Ying Wang*

Formulaic sequences signalling discourse organisation in ELF academic lectures: a disciplinary perspective

https://doi.org/10.1515/jelf-2018-0017

Abstract: Formulaic sequences (e.g. *on the other hand, for example, at the same time*) are pervasive in natural language use and play an important role in differentiating socially situated practices. This paper examines formulaic sequences signalling discourse organisation in academic ELF lectures from a disciplinary perspective. Most previous studies of this kind employ a frequency-based approach; however, the inherent limitations of the methodology (e.g. arbitrary operational criteria, difficulty in handling discontinuous units) mean that a great deal may have been overlooked. This may be particularly relevant to ELF communication, which involves a high degree of flexibility and adaptability. The present study aims to address this gap by taking a manual approach in the identification of formulaic sequences, continuous or discontinuous, in context. The results provide further evidence for disciplinary differences and variability in the use of formulaic language to signal discourse organisation by lecturers in academic ELF settings.

Keywords: formulaic sequences, discourse organisation, discipline, academic lectures, ELF, manual identification

语块 (如: on the other hand, for example, at the same time) 在语言使用中普遍存在, 同时也是区分不同社会实践活动的重要话语标识。本文拟从跨学科的视角来研究 ELF 学术讲座中标记语篇结构的语块。 目前语块研究主要 采用频率的方法, 但是, 由于该研究方法仍然存在一定的弊端 (比如识别标准的任意性, 以及处理非连续性语块的难度), 导致研究结果不够全面。在ELF语境中, 由于语言使用的灵活性更高, 基于频率的研究方法的不足尤为显著。为了弥补这方面的不足, 本文采用人工标识的方式对语篇中连续或不连续的语块做出了更系统的分析, 其研究结果为不同学科领域中语块使用的差异性提供了进一步的佐证。另外, ELF学术讲座中标记语篇结构的语块由于使用者的不同 (来自不同母语背景) 而体现出更丰富的变异性。

关键词: 语块, 语篇结构, 学科, 学术讲座, ELF, 人工标识

***Corresponding author: Ying Wang,** Department of English, Stockholm University, S-106 91, Stockholm, Sweden, E-mail: ying.wang@english.su.se

1 Introduction

The rapid spread of English as a lingua franca (ELF) in the last decade and the fact that non-native speakers of English have now greatly outnumbered native speakers of English call for a revision of the role of non-native speakers in shaping the language as a means of intercultural communication (Seidlhofer 2005; Formentelli 2017). One of the issues that has received increasing attention concerns formulaic language (FL) in ELF contexts (e.g. Kecskes 2007; Seidlhofer 2009; Mauranen 2009; House 2009; Wang 2017).

Formulaic sequences (FSs) refer to words that have "an especially strong relationship with each other in creating their meaning" (Wray 2008: 9) such as *by and large, of course, on the other hand*. Corpus studies have revealed that such sequences are pervasive in natural language use and play an important role in differentiating socially situated practices (e.g. Biber et al. 1999, Biber et al. 2004; Hyland 2012). In particular, formulaicity in academic discourse has been a topic of extensive discussion, given its role in achieving fluency, facilitating comprehension, and identifying membership in different disciplinary communities. However, much attention has hitherto been given to written academic discourse (e.g. Hyland 2008, Hyland 2012; Ädel and Erman 2012). A small number of studies on spoken academic discourse focus primarily on English as a native language (ENL) (e.g. Nesi and Basturkmen 2009; Kashila and Heng 2014; Schnur 2014). Disciplinary practices, which prove to be an important variable in written academic language production, have seldom been a consideration in studies of spoken academic discourse (Ädel 2008; Schleef 2008).

The predominant trend in FL research is to take a frequency-based approach (e.g. lexical bundles, *n*-grams), relying on the computer to automatically identify frequently co-occurring word sequences in a given corpus. While this approach has the advantage of being methodologically straightforward, its inherent limitations mean that the picture of formulaicity may not have been fully revealed (Ädel and Erman 2012; Wang 2018). Among others, it disregards discontinuous and infrequent FSs, which can amount to a substantial proportion (Schneider et al. 2014). Taken together, as Wang (2018) shows, even those seemingly idiosyncratic choices may reveal important functional and formulaic features that characterise a particular community. In addition, different operational criteria regarding the rate of occurrence (i.e. the number of times a word sequence must occur to count as "frequent"), distribution across different texts, and the way of dealing with overlapping sequences can lead to grossly different conclusions to be made about the same data (see Ädel and Erman [2012] for a more detailed discussion). To conclude with Biber's (2009) suggestion, as FL is such a complex phenomenon, of which much still remains unknown, there is a need to embrace new and complementary

methodological approaches. The present study is an attempt in that direction in combining manual identification and annotation of FSs with a frequency-based approach to allow a fuller investigation of various forms of FSs and their functions in ELF communication. Using this approach, the present study focuses on FSs signalling discourse organisation in academic ELF lectures and aims to answer the following research questions:

1. What are the most common FSs signalling discourse organisation in the three disciplinary groups selected, namely Social Sciences, Natural Sciences, and Medicine? Are there any differences across these disciplines?
2. Are there any usage patterns that are shared by the lectures involved?
3. How many and what types of FSs that are detected by manual analyses tend to be overlooked by frequency-based identification?

The remainder of the paper is organised as follows. Section 2 offers a brief overview of the role of formulaicity in communication and in academic lectures in particular. Section 3 introduces the data for the investigation (3.1) as well as the selection criteria used to decide whether a word sequence is formulaic or not (3.2), and a functional taxonomy that is based on Systemic Functional Linguistics (SFL) for analysing the communicative functions of FSs (3.3). The results are presented and discussed in Section 4. The paper ends with a summary of the main findings and their implications in Section 5.

2 Formulaicity in academic lectures

FSs contribute to fluency in language production and communication. For native speakers, a large number of FSs are internalised through the natural process of acquisition. When the need arises, they are invoked readily as single entities instead of being composed from their constituent parts according to grammatical and semantic rules, thereby reducing cognitive processing demands on both the speaker and the listener (Sinclair 1991; Wray and Perkins 2000).

