Vilkaitė, L. Formulaic language is not all the same: comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji kalbotyra* 2016 (8), www.taikomojikalbotyra.lt

Formulaic language is not all the same: comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs

Laura Vilkaitė

University of Nottingham

laura.vilkaite@nottingham.ac.uk

Abstract

Formulaic language is widely acknowledged to be a central part of a language. However, it is heterogeneous in nature, made up of various formulaic categories with their own characteristics and behaviour. A first step towards systematically describing the relationship between these categories is to describe their distribution in language. This study investigated the frequency of occurrence of four categories of formulaic sequences: collocations, phrasal verbs, idiomatic phrases, and lexical bundles. Together the four categories made up about 41% of English, with lexical bundles being by far the most common, followed by collocations, idiomatic phrases and phrasal verbs. There were differences in the frequencies of each category in the overall corpus, and also in the four registers analysed (academic prose, fiction, newspaper language, and spoken conversation). Language mode (spoken/written) had a substantial effect on the frequency distribution of the categories as well.

Keywords: vocabulary studies, formulaic language, collocations, phrasal verbs, idiomatic phrases, lexical bundles.

1. Introduction

One of the most interesting discoveries in applied linguistics in the last three decades, driven by corpus linguistic analysis, was establishing that language is not typically generated word-by-word, but rather it is largely formulaic in nature (Barlow 2011). Starting with early discussions by scholars such as Pawley and Syder (1983) and Sinclair (1991) (open-choice principle vs idiom principle), the area of formulaic language[1] has exploded, generating a vast number of research studies and books (e.g. Schmitt 2004; Barfield & Gyllstad 2009; Lewis 1997; Nattinger & DeCarrico 1992; Wray 2002, 2008). There is now a broad consensus on the importance of

[1] In this study, following Schmitt (2010), the term *formulaic language* is used as the cover term for the phenomenon of formulaicity, and *formulaic sequences* for the individual phrasal items.

formulaic language. Pawley and Syder (1983) claimed that native speakers have thousands of memorized sequences in their mental lexicon. Formulaic sequences have been shown to realize various functions in language: they are very important in organizing discourse (Nattinger & DeCarrico 1992), aid and speed up language processing both for speaker and for hearer (Siyanova-Chanturia & Martinez 2015), help to convey very precise meanings (Schmitt 2010), and show one's willingness to identify themselves with a certain speech community (Wray 2008). Therefore it is not surprising that formulaic language is ubiquitous, with estimates of its scope ranging up to over 50% (Erman & Warren 2000).

However, despite the widespread acknowledgement of its importance, formulaic language has proven to be one of the more unruly concepts in applied linguistics. Wray (2002: 9) famously found over 50 terms to describe the notion that recurrent multi-word lexical items can have a single meaning or function. Some highlight the characteristic of multiple words (*multi-word units, multiword chunks*), others the fixedness of the items (*fixed expressions, frozen phrases*), some the recurrent phraseology (*phrasal vocabulary, routine formulas*), while still others focus on the psycholinguistic notion that these multi-word lexical items are stored and processed in the mind as wholes (*chunks, prefabricated routines*). Clearly, researchers are conceptualizing and defining formulaic language in different ways, even though they are studying the same phenomenon. Many also have the habit of using the various terms interchangeably. Various attempts to categorize formulaic sequences have been made (e. g. Granger & Paquot 2008; Moon 1998a, 1998c; Nesselhauf 2003), but different interpretations exist and different researchers make their claims based on different definitions. This leads to the problem raised by Wray (2012: 237): "[t]here is enough contrast in the research already described for it to be evident that different researchers are not all talking about precisely the same thing". Therefore, she warns against introducing new terms to define the same phenomena and overgeneralizing the studies which focused on a certain type of sequence to formulaic language overall.

One way to begin addressing this untidy state of affairs is to start systematically looking at formulaic language as a heterogeneous phenomenon. Sequences that can be defined as formulaic vary considerably in nature; they vary in frequency, fixedness, semantic transparency, length, and many other criteria. They realize different meanings and serve different functions. It also seems that they might even be processed and acquired differently (Siyanova-Chanturia, Conklin, & van Heuven 2011). Even though *formulaic language* as an umbrella term is very useful for making the distinction between language that is phrasal/pre-constructed and language that is newly created, it seems that in order to better understand formulaicity, we have to focus on the fact that formulaic language is not all the same. In fact, formulaic language consists of a number of categories (e. g. idioms, clichés, proverbs, collocations, lexical bundles, and phrasal verbs) each with their own characteristics and behaviour.

The first step in better understanding the nature of formulaic language seems to be a descriptive study suggesting the general trends of distribution of the various categories of formulaic sequences. This study aims to provide this description, by focusing on four categories of formulaic language which have been widely discussed in previous research: collocations, phrasal verbs, idioms, and lexical bundles. It will involve an intensive analysis of these four categories in order to estimate their distribution in English overall, and also in various registers.

2. Estimates of the percentage of formulaic language in English

Various studies have tried to estimate the overall percentage of formulaic language in English. One of the most widely-cited studies was carried out by Erman and Warren (2000). They manually analysed 19 extracts of 100-800 words and estimated that formulaic sequences (‘prefabs’) make up to 58.6% of spoken texts and 52.3% of written texts. More recently, Wei and Li (2013) suggested an improved MI measure to extract collocations and longer n-grams statistically from corpora. They tested this new method on a corpus of academic English and found that 2–6 word sequences extracted from corpus covered 58.75% of their corpus. This figure is surprisingly similar to Erman and Warren’s, even though Wei and Li only studied academic language and extracted the sequences automatically. Other studies have also tried to determine the percentage of formulaic language in English (see, for example, Altenberg 1998; Foster 2001; Sorhus 1977), suggesting various estimates, which varied due to different definitions being adopted. The general tendency seems to be that “at least one-third to one-half of language is composed of formulaic elements” (Conklin & Schmitt 2012: 72). Hence, the overall percentage of formulaic language has been established to be high. However, what all these studies have in common is that they do not try to break the percentage of formulaic language into its component parts, and suggest only an overall percentage of formulaic language based on definitions which vary from study to study. This study sets out to address this limitation by estimating the percentages of four types of formulaic sequences.

3. Register and formulaicity

It has often been claimed that the use of formulaic sequences varies based on mode, topic, genre and individual speaker (Schmitt 2010). However, the studies that tried to estimate the overall percentage of formulaicity in language, typically did not take register information into account (apart from the distinction between spoken and written language), which may have obscured any register-specific trends. Indeed, studies looking into the use of individual categories (Biber et al. 1999; Liu 2011) have demonstrated decided register differences. Therefore, this paper will address the question of how registers vary in terms of their formulaicity. The definition of register chosen for the paper is “any language variety defined in terms of a particular constellation of situational characteristics” (Biber & Conrad 2001: 3). Following Biber (2006) the term ‘register’ will be used in a very broad sense, without implying any specific distinction between ‘genre’ and ‘register’. Four major registers will be analysed: academic prose, fiction, newspaper language and spoken conversation. These four registers have previously been the focus of many studies looking at grammatical and lexico-grammatical variation in language (Biber, Conrad, & Cortes 2004; Biber et al. 1999; Conrad & Biber 2004), and this study will contribute a formulaic element to that knowledge base.

