Rita Juknevičienė

ENGLISH PHRASEOLOGY AND CORPORA

An introduction to corpus-based and corpus-driven phraseology

VU 1575
Vilniaus universiteto leidykla

VILNIAUS UNIVERSITETAS

Rita Juknevičienė

ENGLISH PHRASEOLOGY AND CORPORA

An introduction to corpus-based and corpus-driven phraseology

VILNIAUS UNIVERSITETO LEIDYKLA

VILNIUS, 2017

Apsvarstė ir rekomendavo išleisti
Vilniaus universiteto Filologijos fakulteto taryba
(2017 m. kovo 3 d., protokolas Nr. 9)

Recenzentai:

Prof. dr. Inesa Šeškauskienė (Vilniaus universitetas)
Dr. Jolanta Kovalevskaitė (Vytauto Didžiojo universitetas)

ISBN 978-609-459-820-3

CONTENTS

ACKNOWLEDGEMENTS

This course book is a result of my work at the Department of English Philology in Vilnius University. I would like to thank my colleagues for letting me enjoy a sabbatical term in autumn 2016. Warm thanks are also due to all my students for being a source of constant inspiration. I hope this course will encourage you to never stop marvelling at the secrets of language.

A very important milestone in the understanding of research in phraseology was my study visit at Cardiff University in February 2015. It was made possible through funding provided by the Research Council of Lithuania. I would like to thank my academic hostesses Alison Wray and Tess Fitzpatrick for their encouragement and support.

I was very lucky to have two wonderful reviewers—my grateful thanks go to Jolanta Kovalevskaitė and Inesa Šeškauskienė for their valuable comments and advice. Many questions and suggestions from my colleagues at the Department also helped me make the course book more accessible to the intended readers.

Lastly, I would like to thank my family for being what they are and never allowing me to lose the sense of reality.

PREFACE

The present course book is a summary of lecture notes compiled for undergraduate students of English Philology at Vilnius University for the course on Phraseology. What started as an optional course on the study of phrases was gradually developed into a seminar on research writing in the field of phraseology. The design of the seminar requires that students gain basic knowledge in phraseological research, develop skills in the use of a number of corpus tools, acquire elementary understanding of several statistical tests relevant for the analysis of corpus data and, last but not least, are introduced to the process of research writing. This course book was thus primarily conceived as a manual for students enrolled on the course and interested in carrying out their own phraseology-based research projects for their term papers. Therefore, it introduces the reader to major research strands in contemporary English phraseology, discusses its terminology and methods as well as offers tasks for individual exploration.

Despite its relatively short history, phraseological research is currently undergoing a rapid growth, so it would be impossible to do justice to a plethora of studies within one academic course lasting for one semester. Therefore the coverage of literature in this course book is inevitably limited. Two major considerations lie behind the selection of studies and scholars referenced here. Firstly, an attempt was made to use for reference the seminal publications which serve as background reading in the field. Secondly, the choice of research articles for discussion is expected to cater for students facing their first academic challenges. Many of them are exemplary works whose research design could be replicated by budding linguists wishing to carry out small-scale individual projects on linguistic data from a different language or with a different set of phraseological expressions. Hopefully, the course will provide a reasonable methodological basis and encourage students to take a step further and start developing their own original ideas for BA and MA theses. As such, the present course book should be regarded as an introductory reader rather than the ultimate resource. Throughout the book, the reader will find not only explanations and illustrations but also references for further reading. More

generally, the course book may be beneficial to anyone willing to get hands-on experience in corpus data extraction and analysis and will thus serve as an introduction to corpus-based and corpus-driven analysis of phraseological units.

Phraseological research overviewed here represents a number of linguistic schools and traditions and, inevitably, makes use of diverse terminology. For the purposes of brevity and clarity and in order to maintain an explicit link to phraseology as a field of linguistic study, I will be using two equivalent terms: *phrase* and *phraseological unit*, of which the former is perhaps more convenient for its brevity. Both terms will refer to any type phraseological unit which is broadly understood as a multi-word sequence that has a (semi-)fixed form and is partly or fully non-compositional. Specific terms, however, will be introduced and used to discuss different types of phrases described in individual studies and defined in a more detailed manner.

Lastly, to make the most of this course book, the reader should register with any website giving access to a corpus of English. The registration is free yet registered users may enjoy many more search options than those who do not register. The two widely known options are the British National Corpus (BNC) which can be accessed on different websites, one of them hosted by the University of Lancaster (UK) at http://bncweb.lancs.ac.uk; its current version was developed by Sebastian Hoffmann and Stefan Evert. The same corpus and, in addition, access to the largest language corpus ever compiled in the world, i. e. the Corpus of Contemporary American English (COCA), are provided by Mark Davies from the Brigham Young University (USA) at http://corpus.byu.edu/coca/. It is these two corpora that have been used for preparation of this course and that would be relevant for individual tasks.