Non-native speakers are said to depend primarily on what Sinclair (1991) calls "open-choice principle," or in other words, compositional processing or semantic analysability (Wray 2002; Kecskes 2007), sometimes giving rise to grammatically correct but unidiomatic expressions (Kjellmer 1991). However, as Wang (2016) shows, EFL (English as a foreign language) learners' dependency on the open-choice principle does not necessarily preclude their adherence to the idiom principle – that is, the use of FSs, in particular those highly fixed or frozen expressions in the target language. The same is said to be true with regard to speakers in ELF settings (Seidlhofer 2009, Seidlhofer 2011). At the same

time, the use of the idiom principle displays its own distinct characteristics among different types of non-native speakers. House (2009), for instance, finds that the expression *you know* is used by speakers in ELF settings with a unique discourse function (a speaker strategy to increase coherence), which differs from ENL usage (social interaction). Mauranen (2009, Mauranen 2012) demonstrates that ELF phraseology may manifest what she calls approximation of conventional forms (e.g. *I'm not **very** sure* instead of *I'm not **quite** sure*). Kecskes (2007) relates this issue to the fact that speakers in ELF settings normally have little in common both culturally and socio-linguistically, and therefore have no shared conventions of established phraseology. Considering interlocutors' needs, speakers in ELF settings may intentionally resort to the open-choice principle, preferring the literal meaning to the semantically less transparent but more idiomatic usage (e.g. *I'll call you **later*** instead of *I'll call you **back***) (see also Bardovi-Halig 2009, Bardovi-Halig 2012).

FSs in academic lectures or classroom interactions have been widely studied, mostly with a pedagogical orientation for the designing and teaching of academic listening comprehension (e.g. Flowerdew and Miller 1997; Thompson 2003; Nesi and Basturkmen 2009; Neely and Cortes 2009; Schnur 2014; Deroey 2015; Formentelli 2017). The most common approach is to compare lexical bundles (i.e. continuous word sequences that occur frequently in a given corpus) used in English for Academic Purposes (EAP) listening materials with those occurring in authentic lectures, mostly in ENL settings. One of the findings is that bundles signalling discourse organisation are heavily represented in academic lectures, particularly in recorded EAP materials for teaching and learning purposes. Nesi and Basturkmen (2009) attribute the prevalence of such lexical bundles in lectures to real-time production constraints which encourage the use of prefabricated chunks. Meanwhile, given the high density of information content in academic lectures, the pedagogical function also requires the connection between propositions to be made clear (see also Björkman 2011; Deroey 2015; Wang 2017). Deroey (2015), for instance, has found that prospective textual markers used to prepare the listeners for the upcoming discourse are particularly prevalent in lectures, suggesting an awareness among lecturers of the need to facilitate processing and note-taking. In addition, the pre-planned nature of mostly monologic lectures means lack of normal opportunities for negotiation of meaning as in conversation, and hence a greater need for discourse structuring devices (Björkman 2011).

FSs play an important role in identifying membership in different speech communities (Wray 2002, Wray 2008; Hyland 2008). In academic discourse, discipline means engaging with others in a community that shares the same domain of knowledge. Staples et al. (2016) argue forcefully that any discussion

of complexity in academic language production has to consider disciplinary (and genre) differences. While this variable has attracted a great deal of attention in written academic discourse (e.g. Cortes 2004; Hyland 2008; Durrant 2017), investigations of its effect on the spoken register have been few and far between. This is probably due in part to the hypothesis that academic speech, being fundamentally different from academic writing, tends to blur disciplinary differences (see Ädel 2008: 89). However, there is increasing evidence that discipline matters in academic speech as well (e.g. Ädel 2008; Schleef 2008; Kashila and Heng 2014; Wang 2017). Both Kashila and Heng (2014) and Wang (2017), for instance, identify important differences in the frequency and types of lexical bundles across academic disciplines. Kashila and Heng (2014), focusing on two disciplines (Politics and Chemistry) in ENL settings, find that directive and referential bundles are more frequent in Chemistry, while Politics features particularly bundles of topic elaboration or clarification. Wang (2017) reveals that lectures in ELF settings are dominated by different sets of four-word bundles across three disciplinary groups: directive bundles (e.g. *if you look at, you can see the*) in Medicine; repeats (e.g. *in the in the, and and and and*) and vague expressions (e.g. *and so on and*) in Social Sciences; referential bundles (e.g. *and this is the*) and procedural markers (e.g. *and then you have*) in Natural Sciences. The contrast between Social Sciences and the other two disciplinary groups may be partly explained in terms of the structure of knowledge within these academic divisions, which has been found to be an important factor influencing language use in ENL academic discourse. Schachter et al. (1991), for instance, find that lecturers in hard science disciplines tend to use a significantly smaller number of filled pauses (*uh, er, um*) than in soft science fields, ascribing the highest structure to Natural Sciences and the lowest to the Humanities (Schachter et al. 1991: 74). While Wang's (2017) finding of the prevalence of the so-called repeats in lectures of Social Sciences echoes Schachter et al.'s (1991) observation about the filled pauses in the Humanities, it is unclear what other lexical options are typically used. In addition, given that ELF is an "open-source phenomenon" (House 2014: 364), which is constantly adapted by its users according to the context (see also Seidlhofer 2011; Cogo and Dewey 2012), it is even more likely that only part of the picture has been revealed – that is to say, the methodological limitations of the frequency-based approach (cf. Section 1) may make it difficult for the computer to capture the "inherent variability" (Firth 2009: 162) characteristic of ELF communication. Through manual identification and annotation of a small subset of the data used in Wang (2017), the present study is aimed to be complementary to the corpus study, shedding new light on formulaicity in ELF academic lectures.

3 Data and methodology

3.1 Data

The data for the present study form part of the lecture subset used in Wang (2017), which in turn was drawn from a one-million-word corpus of transcribed spoken academic lingua franca English (ELFA) (Mauranen 2008). Specifically, the dataset used for the present study comprises nine lectures of about ten hours' transcribed material, covering three broad disciplinary domains: Social Sciences (SS), Natural Sciences (NS), and Medicine (Med). Each lecture is of a distinct topic, given either by an invited speaker or the main lecturer of a course module. Questions from the audience are minimal and are excluded from the investigation because the study deals only with the lecturers' output. Table 1 presents the word counts of each file involved.