4. Categories of formulaic language: Definitions and behaviour

There are a number of overlapping categories of formulaic language. Some have been mentioned in passing in discussions of formulaic language, but have received very little principled research from the applied linguistics perspective. Clichés and proverbs seem to fall into this group. Clichés could be almost any kind of (usually transparent) repeated phrase, with the key distinguishing characteristic being that they have been used so often to reflect an idea that they become unfashionable. The notion of ‘overuse’ is difficult to establish empirically, so it is not surprising that little research has focused on this category. Proverbs exemplify aspects of folk wisdom. There is no definitive study of the number of proverbs in English (although the *Oxford Dictionary of Proverbs* (Simpson & Speake 2008) lists over 1,100 of the most widely-used ones), and little research into how and when they are used in practice.

Other categories of formulaic language have benefitted from limited amounts of research. For example, lexical phrases were highlighted by Nattinger and DeCarrico (1992) as formulaic sequences with pragmatic functions. But their formal criteria were so broad that the category is hardly definable beyond the functional element (e.g. *How do you do?, I'll say, what on earth, a* time marker *ago, that reminds me of X* were all considered lexical phrases), and so *lexical phrase* often came to be used as an umbrella term for formulaic language rather than for a functionally-based category. Binomials have been a useful lexical target for psycholinguistic studies into the holistic storage of formulaic language (e.g. Siyanova-Chanturia, Conklin, & van Heuven 2011), but we know little about the extent and behaviour of binomials in general language use.[2] There are also more variable expressions considered to be formulaic and addressed in research. Renouf and Sinclair (1991) suggested the idea of collocational frameworks, which are patterning of grammatical words, such as *a ..... of* or *too ... to*. While this idea was shown to be relevant to other languages (e.g. Butler (1998) studied collocational frameworks in Spanish) and in specific registers (e.g. Marco (2000) analysed collocational frameworks in medical research papers), with the exception of these few studies, there has not been much research on this category. More recently Cheng et al. (2009) suggested the idea of meaning shift unit, which is based on the idea that a co-selection is meaningful and that the co-selected words get a new meaning. While meaning shift units were addressed in a few corpus studies (e.g. Cheng et al. 2006; Cheng 2009; O'Donnell et al. 2012), we still do not have evidence about their psychological validity.

There have been, however, a small number of categories which have been the objects of substantial research, with the main ones being collocations, phrasal verbs, idioms, and lexical bundles. As the present study will feature these four better-researched categories, they will be discussed briefly below.

4.1. Collocations

"Collocations (...) are associations between lexical words, so that the words co-occur more frequently than expected by chance" (Biber et al. 1999: 998). They have been identified by two distinct approaches: the phraseological approach requires a collocation to meet certain meaning or restrictiveness criteria, while the statistical approach measures frequency of co-occurrence and statistically-measured mutual expectancy (Barfield 2012). The approach adopted obviously leads to different items being classified and counted as collocations. In this study, a corpus-based statistical approach of defining collocations as frequent co-occurrences, without any additional semantic

[2] Binomials can also be considered collocations. This study captured a number of binomials that met the statistical criteria for collocations, i.e. they were frequent enough and had a high MI score.

criteria is adopted. Examples of such collocations could be *take care, last night, learning difficulties*.

Collocations in English are considered to be a very frequent, and therefore important, part of language. McCarthy even claimed that "collocation deserves to be a central aspect of vocabulary study" (1990: 12). Shin and Nation (2008) showed that some collocations are frequent enough to be included in the most frequent 2000 words lists. Although considered frequent, there is yet no research which explores how estimate of the percentage of collocations varies across registers.

4.2. Phrasal verbs

Liu (2011) noticed that definitions of phrasal verbs vary, with some of them focusing only on formal criteria (verb occurring together with an adverbial particle), while others having a certain meaning criteria as well. Biber et al. (1999) made a distinction between phrasal verbs (sequences having an idiomatic meaning) and prepositional verbs (verbs used with a preposition without acquiring any extra meaning). In this study a meaning criterion was adopted, defining phrasal verbs as sequences of verb and adverbial particle which carry a single meaning, for example, *give up, get up*. It has been estimated that phrasal verb dictionaries list 4,500-6,000 different phrasal verbs (Liu 2011), so they seem to be a very numerous category. However, the percentages of text coverage suggested by previous research are not that high: phrasal verbs and prepositional verbs together were suggested to cover about 2% in conversation and 1% in academic prose (Biber et al. 1999).

4.3. Idioms

Despite the fact that idioms have received a lot of attention in research, it has been noted that idioms are usually not well defined in literature (Grant & Bauer 2004) and the term is used for different sequences, such as metaphorically used phrases or formulaic phrases in general sense (Moon 1998a). In this paper idioms are defined as two or more word sequences which have a non-compositional meaning, that is, the meanings of their parts do not sum up to create the meaning of the idiom, as in the sequences like *spill the beans*.

Idioms are very numerous as a category, with various idiom dictionaries listing thousands of items (Liu 2003). However, individual idioms are rather rare in discourse. For instance, Moon (1998a) estimated that most of the items from her list of fixed expressions including idioms were used less than once per million words. Even if individual idioms might be infrequent, considering the vast number of idioms, together they can account for an important part of English. As far as the distribution of idioms in various registers is concerned, Biber et al. (1999) claimed that idioms are the most frequent in fiction, where they are adopted to represent spoken language, even if idioms do not seem to be frequent in spoken language. Overall, although there are no clear estimates of the percentage of idioms in language, the sense is that generally idioms are not used very frequently.

4.4. Lexical bundles

"Lexical bundles can be regarded as extended collocations: bundles of words that show statistical tendency to co-occur" (Biber et al. 1999: 989). As their extraction is purely frequency-based, there are no additional criteria for well-formedness or meaning independence, so sequences as *in terms of* and *I think I* would both be defined as lexical bundles. However, an analysis of the sequences extracted showed that lexical bundles often have functional purposes, such as expressing stance, organizing discourse, and expressing referential meaning (Biber et al. 2004). Therefore, they seem to be more than just incidental co-occurrences of words. There are two main differences distinguishing lexical bundles from collocations. Firstly, lexical bundles are longer (consisting of three or more words) and completely fixed sequences. And secondly, they are defined by their frequency of occurrence alone without any specific requirements for mutual expectancy of the words in the bundle.

Lexical bundles were shown to cover about 30% of conversation and 21% of academic discourse (Biber et al. 1999). As these sequences are extremely frequent, studies that focus only on idiomatic phrases will miss a large part of formulaic language (Conrad & Biber 2004). Therefore the present study included lexical bundles and adopted the corpus-driven criteria to extract them (Biber et al. 2004, 1999; Conrad & Biber 2004).

In sum, although there have been attempts to quantify the use of different categories of formulaic language before, they were based on different corpora and different methods, and so are hardly comparable. Also, previous studies have looked at single categories, with definitions that often overlapped with definitions from other categories in other studies. Furthermore, it was also noted that "[t]here seem to be decided genre preferences for phrasal lexemes in general, as well as for individual expressions (…)" (Moon 1998b: 100). Hence, previous results that do not take into account register information are not easily comparable either. In order to gain a better understanding of the heterogeneous nature of formulaic language, four categories concurrently using the same corpus and mutually-exclusive definitions will be investigated, leading to the two research questions of this study:

1. What is the distribution of collocations, phrasal verbs, idiomatic phrases, and lexical bundles in discourse?
2. Does this distribution depend on register?