The body of the course book is arranged in four chapters. Each chapter starts with a brief explanation of relevant terminology and definitions, an overview of the most important methodological issues and/or relevant approaches to the study of phrases in question, and then invites the reader to embark on a journey of phraseological discoveries by doing a set of study tasks.

Let it be an inspiring experience!

RJ
January 2017

CHAPTER 1.
PRELIMINARIES

The term *phraseology* is defined in the *Oxford English Dictionary* (www.oed.com) as:

1. a. *The selection or arrangement of words and phrases in the expression of ideas; manner or style of expression; the particular language, terminology, or diction which characterizes a writer, work, subject, language, place, etc.*
   b. Music. *Arrangement or construction of musical phrases.*

†2. *A collection or handbook of the phrases or idioms of a language; a phrasebook.*

†3. *The use of phrases in speech*.

In recent years, however, linguists use *phraseology* (from Greek *phrasis* 'utterance' + *logos* 'science') to refer to an interdisciplinary research field which deals with a broad variety of fixed and semi-fixed expressions. Such expressions function in language as whole units and are increasingly seen by scholars as independent constituents of vocabulary. In its broadest sense, the notion of fixed expressions covers such multi-word units as collocations, formulaic sequences, idioms, lexical bundles, n-grams and many others which will be jointly referred to in this course book as *phrases* or *phraseological units*. Hence *phraseology* is understood as the study of phrases.

Recent developments in corpus linguistics and vocabulary studies have provided ample evidence of phrasal, or *formulaic*, nature of naturally produced language. Apparently, a large part of what we say or write is not made up of discrete words as a glance at an alphabetically arranged dictionary page might suggest, but rather consists of multi-word units, or phrases. According to different estimates of the English vocabulary, most words in speech or written language belong to some type of phrase. The estimates of the level of *formulaicity* of language (Table 1.1), however, are rather different.

Variation of the estimates of formulaicity arises from the choice of items, or types of phrases considered, methods of calculation, and the type of language examined (spoken or written). Moreover, most counts exemplified in Table 1.1, except for

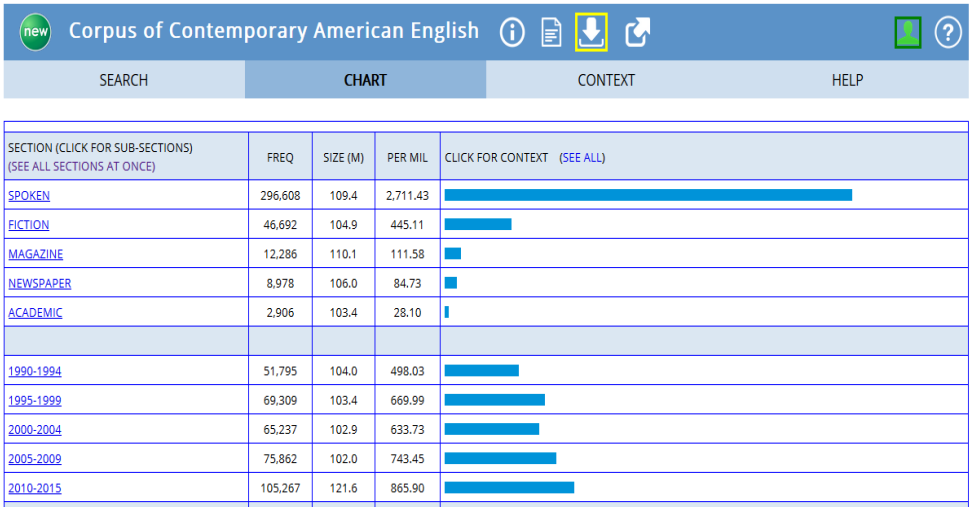
Table 1.1 Estimates of degree of formulaicity of English