Table 1: Data used for the study.

Social Sciences		Natural Sciences		Medicine	
File	No. of words	File	No. of words	File	No. of words
01A	6,004	080	7,050	150	6,503
020	9,704	090	9,504	180	6,290
030	11,136	160	11,773	23A	6,386
Total	26,844		28,327		19,179

In total, a corpus of 74,350 words was manually examined to identify and annotate FSs signalling discourse organisation, using the UAM Corpus Tool (O'Donnell 2013).

It should be mentioned that with such small quantity of data, it is difficult to prove whether a certain characteristic is disciplinary or unique to a specific lecturer. However, given that disciplinary variation has been ascertained in Wang (2017) with more corpus data, including these nine lectures, the selection of data in the present study was based on the topic (distinctive of a particular discipline) and the structure (largely monologic) as representing the disciplines involved for the sake of a more qualitative analysis.

3.2 Identification criteria

The present study aims to be as inclusive as possible when it comes to the identification of FSs in the data. Therefore, mixed criteria drawn from previous FL research

were adopted in this study to decide whether or not a word sequence is formulaic, based on the rationale that "most examples will be captured one way or another" (Wray 2008: 110). If a multi-word sequence satisfies any one of these criteria below, it is regarded an FS (see Wang [2018] for a more detailed explanation).

Grammatical irregularity and/or semantic opacity (Wray 2008; Martinez and Schmitt 2012; Schneider et al. 2014): 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 an FS, e.g. *on the other hand, at least, for example, first of all*. In order to ensure that this criterion is applied consistently in the study and replicable results can be produced by other researchers, dictionaries (primarily *Oxford Advanced Learner's Dictionary*) and the list of phrasal expressions provided by Martinez and Schmitt (2012) were constantly consulted. If a word sequence is highlighted in the dictionaries (either as a separate entry or emphasised in boldface) or occurs on the list, it was considered to contain some kind of irregularity and therefore an FS.

Underlying frame (Wray 2008): this refers to a formulaic frame that links parallel structures with open slots to be filled, often by items of similar characteristics, e.g. *not only ... but also, as ... as*.

Situation/register/genre-specific formula (Wray 2008; Buerki 2016): what is idiomatic about this type of FSs is not their internal semantics or syntax, but the fact that they are the common ways (judged by frequency of occurrence) of saying things in a particular situation, e.g. *as i said, which means that, i will show you*. In this regard, IDIOM Search (Colson 2016), an online tool for the extraction of multi-word phrases, was used. Also included in the analysis are expressions that were not identified by the program but nevertheless share an underlying pattern with those that have been recognised. As will be seen in Section 4.2, such instances are key to our understanding of formulaicity in ELF communication and therefore need to be included.

3.3 A functional taxonomy

The classification of functions for the present study is based on Systemic Functional Linguistics (SFL), which focuses on the underlying communicative functions of language and the systemic choices that are made available by the language system (Halliday 2014; see Wang [2018] for justification of using this framework). Key to the use of SFL is the notion of metafunction, which refers to three separate strands of meaning: ideational, textual, interpersonal. The present study focuses on the textual metafunction that is related to expressions whose primary purpose is to tell the listener how the discourse is organised and to create cohesion as it moves along. The three options within the system of textual metafunction are *elaboration*,

extension, and *enhancement* (Halliday 2014), all serving to link a clause to its surrounding context. *Elaboration* means making an element more precise by means of restatement, summarising, or in some other way for discourse purposes. *Extension* expresses an additive, contrasting, or replaceable relationship between prior and coming discourse. *Enhancement* qualifies the context by reference to the physical environment, topics of discussion, logical connection, or manner. Each of the three categories gives entry to a more specific system with its own options, which are presented with illustrative examples in Table 2.

Table 2: Classification of textual functions.

Category	Sub-category	Function	Examples
Elaboration	Exposition	Restatement or re-presentation of an element	*in other words* *i mean* *which means that* *you know*
	Exemplification	Providing more details by examples	*for instance* *let's say*
	Clarification	Making the elaborated element more precise by means of particularising, summarising, evaluating, etc.	*in particular* *at least* *all in all* *in fact*
Extension	Additive	Adding new information related to the current information	*in addition* *neither … nor* *apart from*
	Adversative	Expressing a contrasting relationship with the given message	*on the other hand* *in contrast*
	Variation	Expressing a replaceable, subtractive, or alternative relationship	*not … but* *either … or*
Enhancement	Spatio-temporal	Signposting the macro-structure of the discourse; referring to physical or abstract entities	*with regard to* *first of all* *i'll show you* *if you look at* *as i said*
	Causal-conditional	Expressing logical connection between adjacent utterances (e.g. causal-resultative, conditional, concessive)	*as a result* *for that reason* *if … then* *despite that* *even though*
	Manner	A statement X is made by means of comparison with another statement Y, or simply via Y.	*as … as* *as well* *by which means*

While most FSs occurring in the data have a distinct function that is fairly easy to identify, some are less straightforward, in particular those that may potentially have more than one function depending on the context. The FS *you know* is a typical example. Rintaniemi (2017) lists ten different functional categories of *you know* in academic ELF interaction, operating on both textual and interactional (a.k.a. interpersonal) levels, but primarily the former, which tallies well with House's (2009) findings. In the present study, the use of *you know* as in example (1) is regarded as serving a textual function as its main purpose seems to be to introduce an explanation. However, one may argue that it is also used to establish rapport with the audience by assuming shared knowledge, and therefore serving an interpersonal function as well.

(1) *the ethics that i have been erm explaining to you in the beginning of my lecture er with justice er* ***you know*** *equality solidarity blah blah blah erm all those erm are vital are very important are very important for the er sustainable development* ... (ULECD030)

In such cases where a FS may have more than one function, as long as it fulfils a textual function, it was included in the analysis. A more refined analytical framework is needed in future research to deal with the issue of multifunctionality of FSs in a systematic way.

4 Results and discussion

4.1 Frequency and distribution of textual FSs across disciplines

Altogether, 2,249 FSs were identified with a discourse organisation function. Table 3 presents the frequencies (both raw and normalised) of such FSs in the sub-corpora.

