Corpus Pragmatics (2017) 1:351–372
DOI 10.1007/s41701-017-0016-9

ORIGINAL PAPER

From *Do You Know* to *I Don't Know*: An Analysis of the Frequency and Usefulness of Lexical Bundles in Five English Language Self-Study Books

Rachel Allan[1]

Received: 15 December 2016 / Accepted: 14 April 2017 / Published online: 25 April 2017

Abstract Knowing which phrases to use in everyday situations is a key part of communicating effectively in English, and increasingly language learning materials are expected to reflect this. This paper presents a corpus analysis of five contemporary self-study books for English language learners, to identify common phrases taught, assess their form and function, and evaluate them against a baseline of lexical bundles (i.e. recurring sequences of words) used in social situations by users of English as a Lingua Franca (ELF). Self-study textbooks aim to equip the learner with enough English to function appropriately in a range of different contexts; they usually present language in the form of dialogues in common everyday situations, often supplemented with exercises, grammar explanations and glossaries. While they may differ in pedagogical approach, it could be anticipated that the lexical bundles found would be broadly similar. However, analysis of this corpus showed a lack of consistency both in the form and number of bundles found in the different publications. Furthermore, comparison with a corpus of ELF conversations extracted from the Vienna-Oxford International Corpus of English (VOICE, version 2.0 XML) (2013) highlighted the underrepresentation of lexical bundles with certain pragmatic functions, such as hedges/stance expressions (I don't know, I think) and vague language (a little bit).

Keywords Self-study books · Lexical bundles · English as a Lingua Franca · Corpus analysis · Language acquisition

✉ Rachel Allan
rachel.allan@miun.se

[1] Department of Humanities, Mid Sweden University, 87170 Sundsvall, Sweden

Introduction

Learning commonly-used phrasal language is an important part of gaining fluency from the very first stages of learning a language. Phrases, rather than individual words, tend to be the primary holders of meaning (Sinclair 2008). Some lexical bundles (i.e. recurrent sequences of words) are just as frequent as common single items (O'Keeffe et al. 2007: 46), and there is evidence that knowing word sequences holistically, as chunks, reduces processing demands for the learner (Siyanova-Chanturia et al. 2011; Tremblay et al. 2011). If we accept the position that language recycling in a text, i.e. repeated exposure to target language in a variety of contexts, is more likely to encourage acquisition (e.g. Webb 2007; Nation 2014), not only at word level, but also at phrase level (Durrant and Schmitt 2010), materials that expose language learners to relevant phrasal language will be more effective. Previous studies have examined lexical bundles taught in English language teaching (ELT) coursebooks (e.g. Koprowski 2005; Meunier and Gouverneur 2007; Tsai 2015). In this study, I add to this body of work by examining lexical bundles present in ELT materials marketed for independent study, which I will refer to as self-study books.

Although there are plenty of web-based options and multimedia courses now available, sales figures suggest that self-study books remain a popular way for the casual learner to brush up on their basic English knowledge. Self-study books are an eclectic mix, some following a grammatical syllabus, with others following a thematic one. However, as they tend to cover many of the same topic areas and communicative functions, their lexical content could be expected to be similar. In this study, I examine the number and range of lexical bundles learners are exposed to in a selection of self-study books, to assess both how much exposure to formulaic language these texts give, and to explore the forms and functions present. To provide a point of reference, lexical bundles used by speakers of English as a Lingua Franca in conversations in a similar range of contexts have been extracted from the VOICE (2013) corpus of spoken English, to examine the form and function of phrases typically used.

Formulaic Language and Lexical Bundles

Formulaic sequences of language are word strings which seem to be processed holistically rather than generated word by word (Wray 2002). Although relatively little is known about how these strings of words are processed (for a recent comprehensive review, see Siyanova-Chanturia and Martinez 2014), it is clear that they are fundamental in language production. Corpus analyses have enabled us to see just how prevalent they are, with one study into their use finding that over 50% of fluent spoken and written discourse consisted of formulaic language (Erman and Warren 2000). From a language learning perspective, studies have indicated that non-native speakers who know formulaic phrases display greater processing

efficiency (Tremblay et al. 2011; Conklin and Schmitt 2012). It is, therefore, highly desirable for learners to acquire formulaic language.

The importance of formulaic language is generally acknowledged, but the definitions and terminology used for it vary. Wray (2002) lists over 50 terms applied to these word strings, some of the most common being formulaic sequences (Wray 2002; Schmitt and Carter 2004), n-grams (Stubbs and Barth 2003), lexical phrases (Nattinger and DeCarrico 1992), lexical chunks (O'Keeffe et al. 2007), multi-word expressions (MWEs) (Siyanova-Chanturia and Martinez 2014b) and lexical bundles (Biber and Conrad 1999). Definitions of these terms differ somewhat, but all refer to sequences of language that occur so frequently as to suggest they function as ready-made units, not requiring processing by the user (Sinclair 1991; Wray 2002). In this paper, lexical bundles are identified as "the most frequent recurring sequences of words" (Biber 2006: 132) occurring in the corpora. Following Biber's (2006) approach, they are automatically extracted by the software, and not edited for pragmatic integrity. This means that the bundles may not be syntactically complete; however high frequency word sequences of this kind tend to work as meaningful units, as I will now discuss.

Lexical Bundles and Their Functions

Although many lexical bundles are incomplete, very often they have clear pragmatic functions. The same bundles recur in different types of spoken discourse, and these tend to revolve around the organization and management of conversation and the speaker-listener relationship (O'Keeffe et al. 2007: 75). In the five-million word CANCODE corpus (a corpus of spoken English in a range of contexts of use including casual conversation, workplace and academic settings among different speaker relationships), O'Keeffe et al. (2007) report the top three three-word bundles as *I don't know*, *a lot of*, and *I mean I*, all of which can contribute to the smooth running of conversation in various ways. *I don't know* has a range of functions as both a discourse and stance marker; it may express genuine uncertainty, but is often employed as a face-saving device, both for the speaker and listener (Tsui 1991), to take the floor and to open up the floor to other speakers (Baumgarten and House 2010). *A lot of* offers the potential for vagueness that characterizes much conversation. *I mean I* frames an utterance or its reformulation, and could be used to mitigate the proposition, as a hedge.