Answering these questions will shed more light on the nature of formulaicity in language overall and in various registers, and at the same time, it will have pedagogical implications by informing teachers which types of formulaic sequences are the most frequent ones, and therefore presumably the most important ones for teaching.

5. Methodology

5.1. Corpus and software

The research questions were addressed by carrying out a corpus study, based on the BNC *Baby* corpus. It is a sample of the British National Corpus (BNC), which is a general corpus of British English. BNC *Baby* consists of about 4 million running words and it is divided into four parts: Academic prose, Fiction, Newspaper language and Spoken conversation, each covering about 1 million words of the corpus. These four registers will be compared to answer the second research question. Two different software programs were used to extract sequences: *WordSmith Tools* (Scott 2012) and *AntConc* (Anthony 2011). Two different software programs were chosen for practicality reasons: *AntConc* has a built-in function for extracting n-grams of various lengths with a minimum input, while *WordSmith Tools* offers better interface for manually checking and grouping examples. The statistical analysis was performed using *R Studio* (R Core Team 2013).

5.2. Extraction of the four formulaic sequence categories

Collocations and lexical bundles were extracted from the corpus based on statistical criteria. Lists compiled by previous research were used to identify the idiomatic expressions and phrasal verbs. The procedure for the extraction of each type of sequence is detailed below.

Collocations. Collocations were defined as two or three word sequences, showing a tendency to co-occur. Only collocations which were adjacent (e.g. *hard work*) or had one word between the two collocates (e.g. *take a break*) were counted. A statistical approach for identifying collocations was adopted, and the following three criteria used:

1. Raw frequency ≥5 (in the entire BNC *Baby* corpus). This threshold was chosen due to warnings that MI scores become problematic with lower frequencies (Dunning 1993; Kilgarriff 2005)
2. MI score of ≥ 3 (Church & Hanks 1990)
3. Composed of at least two lexical words

All the bigrams and trigrams were extracted from the entire corpus (in order to capture collocations with the required span). The MI scores were calculated and the candidates for collocations to obtain a list of sequences that show mutual expectancy (MI≥3). Then remaining collocations were filtered based on their part of speech and collocations of function words were excluded from further study.

The list of collocations obtained with these criteria was checked for the sequences containing proper names and those were excluded. The final list of collocations in corpus consisted of 9,913 two-word and 3,262 three-word items. Three word items included non-adjacent collocations with one intervening word (*taken into account*), some binomials (*black and white*), and some extended collocations (*indexed sequential file*). These cases were counted as 3-word sequences, in order not to ignore these 3-word examples and underestimate the collocation percentage, and also because they seemed to be relatively fixed as units.

Phrasal verbs. Phrasal verbs are also very numerous as a category, with Liu (2011) analysing 8,847 different ones in his study. However, there have already been corpus studies that tried to identify the most frequent ones so pre-constructed lists for identification of phrasal verb could be adopted. Gardner and Davies (2007) compiled a list of the most frequent 100 phrasal verbs, while Liu (2011) expanded this list to 150 items, which covered about 63% of the total 512,305 phrasal verb occurrences in the full BNC. As Liu's list was the most recent and comprehensive, compiled based on information from both the BNC and COCA, it was used as the basis for identifying phrasal verbs in this study.

All of the phrasal verbs on the list were checked if they allowed an insertion, and if so, they were listed separately and searched for with wildcards (to capture sequences such as *figured it out*). The concordance lines of phrasal verbs with insertions were then checked manually to exclude any noise.

Idiomatic phrases. Due to the large number of idioms and their numerous variations, a manual search for all the existing idioms is a very laborious process. From various previously compiled lists of frequent idiomatic language, the following two were chosen, because they were the most comprehensive, included the most frequent items, and therefore seemed to be the most suitable for this study:

1. List of the most frequently used spoken American English idioms (Liu 2003)
2. Phrasal expressions list (Martinez & Schmitt 2012)

These lists are quite different in terms of the methodology they are based on. Liu's list is corpus-informed: he collected 9,683 idioms from seven dictionaries and manually searched for them in three corpora of spoken American English. The frequency cut-off point of 2 per million adopted in his study limited the list to the 302 most frequent idioms. Hence Liu's list seemed to be a very good starting point as it provided a ready-made list of the most frequent idioms. However, it had a very obvious limitation for this particular study: it was based on spoken American corpora. As Liu claims (2003), the list might have been different if British English or written corpora was used.

To account for this limitation, the Phrasal Expressions List (PHRASE List) (Martinez & Schmitt, 2012) was added. The PHRASE List is a corpus-driven list: the most frequent n-grams were extracted from the BNC and then checked manually. Only the sequences that had non-transparent meanings were included in the list. As this list is based on the BNC (and the BNC *Baby* used for this study is a sample of the BNC), we can assume that this list covers the most frequent non-transparent sequences in the study corpus.

Even if the methodologies of compiling the lists were different, frequency was the key criteria for both of them. In Liu's list frequency cut-off point was 2 per million, while Martinez and Schmitt's phrasal expressions list included items that would be included in the 5,000 most frequent word families list (which means their frequency was at least 7 per million). The two lists were combined and duplicate items deleted, leaving a total of 554 items. Although the lists followed different methodologies, there was a considerable overlap between them, which was reassuring, as it shows that despite the methodological differences, the lists still capture similar sequences. Combined together these lists seemed to provide a good resource of frequent idiomatic phrases.

As not all the idioms were covered in the study, this category will be referred to as *idiomatic phrases*, especially because Martinez and Schmitt (2012) do not define the items in their list as idioms, but as phrasal expressions. Thus idiomatic phrases in this study are frequent, non-transparent sequences of two or more words.

The phrasal verbs that were included in these lists were deleted and added to the list of phrasal verbs (see below) to keep these two categories as distinct as possible. In addition, ten phrases from the list were deleted because it was decided that their meanings were too transparent[3]. Then all of the forms of the idiomatic phrases were listed (including insertions and variations) and searched for in the BNC Baby. The concordance lines of idiomatic phrases that seemed to be the most problematic (i.e. very often used literally) – such as *a good*, *they say* – were checked manually to adjust their frequencies, and exclude the cases where the phrase was used literally.

Lexical bundles. Lexical bundles were extracted based on the criteria adopted in previous research: frequency and dispersion of the sequence were considered when deciding if to include an n-gram in the list of lexical bundles (Biber et al., 2004; Biber, Conrad, & Reppen, 1994; Biber et al., 1999). The cut-off points varied based on the length of the n-gram and are summarized in the Table 1.

Table 1. Criteria for identifying lexical bundles

Length of the bundle	Frequency	Range
3-gram 4-gram	10 per million	min 5 texts
5-gram 6-gram 7-9gram	5 per million	min 5 texts

When lexical bundles are automatically extracted from a corpus, some of the shorter bundles are inevitably also used as parts of longer bundles (Gries, 2010). To account for this, the lists of the lexical bundles were checked and the frequencies of the shorter bundles were adjusted so that the

[3] The phrases excluded from the study, because they were judged to be too literal, were the following: *add to, amount to, care to, heard to, look like, look to, this stage, those who, to blame*.

same sequence would not be counted twice. Lexical bundles were extracted from each sub-corpus separately and those counts of the sub-corpora were added together to estimate the percentage of lexical bundles in the overall corpus.

