to their structural make-up.

A few observations can be made from Table 5.5. To begin with, the FSs associated with material processes are made up of three main structural types: verb + preposition, verb + noun, and nominalisation + of. With regard to the first type, there are clearly more verb + preposition combinations, or phrasal/prepositional verbs, in the student corpus than in the expert counterpart. As can be seen in Table 5.5, some of the phrasal/prepositional verbs are shared by both corpora, e.g., *deal with*, *carry out*, *find out*, which are often used in academic writing to introduce a research topic, procedure, or a finding. However, the majority of the phrasal/prepositional verbs occur exclusively in the student corpus. Many of them seem to involve some kind of bodily movement and/or a figurative sense (e.g., *run away*, *storm out*, *fiddle with*, *trawl through*). As illustrated in the following examples, the use of such multi-word expressions is often associated with a narrative approach taken by the students in their essays.

(1) *The camera also **zooms out** to offer a wide shot of the four women, this serves to show how Miranda is surrounded and cornered by the others.* (BAWE_3160b)

(2) *Having **trawled through** the archives the historian's next task according to him was to corroborate and compose a critique of the evidence at hand.* (BAWE_0255h)

(3) *However the difficulties with complex structures could be related to the suggestion that Broca's, and other non-fluent, aphasics **struggle with** comprehension of unfamiliar, less frequent and longer word retrieval,…* (BAWE_6206c)

Multi-word lexical verbs are more commonly seen in conversation and fiction than in academic prose (Biber et al. 1999: 409); the frequent occurrence of such verbs in the student corpus may thus also be taken as suggesting an informal

Table 5.5: Examples of FSs representing material processes

	Student corpus	Expert corpus
Verb + Preposition	*expand on, look for, find out, deal with, trawl through, strip down, storm out, run away, go after, dress up, waltz into, fiddle with, move away, engage in, carry out, cover it with, break into, work on, sweep out, cut down, interfere with, trick sb into, force upon, suffer from, prevent/protect sb from, benefit from*	*delve into, build on, deal with, find out, engage explicitly with, was (calmly) engaged in, search for, work with, bring about, carry out, set back, interfere with, prevent sb from*
Verb + Noun	*make a detailed analysis, research conducted into, take a quick look at, tackle the problem, commit crimes against, commit an (earlier) error, wage wars, launch a media campaign, make some changes to, make more sales, make profit*	*the original research undertaken, overcome barriers, gain momentum, the murders/crimes/errors committed, wage wars, data was collected, take parental leave, take care of, meet their/the buyers' needs, impose limitations on, restrictions imposed by, further restrictions are imposed on*
Nominalisation + of	*an examination of, (the/a) study of, the development of, an/the analysis of, a wide shot of, the/an engagement of*	*scientific assessment of, the comprehensive collection and analysis of, the return of, a/the (thorough) development of, his engagement with*
Adjective + Noun	*scientific research, sexual abuse, marketing efforts, human endeavour*	*empirical study, further investigation/research, recent developments, scientific discoveries/advances, genetic modification, (in) previous research*
Other	*bought and sold, distribution and promotion, cooking and heating*	*widely used, newly generated, fully developed, finely tuned, easily overcome, collecting and analysing, data collection, video recording, fight and kill*

style, which has been attested as a feature of student writing in general, regardless of L1 background (Granger & Rayson 1998; Gilquin & Paquot 2008).

When it comes to verb + noun collocations, there seems to be a great deal of similarity between the two corpora. Some of them (e.g., *wage wars*, *commit crimes*) occur in both corpora, prompted by the same topic or subject area. Other topic-related collocations were also found, such as *take parental leave*, *meet someone's needs*, *impose restrictions on* in the expert corpus and *launch a media campaign*, *make more sales*, *make profit* in the student corpus. What remains are research-related collocations (e.g., *conduct + research*, *collect + data*, *make + analysis*), which, again, can be found in both corpora. An additional point to be made here is that verb + noun collocations often show a great deal of formal variability in terms of word order and intervening elements (e.g., *impose limitations on*, *further restrictions are imposed on*). Such formal variations mean that the core lexical items are not always contiguous and therefore are likely to be missed by automatic retrieval methods; in other words, for both methodological and theoretical reasons, this is an area that is worth further exploration using large corpora.