Lexical bundles may have a certain degree of consistency across genres, but they can also illustrate features specific to a certain type of discourse. In university spoken discourse, for example, Biber (2006) found that stance bundles accounted for over half of all lexical bundles in all spoken registers except classroom teaching (Biber 2006: 150), where discourse organisers and referential expressions were almost equally common. Furthermore, epistemic bundles (i.e. those commenting on knowledge about the following utterance, e.g. *I don't know if*) were found to be twice as common in conversation than in service encounters, where obligation bundles (e.g. *you have to, you need to*) are most frequent.

There is a growing body of research on use of lexical bundles by speakers of English as a Lingua Franca (ELF). Carey's (2013) study into formulaic organizing

chunks in academic ELF compared data from the English as a Lingua Franca in Academic Settings (ELFA) (2008) corpus of spoken English with that of the Michigan Corpus of Academic Spoken English (MICASE). He found that ELF users appear to store and retrieve interaction-organizing chunks in the same way as English as a native language (ENL) users do, and in fact use the most common bundles with more frequency than native speakers. He also finds some evidence of approximation of form (e.g. *so to say* in place of *so to speak*) in less frequent chunks (Carey 2013: 226). Vague expressions (such as *and so on*, *some kind of*) used in spoken language in an academic context have also been examined (Metsä-Ketelä 2012), with the finding that, overall, vague expressions are used almost twice as frequently by ELF users as by native speakers, although with a narrower range of expressions (Metsä-Ketelä 2012: 263). Similarly, when comparing the pragmatic markers *I don't know* and *I think* in general spoken discourse by ELF speakers and native speakers, Baumgarten and House (2010) found that ELF speakers tended to use *I don't know* in a more literal sense, i.e. for lack of knowledge about something, while L1 speakers used it to mark stance, expressing uncertainty, avoidance, neutrality and non-commitment. *I think,* on the other hand, was the most common way of expressing stance for L2 speakers, much more so than for native speakers.

In a business context, Allan's (2016a) study into lexical bundles in ELF in business meetings using VOICE (2013) found parallels with Handford's (2010) ENL study. Many of the same phrases were identified, but it was the simpler pragmatic markers (e.g. *I don't know, we have to, you can see*) that were identified as key features, while more idiomatic language and opaque hedges found in ENL meetings (e.g. *have a look, in terms of*) were avoided. Echoing the findings of Carey (2013) and Metsä-Ketalä (2012), there was also evidence of a higher reliance on core bundles by ELF users.

To summarise, research to date points to lexical bundles being used as building blocks to structure discourse for ELF users, perhaps to an even greater degree than they are for native speakers. Semantically-transparent bundles are favoured over more idiomatic constructions, but these phrases are, in many cases, identical to those used by native speakers. There is evidence, however, that they tend to be used in a more literal sense, and as such, may not always be used with the same function.

Lexical Bundles in ELT Materials

Although English is taught as a compulsory subject in schools in many parts of the world, many students do not acquire the practical skills needed for everyday communication in an English-speaking environment and turn to self-study for this. Self-study textbooks offer a combination of basic grammar and useful phrases set in everyday contexts, and are a potentially useful resource for such purposes, alongside newer web-based language learning applications. As a kind of hybrid of a phrase book and classroom textbook, they include formulaic phrases for engaging in service encounters, opening and closing conversations, agreeing and disagreeing and other speech acts. English language teaching materials in general have been criticized for not placing phrasal vocabulary in a central role, and failing to introduce it systematically (Koprowski 2005; Gouverneur 2008). Similarly, learner

dictionaries have been criticized for their approach to phraseology, focusing on semantically non-transparent items, while more useful transparent items are often neglected (Siepmann 2008). These transparent bundles often have important pragmatic functions, and tend to be under-used by learners; they have the potential to help learners manage interactions in English. Self-study books may fare better, given that they are likely to be based around transparent, functional phrases; this is what the present study will explore.

Self-study books are produced both by specialist English language teaching publishers and generalist publishers as part of "how to" or "teach yourself" series. This may have implications for the target language they contain, given that the former use corpora to inform their content, while the latter rely on the intuition of the textbook writer. However, it may make little difference at lower proficiency levels, given that it has been found that native speakers have good intuition for selecting frequent lexical items (Ringeling 1984; McCrostie 2007), not only for individual words but also for phrasal language (Siyanova-Chanturia and Spina 2015), although this diminishes at mid-frequency levels.

Research Questions

The specific aim of this study is to identify the lexical bundles learners are exposed to when using self-study books to learn English through the analysis of five contemporary textbooks. These contain both lexical bundles of the target language (*TL bundles*) and bundles associated with pedagogy (*instructional bundles*). The proportions of the two bundle types will be examined. The main focus, however, is on TL bundles, with the following questions explored:

- How many TL bundles occur in these books?
- To what extent are the TL bundles consistent, in terms of form and function, across this range of self-study books?
- How far do the TL bundles used reflect the form and function of common phrases used in similar interactions amongst speakers of English as a Lingua Franca (ELF)?

The first two questions are straightforward; the third requires some explanation. Self-study books contain language that is simplified and has been cleaned up; even when based on corpus data, the dialogues do not demonstrate the reality of co-constructed interaction, which is messy and often difficult to interpret (Clancy and McCarthy 2014). However, the target language has been selected to encourage successful interactions in English, so it should still contain the kinds of bundles learners need to use. To see what these are, I identified bundles used in similar interactions by speakers of ELF i.e. English used as a common means of communication among speakers from different first-language backgrounds. ELF is considered to be the most wide-spread contemporary use of English throughout the world. It is difficult to quantify this usage, but a frequently-cited statistic is that only one of every four users of English is a native speaker (Crystal 2003). Thus, it is

highly likely that users of the self-study books will find themselves communicating with other ELF speakers more frequently than native English speakers. A further justification for using ELF rather than native speaker data as a point of comparison is that ELF tends to be more transparent and accessible. For the target learners of these self-study books, I would argue that this is more useful than exposure to the more idiomatic bundles found in real-life native speaker interactions, although at a higher level of proficiency this may not be the case. Examining real-life interactions among ELF users shows us how they are co-constructed, and gives insights into the linguistic devices used, as I will demonstrate in "ELF Interactions" Section. The bundles extracted from ELF were taken as a point of reference to offer a perspective on the functions needed by speakers to manage similar situations effectively in an English speaking environment, and give a picture of how these tend to be operationalized.

Materials

Self-Study Books

Five self-study books were selected for analysis according to the following criteria. First, the books should be principally for self-study to learn English, focusing on spoken language in use, i.e. not simply a grammar practice book, and they should include some language instruction, making them more than a phrasebook. As far as possible, they should be stand-alone courses, with target language included in the text (although accompanying recorded material may be available). Their target audience should be adults with some knowledge of English, and their instructional language should be in English only. Finally, they should be reasonably up-to-date. This yielded a small collection of books, as outlined in Table 1. All of these books covered essential topics such as getting directions, shopping, and day to day interaction with friends and strangers.