5.3. Adjustments

When the sequences were extracted from corpus, there was some overlap between the different categories. In order to avoid overestimation of the counts of the formulaic sequences, two types of adjustments were made:

1. If the same sequence was included in resulting lists of two different categories (for example, *on the other hand* listed both as an idiomatic phrase and as a lexical bundle), it was counted as a part of the category which was more strictly defined (in this case as an idiomatic phrase).
2. If a shorter sequence was a part of the longer sequence (e.g. a 3-gram was a part of a 4-gram or an idiomatic phrase used as a part of a longer lexical bundle), it was counted as a longer sequence and the count of the shorter sequences was adjusted accordingly.

This procedure led to estimation of the frequencies of collocations, phrasal verbs, idiomatic phrases, and lexical bundles in all four registers (academic prose, fiction, newspaper language, and spoken conversation). These estimates were then added up to estimate the number of formulaic sequences in the entire corpus.

6. Results

The purpose of the study was to gain a better understanding of the nature of formulaic language overall by exploring and describing the occurrence of four major formulaic categories concurrently. The summary results for the entire BNC *Baby* and the sub-corpora are presented in Table 2, which provides both the raw number of occurrences of each category and their proportion in the corpus. It shows the four categories addressed in this study covered 40.55% of the language in the corpus. This percentage varied from register to register, with the spoken conversation being the most formulaic (68.81% covered by the four categories addressed) and the newspaper language being the least formulaic (only 24.46%). It is worth noting that the percentages of formulaic sequences in the corpus and in each sub-corpus consist only of the frequencies of collocations, phrasal verbs, idiomatic phrases, and lexical bundles, so these figures will underestimate the total formulaicity of the corpus to some extent, as sequences belonging to other categories are not captured, and some infrequent phrasal verbs and idioms might have been used in the texts, but not included in the lists used for the study. However, the figures probably capture the main part of the formulaicity; at least as far as adjacent sequences without variable slots are concerned.

Table 2. Frequency of four formulaic categories in the BNC *Baby* corpus

	Collocations	**Phrasal verbs**	**Idiomatic phrases**	**Lexical bundles**	**Formulaic language**
Academic prose	97,406 (9.72%)	5,321 (0.53%)	29,288 (2.92%)	193,092 (19.27%)	325,107 (32.44%)
Fiction	61,375 (6.04%)	12,405 (1.22%)	26,314 (2.59%)	261,601 (25.75%)	361,695 (35.60%)
Newspaper language	72,261 (7.52%)	8,513 (0.89%)	21,999 (2.29%)	132,157 (13.76%)	234,930 (24.46%)
Spoken conversation	69,102 (6.82%)	13,319 (1.31%)	24,216 (2.39%)	590,770 (58.29%)	697,407 (68.81%)
BNC *Baby*	300,144 (7.52%)	39,558 (0.99%)	101,817 (2.55%)	1,177,620 (29.49%)	1,619,139 (40.55%)

Figure 1 illustrates the distribution of the four categories of formulaic language in the BNC *Baby* corpus. It is obvious from the figure that the various categories are very unevenly distributed. Lexical bundles are the most frequent category, making up 29.49% of the BNC *Baby* corpus. Collocations make up 7.52% of the corpus, while idiomatic phrases and phrasal verbs make up only 2.55% and 0.99% respectively. Thus the various categories of formulaic language are quite different in their patterns of occurrence, with lexical bundles being as much as 10 times more frequent than idiomatic phrases, and nearly 30 times more frequent than phrasal verbs. Even collocations, which are usually considered a frequent and important part of language, occur as a category only about one-quarter as often as lexical bundles.

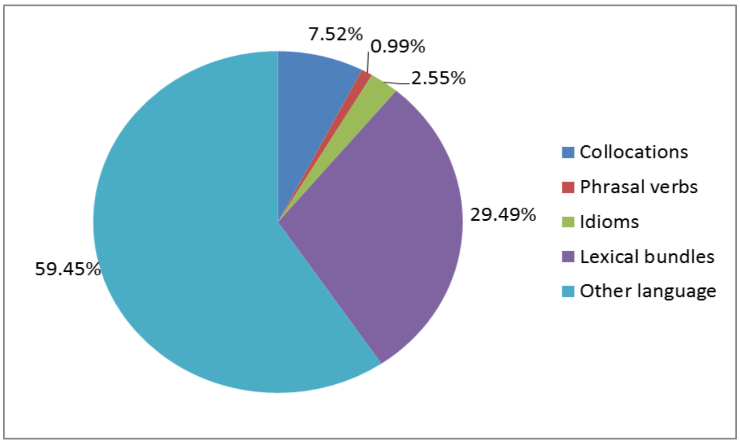

Figure 1. Distribution of formulaic language categories in the BNC *Baby*

In the research on formulaic language, many scholars have noticed that register variation influenced the use of formulaic sequences. This was true for lexical bundles (Biber, Conrad, & Cortes, 2004, Cortes, 2004), phrasal verbs (Liu, 2011), idioms (Moon, 1998a) and collocations (Partington, 1998). However, only the studies on lexical bundles and phrasal verbs tried to quantify this variance. In order to more fully analyse how registers influence the use of formulaic sequences, the four registers were compared and the results illustrated in the pie-charts in Figure 2.