Nominalisations are a well-established feature of academic writing, used to pack more information into a single sentence. In the present study, the frame nominalisation + *of*, with or without an article *a/an* or *the* before the combination, is fairly common in both corpora. However, as shown in Table 5.5, nominalization + *of* constructions in the expert corpus often also contain adjectives (e.g., *scientific assessment of*). While the *of*-frame represents a grammatical construction, which is considered formulaic on the grounds of its high frequency, there is a strong collocational tie between the two core lexical items involved. A similarly strong collocational link is also apparent in most FSs of the next two categories drawn from the expert corpus (e.g., *empirical study*, *widely used, fully developed*), many of which are associated with research processes. The student corpus, in sharp contrast, is still dominated by processes related to subject areas (e.g., *sexual abuse, bought and sold*).

Moving on to FSs associated with mental processes, although the two corpora display no statistically significant difference in terms of frequency, a close examination of the FSs themselves provided some interesting insights. As can be seen in Table 5.6, which contains examples identified from both corpora, most of this group of FSs involve two or three key components, which, again, are not always contiguous. Apart from verb + noun and adverb + verb collocations, most of the FSs involve a combination between a noun/adjective/verb and a preposition. Semantically, a great number of the FSs in both corpora are associated with awareness, understanding, decision-making, and opinion. However, the expert

corpus yielded more FSs representing a reasoning process (e.g., *derive from*, *draw conclusion*, *make observation*, *the verification of*).

In contrast, the students seemed more inclined to employ another type of FSs, associated with an emotional state, as illustrated in the following examples.

(4) *The Führer* ***was*** **only satisfied with** *forming a Protectorate rather than outright annexation when Hàcha unexpectedly co-operated.* (BAWE_0318e)

(5) *She* ***is*** **anxious to** *hear Nicholas say she looks beautiful and forces him to say so, this infantile behaviour matches her personality and role as a Gothic heroine.* (BAWE_3160b)

(6) *Exporters need to* ***be*** **wary of** *using the same promotional strategy in the UK as in their home country.* (BAWE_0222a)

This tendency seems to mirror the students' use of multi-word lexical verbs associated with material processes as discussed earlier, evincing characteristics of a narrative approach and everyday language in the student essays.

As in the case of FSs associated with mental processes, quantitatively, there is no statistical difference between the two corpora with regard to verbal FSs. Yet a few comments need to be made about the particular FSs involved. Table 5.7 gives a list of examples from the dataset. What the two corpora have in common is the use of FSs to offer an explanation or to raise or answer a question, particularly in the expert corpus, with a range of lexical and syntactic variations (e.g., *answer the question, answer 10 blocks of questions, an answer to the question, the questions asked, ask objective-knowledge questions, ask a follow-up question, ask him a question*). In addition, topic-related FSs can be found in both corpora (e.g., *give + consent*).

The main difference between student and expert writing in this regard can be seen in the number of FSs associated with arguments and debates as well as elaboration in the expert corpus (e.g., *the justification for, an objection against, elaborate on*) versus that of FSs expressing actual verbal behaviour in the student corpus (e.g., *cheer someone up, laugh at, raise one's voice*). As Example (7) shows, the latter, most of which are highly idiomatic (e.g., *take/hold the floor*), seem to be prompted by, again, a need to narrate what is being analysed – a conversation in this case.