As can be seen in Table 3, such FSs are most frequent in the sub-corpus of Natural Sciences (NS) and least in Social Sciences (SS), with Medicine (Med) lying in between. No outliers were identified in each disciplinary group. The log-likelihood test was carried out throughout the study, and the results here show that the differences across the three disciplinary groups are all statistically significant. However, as mentioned in Section 3.1, the small quantity of data involved needs to be taken into consideration while interpreting these results.

Figure 1 shows the distribution of the functional categories involved in the dataset. FSs as spatio-temporal markers form the largest group. Table 4 presents

Table 3: Distribution of textual FSs across disciplines (normalised frequency per 1,000 words in brackets).

Social Sciences		Natural Sciences		Medicine	
File	Freq.	File	Freq.	File	Freq.
01A	111 (18.5)	080	153 (21.7)	150	160 (24.7)
020	241 (24.8)	090	342 (36.0)	180	181 (28.8)
D030	340 (30.5)	160	486 (41.3)	23A	235 (36.8)
Total	692 (25.8)		981 (34.6)		576 (30.0)

SS vs. NS: G2 = 35.84, $p < 0.0001$.
SS vs. Med: G2 = 7.29, $p < 0.01$.
NS vs. Med: G2 = 7.45, $p < 0.01$.

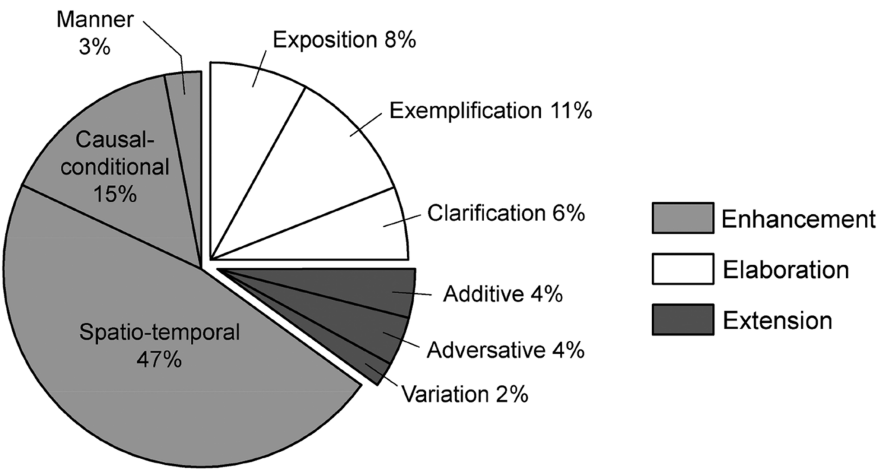

Figure 1: Distribution of FSs of different functional categories in the data.

Table 4: Frequencies of spatio-temporal FSs across disciplines (normalised frequency per 1,000 words in brackets).

Social Sciences	Natural Sciences	Medicine
202 (8)	552 (19)	277 (14)

SS vs. NS: G2 = 150.60, $p < 0.0001$.
SS vs. Med: G2 = 50.48, $p < 0.0001$.
NS vs. Med: G2 = 17.05, $p < 0.0001$.

the frequencies of such FSs across the disciplines. Again, disciplinary differences can be seen – spatio-temporal FSs occur most frequently in the NS sub-corpus and least in SS, with Med in between.

Table 5 lists the most frequent textual FSs in each disciplinary group. Overall, we can see that the majority of these frequent textual FSs are two- or three-word sequences, whereas four-word sequences, which seem to have attracted most attention in the literature, are scarce with only six types (e.g. *on the other hand, you can see here*). At the same time, many of the textual FSs are fixed expressions (e.g. *for instance, and so on, at least, in fact, because* of), unlike in the case of four-word lexical bundles as found in Wang (2017), where most are so-called genre-specific formulas (e.g. *as you can see, what do you think, is one of the*).

Table 5: Most frequent textual FSs across disciplines.

Social Sciences	Natural Sciences	Medicine
for instance (54)	*and then* (83)	*and then* (46)
you know (43)	*this is (the)* (60)	*you (can) see (here/that)* (29)
i mean (37)	*you (can) see (here/that)* (43)	*this is (the/a)* (20)
and so on (24)	*and so on* (31)	*at least* (18)
for example (20)	*but still* (21)	*for instance* (16)
(but) on the other hand/way (18)	*for example* (18)	*so that* (13)
and then (15)	*and also* (18)	*in this case* (9)
which means (that) (13)	*as i said* (17)	*not only … but also* (9)
let's say (12)	*you know* (17)	*not … but* (8)
because of (that) (11)	*of course* (16)	*in fact* (6)
if you (26)	*that's all* (13)	*because of* (5)
et cetera (10)	*if … then* (12)	*et cetera* (5)
so that (9)	*at least* (11)	
as long as (8)	*i would like to show/talk about* (11)	
that is (7)	*i will show you* (10)	
and also (7)	*on the X side* (10)	
that's why (6)	*in this case* (10)	
	last time (6)	
	so that (6)	

The NS sub-corpus yielded not only the highest frequency of textual FSs, but also a greater variety of different FSs, some of which have a high repetition rate, suggesting that the lecturers in this group relied on discourse-structuring devices to a greater extent than those in the other two groups. The top three most frequent FSs in the NS sub-corpus are shared with the Med counterpart. Two of them, *this is* (*a/the*) and *you* (*can*) *see* (*here/that*), suggest that the

lecturers in these two disciplines made use of visual aids, corroborating Wang (2017), which is based on a larger dataset. Apart from these, the Med sub-corpus also shares some FSs with the SS sub-corpus (e.g. *so that, because of, et cetera*), while the NS and SS sub-corpora display distinctly different preferences.

In the NS sub-corpus, we see in Table 5 a large number of spatio-temporal FSs, used to either introduce the macro-structure of the lecture or to direct the audience's attention to a particular entity (e.g. *as i said, that's all, i would like to show/talk about, i will show you, on the X side*). The same observation is made by Kashila and Heng (2014), who explain that hard science fields often deal with a variety of concrete entities such as a particular procedure, materials, and instruments, which need to be identified and explained in the lecture (see also Schleef 2008). A wide range of expressions such as those mentioned above may thus have been adopted to meet this need, and have gradually acquired the status of formulas in these disciplines because of their high frequency of occurrence.