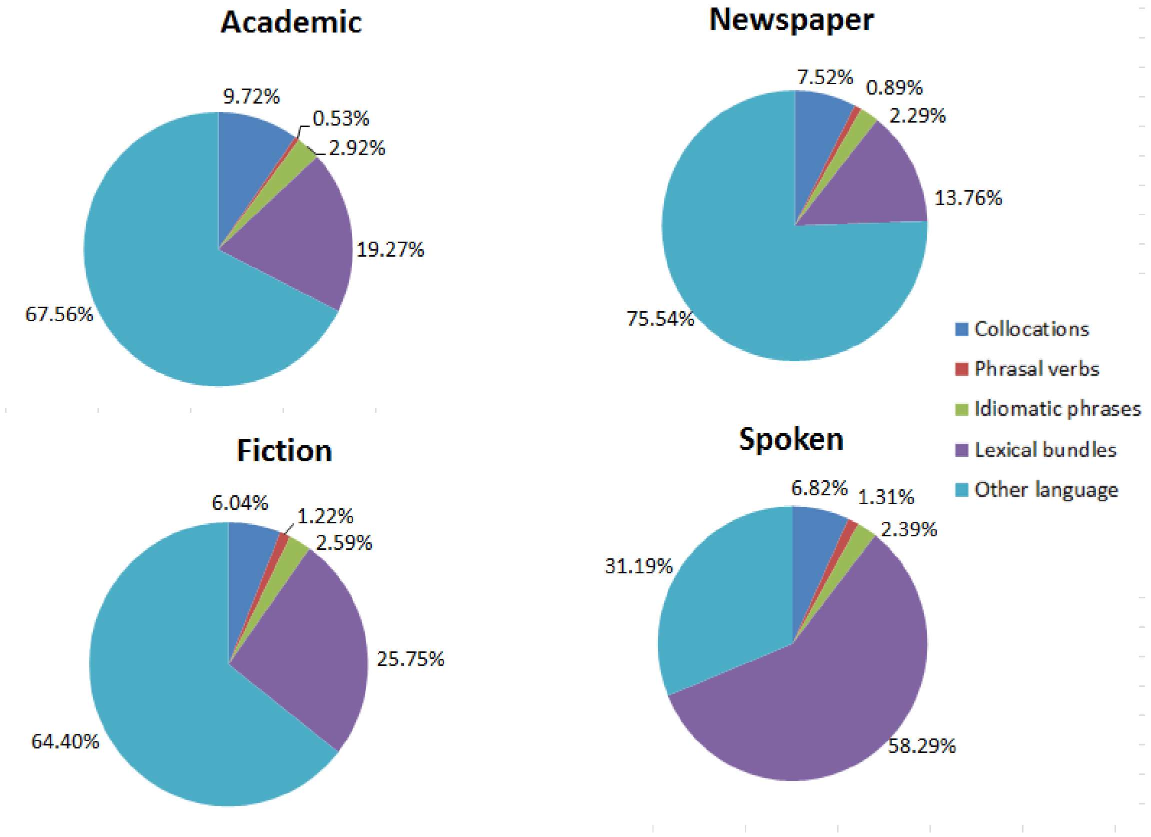

Figure 2. Comparison of the distribution of formulaic categories across registers

It is clear from Figure 2 that the spoken sub-corpus is by far the most formulaic, with about 68.81% of it made up of the formulaic sequences analysed here. Fiction is the second most formulaic sub-corpus, with about 35.60% covered by formulaic sequences. The main difference between fiction and spoken conversation seems to lay in the use of lexical bundles. Academic prose seems to show very similar tendencies to fiction, while newspaper language is the least formulaic with only 24.46% of the sub-corpus consisting of the four categories of formulaic sequences analysed.

To go beyond simply visual comparison and to confidently reject the null hypothesis that there is no significant difference of the distribution of formulaic sequences across the registers, the chi-square analysis was carried out. The chi-square analysis was chosen as it does not assume normal distribution of the data (Manning & Schütze, 1999) and it was carried out using *R studio* software (R Core Team, 2013). The statistical procedures were chosen following Gries (2014), who has provided guidelines for comparing frequencies in different corpora. Statistical significance was tested and the effect size calculated. A 4 x 4 chi-square analysis revealed that there was a significant relationship between the register and the distribution of different categories of formulaic language (χ^2 = 121,971.1, *df* = 9, *p* < .000). However, the register difference had only a weak effect (Cramer's *V* =.16) on the distribution formulaic language categories.

As spoken conversation seemed to be different from the other registers (which was not surprising as all the other registers were written), a comparison between written and spoken language was also carried out. All the written sub-corpora were added together and a 2 x 4 chi-square analysis ran. The chi-square and significance value remained similar as in the previous analysis (χ^2 = 90,250.14, *df* = 3, *p* < .000), but the effect size increased showing that there was a moderate effect (Cramer's *V* = .24). This shows that even if the register has an effect on the distribution of various formulaic categories, the mode of language production (spoken versus written) has a larger effect.

Hence from Table 2 and the Chi-square analysis two main conclusions can be drawn:

1. The pattern of the distribution is always the same between the registers: lexical bundles are always the most frequent category, followed by collocations, idiomatic phrases and phrasal verbs;
2. The quantity of sequences of each category differs significantly from register to register.

7. Discussion

As argued in the Introduction, formulaic language needs to be viewed as a heterogeneous phenomenon, with an understanding of how its various categories relate to each other. To my best knowledge, this study is the first one to estimate the frequencies of use of four different categories of formulaic sequences using the same corpus and mutually exclusive definitions, making the percentages of the categories comparable.

The results showed that about 41% of the BNC *Baby* corpus was covered by the four categories of formulaic language addressed in this study (at least according to the definitions used). As there are other categories (such as clichés, proverbs, lexical phrases, etc.) that were not included in the study, the overall percentage of formulaic language would be higher to some extent. These results are in line with Conklin and Schmitt's (2012) conclusion that formulaic sequences make up at least 30-50% of language overall.

Our results clearly show that the various categories of formulaic language have very different distributions of use, with lexical bundles being the most frequent category, followed by collocations, idiomatic phrases, and phrasal verbs.

In some cases, the percentage findings are in line with previous research. Phrasal verbs accounted for 0.5-1.3% in the current study, which is congruent with Biber et al. (1999). Similarly, this study showed that lexical bundles in academic prose covered about 19%, compared to 21% in Biber et al. (1999). However, lexical bundles in spoken conversation made up 58.29% in this study, far higher than the percentage (30%) suggested by Biber and his colleagues (Biber et al. 1999; Conrad & Biber 2004). This might be due to corpus differences, as exactly the same criteria were used for extraction of lexical bundles in both studies[4]. This study also produced an estimate of the extent of collocations in English: about 8 %. While this percentage would clearly change with different definitions, it is interesting to note that while collocations are widely considered very frequent and useful, the estimates of the study indicate that lexical bundles make up a much higher percentage of English (about 4 times).

As far as different registers are concerned, spoken sub-corpus was the most formulaic one, which is probably due to the nature of the data in the sub-corpus: it consisted of casual conversations, which put a time pressure on the speaker leading to a reliance on formulaic language in order to minimize the chances of misunderstanding (Wray 2002). It is interesting to note that fiction was the second most formulaic register. It seems that fiction should by definition employ a very creative language use. However, this result probably only confirms that formulaic language is a default choice in our

[4] Even if methodologies of the studies were the same, there was a difference in the treatment of contractions. Conrad and Biber's study (2004) counted contractions as one word, while in this study they were treated as separate words (for example, both *I don't know* and *I do not know* would be counted as 4 word sequences). However, a closer analysis revealed that even if contractions were counted as one word, the percentage would drop only about 6%, but would still remain considerably higher than in Conrad and Biber's study.

language repertoire (Wray, 2008). Newspaper language appears to be the least formulaic. In future, it would be interesting to investigate this trend further and see why this is the case. It might be that newspaper language uses more variable expressions with open-slots which were not captured in the study.

It is also interesting to note that the mode of language production (written/spoken) seems to have a greater influence over the use of formulaic language than register does. This is in line with the results of Erman and Warren (2000), who also found that spoken discourse was more formulaic than written discourse, even if the difference in their study was rather small (58.6% vs. 52.3%, respectively). One thing that has to be noted, though, is that spoken conversation was clearly different in terms of extensive use of lexical bundles and included more phrasal verbs, but collocations and idiomatic phrases were more frequent in written registers.

Our analysis of registers provides additional information about the nature of formulaic language. The same rank ordering of categories (lexical bundles>collocations>idiomatic phrases>phrasal verbs) was found in each of the four registers studied. This means the hierarchy is applicable not only to English overall, but seems to apply consistently to more specific registers as well. However, even if rank ordering remains the same, the quantity of formulaic sequences significantly differs from register to register. It also seems that each category has its own patterns of use in different registers.