(7) *At line 11, B* ***makes a closing kind of statement****. It is not very meaningful to the discussion and B is therefore indicating that she has nothing further to add. Speaker A and C both respond with a backchannel, and even though C's is quite long, (line 13), neither* ***take the floor****.* (BAWE_6009b)

Table 5.6: Examples of FSs representing mental processes

	Student corpus	Expert corpus
Nominalisation + Preposition	*the same commitment to, one's (a full) understanding of* (F), *one's conception of, the feeling of, sb's thoughts about, an awareness of, his view(s) on/about/towards, intentions towards*	*an/the (full) understanding of, awareness of, confidence in, the views of, greater attention to, the thought of, sb's thoughts on, the perception of, the comprehension of, the verification of, be seen from*
Be + Adjective + Preposition/ *that/to*-infinitive	*be (un)aware of, be reluctant to, be inclined to, be anxious to, be interested in, be very conscious about, be mindful that, be highly appreciated by, be wary of, be expected from*	*be (more/not) aware of/that, be opposed to, be concerned with, be prepared to, indifferent to*
Verb + Preposition	*conceive of, take into account, extend to, come up with, rule out, from this emotion we derive*	*derive from, know about, wonder about, hope for, take into consideration/account*
Verb + Noun	*have sympathy for, make judgements, make sense (of), have no sense of, get some sense of, get an/the rough idea of, take the decision to, decisions were taken, the decisions taken, make informed decisions for, make strategic decisions, decisions regarding which market segments to target can be made, generalisations made, bring to light, give a proof (of)*	*have little (to no) knowledge about/of, the decision should be made, make sense of, take stock of, the choices parents make, the major conclusions that can be drawn from, come to these conclusions, make that/this/more final observation(s)*
Other		*considered carefully, well understood, easily overlooked, better understand, feelings and thoughts*

Table 5.7: Examples of FSs associated with verbal processes

Student corpus	Expert corpus
a direct answer to the question, ask the question of, *a more plausible explanation for, be explained by, a plausible explanation, an explanation for, have an explanation of,* *the narration of, a discussion about, be said bout, an excellent/objective account of,* *be called upon to, be accused of, speech made, cheer sb up, talk about/to, laugh at, raise one's voice, hold/take the floor, make a closing kind of statement, consent to, give informed/full consent*	*answer a list of questions, the questions asked,* *(the) argument(s) for, argue directly against, the justification for, the postulation of, claims that he set out, objections to,* *three objections that have been raised against, a final objection against,* *explanations of, some/an explanation of, have no plausible naturalistic explanation, any deeper explanation of, account for,* *give a plausible account of, elaborate on, a summary of, a description of, a brief overview to, go into the fine details of, research reports,* *give their informed consent to, commonly called, point out, enquire about, talk about, the repetition of, the utterances of, verbal communication, science communication*

Table 5.8 provides some examples of FSs representing the most common circumstantial sub-category, namely manner. Most of such FSs are prepositional phrases. As can be seen in Table 5.8, the expert corpus yielded a more limited range of FSs, mostly in association with the way (manner, fashion, means) in which a process takes place, than did the student corpus. Some of the FSs occurring exclusively in the student corpus, again, involve emotional states such as *in admiration, with tolerance, in anger, without any major headaches.*

Table 5.8: Examples of FSs associated with manner

Student corpus	Expert corpus
in such strict dichotomy, with tolerance, in isolation, in Nazi rhetoric, in such a way that, by chance, in the same fashion, in a straightforward manner, the detail in which, in detail, quickly and easily, in anger, in admiration, at rest, in the form of, without any major headaches, positively or negatively, with difficulty, long/short term, in equilibrium, on the macro scale, at this fundamental level, at resonant specific frequencies, in 26 space-time dimensions, at a speed, at 100%, to a minimum	*in an existential manner, in an easy-to-read and understandable manner, in this manner, in a somewhat Hobbesian fashion, the ways in which, by way of, in this strange way, a political means through which, at the global level, at the macro level, in the conventional form, in detail, under the guidance of, in abstract/ADJ terms*

We have thus far witnessed a tendency, which is distinctive of the student corpus, to involve FSs related to emotion as well as verbal and bodily behaviour, regardless of discipline. Together, they may suggest that we are dealing with two different genres here: narrative versus argumentative. However, given that the student essays are academic assignments given to final-year university students in the UK and that they are structured in a similar way to that of the published papers, it may be fair to say that the students at this stage are expected to produce work of a similar genre, albeit limited in scope and depth in comparison to published ones. Or, to put it in another way, they can be regarded as novice writers in training. Indeed, bearing in mind that the two corpora also share a great number of FSs, it is unlikely that they represent two completely different genres of writing. Rather, a more reasonable explanation for the differences observed

between the two corpora may be put down to the students' lack of awareness of genre conventions in terms of the way disciplinary knowledge is constructed and the style of delivery.