As mentioned in Section 2, using a frequency-based approach, Wang's (2017) study yielded a number of fairly frequent sequences that contain repetitions of a single word or sound (e.g. *in the in the, and and and and*) in Social Sciences, which in turn can be explained in terms of the structure of knowledge in different academic divisions. Through manual identification, however, a clearer picture emerged with regard to the other "real" lexical options that may be typical of this disciplinary group. As shown in Table 5, the SS sub-corpus features in particular phrases of *elaboration* (e.g. *i mean, which means* (*that*), *that is, let's say, for example/instance*), logical connection (e.g. *because of, that's why, as long as*), and transition (e.g. *on the other hand*). Again, this observation is consistent with Kashila and Heng (2014) (cf. Section 2). One possible explanation for the prevalence of such FSs in SS is related to the need for more discussion of ideas in this discipline. It should be noted that most previous studies such as Kashila and Heng (2014) and Schleef (2008) are based on ENL settings; the results of the present study as discussed above suggest that lecturers in ELF settings follow more or less the same patterns distinctive of their own disciplinary practices. As Wray (2002) claims, FSs develop to serve important communicative needs of a given discourse community. The lecturers' preferences for different FSs in the NS and SS sub-corpora may thus be intrinsically bound up with different needs that arise from conveying different domains of knowledge. In disciplines such as SS where one uses language to construct research findings, as opposed to disciplines such as NS where language is used to report findings, the formulaic structures are bound to be different.

The following extracts (2)–(4) from the data should give a taste of the contrast in the use of textual FSs between SS and NS lectures. Example (2) was drawn from an SS lecture, where coherence is achieved as the topic, which

involves several related ideas, progresses seamlessly in the stretch of discourse, through the use of a range of textual FSs. To start with, a formulaic frame (*not … but*) is used to link two parallel ideas, with the emphasis (new information) being placed on the latter part, which is then followed by a comment (*that is important*) and an explanation led by an exposition marker *that means that*. In the explanation, logical relationships between segments are made explicit by the use of expressions such as *because* and *so that*. The clarification marker *let's say* would be superfluous in written language as it adds no meaning to the sentence. The use of it here in the lecture, however, may serve to give the lecturer time to think of the upcoming proposition – a specific time period in this case (*since the eighteenth century*), followed immediately by another clarification led by *at least*. The connector *and also* introduces another two related ideas, with *i mean* serving a similar purpose as *let's say* in the preceding sentence, and the whole stretch of discourse ends with a brief summary (clarification), led by *so to sum up*.

(2) *and modernisation is* ***not*** *just a change of social structures, towards what we have now* ***but*** *i would emphasise that modernisation includes also change in the ways of thinking, and that is important, and* ***that means that*** *during the past 150 years people have understood that the society is changing, and they have wanted to change things and change as such has been evaluated as a positive matter,* ***because*** *change has been seen as a promise of better future,* ***so that*** *people could have better future on earth not only in heaven, and this way of thinking has changed rapidly* ***let's say*** *since the eighteenth century, also in finland people,* ***at least*** *in the first half of the nineteenth century still believed that things go better if the if things (won't) change.* ***and also*** *two major ideologies of the of the twentieth century* ***i mean*** *capitalism and socialism both have been aware very future oriented ways of thinking they both give promises from change towards a better society,* ***so to sum to sum up*** *clo- er shortly er …* (ULEC020)

Wood's (2006) discussion of how FSs can facilitate fluency in speech may be relevant here. The use of the textual markers as delineated above, mostly multiword sequences, allows a continuing flow of speech to occur while giving time to both the speaker and the listener for mental processing of other aspects of speech. On the part of the speaker, these markers can help shorten the processing route of speech by bypassing the need for assembling components so that the conscious mind can be focused elsewhere during speech (e.g. planning and formulating the following utterance); on the part of the audience, these markers serve as important signposts to prepare them for what they need to understand, and in so doing help ease their processing load as well.

(3) ***i'll talk about*** *the synthesis of 2-methylanthraquinone is this thing commercially available* ***then i'll talk about*** *some of the colour problems in the early stages of bleaching* ***and then i'll talk about*** *lignin condensation which appears to be w- the root of the problem, er our route of making synthesising 2-methylanthraquinone is to replace benzine with toluene* ***and i'll show you*** *that reaction* ***in a little while*** *...* (ULEC080)

(4) *the way we got the pure product is just by washing it with hot water* ***and*** *this thing with the carboxylic group wash washes out* ***so*** *in a commercial synthesis you could actually wash the crude-product the mixture of this and this with hot water and then vacuum dry your evaporate your water and get this product and recycle it back er to the start of the cyc- cyclisation stage the ring-closing stage* ***so*** *if you if you recycle the unconverted product you could get a theoretical yield of a 100 per cent for this stage of the reaction* ***and then*** *the yield of the overall reaction would be close to the 92 per cent that for the intermediate,* ***so*** *basically er w- you don't get any degradation the only product you get from this reaction is this product* ***and er*** *you get some unreacted material* ***and*** *all you'd have do is rec- keep on recycling it* ***and*** *the only thing that would leave is this product it's a very simple reaction* ***and*** *water is the only solvent er you use ...* (ULEC080)

Examples (3) and (4) were taken from a NS lecture. Example (3) demonstrates how spatio-temporal FSs are typically employed in the NS sub-corpus. Most of them (*i'll talk about, i'll show you, in a little while*) are what Flowerdew and Miller (1997: 38) call macro-markers, which outline the structure of the lecture, announce discourse goals, or refer to outside the lecture. As we can see in example (3), the signposts are arranged purely in chronological order, linked by the connectors such as *then, and then, and*. When it comes to the description of the actual research procedure itself, as we can see in example (4), it is mainly the connectors, mostly single-word expressions (such as *and, so*) that are used to structure the narrative in a linear way, as a response to how the sequence of actions is arranged, namely in chronological order.