Table 1 Self-study books used, number of words (tokens) in relevant text per book, and sales rank in all books on Amazon.co.uk 4/8/16

Title	Word count	Sales rank
Learning English as a Foreign Language for Dummies (Dudeney and Hockly 2009)	73,465	84,985
Teach Yourself Complete English as a Foreign Language (Stevens 2010)	58,242	78,956
Colloquial English (King 2004)	51,946	3,384,697
Easy Learning English Conversation (Collins Dictionaries 2011)	34,465	54,114
Speaking B1+ Collins English for Life: Skills (Pelteret 2012)	20,867	178,392
Total	238,985	

Although the self-study books have the criteria above in common, they are quite diverse in terms of length and content. Two of the publications are from a series of general self-study books, *Learning English as a Foreign Language for Dummies* (Dudeney and Hockly 2009) (hereafter *EFL for Dummies*), and *Complete English as a Foreign Language* (Stevens 2010), which is part of the *Teach Yourself* series (hereafter *Complete EFL*). *Colloquial English* (King 2004) is also part of a series, but this is restricted to language instruction. *Easy Learning English Conversation* (Collins Dictionaries 2011) (abbreviated to *Easy Conversation*) and *Speaking* (Pelteret 2012) are both published by Collins, which has a long tradition of English language reference publishing. These two books are advertised as being informed by the 4.5 billion word Collins Corpus, while none of the others make any reference to being corpus-informed. In terms of length, as Table 1 shows, *EFL for Dummies* is the longest. This contains more discussion of social and cultural issues, in addition to specific language input. The two shortest books are the Collins publications, which focus principally on speaking. These are advertised as being suitable for both self-study and classroom use.

The books that were identified are not for beginner learners; they target students in the intermediate range i.e. B1–B2 of the Common European Framework of Reference (CEFR) (Council of Europe 2001). *Easy Conversation* states that it "has been specially designed for intermediate-level learners who want to communicate successfully and with confidence in everyday situations, at work, or when travelling or studying", and is labelled B1–B2. *Speaking* is one of a series of books; this one was identified as being for B1 + level, taking the student to the next in the series at B2 level. Similarly, *Complete EFL* and *Colloquial English* claim to take the learner to B2 level of the CEFR by the end of the course. *EFL for Dummies* is the only book that does not make specific reference to level, but claims it is for learners who know "a bit of English".

The books used were all published in Britain and use British English, although they are available worldwide. Their alignment with the CEFR suggests, however, that they are mainly used in Europe. It is difficult to accurately assess the popularity of these books, but their overall sales rank on the Amazon.co.uk, shown in Table 1, gives a snapshot of their sales levels at a particular point in time.[1] The highest-ranking book at this point was *Easy Conversation*, which is the lowest priced of the books, however, *Speaking,* another relatively inexpensive publication, had a much lower sales rank. One of the books is ranked significantly lower than the others, *Colloquial English,* perhaps partially because it was updated with a 2nd edition in 2016 (ranked at 663,542).

The books were scanned and the relevant text put into electronic format using word recognition software, with the resulting files manually checked for error. All language within the core instructional chapters was considered relevant and included, but reference material in other parts of the book, such as indexes, grammar

[1] All books had higher rankings in the *English as a foreign language* category, but were categorised differently making cross-comparisons misleading. A review of the figures in December 2016 showed significant changes to the figures due to seasonal sales, although overall rankings were similar.

resource sections and glossaries were excluded. Total word counts of the text included are shown in Table 1, with contractions counted as a single token.

ELF Interactions

The ELF interactions were extracted from VOICE (2013), a collection of different types of authentic spoken discourse by ELF speakers in different domains. A subcorpus of 36 conversations and 11 service encounters from the three different domains, leisure, professional and business was constructed, making a corpus of 172,969 words. The number of individual speech events and their domain, together with the number of different speakers involved, is shown in Table 2.

These speech event types and domains broadly reflect the types of context represented in the self-study texts, and both contain interactional, conversation-type dialogues and transactional, service-oriented exchanges, although there is a bias towards the former in the VOICE subcorpus. Extracts from the two different source corpora are shown in the examples below. Example (1) is from *Colloquial English*, and (2) from a VOICE service encounter in the educational domain; these show direct transactional discourses. Example (3), from *EFL for Dummies*, is more interactional in nature, as is (4), from a VOICE conversation in the leisure domain.

(1) Nina: Excuse me - could you tell me how to get to the tourist information office?
Passer-by: Hang on… let's see now. Right, go back to the post office and turn right. And then go along the road till you get to a big supermarket. The tourist information office is opposite.
Colloquial English

(2) S2: can you tell me where is er hirschengasse because my my chef lives here i suppose
S1: er that's in the sixth district
[VOICE 2013: EDsve423]

(3) Jacques: It took me an hour and a half to get home last night. The traffic was terrible!
Gill: Tell me about it! I stopped taking the car and now I use public transport - it's much quicker.
EFL for Dummies

(4) S1: to get home now I have to take the night bus or something er so erm difficult
S2: er h-
S2: how do you get home from the city center do you take a u bahn or a tram or
S1: er well er during the daytime now I took to t- er when I came here now I took the number five from er praterstern
[VOICE 2013: LEcon228:92-5]

These examples illustrate some of the differences between scripted dialogue and co-constructed dialogue. In (1), hesitation is expressed using clear phrases, *Hang*

Table 2 Sub-corpus extracted from VOICE. All information taken from the VOICE website (https://www.univie.ac.at/voice/stats/voice20_domains_spets#voice20_spets)

Speech event (type)	Domain	Speech event (N)	Speakers	Words
Conversation	Educational	4	36	28,543
	Leisure	21	101	98,100
	Professional business	1	4	2212
	Professional organizational	3	26	15,360
	Professional research and science	23	23	13,860
Service encounter	Educational	5	30	11,698
	Professional business	6	13	3196
Total		47	233	172,969

on/let's see now, whereas in (2), *er* is used, and in (4), hesitation is evident in the fillers (*well er*), vague language (*or something),* a false start and reformulation (*now I took to t- er when I)*. In addition, the speakers do not clearly end turns, one speaker picks up when the other trails off (*I suppose* in (2)) or invites a turn (*or* in (4)). In contrast, examples (1) and (3) demonstrate the speakers using sentences with clear boundaries between each turn.