The students' lack of awareness of genre conventions can also be detected elsewhere. Take, for instance, FSs containing the word *way*. Altogether, 25 tokens with 16 different types were found in the expert corpus, and 42 tokens with 36 different types in the student corpus. Some examples are given in Table 5.9.

A few points can be made here. First of all, again, there is a great deal of variability in form, with fixed and variable slots occurring in a particular order. Three main patterns emerged from the examples about the use of FSs containing the word *way* in expressing means. All of them involve pairings of function words (prepositions *in* and o*f*) with at least one variable slot: *(a) X way of/to*, *in a X way*, (*the*) *X way in which*. As noted by Biber (2009), while conversation prefers continuous fixed sequences, written discourse prefers FSs with internal variable slots. As can be seen here, more often than not, the fixed elements are not adjacent to each other. This is obviously another area where the automatic retrieval methods may be of limited use and which would benefit from a more systematic investigation involving a larger dataset to generate possibly new understanding of features of formulaicity in language use in general and in academic discourse in particular.

As Table 5.9 shows, the student writers appeared to be less restrained in filling the variable slots than were the expert writers. The same can be said of the student writers' use of FSs associated with quantification. As shown in Table 5.4, the student writers employed this category of FSs more frequently than did the expert writers, the difference between the two corpora being statistically significant. Table 5.10 gives some examples of such FSs, which show a wide range in the student corpus, in contrast to a limited set in the expert counterpart.

Some of the expressions, which occur exclusively in the student corpus such as *a bit more* in Example (8), testify again to an informal register that is said to be typical of learner writing as a whole, including both L1 and L2 writing. The use of *harmless* in Example (9) illustrates a trend that has been observed throughout the current study, namely the extent of liberty or "creativity" that the student writers seemed to assume in filling the internal variable slots of FSs, without realising that some of them may be subject to certain restrictions in a given discourse community.

(8) *Poincare duality follows after* ***a more*** *work.* (BAWE_0049b)

(9) *The patient inhales* ***a small harmless amount of*** *radioactive gas which then attaches itself to red blood cells in the blood…* (BAWE_6206c)

Table 5.9: FSs with the key word *way*

Student corpus	Expert corpus
a quicker way to	*the way in which*
a simple way of	*its/his/the nurse' way of*
an invasive way of	*in a way that*
its own way of	*by way of*
the German way of thinking	*his way of*
the most effective way of	*in this strange way that*
the most suitable way to	*this way*
the only way to	*a different way to do*
the ways of	*one way that*
in a better way	*in this way*
in a mechanical way	*the way in which*
in a purely mathematical way	*in such a way that*
in a rather abstract way	*in the same way*
in a similar way	*in a natural way*
in a simple way	*no way of doing*
in a sustainable way	*in a deterministic way*
in a very physical way	
in a way	
in an unsustainable way	
in complex ways	
in quite a simple way	
in this way	
in the way	
in the way of	
in such a way that	
through its unobstrusive way of	
different ways in which	
one of the ways in which	
the way in which	

Table 5.10: Examples of FSs associated with quantification

Student corpus	Expert corpus
(quite) a few, lot of, a bit more, a pair of, a piece of, a (very small) number of, a great swathe of, one/some/all/none of, a vast amount of, a piece of, a collection of, a wide array of, a series of	*a (limited/small/large) number of, a wide range of, some/many/most/either/one of, a multitude of, the vast majority of*