7.1. Issues with categorization of formulaic sequences

Even if this study started from clear definitions of each category, additional decisions were necessary to keep the categories separate and that even then, some issues remained. For example, there were still collocations and phrasal verbs which were counted as parts of lexical bundles, and thus excluded from the counts of collocations and phrasal verbs. While this was necessary in order not to overestimate the counts of formulaic language in the study, it is debatable if a collocation ceases to be a collocation when it is used as a part of a longer sequence, making it difficult to draw a boundary between two individual sequences or two categories. More generally, it seemed that the overlap occurred mostly because lexical bundles and collocations were extracted based on frequency and co-occurrence criteria, while the extraction of phrasal verbs and idiomatic phrases was based on their non-transparency, hence included semantic criteria as well. This illustrates the different nature of sequences: some of them characterized by their meaning, while others by their usage. However, although categorization was sometimes problematic, not that much overlap occurred, making it possible for us to classify the sequences.

This study investigated the English language, but looking at the behaviour of different categories is important for other languages as well. Even though there are claims that formulaic sequences are also ubiquitous in other languages, performing similar functions and having similar characteristics (Nattinger & DeCarrico, 1992), these claims should be addressed systematically and empirically. When languages with different structures are analysed, different categories of formulaic language might have different frequencies and somewhat different characteristics. For example, Granger (2014) analysed lexical bundles in parliamentary debates and newspaper editorials both in French and in English, and listed certain challenges when the categories based on English research were applied to French. If certain categories are more relevant to some languages than others, the nature of formulaic language will be language specific to some extent. For this reason, if we want to investigate results across languages, we might do well to focus on individual categories rather than formulaic language overall.

7.2. Pedagogical implications

While not conclusive in their own right, these findings provide some useful information which adds to the discussion about what formulaic language to teach. For example, one pedagogical approach would be to teach the categories that offer the most coverage in language on the grounds that they are the most useful for the learner. However, this approach would have to be used in conjunction with other factors, in particular feasibility. Table 3 provides a summary of the number of individual sequences considered in this study.

Table 3. Number of items considered from each category

Type of sequences	Individual items analysed	Percentage covered
lexical bundles	~ 17,000	29.49
collocations	~13,000	7.52
idiomatic phrases	~ 550	2.55
phrasal verbs	~ 200	0.99

Based on the present corpus analysis, lexical bundles would provide the largest coverage for language learners and thus would presumably be the category that should get the most of classroom time. However, as Table 3 shows, for the purpose of this study, more than 17,000 different lexical bundles were extracted. Therefore, even using a relatively small corpus, the numbers of different lexical bundles are very high. Obviously not all of these sequences can and should be taught directly. The selection of items for explicit attention could be guided by another factor: the level of mastery required. For receptive purposes (listening and reading), it is probably not necessary to teach lexical bundles, as they are compositional and transparent in meaning and should not create difficulty for understanding. However, if teachers want their learners to use lexical bundles productively (in speaking or writing), they are faced with a challenge of selecting the lexical bundles to teach. This is not an easy task, as most sequences classified as lexical bundles are structurally incomplete and also not salient for learners (Biber et al. 1999). Furthermore, there is no one list of lexical bundles of general language use or ready guidance for selection. An option could be relying of some statistical measures to select the most useful sequences to teach. Researchers have tried to provide computational measures to guide the selection (e.g. *formula teaching worth*, which tries to account for native speakers intuitions in ranking the sequences extracted from corpora (Simpson-Vlach & Ellis, 2010). Using this method, Simpson-Vlach and Ellis developed their Academic Formulas List. It could be a starting point for choosing academic lexical bundles, but it remains to be established how teachable the items in this list are.

The category offering the second most coverage was collocations, suggesting that in terms of scope, they would also be useful for language users. However, as it was also a very numerous category, the number of items would simply be overwhelming for a learner. Thus, the situation seems to be very similar as for the lexical bundles. Thinking about the same distinction between productive and receptive knowledge, for receptive knowledge it is probably not necessary to teach all the collocations, as they are mostly transparent in meaning. For language production, though, collocations pose serious difficulties and even advanced learners tend to make mistakes (e.g. Bahns & Eldaw 1993; Laufer & Waldman 2011). As it is virtually impossible to teach all collocations, we are faced with the problem of selecting individual items to prioritize.

If we turn to lists for guidance, there are several candidates. Shin and Nation (2008) created a list of the most frequent collocations, which would be a part of the most frequent 2,000 words. Even though this list has limitations (e.g. used only the Spoken Part of the BNC, treated each word as a type rather than lemmatizing them), it could be a good starting point. However, it only covers the most frequent collocations, so it might only be useful with learners at the beginner level. For more advanced learners, Durrant (2009) looked at the viability of a list of academic collocations. After analysing a corpus of academic language he concluded that the most frequent collocations are grammatical ones. Even though grammatical collocations can be “legitimate learning targets” (Durrant 2009: 165), his list does not provide much guidance for the lexical collocations to teach.

Overall, it seems that both lexical bundles and collocations provide good coverage in language, but there are no clear best guidelines of which individual items to teach from these categories. For idiomatic phrases and phrasal verbs, the situation seems to be the opposite. In this study, idiomatic phrases and phrasal verbs together covered about 3.2% of language. Even if this percentage seems to be rather low, there is still value in teaching these categories, as the meanings of the items of these categories are non-transparent and thus could cause problems for learners both receptively and productively. Still, considering the relatively low coverage percentages, it is probably important to focus precious teaching time on only the most essential idiomatic phrases and phrasal verbs, i.e. the most frequent ones. A good starting point would be to refer to published lists which are strongly pedagogically-based. One example for phrasal verbs is the PHaVE List (Garnier & Schmitt 2015), which includes only the most frequent phrasal verbs, together with their most frequent meanings. For idioms, there are a few different lists available. Grant and Nation (2006) tried to list ‘core idioms’ (opaque and figurative phrases), but they found that only 7 of their selected core idioms would be included into a list of the 7,000 most frequent words, while others were much less frequent. Therefore they concluded that “[c]ore idioms are not frequent and do not deserve classroom attention” (Grant & Nation 2006: 15). The idiomatic phrases that were analysed in this study, though, are much more frequent and hence more useful for learners. Therefore, the two lists that were used for their identification (Liu 2003; Martinez & Schmitt 2012) could be used as a starting point for pedagogy.

7.3. Limitations

There are inevitably limitations of this study. The BNC *Baby* corpus consists of only 4 million words, which is a rather small corpus by contemporary standards and obviously cannot reliably represent all the English language. However, this was inevitable given the technical limitations of extracting the lexical bundles and collocations, and also the amount of manual analysis required. Also, as the study did not investigate the use of individual sequences, but rather general trends of use, this corpus seemed to be a good starting point. Despite the limited size of the corpus analysed, this study can be argued to be an important first step in the exploration of formulaic language as a heterogeneous phenomenon. Further research, using larger corpora or replicating this study with different corpora, perhaps while also covering different registers (such as academic spoken language or online language), would be an interesting next step.

Another limitation is that the extraction of phrasal verbs and idiomatic phrases was based on lists and therefore did not cover all the instances of their use. Nevertheless, the most frequent individual sequences in these categories were included, so the suggested estimates, although not definitive, should be reasonably accurate reflections of these categories. It should be stressed that the figures reported here are *estimates*, based on chosen definitions, and other definitions would presumably lead to slightly different figures.