Analysis

For the analysis, the effects of length and vocabulary range were first explored for the the self-study books. The top three-word lexical bundles were then extracted from each corpus using AntConc (2016). Three words was considered the most appropriate length to generate sufficient bundles, as four-word bundles are limited in number when text length is restricted, while with two-word bundles it can be difficult to identify functions.

Analysis of Self-Study Books

First the self-study books were analysed. I began with an overview of the lexical content overall, before examining the target language in detail. As indicated in Table 1, the textbooks were of different lengths. They also used different ranges of vocabulary. Although we can normalise figures to make comparisons between the books, it is important to consider the impact these differences have on the bundles in each book. Figure 1 shows the total number of words (tokens) and the total number of different vocabulary items (types) used in each book. The longest book used the greatest range of vocabulary, *EFL for Dummies*, with 5843 different types. This contrasts with *Complete EFL*, which is the second longest book, but uses half the number of types (2866). *Colloquial English* used almost 4000 word types, while the two shorter books had a more limited range of vocabulary, at around 2500.

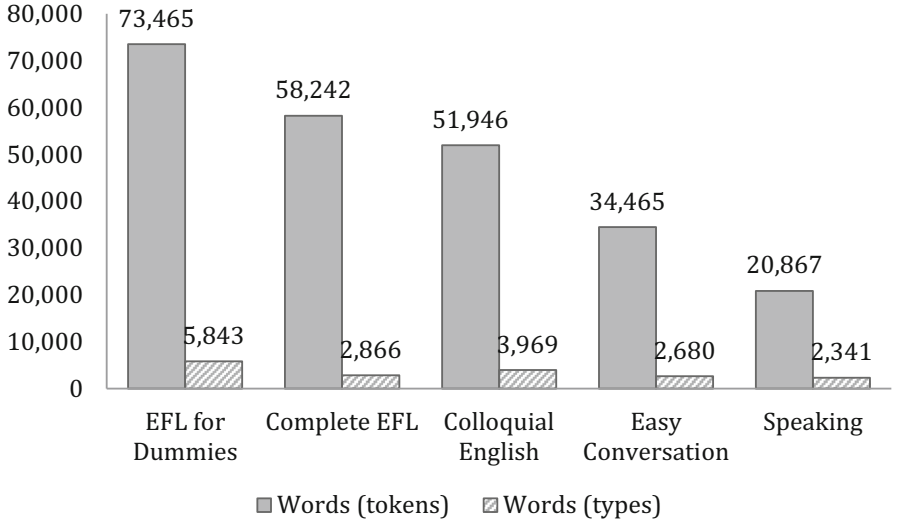

Fig. 1 Word tokens and types in each self-study book

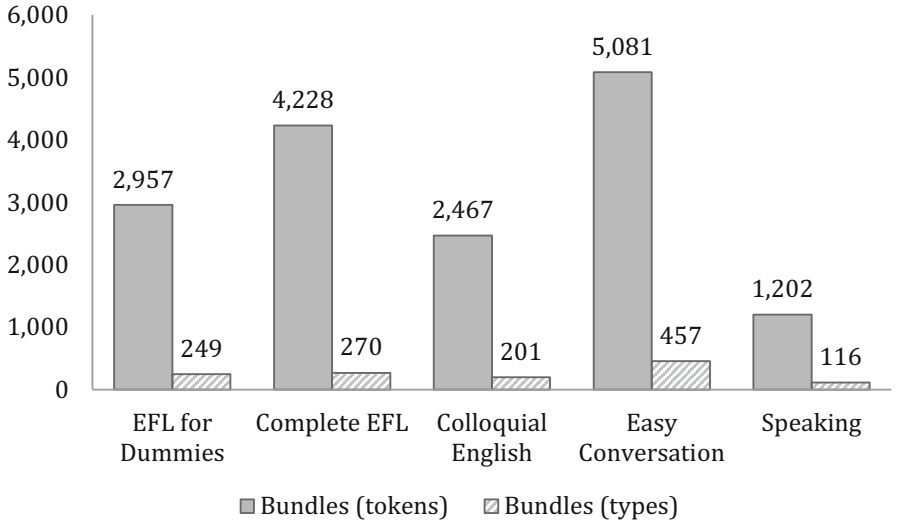

Fig. 2 Raw frequencies of bundle tokens and bundle types in each self-study book

To get an overview of the quantity and range of bundles present, I first identified all three-word bundles occurring at least five times in the book to find the total number of bundles, and the number of types of bundle. Figure 2 displays this information, and, when compared with Fig. 1, clearly shows the effect of text range

on bundle frequency. Two of the books with the more restricted vocabulary, *Complete EFL* and *Easy Conversation*, include more bundles and, interestingly, a wider range of bundles than the other three books. The range of vocabulary included in *EFL for Dummies* means that there is less likelihood of phrases recurring, and there are fewer bundles both in terms of total bundles and types of bundle. *Speaking* has the fewest bundles and types of bundle; this is due to the limited length of the text in comparison to the others. This suggests, then, that for optimal exposure to common bundles, limiting the range of vocabulary used is useful. This is in line with findings on bundle use in graded readers (see Allan 2016b).

To enable comparison of the bundles across the different books and corpora, raw frequencies were normalized to frequency per 10,000 words (f/10,000) of text, and only those bundles occurring at a rate equal to or greater than 2 f/10,000 were then considered. This relatively high cut-off point was used because text length was limited; to achieve this rate of frequency in the shortest of the books, the bundle only had to occur five times. These were sorted into three categories, *TL, instructional,* and *TL* + bundles, using the concordance lines for each phrase to assess their use and function. *TL bundles* are those phrases used specifically as language input, in sample conversations for example. *Instructional bundles* refer to those phrases that were consistently used for instructions, e.g. *check your answers,* in headings, e.g. *Words to know* or in explanations e.g. *in this unit* or *of the verb*. Some bundles had a dual function, as both input and for instruction, and these were classified as *TL* + *bundles*. An example is shown in selected concordance lines for *end of the* in Fig. 3, in which some uses are explanatory (instructional), e.g. "add –ed to the *end of the* verb", and other uses illustrate language relating to directions (TL), e.g. "Go to the *end of the street*". For the TL + bundles, frequencies for TL and instructional use were also recorded at this stage. For example, in Fig. 3, ten of the uses relate to instructions, directions or metalanguage (TL+), while five of them relate to conversations giving directions (TL).