It is generally accepted that successful academic writing is marked by a high degree of formulaicity, but what is perhaps less well recognised is that even those seemingly transparent and syntactically flexible word sequences may have established particular patterns of usage that are adhered to, consciously or not, by the members of the discourse community (Pérez-Llantada 2014; Wang 2018). In this case, even though the use of *harmless* is not semantically or grammatically deviant, in academic prose at least, it is not common to have another intervening adjective together with *small* in the FS *a X amount of*.[4] Although native speakers have available to them a large repertoire of everyday formulaic language (Sinclair 1991), the degree of liberty that the student writers seemed to take here, and in many other cases as shown in the study, suggests that the restrictions such FSs are subject to in academic prose may not be readily accessible to L1 students.

4 Conclusion

The present study set out to explore the potential of a computer-assisted manual approach in identifying and annotating formulaic language in academic writing, with a focus on ideational, or research-oriented, FSs. The first important finding is that ideational FSs account for 70% of all the FSs identified, a considerable proportion that would certainly warrant more serious attention than they have hitherto received. Most of such FSs contain two core lexical items or one lexical and one functional item in fixed slots, with the possibility of variable slots in between and change of word order, making it a particularly challenging task to automatically identify them. However, given their importance in understand-

[4] A search of *small + amount of* in the academic subset of British National Corpus (BNC) returned no instance involving any other adjective in between.

ing the nature of formulaicity in language use, these are the areas that would certainly benefit from a more vigorous investigation in future research.

Both student and published papers were found to be highly formulaic, particularly in quantitative terms. Indeed, the main differences between the two corpora are of a qualitative nature – that is, the two sets of texts seem to be formulaic in different ways. To start with, FSs associated with research and reasoning processes are conspicuously abundant in the expert corpus, whereas those expressing emotional states as well as verbal and bodily behaviour stand out in the student counterpart, suggesting the students' lack of awareness of genre conventions in terms of knowledge construction and language style.

Throughout the analysis, we also saw that the student writers seemed to be less restrained in filling the variable slots than were the expert writers. The results suggested that academic writing may not be as "creative" linguistically as the students might have assumed. Rather, many seemingly transparent and syntactically flexible word sequences may have their preferred or conventional patterns of usage in academic discourse, just as members of a particular speech community have preferred ways of saying things (Wray 2002; Kecskes 2016). It was argued that knowledge of such patterns of usage, which are probably not psychologically salient enough, may not be readily accessible to native speakers, echoing the claim that success in academic writing is "never guaranteed by generics or birth right alone" (Rajagopalan 2004: 116), but "is acquired rather through lengthy formal education" (Ferguson et al. 2011: 42) (see also Hyland 2016).

The SFL framework for the classification of FSs has proved particularly useful in pinpointing areas of difference between student and expert writing. From a pedagogical point of view, these areas of difference would benefit from more targeted awareness-raising activities in the training of novice writers.

To conclude, through capturing and addressing discontinuous and less frequent - but nevertheless formulaic - FSs that have been largely overlooked in previous research, the approach taken in the present study clearly has potential to contribute to both the understanding and the teaching of FSs in disciplinary writing. However, more data are needed in order to draw more informative and definitive conclusions. As manual identification and annotation can only be carried out to a certain extent, to proceed, there is a need to explore the possibility of at least semi-automated methods for recognising and annotating entities in a large text corpus. Given that most of the ideational FSs identified in the present study involve two core node words, it may be promising to start from individual lexical items, either through a keyword analysis (see, for instance, Wang & Soler 2019) or with a list of pre-selected node words (see Römer 2019), to retrieve FSs and their recurrent usage patterns in an effective and consistent way.

References

Ädel, Annelie & Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundle approach. *English for Specific Purposes* 31(2). 81–92.

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

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Buerki, Andreas. 2016. Formulaic sequences: A drop in the ocean of constructions or something more significant? *European Journal of English Studies* 20(1). 15–34.

Caines, Andrew & Paula Buttery. 2017. The effect of task and topic on opportunity of use in learner corpora. In Vaclav Brezina & Lynne Flowerdew (eds.), *Learner corpus research: New perspectives and applications*, 5–27. New York: Bloomsbury.