4.2 Congruency with frequency-based identification

One of the purposes of the present study is to give an idea of how many and what types of FSs tend to be overlooked by frequency-based methods (cf. Section 1). As introduced in Section 3.2, IDIOM Search was used as a pointer to potential FSs on account of frequency of occurrence. The identified FSs were subsequently coded according to whether or not they were recognised by IDIOM Search. In

some cases such as *in another occasion* in example (5), only part of it (*another occasion*) was identified by the frequency-based program. Sometimes the sequences captured by the program may contain elements adjacent to a complete formulaic unit such as *you* in *so that you* and *have* in *have at least* in example (6). These cases were coded as "Partly."

(5) *... and maybe if there is not enough discussion now we can do it later* ***in another occasion****, ...* (ULEC01A)

(6) *... and er i understand that two or three dis not get it yet so er will bring i hope before ten o'clock* ***so that*** *you have* ***at least*** *when he begins, ...* (ULEC01A)

The remainder of this section is divided into two sub-sections. Section 4.2.1 provides an overall picture, followed by a discussion of common features found in the data with regard to the use of some textual FSs. Section 4.2.2 is devoted to a close look at the types of FSs not identified by IDIOM Search.

4.2.1 An overall picture

Table 6 presents the proportions of identified textual FSs in relation to whether or not they were recognised by the frequency-based program.

Table 6: Congruency with the IDIOM Search results.

Type	Yes	No	Partly	Total
No. of occurrences	615	1123	510	2,249
(Proportion)	(27%)	(50%)	(23%)	

As can be seen, only 27% of all identified textual FSs are strictly identical to the results of automatic retrieval, whereas 50% were missed completely by the computer. In other words, relying on the recurrence of uninterrupted linguistic forms would result in a large proportion of FSs being overlooked. Taking a lexical bundles approach, Wang (2017) finds frequent occurrence of what Biber et al. (1999) call "repeats," such as *mhm hm mhm hm* and *in the in the*, in the same data, providing further evidence for a well-observed phenomenon in ELF research, namely that repetition or hesitation serves as a communicative strategy in ELF interactions (e.g. Björkman 2011, Björkman 2014; Cogo and Dewey 2006; Cogo and House 2017). One of the reasons why such a substantial proportion of FSs were missed by the frequency-based program may thus be related to hesitation markers

of different kinds, which can break up otherwise continuous FSs, making them difficult for the computer to identify. However, this was found not to be the case, in particular with those highly fixed FSs. As examples (7) and (8) illustrate, hesitation markers, if they do occur, often stay outside, rather than inside, an FS.

(7) *... **so to conclude erm**, the long-range temporal correlations of the DFA exponent that we use to index there erm this has a fairly high heritability and er it's independent of amplitude so ...* (ULEC180)

(8) *er we have done already a lot of work still doing by using er lentiviral vectors or adenoviral vectors to transfuse the cess in vitro, **er with er for instance er** amylase-cre er reporter system ...* (ULEC150)

The fact that these FSs seem to be unbreakable suggests that they may be treated as single units in the speaker's mental lexicon – in other words, Sinclair's idiom principle applies (Sinclair 1991; see also Wang [2016, Wang 2017] for a brief introduction). In addition, as can be seen in examples (9) and (10), apart from signaling discourse organisation, some of such FSs seem to be used as a means of gaining time for online processing at the same time.

(9) *... and X er axis does not point directly at sun because Z-axis is parallel to north magnetic pole **i will show you i will show you er i think** the yes this is it this is the the picture of all coordinate systems* (ULEC160)

(10) *ere r i i i when i lecture as the matter of fact these topics i erm i'm a bit emotional **er because of for instance i don't know** maybe it has something to do with the with my background where i am from originally* (ULEC030)

Example (9) presents a clear case of repeating a formula or variations of the same formula within a run. Instead of pausing to look for *the picture*, the use of these formulas allows the speaker to keep the rhythm of the narrative moving and hence achieving fluency in the sense of avoiding long pauses or breakdown. The FS *i think* (which was not considered in the present study) may be seen as one example of what Wood (2006: 30) calls "filler formulas," or "lexical fillers" to be more specific, in contrast to early stages of fluency when hesitation or non-lexical fillers are the dominant means of gaining time to complete processing.

Likewise, in example (10), a number of FSs are strung together. In this case, the first FS (*because of*) hints at an upcoming explanation (why the speaker became *a bit emotional* when lecturing on the given topic). But he was obviously trying to formulate the explanation on spot, maybe trying to think of a way to weaken the

force of the statement: instead of saying outright (*because of my background*), some hedges were employed in the actual utterance (*maybe it has something to do with my background*). While *i don't know* may serve as a hedging device, the use of *for instance* seems to be completely irrelevant, functioning mainly as a lexical filler to allow the speaker to formulate the subsequent speech segment.

4.2.2 FSs not identified by IDIOM search

Table 7 gives some examples of the category ("No"), which can be further divided into a number of main sub-categories, providing indications of what types of FSs tend to be overlooked by frequency-based methods.

Table 7: Types of FSs that are not identified by IDIOM Search.

Type	Examples
Continuous FSs	*as well, look up, for instance, in reality, that's all, you know*
Deviations	*in other word, on one way* (meaning "hand"), *on conclusion, so on, in the same time*
Variations	*i'll just show, i showed you, as you see, you can immediately see, you see here/there*
Intervening elements	*go a little bit into, go this through, make a couple of comments about*
Underlying frames	*if … then, not only … but also, either … or, whether … or not, as … as*

Only a small number of continuous FSs failed to be captured by the frequency-based program. They are mostly two-word sequences, including some highly fixed FSs such as *as well* and *in reality*, and those that are more transparent in meaning but with a distinct discoursal function such as *you know*, *i mean*, and *that's all*, most of which are typically used in spoken language. The next two categories consist of either deviations or variations of a recognised form. Deviations from standard usage (e.g. *in other word* instead of *in other words*, *on conclusion* instead of *in conclusion*) in academic ELF communication are discussed in detail in Mauranen (2009, Mauranen 2012), who comes to the conclusion that ELF phraseology is at the interface of linguistic convention and creativity (see also Seidlhofer 2009; Pitzl 2012). While the examples given in Mauranen's studies (e.g. ***in*** *my point of view*, *as* ***the*** *matter of fact*) are taken as representing the group of ELF speakers under investigation, the deviations found in the present study are mostly restricted to individual speakers, probably due to the small size of data involved.