A breakdown of the bundles into these three categories is given in Fig. 4. In each book, instructional bundles make up the majority of the phrases identified, with TL

```
, who) the pitch usually goes down at the    end of the sentence: What are you
       (a form of 'to be') moves to the      end of the question in the
  is very easy: simply add 'ed' to the       end of the verb. Watch out for
  above, the 'ed' that you add to the        end of the word is pronounced
  today.  Often you add a question tag to    end of the weather sentence. So
and predicting what the person on the other  end of the telephone may
       and what the person at the other      end of the phone line may say to
main reason for the phone call. Signal the   end of the conversation, for
  go into the centre of town from the        end of the road.  Akira: Okay,
bring everything over when it's ready. [At the  end of the evening, Piotr
ccommodation available in the UK. At the cheapest  end of the scale are youth
        right. Go past the bus station to the  end of the road and it's the
  [first / second / third] left. Go to the   end of the road. Go [over
  about a five-minute walk. Go to the        end of the street and turn left
       Oscar:  Okay, so first I go to the    end of the street and turn right.
```

Fig. 3 Selected concordance lines on *end of the* from *EFL for Dummies*

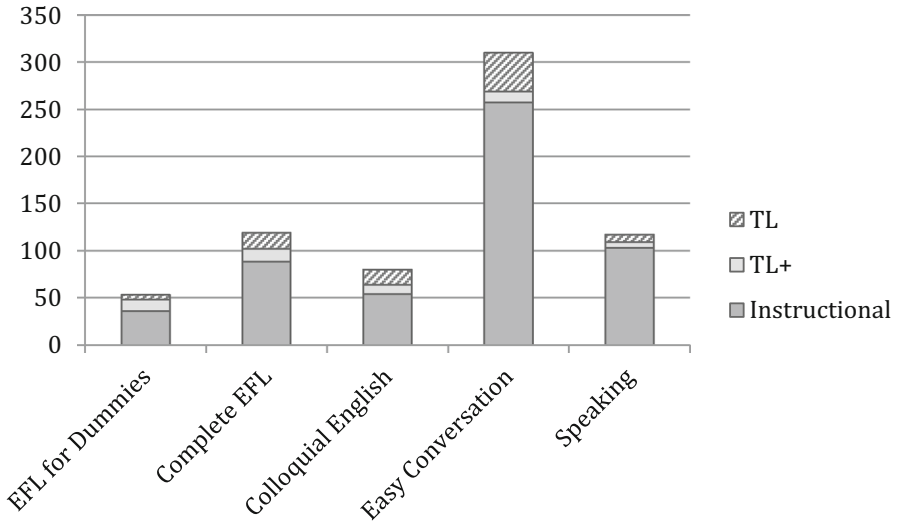

Fig. 4 Proportion of TL, Instructional and TL+ bundles in each self-study book

and TL+ bundles representing only a fraction of the bundles found in the books. It can be seen that *Easy Conversation* has the greatest number of bundles overall, containing 310 bundles in total, 43 of which were TL and TL+ bundles, compared to only 14 in *Speaking* (largely due to the limited length of the text and the cut-off point) and 53 in total (17 TL/TL+) in *EFL for Dummies*, the longest book. A total of 80 bundles are used in *Colloquial English,* 26 of which relate to input, and 119 are found in *Complete EFL*, 31 of these are TL/TL+ bundles.

From this point on in the analysis, the instructional bundles are disregarded and only the TL uses of TL+ bundles are considered. Although learners might benefit from additional exposure to the TL bundles due to their inclusion in an instructional capacity, the main aim of this study is to establish what the self-study books use as input for spoken language. This reduces the number of TL bundles, as some of the TL+ bundles do not meet the frequency criteria when only the input uses are considered. Table 3 shows the 20 most frequent TL bundles occurring when the five self-study books are treated as a whole. This shows that *would you like* at the top of the list, occurring at a much higher frequency than all other bundles, 8.9 f/10,000, followed by *you tell me* at 5.8 f/10,000. The remaining bundles occur at similar levels, ranging from *have you got* at 3.9 f/10,000 to *you going to* at 2.4 f/10,000. Before discussing these further, however, it is important to take note of the individual differences between the five individual books.

Table 4 shows the top bundles of input language at the required frequency level in each of the self-study books, limited to ten where there were more than this. As previously shown in Fig. 4, there is a different number of bundles in each book, and once only the input TL bundles are considered, these gaps widen. *Easy Conversation* contains most at 44, while *EFL for Dummies* and *Speaking* contain only six and ten respectively. Between these extremes are *Colloquial English* (17)

Table 3 Top 20 TL bundles occurring in the self-study books overall

Bundle	Frequency per 10,000 words
would you like	8.9
you tell me	5.8
have you got	3.9
I'm going to	3.8
go to the	3.7
could I have	3.5
you like to	3.3
I'd like to	3.1
what do you	3.1
do you have	3.0
do you know	3.0
are you going	2.9
could you tell	2.9
do you like	2.8
to go to	2.7
I have a	2.6
would like to	2.4
do you think	2.4
I need to	2.4
you going to	2.4

and *Complete EFL* (19). The range of frequencies varies a great deal between the different books. The bundles in *Easy Conversation* occur with much greater frequency than all the other books, with the greatest range (between 16.8 and 2.6 f/10,000) and highest mean (6.3 f/10,000). The book with bundles occurring least frequently is *EFL for Dummies,* ranging between 4.1 and 2, with an overall mean of 3.1 f/10,000. The other books contain bundles with frequencies at various points within this range.

On examining the forms of the bundles, the first striking finding is that no single bundle is common to the top ten bundles in each self-study book, and if all the input bundles at the required frequency level in each book are considered, only three (*would you like, go to the* and *I'm going to*) are common to four of them. Only two bundles are common to three of the books (*do you like* and *what do you*), whereas 14 of the same bundles occur in two books. Furthermore, each of the books contains a number of bundles unique to it, ranging between four and 32 bundles. There is, then, a very limited degree of consistency across the different books in terms of the TL bundles. The notable difference between the corpus-informed and non-corpus informed books is that one of the former, *Easy Conversation,* has a higher number of TL bundles, occurring with higher frequencies. It is difficult to evaluate the bundles in *Speaking* due to the limitations stemming from text length; most of the bundles are based around the set phrases *I'm really sorry* and *I know what you mean*, which occurs eight times in this short book, giving it a frequency of 3.8 f/10,000.