8. Conclusions

This study was intended as a first step towards systematically looking at formulaic language as a heterogeneous phenomenon, viewing formulaic categories as having their own characteristics and behaviour. The results revealed that there is a substantial difference in the percentages of various categories of formulaic sequences in language, showing the importance of treating formulaic language as a heterogeneous phenomenon and non-overgeneralizing the findings based on one category of formulaic language to other categories. However, it seems that these results are merely the tip of the iceberg, and further investigations into how the various categories are acquired and used may reveal important new insights into the nature of formulaic language.

9. Acknowledgements

I am very grateful to my supervisor Norbert Schmitt for his very useful comments on designing the study and on the first drafts of this paper.

10. References

Altenberg, B. 1998. On the phraseology of spoken English: the evidence of recurrent word-combinations. *Phraseology : Theory, Analysis, and Applications*. A. P. Cowie (ed.). Oxford: Clarendon Press. 101–122.

Anthony, L. 2011. *AntConc* (Version 3.2.4.). Japan: Waseda University.

Bahns, J. & M. Eldaw. 1993. Should we teach EFL students collocations? *System* 21 (1), 101–114.

Barfield, A. 2012. Lexical collocations. *The Encyclopaedia of Applied Linguistics*. C. A. Chapelle (ed.). Oxford: Blackwell Publishing Ltd. Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/9781405198431.wbeal0690/abstract. Accessed 20 January 2014.

Barfield, P. A. & H. Gyllstad. 2009. *Researching Collocations in Another Language: Multiple Interpretations*. New York: Palgrave Macmillan.

Barlow, M. 2011. Corpus linguistics and theoretical linguistics. *International Journal of Corpus Linguistics* 16 (1), 3–44. http://doi.org/10.1075/ijcl.16.1.02bar

Biber, D. & S. Conrad. 2001. *Variation in English: Multi-Dimensional Studies* (1st edition). Harlow, England / New York: Routledge.

Biber, D., S. Conrad & V. Cortes. 2004. If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25 (3), 371–405. http://doi.org/10.1093/applin/25.3.371

Biber, D., S. Conrad & R. Reppen. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15 (2), 169–189. http://doi.org/10.1093/applin/15.2.169

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

Butler, C. S. 1998. Collocational frameworks in Spanish. *International Journal of Corpus Linguistics* 3 (1), 1–32. http://doi.org/10.1075/ijcl.3.1.02but

Cheng, W. 2009. Income/interest/net: Using internal criteria to determine the aboutness of a text. *Corpora and Language Teaching*. K. Aijmer (ed.). Amsterdam/Philadelphia: John Benjamins Publishing. 157–178.

Cheng, W., C. Greaves, J. M. Sinclair & M. Warren. 2009. Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams. *Applied Linguistics* 30(2), 236–252. http://doi.org/10.1093/applin/amn039

Cheng, W., C. Greaves & M. Warren. 2006. From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics* 11 (4), 411–433. http://doi.org/10.1075/ijcl.11.4.04che

Church, K. W. & P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16 (1), 22–29. http://doi.org/10.3115/981623.981633

Conklin, K. & N. Schmitt. 2008. Formulaic sequences: are they processed more quickly than non-formulaic language by native and non-native speakers? *Applied Linguistics* 29 (1), 72–89. http://doi.org/10.1093/applin/amm022

Conklin, K. & N. Schmitt. 2012. The Processing of formulaic language. *Annual Review of Applied Linguistics* 32, 45–61. http://doi.org/10.1017/S0267190512000074

Conrad, S. & D. Biber. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20, 56–71. http://doi.org/10.1515/9783484604674.56

Cortes, V. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23 (4), 397–423. http://doi.org/10.1016/j.esp.2003.12.001

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19 (1), 61–74.

Durrant, P. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* 28 (3), 157–169.

Erman, B. & B. Warren. 2000. The idiom principle and the open choice principle. *Text—Interdisciplinary Journal for the Study of Discourse* 20 (1), 29–62.

Foster, P. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. M. Bygate, P. Skehan, & M. Swain (eds.). Harlow, England: Longman. 75–93.

Garnier, M. & N. Schmitt. 2015. The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research* 19 (6), 645-666. 10.1177/1362168814559798.

Gardner, D. & M. Davies. 2007. Pointing out frequent phrasal verbs: a corpus-based analysis. *TESOL Quarterly* 41 (2), 339–359. http://doi.org/10.1002/j.1545-7249.2007.tb00062.x

Granger, S. 2014. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast* 14 (1), 58–72. http://doi.org/10.1075/lic.14.1.04gra

Granger, S. & M. Paquot. 2008. Disentangling the phraseological web. *Phraseology: An Interdisciplinary Perspective*. S. Granger & F. Meunier (eds.). Amsterdam;/Philadelphia: John Benjamins Publishing. 27–46.

Grant, L. & L. Bauer. 2004. Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics* 25 (1), 38–61. http://doi.org/10.1093/applin/25.1.38

Grant, L. & P. Nation. 2006. How many idioms are there in English? *ITL International Journal of Applied Linguistics* 151, 1–14.

Gries, S. T. 2010. Useful statistics for corpus linguistics. *A Mosaic of Corpus Linguistics: Selected Approaches*. A. Sánchez & M. Almela (eds.). Frankfurt am Main: Peter Lang. 269–291

Gries, S. T. 2014. Frequency tables, effect sizes, and explorations. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. D. Glynn & J. Robinson (eds). Amsterdam/ Philadelphia: John Benjamins Publishing. 365–389.

Kilgarriff, A. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1 (2), 263–276. http://doi.org/10.1515/cllt.2005.1.2.263

Laufer, B. & T. Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners’ English. *Language Learning* 61 (2), 647–672.

Lewis, M. 1997. *Implementing the Lexical Approach: Putting Theory into Practice* (1st edition). Hove, England: Cengage Learning.

Liu, D. 2003. The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly* 37 (4), 671–700. http://doi.org/10.2307/3588217

Liu, D. 2011. The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly* 45 (4), 661–688. http://doi.org/10.5054/tq.2011.247707

Manning, C. D. & H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Marco, M. J. L. 2000. Collocational frameworks in medical research papers: A genre-based study. *English for Specific Purposes* 19 (1), 63–86. http://doi.org/10.1016/S0889-4906(98)00013-1

Martinez, R. & N. Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33 (3), 299–320. http://doi.org/10.1093/applin/ams010

McCarthy, M. 1990. *Vocabulary*. Oxford/ New York/ Toronto: Oxford University Press.

Moon, R. 1998a. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Clarendon Press.

Moon, R. 1998b. Frequencies and forms of phrasal lexemes in English. *Phraseology : Theory, Analysis, and Applications.* A. P. Cowie (ed.). New York: Oxford University Press. 79–100.

Moon, R. 1998c. Vocabulary connections: multi-word items in English. *Vocabulary: Description, Acquisition and Pedagogy*. M. McCarthy & N. Schmitt (eds.). Cambridge/ New York: Cambridge University Press. 40–63.

Nattinger, J. R. & J. S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford England ; New York: OUP Oxford.