Callies, Marcus. 2015. Learner corpus methodology. In Sylviane Granger, Gaetanelle Gilquin & Fanny Meunier (eds.), *The Cambridge handbook of learner corpus research*, 35–55. Cambridge: Cambridge University Press.

Chen, Yu-Hua & Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology* 14(2). 30–49.

Colson, J. P. 2016a. Set phrases around globalization: An experiment in corpus-based computational phraseology. In Almeida Alonso, Ortega Barrera Ivalla, Toledo Elena Quintana & Cuervo Margarita Esther Sánchez (eds.), *Input a word, analyze the world: Selected approaches to corpus linguistics*, 141–152. Newcastle: Cambridge Scholars Publishing.

Colson, Jean-Pierre. 2016b. *IdiomSearch*. http://idiomsearch.lsti.ucl.ac.be.

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

Durrant, Philip. 2017. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics* 38(2). 165–193.

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

Ferguson, Gibson, Carmen Pérez-Llantada & Ramón Plo. 2011. English as an international language of scientific publication: A study of attitudes. *World Englishes* 30. 41–59.
Gilquin, Gaëtanelle & Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1). 41–61.
Gledhill, Christopher. 2011. The "lexicogrammar" approach to analysing phraseology and collocation in ESP texts. *ASP. la revue du GERAS* 59. 5–23.
Granger, Sylviane & Paul Rayson. 1998. Automatic profiling of learner texts. In S. Granger (ed.), *Learner English on computer*, 119–131. London: Longman.
Halliday, Michael. 2014. *Halliday's introduction to functional grammar*. 4th edn. Revised by C. M. I. M. Matthiessen. Oxen: Routledge.
Herbst, Thomas. 2015. Why construction grammar catches the worm and corpus data can drive you crazy: Accounting for idiomatic and non-idiomatic idiomaticity. *Journal of Social Sciences* 11(3). 91–110.
Hyland, Ken. 2008a. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18(1). 41–62.
Hyland, Ken. 2008b. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27. 4–21.
Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32. 150–169.
Hyland, Ken. 2016. Academic publishing and the myth of linguistic injustice. *Journal of Second Language Writing* 31. 58–69.
Kecskes, I. 2016. Deliberate creativity and formulaic language use. In K. Allan, A. Capone & I. Kecskes (eds.), *Pragmemes and theories of language use, perspectives in pragmatics, philosophy & psychology 9, pp*, 3–20. Cham, Switzerland: Springer International Publishing.
Martinez, Ron & Norbert Schmitt. 2012. A phrasal expression list. *Applied Linguistics* 33(3). 299–320.
Nesi, Hilary & Sheena Gardner. 2012. *Genres across the disciplines: Student writing in higher education*. Cambridge: Cambridge University Press.
O'Donnell, Mick. 2013. *UAM corpus tool*. Version 3.0.
Pérez-Llantada, Carmen. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes* 14. 84–94.
Rajagopalan, Kanavillil. 2004. The concept of "World English" and its implications for ELT. *ELT Journal* 58(2). 111–117.
Römer, Ute. 2019. A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics* 24(3). 268–290.

Schneider, Nathan, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael Mordowanec, Henrietta Conrad & Noah Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the 9th Linguistic Resources and Evaluation Conference*. Reykjavík.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, Anna. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon* 8(2). 245–268.

Wang, Ying. 2018. As hill seems to suggest: Variability in formulaic sequences with interpersonal functions in l1 novice and expert academic writing. *Journal of English for Academic Purposes* 33. 12–23.

Wang, Ying. 2019. A functional analysis of text-oriented formulaic expressions in written academic discourse: Multiword sequences vs single words. *English for Specific Purposes* 54. 50–61.

Wang, Ying & Josep Soler. 2019. What gets published in predatory journals: A corpus-based comparison of two journals in political science. *Learned Publishing* 32. 259–269.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2008. *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20(1). 1–28.