Table 4 Top bundles (limited to 10, total number of bundles occurring shown in parentheses) in each of the self-study books, with f/10,000 reported and bundles most common across the books shown in bold

EFL for Dummies (6)		Complete EFL (19)		Colloquial English (17)		Easy Conversation (44)		Speaking (10)	
I'd like to	4.1	would you like	20.6	would you like	8.7	do you have	16.8	I'm really sorry	5.3
would you like	**3.9**	you tell me	10.8	have you got	4.6	**I'm going to**	**12.2**	what do you	5.3
I help you	3.4	do you know	5.8	do you like	3.9	I have to	11.6	what you mean	4.8
can I help	2.9	**go to the**	**5.8**	**go to the**	**3.9**	can I have	11.3	I know what	4.3
do you have	2.3	would like to	5.7	I don't know	3.9	are you going	10.2	know what you	4.3
go to the	**2.0**	have you got	5.3	**I'm going to**	**3.9**	I need to	10.2	**I'm going to**	**3.4**
		you would like	5.3	I don't think	3.7	you going to	9.9	have you been	2.9
		do you like	4.5	why don't you	3.7	I'd like to	9.3	do you know	2.4
		I don't mind	3.6	do you think	3.1	could I have	9.0	you want to	2.4
		you like to	3.6	do you want	3.1	do you think	9.0	don't want to	2.0

Among the non-corpus-informed books, there is greater consistency in the TL bundles appearing at the top of the list, e.g. *would you like, go to the;* however it should be noted that these phrases occur in *Easy Conversation* with equal or greater frequency, but are ranked lower on the list. A number of other bundles ranked near the top of the lists for the other books are similarly ranked lower in *Easy Conversation,* although they occur with similar frequencies. This indicates some degree of consensus among the writers regarding what students should know, whether corpus-informed or not.

Most of the bundles in the self-study books occur as part of the routines of offering and requesting, and expressing intention and desire. Some of them have one clear function, e.g. *I'd like to* to express desire and *do you like* to ask direct questions, while *go to the* has a largely grammatical role (although it has a pragmatic function when used as an imperative). Some of the phrases, however, are used in a range of contexts, as Table 5 illustrates, and are extended beyond a single form to function relationship. The bundles included in Table 5 are high frequency

Table 5 Selected examples showing contexts and functions of common bundles

Bundle	F/ 10,000	Sample contexts of use	Function in given context
would you like	8.9	*Would you like a drink/something to eat*	Offering/inviting
		When would you like travel?	Clarifying
		Would you like an aisle, middle or window seat?	
		Would you like to pay?	Asking politely
have you got	3.9	*Have you got the time/a second/any change?*	Asking for information
		What have you got to lose?	Encouraging
I'm going to	3.8	*I'm going to tell him I can't come*	Expressing intent
		Sorry, I'm going to be late	Leave-taking
		Right, I'm going to need to see two forms of identification for that	Mitigating obligation
what do you	3.1	*What do you think/reckon/recommend?*	Asking opinion
		What do you mean?	Clarifying
do you have	3.0	*Do you have to rush off quite so soon?*	Asking about obligation
		Do you have the extension number?	Asking for information
		Do you have any kids, then?	Developing conversation
do you know	3.0	*Do you know Chris?*	Introducing people
		How do you know Jo?	Developing conversation
		Do you know what's on?	Asking directly
		Do you know how much this is, please?	Asking indirectly
do you think	2.4	*What do you think of Helen?*	Introducing topic/asking opinion
		What do you think about animal testing?	
		Do you think we should leave a tip?	
		Do you think you could not smoke in the house?	Asking politely
		Do you think I could see one of the doctors?	

(i.e. occur in Table 3) and occur in at least two of the self-study books. Where there is overlap, e.g. *would you like* and *you like to*, the bundle used with a broader range is shown. Phrases are used to ask indirect questions, with *do you think* and *do you know*, and the same phrases are used as frames for opening conversations, with *what do you think/how do you know*. *I'm going to* demonstrates intent, but is also used for indirect leave-taking (*sorry, I'm going to be late*) and also more unusually as a downtoner in *I'm going to need to see your passport,* mitigating the force of the demand. Idiomatic use of language is included with *have you got,* both in genuine questions *have you got the time?* and the fixed idiom *what have you got to lose?* There is, then, a range of uses of these phrases, but it can be seen that they are principally used to ask questions, to offer or suggest and to start and develop conversation.

Comparison with ELF Bundles

For the VOICE subcorpus, the same frequency requirement was applied (2 f/10,000), and a range restriction was implemented, that the bundle should occur in at least 12 (approximately 25%) of the interactions to ensure that it was not domain specific, or due to a particular speaker. Bundles which included repetitions (*yeah yeah yeah)* and words that clearly demonstrated hesitations (*er, erm*) were excluded, as although these clearly have a pragmatic function, they are not lexical items that would be considered for teaching purposes. This resulted in 15 bundles at the required frequency, as shown in Table 6. At the top of the list, *I don't know* occurs

Table 6 Lexical bundles found in VOICE subcorpus, cross-referenced with those occurring in the self-study book corpus (SSB corpus) and in each individual self-study book

Lexical bundle in ELF	f/10,000	Range	SSB corpus 2 f/10,000	Occurs in self-study book at 2 f/10,000
I don't know	12.3	39	x	Colloquial English, Complete EFL
you have to	7.2	34	x	
a lot of	4.9	22	x	
and then you	3.3	24	x	
I have to	3.3	22	x	Easy Conversation
do you know	2.8	19	✔	Complete EFL, Speaking
I think it's	2.5	25	x	
do you have	2.4	23	✔	EFL for Dummies, Easy Conversation
a little bit	2.2	24	x	
go to the	2.2	22	✔	Colloquial English, Complete EFL, Easy Conversation, EFL for Dummies
I don't think	2.2	17	x	Colloquial English, Easy Conversation
there is a	2.2	24	x	
to go to	2.1	22	✔	Easy Conversation
you want to	2.1	22	x	Colloquial English, Speaking
and then I	2.0	21	x	

at a comparatively high frequency level (12.3 f/10,000), with the next on the list, *you have to* occurring at a much lower rate (7.2 f/10,000), a further drop with *a lot of* (4.9 f/10,000) and the remaining bundles occurring close to the cut-off point. These are typical of the kinds of semantically transparent bundles found in other studies of ELF discourse. Comparing these with the most frequent bundles in the self-study books viewed as a single corpus, there are four phrases in common: *do you know, do you have, go to the,* and *to go to.* These bundles are used in much the same way in both discourse types, except for an additional pragmatic use for *do you know* where there are examples in the ELF data of it having a discourse-marking function, in *do you know what I think/why I think* and *do you know what I mean.* If we look at the text of each individual self-study book, each one has different bundles in common with the ELF bundles, with *English Conversation* having the most bundles (5), and *Speaking* having least (2), as shown in the right-hand column of Table 6.