Variations, in contrast, are a more common feature of the group of lecturers under study and tend to manifest themselves in what seem to be

semantically transparent and syntactically flexible sequences; for instance, while *as you can see* and *here you can see* are identified by IDIOM Search, variability in form involving the choice of modal verbs or change of word order as in *as you see, you can immediately see*, and *you see here/there* may result in the sequences being missed out by the computer. Similarly, while the form *i will show you* is frequent enough to be automatically retrieved, variations produced by the lecturers such as *i'll just show* and *i showed you* would be excluded from the picture, should a purely frequency-based approach be applied. The majority of the identified FSs coded as "Partly" also involve variations of different kinds (e.g. *i'm just trying to show that*, *i'll show you*, *what you can see is that, we saw earlier that*). Using the same manual approach, Wang (2018) detects a great degree of formal variability in the same kind of what appear to be transparent and flexible sequences in novice academic writing, whereas expert writers display preferred patterns of usage, providing empirical evidence for the view that academic writing is marked by formulaicity. Given that orality is seen as a feature of novice writing (Granger and Rayson 1998; Wang 2016: 37–38), the variations in form as shown in the present study may be understood as a feature characteristic of spoken language, which in turn can be transferred into writing by novice writers. Another possible interpretation is related to the speakers involved, namely that this may be a feature typical of ELF communication. As ELF interactions are asymmetric by nature, speakers have to work proactively for communicative effectiveness (Björkman 2014: 129). Among others, transparency is a prominent feature of ELF communication, which has been observed in previous studies including Seidlhofer (2009) and Pitzl (2012), which look at the use of idioms and metaphors, and Björkman (2011, Björkman 2014) on morphosyntactic use. As mentioned in Section 2, non-native speakers in general are said to depend primarily on the open-choice principle. The sequences such as *as you see* and *i showed you* that were produced by the lecturers, who happened to be all non-native speakers in this study, are grammatically correct and semantically felicitous, and therefore may be seen as the open-choice principle in operation. Given the manifold contexts and speaker constellations in ELF communication, as argued by Seidlhofer (2009), both native and non-native speakers may sometimes intentionally resort to the open-choice principle for the sake of transparency, which necessarily involves variability of linguistic forms.

The other two categories both involve discontinuous sequences, those that contain intervening elements and those that have an underlying formulaic frame with one or more open slots to be filled. In the former category, the intervening elements can be another self-contained FS, such as *a little bit* in *go a little bit into*, and *a couple of* in *make a couple of comments about*. Underlying formulaic

frames are limited to a few quite well-established phrases, normally used to link two parallel structures or elements, such as *not only ... but also* and *either ... or*. The list can serve as an index to be incorporated in automatic identification methods for large-scale corpus research.

5 Conclusion

Through careful manual identification and annotation of FSs in context, the present study was able to avoid the limitations of frequency-based methods in FL research and uncovered some previously unrecognised or neglected aspects of formulaicity in ELF academic lectures. First of all, the majority of the most frequent FSs in the data turned out to be highly fixed, two- or three-word sequences, which seem to have been sidelined in previous research by four-word lexical bundles, many of which are incomplete semantic and/or syntactic units. More importantly, the study gave further evidence for disciplinary variation in spoken academic ELF. Despite the small quantity of data involved, the study threw light on the lecturers' preferences for different FSs, which were potentially associated with different disciplines. The lecturers in NS, for instance, relied on discourse-structuring devices to a greater extent than those in the other two groups, in particular spatio-temporal lexical resources. The SS counterparts used frequently phrases of elaboration and logical connection. The observation goes to show that as in ENL settings, FSs in ELF settings also develop to serve important communicative needs of a speech community associated with the same domain of knowledge. While disciplinary variation has attracted a certain amount of attention in previous research on written academic prose, more needs to be known about the impact of this factor on spoken academic discourse.

Apart from the FSs that were indicative of different disciplinary practices, the study also revealed some usage patterns that were common to the whole group of speakers under investigation. Some rather fixed FSs seemed to have been internalised in the speakers' mental lexicon and were called upon when the need arose (as discourse-structuring devices or lexical fillers). Variations in form were more evident in semantically transparent and what may also appear to be syntactically flexible formulas, suggesting the operation of the open-choice principle, which in turn may be a strategy used by speakers in ELF settings for the sake of communicative effectiveness. These findings corroborate what has been observed as a typical feature of most ELF situations, namely the co-existence of "the variability of linguistic forms and the use of non-attested features alongside codified ones" (Pitzl 2012: 39).

Another important contribution made by the study was related to the methodological issue in FL research. The results suggested that frequency-based methods may be particularly problematic when dealing with spoken and/or ELF data, which inevitably involve irregularities of different kinds. The study identified some main types of FSs that were missed by the frequency-based program including deviations, variations in form, those containing intervening elements, and formulaic frames, some of which may be considered in the future development of automatic retrieval tools for FL research.

References

Ädel, A. 2008. *What uh the folks who did this survey found*: Expert attribution in spoken academic lectures. *Nordic Journal of English Studies* 7(3). 83–102.

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

Bardovi-Halig, K. 2009. Conventional expressions as a pragmalinguistic resource: Recognition and production of conventional expressions in L2 pragmatics. *Language Learning* 59. 755–795.

Bardovi-Halig, K. 2012. Formulas, routines, and conventional expressions in pragmatics research. *Annual Review of Applied Linguistics* 32. 206–227.

Biber, D. 2009. A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, 14(3). 275–311.

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

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Björkman, B. 2011. Pragmatic strategies in English as an academic lingua franca: Ways of achieving communicative effectiveness?. *Journal of Pragmatics* 43(4). 950–964.

Björkman, B. 2014. An analysis of polyadic English as a lingua franca (ELF) speech: A communicative strategies framework. *Journal of Pragmatics* 66. 122–138.

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

Cogo, A. & M. Dewey. 2006. Efficiency in ELF communication: From pragmatic motives to lexico-grammatical innovation. *Nordic Journal of English Studies* 5(2). 59–93.

Cogo, A. & M. Dewey. 2012. *Analysing English as a Lingua Franca: A Corpus-Driven Investigation*. London: Continuum.