Nesselhauf, N. 2003. The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics* 24 (2), 223–242. http://doi.org/10.1093/applin/24.2.223

O'Donnell, M. B., M. Scott, M. Mahlberg & M. Hoey. 2012. Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory* 8 (1), 73–101. http://doi.org/10.1515/cllt-2012-0004

Partington, A. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam/ Philadelphia: John Benjamins Publishing.

Pawley, A. & F. H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. *Language and Communication*. J. C. Richards & R. W. Schmidt (eds.). London: Longman. 191–227.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Renouf, A. & J. M. Sinclair. 1991. Collocational frameworks in English. *English Corpus Linguistics* K. Aijmer, B. Altenberg, & H. Longman (eds.). New York: Longman. 128–143.

Schmitt, N. 2004. In *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: John Benjamins Publishing.

Schmitt, N. 2010. *Researching vocabulary: a vocabulary research manual*. Houndmills, Basingstoke, Hampshire/ New York: Palgrave Macmillan.

Scott, M. 2012. *WordSmith Tools* (Version 6). Liverpool: Lexical Analysis Software.

Shin, D. & P. Nation. 2008. Beyond single words: the most frequent collocations in spoken English. *ELT Journal* 62 (4), 339–348. http://doi.org/10.1093/elt/ccm091

Simpson, J. & J. Speake. 2008. *A Dictionary of Proverbs* (5th edition). Oxford: Oxford University Press.

Simpson-Vlach, R. & N. C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31 (4), 487–512.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A., K. Conklin & W. van Heuven. 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. http://doi.org/10.1037/a0022531

Siyanova-Chanturia, A. & R. Martinez. 2015. The idiom principle revisited. *Applied Linguistics* 26 (5), 549–569. http://doi.org/10.1093/applin/amt054

Sorhus, H. B. 1977. To hear ourselves - implications for teaching English as a second language. *English Language Teaching Journal* 31 (3), 211–221. http://doi.org/10.1093/elt/XXXI.3.211

Wei, N. & J. Li. 2013. A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics* 18 (4), 506–535. http://doi.org/10.1075/ijcl.18.4.03wei

Wray, A. 2002. *Formulaic Language and the Lexicon*. New York: Cambridge University Press.

Wray, A. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford/ New York: Oxford University Press.

Wray, A. 2012. What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics* 32, 231–254. http://doi.org/10.1017/S026719051200013X

Formulaic language is not all the same: comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs

Laura Vilkaitė

Summary

Formulaic language is a widely acknowledged central part of a language. However, it is heterogeneous in nature, made up of various types of formulaic sequences with their own characteristics and behaviour. The first step towards systematically describing the relationship between these different categories is to describe their distribution in language. This study investigated the frequency of occurrence of four categories of formulaic sequences: collocations, phrasal verbs, idiomatic phrases, and lexical bundles.

Corpus analysis was carried out to investigate the distribution of the four categories of formulaic language. The BNC Baby general corpus (4 million running words) was analysed using *AntConc* and *WordSmith Tools* software. Collocations and lexical bundles were extracted from the corpus automatically, while the extraction of the idiomatic phrases and lexical bundles was based on the pre-selected lists of the most frequent items of these categories. Later those extracted sequences were checked to avoid any overlapping and to ensure that the counts provided are as accurate as possible.

The results show that taken together the four categories of formulaic sequences made up about 41% of English. Lexical bundles were by far the most frequent category and covered about 29.5% of the corpus. Collocations were the second most frequent category, used much less (7.5% of the corpus). Idiomatic phrases and phrasal verbs were even less frequent (2.6% and 1% accordingly). There were differences in the frequencies of each category in the overall corpus, and also in the four registers analysed (academic prose, fiction, newspaper language, and spoken conversation). Spoken language turned out to be the most formulaic part of the corpus, followed by fiction and academic prose. Newspaper language was the least formulaic. Importantly, even if the registers differed in terms of their formulaicity, language mode (spoken/written) had a more substantial effect on the frequency distribution of the categories, with spoken language being more formulaic than written language.

Pastovieji žodžių junginiai nėra vienodi: idiomatinių frazių, kolokacijų, leksinių samplaikų ir frazinių veiksmažodžių dažnumo kalboje palyginimas

Laura Vilkaitė

Santrauka

Dabar jau plačiai pripažįstama, kad pastovieji žodžių junginiai – labai svarbi kalbos dalis. Tačiau tai gana heterogeniška leksinių vienetų grupė, sudaryta iš įvairių, tarpusavyje besiskiriančių pastoviųjų žodžių junginių tipų. Norint aiškiau suvokti pastoviųjų žodžių junginių tipus, pirmiausiai reikėtų aprašyti jų dažnumą ir pasiskirstymą kalboje. Šis tyrimas analizavo keturių pastoviųjų žodžių junginių tipų (kolokacijų, frazinių veiksmažodžių, idiomatinių frazių ir leksinių samplaikų) dažnumus.

Norint aptarti šių keturių pastoviųjų žodžių junginių pasiskirstymą kalboje buvo atlikta tekstyno analizė. *BNC Baby* bendrasis tekstynas (4 milijonai žodžių) buvo analizuotas *AntConc* ir *WordSmith Tools* programomis. Kolokacijos ir leksinės samplaikos buvo išskirtos iš tekstyno automatiškai, o idiomatinių frazių ir frazinių veiksmažodžių buvo ieškota pagal anksčiau sudarytus dažniausių šių tipų leksinių vienetų sąrašus. Vėliau visi iš tekstyno išskirti pastovieji žodžių junginiai buvo sutikrinti, kad nesikartotų tarp atskirų tipų, siekiant pateikti kuo tikslesnius skaičiavimų duomenis.

Tyrimo rezultatai rodo, kad drauge šie keturi pastoviųjų žodžių junginių tipai sudaro apie 41 % anglų kalbos tekstų. Leksinės samplaikos buvo pats dažniausias tipas, kuris sudarė apie 29,5 %

tekstyno. Kolokacijos buvo antrasis pagal dažnumą pastoviųjų žodžių junginių tipas, vartotas gerokai mažiau (7,5 % tekstyno). Idiomatinės frazės ir fraziniai veiksmažodžiai buvo dar retesni (atitinkamai 2,6 % ir 1 %). Visų tipų vartojimo dažnumas skyrėsi visame tekstyne, bet taip pat ir atskiruose registruose (akademinėje prozoje, grožinėje literatūroje, publicistikoje ir sakytinėje kalboje). Sakytinė kalba paaiškėjo besanti pati fraziškiausia tekstyno dalis. Kiek mažiaus pastoviųjų žodžių junginių buvo grožinėje literatūroje ir akademinėje prozoje. Publicistika buvo mažiausiai fraziška. Tačiau svarbu paminėti, kad nors registrai ir skyrėsi tarpusavyje, vis dėlto, pastoviųjų žodžių junginių vartojimo dažnumu sakytinė ir rašytinės tekstyno dalys skyrėsi gerokai labiau – sakytinei kalbai pastoviųjų žodžių junginių vartojimas buvo daug būdingesnis.

Raktiniai žodžiai: leksikos tyrimai, pastovieji žodžių junginiai, kolokacijos, fraziniai veiksmažodžiai, idiomatinės frazės, leksinės samplaikos

Submitted: November 2015

Published: January 2016