Moving on to examine the functions of the bundles found to be more common in ELF, while the principal function may be the same as in the self-study books, there are additional uses, not found or found only to a very limited degree in the books. *I don't know,* for example, is only used in the sense of lacking knowledge or certainty in the self-study books, and while this may also be its main use in ELF (c.f. Baumgarten and House 2010), the examples in Table 7 show that it is also used to allow thinking time, to hedge an utterance and to add further vagueness to a

Table 7 Functions of selected top bundles from ELF interactions

Function	Examples of uses found in VOICE
Reformulating	*you can do that but then you're never yeah then you have to have a mino-* ***I don't know*** *a min- you have a minority government you can have that*
	yeah but ***I think it's I think it's*** *nice there it's new I think*
Hedging	*I am not going to fail* ***I don't know*** *about you guys*
	w- well the thing is and er and er Finland has been complaining about it ***a little bit*** *it is diffi- too difficult to find er hh good applications of this er disk formulas there're not so many*
	S1: I will change your mind that's my mission
	S2: @@@@ xxxxx but ***I don't think*** *you'll succeed*
	I think it's *what what I thought was dulce de le- dulce de leche*
Vagueness and approximation	*exactly and they had* ***a little bit*** *of like green stuff on top i guess they put some er coriander or something on top just to make it a bit*
	which have erm eight per cent which is ***a little bit*** *less than the er a- about the same now as the german greens*
	our conservative wants to liberalize ***a lot of*** *stuff but the liberals want to just be bigger and to like vote us*
	sounds the same we have a we can understand each other but we have ***a lot of*** *things er so different*
	made with some sort of sauce I think well ***I don't know*** *anyway*
Discourse marking	*that's a bit silly in Italy you have to go through three years training* ***and then you*** *sit the final exam* ***and then you*** *become an architect*
	and went inside ***and then I*** *thought er shit it's not the guy I saw yesterday and he actually went into the building and I thought*

suggestion. Other such VOICE bundles are classified and exemplified in Table 7. These uses work to promote the smooth running of the interaction and avoid it breaking down; the phrases allow time for recall, and offer ways of hedging and approximating to protect face. Simple discourse markers like *and then you/I* enable longer turns and facilitate the recounting of a narrative; they reflect the nature of unscripted spoken discourse, phrases strung together with simple connectors.

These uses, and in some cases, the bundles themselves receive comparatively little attention in the self-study books. This can be accounted for in various ways. Some functions are realised in different ways; hedging, for example, is present in the indirect question frames *do you know/do you think* found at high frequency in the self-study books. This is a reminder of the earlier point made about language register; it is important to acknowledge the more casual nature of many of the conversations in VOICE. Other features of real-life interaction, like the phrases used to facilitate hesitation or to be vague, traditionally have not been taught and may be considered undesirable as target language. The discourse markers found, used in ELF to string together narratives, are infrequent in the self-study books perhaps because there are fewer longer "spoken" turns represented in the books, but largely because the turns are portrayed as sentences rather than utterances.

Discussion

Returning to the research questions posed, the first question asked how many lexical bundles were found in self-study books. This was found to be dependent on text length and complexity, and consequently the results from a range of individual self-study books, all aimed at a similar level, varied greatly. One of the two corpus-informed books stood out as having many more bundles than others; limiting text length to some degree and restricting the number of word types contributed to this. Severely restricting text length, even with a narrow vocabulary range, does not provide sufficient opportunities to recycle the language, as the other corpus-informed book illustrates. At the other end of the spectrum, the longest book containing the widest-ranging vocabulary contained the fewest bundles.

The second question referred to the degree of consistency of the TL bundles across the different books. First, in each of the books, a large majority of the bundles found were connected with instruction-giving. Among the bundles specifically connected with providing target language, there was some consensus on a core set of bundles, but overall there were more differences than similarities. No single bundle was found to be common to all of the books, and all of the books contained bundles unique to it. Functionally, however, there was relatively little difference in the bundles identified. These covered a range of functions, mainly connected with basic conversational routines in semi-formal situations, such as asking for information or opinions, and initiating and developing social interactions, with occasional more idiomatic uses identified.

Finally, I considered the bundles in relation to ELF language in similar social interactions. The results showed overlap in both form and functions of bundles, but the overall impression created by the most frequent bundles is quite different. The

ELF bundles mainly demonstrate interactive features of conversation, which are largely absent from the self-study books. As noted in "Research Questions" Section, the ELF bundles were identified as a reference point only; it would be simplistic to suggest that they represent the bundles that should be taught in the books. However, I would argue that they may be useful for informing the writers of self-study books. When we look beyond the hesitations and repetitions, there are certain bundles that ELF speakers seem to rely on for pragmatic functions such as hedging and vagueness. It would be useful for language learners to be made aware of these bundles, since they represent another way of being indirect. Adding these to the more structured forms of indirectness already included in the self-study books would give a fuller picture of how politeness is realised.

If the premise that exposure to and recycling of target language have a positive effect on acquisition is accepted, the present study has clear implications for self-study materials. Lexical bundles are found at higher frequency levels in those books that restrict vocabulary range and text length to some degree. There is a point at which restriction limits bundles (see also Allan 2016b), and it would be useful to determine clear guidelines on this. As a general rule, however, self-study books could follow the same principle as graded readers, taking a structured approach to limiting vocabulary size and range in order to give higher exposure to target lexical items, including lexical bundles.

The perspective offered by ELF bundles relates to the types and functions of lexical bundles included. The self-study books examined here tend to focus on bundles which err on the more formal side; this is not necessarily a criticism as the kinds of polite and indirect phrases included are undoubtedly useful in helping speakers interact effectively. However, these could be balanced with some input on those bundles that ELF speakers rely on, as these are another important way that issues of face and politeness are addressed. Raising awareness of bundles used for features like hedging and vagueness would not only offer learners an insight into the nature of co-constructed interaction but also give them a concrete means of negotiating it. In other words, it would equip them with the same kind of building blocks that ELF users are known to use in their discourse. Most recently-published ELT texbooks for the classroom now include a focus on real-life language features like these, and some newer courses for self-study are following the same trend[2]. It is hoped that self-study books of a more traditional type, as analysed here, will take a similar approach.