Cogo, A. & J. House. 2017. Intercultural pragmatics. In A. Barron, Y. Gu & G. Steen (eds.), *The Routledge Handbook of Pragmatics*, 168–183. London & New York: Routledge.

Colson, J-P. 2016. IDIOM Search. http://idiomsearch.lsti.ucl.ac.be/index.html.

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

Deroey, K.L.B. 2015. Marking importance in lectures: Interactive and textual orientation. *Applied Linguistics* 36(1). 51–72.

Durrant, P. 2017. Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics* 38(2). 165–193.
Firth, A. 2009. The lingua franca factor. *Intercultural Pragmatics* 6(2). 147–170.
Flowerdew, J. & L. Miller. 1997. The teaching of academic listening comprehension and the questions of authenticity. *English for Specific Purposes* 16(1). 27–46.
Formentelli, Maicol. 2017. *Taking Stance in English as a Lingua Franca: Managing Interpersonal Relations in Academic Lectures*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Granger, S. & P. Rayson. 1998. Automatic profiling of learner texts. In S. Granger (eds.), *Learner English on Computer*, 119–131. London: Longman.
Halliday, M.A.K. (revised by Christian M.I.M. Matthiessen). 2014. *Halliday's Introduction to Functional Grammar*, 4th edn. Oxen: Routledge.
House, J. 2009. Subjectivity in English as Lingua Franca discourse: The case of *you know*. *Intercultural Pragmatics* 6(2). 171–193.
House, J. 2014. English as a global lingua franca: A threat to multilingual communication and translation?. *Language Teaching* 47(3). 363–376.
Hyland, K. 2008. *As can be seen*: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1). 4–21.
Hyland, K. 2012. *Disciplinary Identities: Individuality and Community in Academic Discourse*. Cambridge: CUP.
Kashila, H. & C. S. Heng. 2014. Discourse functions of formulaic sequences in academic speech across two disciplines. *Journal of Language Studies* 14(2). 15–27.
Kecskes, I. 2007. Formulaic language on English Lingua Franca. In I. Kecskes & L. R. Horn (eds.), *Explorations in Pragmatics: Linguistic, Cognitive, and Intercultural Aspects*, 191–218. Berlin: Mouton de Gruyter.
Kjellmer, G. 1991. A mint of phrases. In K. Aijmer & B. Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, 111–127. London: Longman.
Martinez, R. & N. Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33(3). 299–320.
Mauranen, A. 2008. *The Transcriptions of the ELFA Corpus, Downloadable Version* [text corpus]. Kielipankki – The Language Bank of Finland at http://urn.fi/urn:nbn:fi:lb-2014052721.
Mauranen, A. 2009. Chunking in ELF: Expressions for managing interaction. *Intercultural Pragmatics* 6(2). 217–233.
Mauranen, A. 2012. *Exploring ELF: Academic English Shaped by Non-native speakers*. Cambridge: CUP.
Neely, E. & V. Cortes. 2009. A little bit about: Analyzing and teaching lexical bundles in academic lectures. *Language Value* 1(1). 17–38.
Nesi, H. & H. Basturkmen. 2009. Lexical bundles and discourse signaling in academic lectures. In J. Flowerdew & M. Mahlberg (eds.), *Lexical Cohesion and Corpus Linguistics*, 23–43. Amsterdam and Philadelphia: John Benjamins.
O'Donnell, M. 2013. UAM Corpus Tool. Version 3.0.
Pitzl, Marie-Luise. 2012. Creativity meets convention: Idiom variation and re-metaphorization in ELF. *Journal of English as a Lingua Franca* 1(1). 27–55.
Rintaniemi, H. 2017. "Maybe we can just you know see how it's relevant" – The use of *you know* as a discourse marker in academic ELF interaction. Unpublished MA Thesis. University of Tampere.
Schachter, S., N. Christenfeld, B. Ravina & F. Bilous. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* 60. 362–367.
Schleef, E. 2008. The "Lecturer's *OK*" revisited: Changing discourse conventions and the influence of academic division. *American Speech* 83(1). 62–84.

Schneider, N., S. Onuffer, N. Kazour, E. Danchik, M. T. Mordowanec, H. Conrad & N. A. Smith (2014). Comprehensive annotation of multiword expressions in a social web corpus. Proceedings of the 9th Linguistic Resources and Evaluation Conference, Reykjavík, Iceland.

Schnur, E. 2014. Phraseological signaling of discourse organization in academic lectures: A comparison of lexical bundles in authentic lectures and EAP listenting materials. *Yearbook of Phraseology* 5(1). 95–122.

Seidlhofer, B. 2005. English as a lingua franca. *ELT Journal* 59(4). 339–341.

Seidlhofer, B. 2009. Accommodation and the idiom principle in English as a Lingua Franca. *Intercultural Pragmatics* 6(2). 195–215.

Seidlhofer, B. 2011. *Understanding English as a Lingua Franca*. Oxford: OUP.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Staples, S., J. Egbert, D. Biber & B. Gray. 2016. Development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33 (2). 149–183.

Thompson, S. E. 2003. Text-structuring metadiscourse, intonation and the signaling of organization in academic lectures. *Journal of English for Academic Purposes* 2. 5–20.

Wang, Y. 2016. *The Idiom Principle and L1 Influence: A Contrastive Learner-Corpus Study of Delexical Verb + Noun Collocations (Studies in Corpus Linguistics, Volume 77)*. Amsterdam and Philadelphia: John Benjamins.

Wang, Y. 2017. Lexical bundles in spoken academic ELF: Genre and disciplinary variation. *International Journal of Corpus Linguistics* 22(2). 187–211.

Wang, Y. 2018. *As Hill seems to suggest*: Variability in formulaic sequences with an interpersonal function in L1 novice and expert academic writing. *Journal of English for Academic Purposes* 33. 12–23.

Wood, D. 2006. Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *The Canadian Modern Language Review* 63(1). 13–33.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: CUP.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford: OUP.

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

Bionote

Ying Wang

Ying Wang is Lecturer in English Language and Linguistics at Stockholm University. Her main research interests lie in areas of corpus linguistics, with a focus on formulaic language use in different genres.