A final point to note is that this study examines only one aspect of the self-study books; it is not an overall evaluation of them. Although the books have a similar overall aim and level, they approach the subject matter in different ways. Some, like *Complete EFL,* include more repetition of the same dialogues, recycling them in exercises, leading to a higher bundle count. In contrast, *EFL for Dummies,* contains fewer bundles quantitatively, but explicitly highlights certain phrases, with a chapter devoted to phrases "that make you sound fluent in English", including such

[2] See, for example, the *Touchstone* series (McCarthy et al. 2014), with a textbook and online course informed by and focusing on natural language in authentic contexts, with explicit attention paid to conversation strategies.

bundles as (*do you) see what I mean, the thing is,* and *you know what.* The present study does not take account of such features, other than including these bundles in the overall processing of the books.

Conclusion

In this study, five self-study books were analysed for their lexical content in general and for lexical bundles in particular. There were notable differences in the quantities of bundles present in the books and in the forms of the lexical bundles found. In all of the books, most bundles related to language instruction, with only a small minority of target language input bundles. These were not consistent across the five books, although the functions of the bundles were broadly similar. The bundles tended to be relatively formal in style and related to everyday interaction; opening, developing and closing conversations, and requesting, offering and suggesting. When bundles from ELF interactions in similar contexts were examined, a number of bundles and related functions stood out as being missing from the self-study books. These were more informal bundles, and largely related to management of conversation and extended turns. The point should be made here that some of the ELF conversations were of a more informal nature, which may account for some of the differences. As I have also pointed out, ELF is not being suggested as a general model for input. Nevertheless, the ELF bundles highlight both the functions needed and the language used by non-native speakers operating in real-life English-medium environments, and it seems reasonable to suggest that self-study books would benefit from being informed by this.

Acknowledgements This study was funded by a grant from Mid Sweden University. I am grateful to the publishers of the self-study texts for their support of this study, and to the reviewers of this paper for their assistance in improving it.

Compliance with Ethical Standards

Conflict of interest The author declares that she has no conflict of interest.

References

Allan, R. (2016a). Lexical bundles in ELF business meetings. *The Linguistics Journal, 10,* 141–163.

Allan, R. (2016b). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59,* 61–72. doi:10.1016/j.system.2016.04.005.

Anthony, L. (2016). *AntConc (Version 3.4.3) [Computer Software]*. Tokyo: Waseda University.

Baumgarten, N., & House, J. (2010). I think and I don't know in English as lingua franca and native English discourse. *Journal of Pragmatics, 42*(5), 1184–1200. doi:10.1016/j.pragma.2009.09.018.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
Biber, D., & Conrad, S. (1999). Lexical bundles in conversations and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.
Carey, R. (2013). On the other side: formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca, 2*(2), 207–228. doi:10.1515/jelf-2013-0013.
Clancy, B., & McCarthy, M. (2014). Co-constructed turn-taking. *Corpus pragmatics: A handbook* (pp. 430–453). Cambridge: Cambridge University Press.
Collins Dictionaries. (2011). *Easy learning English conversation*. Glasgow: HarperCollins Publishers.
Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32,* 45–61. doi:10.1017/S0267190512000074.
Council of Europe. (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
Crystal, D. (2003). *English as a global language* (2nd ed.). Cambridge: Cambridge University Press.
Dudeney, G., & Hockly, N. (2009). *Learning English as a foreign language for dummies*. Chichester: Wiley.
Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163–188. doi:10.1177/0267658309349431.
ELFA. (2008). The Corpus of English as a Lingua Franca in academic settings. Director: Anna Mauranen. http://www.helsinki.fi/elfa/elfacorpus.
Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text - Interdisciplinary Journal for the Study of Discourse, 20*(1), 29–62. doi:10.1515/text.1.2000.20.1.29.
Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 223–246). Amsterdam: John Benjamins.
King, G. (2004). *Colloquial English: A course for non-native speakers*. Abingdon: Routledge.
Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal, 59*(4), 322–332. doi:10.1093/elt/cci061.
McCarthy, M., McCarten, J., & Sandiford, H. (2014). *Touchstone* (2nd ed.). Cambridge: Cambridge University Press.
McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *RELC Journal, 38*(1), 53–66. doi:10.1177/0033688206076158.
Metsä-Ketelä, M. (2012). Frequencies of vague expressions in English as an academic lingua franca. *Journal of English as a Lingua Franca, 1*(2), 263–285. doi:10.1515/jelf-2012-0019.
Meunier, F., & Gouverneur, C. (2007). The treatment of phraseology in ELT textbooks. In H. Encarnacion, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom. Selected papers from the sixth international conference on teaching and language corpora (TaLC6)* (pp. 119–139). Amsterdam: Rodopi.
Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9000 words? *Reading in a Foreign Language, 26*(2), 1–16.
Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.
Pelteret, C. (2012). *Speaking: B1+ (Collins English for Life: Skills)*. London: HarperCollins Publishers.
Ringeling, T. (1984). Subjective estimations as a useful alternative to word frequency counts. *Interlanguage Studies Bulletin, 8*(1), 59–69.
Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 1–22). Amsterdam: John Benjamins.
Siepmann, D. (2008). Phraseology in learners' dictionaries. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 185–202). Amsterdam: John Benjamins.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Sinclair, J. (2008). The phrase, the whole phrase, and nothing but the phrase. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407–410). Amsterdam: John Benjamins.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: the role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 37*(3), 776–784. doi:10.1037/a0022531.

Siyanova-Chanturia, A., & Martinez, R. (2014). The idiom principle revisited. *Applied Linguistics*. doi:10.1093/applin/amt054.

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning, 65*(3), 533–562. doi:10.1111/lang.12125.

Stevens, S. (2010). *Complete English as a foreign language: Teach yourself*. London: Hodder & Stoughton.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language, 10*(1), 61–104.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). processing advantages of lexical bundles: evidence from self-paced reading and sentence recall tasks. *Language Learning, 61*(2), 569–613. doi:10.1111/j.1467-9922.2010.00622.x.

Tsai, K.-J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research, 19*(6), 723–740. doi:10.1177/1362168814559801.

Tsui, A. B. M. (1991). The pragmatic functions of I don't know. *Text, 11*(4), 607–622.

VOICE. (2013). The Vienna-Oxford International Corpus of English. (version 2.0 XML).

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65. doi:10.1093/applin/aml048.

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