Source	Reported evidence
Altenberg (1998, pp. 101–102)	Altenberg reports data from the London-Lund Corpus of spoken English which consists of ca. 500,000 words: "A rough estimation indicates that over 80 per cent of the words in the corpus form part of a recurrent word-combination in one way or another."
Biber et al. (1999, p. 995)	"In conversation, about 30% of the words occur in a recurrent lexical bundle; if two-word contracted bundles are also considered, almost 45% of the words in conversation occur in a recurrent lexical bundle. In academic prose, about 21% of the words occur in a recurrent lexical bundle."
Erman and Warren (2000, p. 37)	In their study, Erman and Warren focused on so-called prefabs defined as combinations of at least two words. They found that in spoken English prefabs account for 58.6% of the language and in writing for 52.3%. So on average 55% of English could be considered to be formulaic.
Juknevičienė (2011, p. 65)	Two-word lexical bundles account for 73.4% of all the words in the LOCNESS corpus which consists of academic essays written by undergraduate students whose mother tongue is English. Three-word lexical bundles, however, cover only 18.2% of the corpus.
Vilkaitė (2016, p. 40)	The study is concerned with coverage estimates of collocations, phrasal verbs, idiomatic phrases and lexical bundles in four registers (academic prose, fiction, newspapers, and conversation) of English as they are represented in the BNC Baby version. The research shows that taken together all four types of "formulaic categories" account for 32% of academic prose; 36% of fiction; 24% of newspaper language; 69% of conversation.

Vilkaitė (2016), are based on the frequencies of one type of phrase (i. e. recurrent sequence, lexical bundle or prefabricated word combination) and disregard any other possible multi-word combinations, for example, collocations or phrasal verbs, which are often non-contiguous and thus much more difficult to capture. Apparently, the given numbers remain rather crude measures of formulaicity yet they demonstrate that phrases, whatever the type, account for an impressive part of naturally produced English.

Another observation from corpus research gives even more prominence to the status of phraseological units in vocabulary. Evidence abounds that in terms of frequency certain phrases compete with single words, which only proves that such phrases have more currency in language than many less frequent words. O'Keeffe et al. (2007, p. 69) found that individual two-word phrases, which are called *chunks* in this book, are more frequent in the CANCODE corpus of spoken English (the size of the corpus is 5 million words) than single words. For example, *you know* occurs 28,013 times in CANCODE, and there are only 33 single words whose frequency is higher in this corpus. Clearly, the phrase is a high-frequency item in this corpus and would be very important to anyone aspiring to describe spoken English as it is represented in this corpus. But is it an equally important phrase in some other varieties of English?

A screenshot of a search output page from COCA (Figure 1.1) contains frequencies of the phrase *you know* in American English.



SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
SPOKEN	296,608	109.4	2,711.43	
FICTION	46,692	104.9	445.11	
MAGAZINE	12,286	110.1	111.58	
NEWSPAPER	8,978	106.0	84.73	
ACADEMIC	2,906	103.4	28.10	
1990-1994	51,795	104.0	498.03	
1995-1999	69,309	103.4	669.99	
2000-2004	65,237	102.9	633.73	
2005-2009	75,862	102.0	743.45	
2010-2015	105,267	121.6	865.90	

Figure 1.1 Distribution of *you know* across different sections of COCA.

The charts provide convincing evidence that *you know* is characteristic of spoken rather than any other register of American English. Its normalised frequencies per million words (see the fourth column in Figure 1.1) allow us to draw straightforward comparisons: *you know* occurs six times more frequently in speech than in fiction[1] and 96 times more frequently than in written academic language. It is also interesting to note that the frequency of that phrase has been increasing in the subcorpora of COCA representing different periods in time. Yet we would need to delve deeper into the corpus data and examine contexts and collocates of the phrase in order to explain trends behind the numbers. Moreover, if we examine the word frequency list of the COCA corpus, i.e. a list of all the words in the corpus reported in the order of their absolute frequency (available at http://www.wordfrequency.info), more unexpected discoveries are to be made. With its total frequency of 367,470 occurrences, the phrase *you know* ranks between the 108th and 109th most frequent words in COCA (Table 1.2), which are *back* (used as adverb) and *any* (determiner), respectively.

Table 1.2 A sample of word frequency rankings in COCA

Rank	Word	Part of speech	Frequency
104	one	pronoun	369,553
105	very	adverb	391,821
106	her	pronoun	397,950
106	even	adverb	361,067
108	back	adverb	367,844
	you know	*phrase*	*367,470*
109	any	determiner	348,100
110	good	adjective	353,973
111	woman	noun	341,422
112	through	preposition	340,921
113	us	pronoun	351,088

So *you know* is more frequent in this corpus, which undoubtedly remains the most representative corpus of American English up-to-date, than such functional words as *through*, *down*, *after*, modal verbs *may*, *should*, *need*, not to mention many simple open-class words, for example, *good*, *work* (verb), *call*, *try*, *ask* etc., all of which appear much lower in the word frequency list.

[1] To calculate how many times one normalised frequency differs from the other, we divide one by the other. For example, spoken vs. fiction (Figure 1): 2,711/445 = 6.1 times; spoken vs. academic: 2,711/28.10 = 96.5 times.

Furthermore, the frequencies of the constituent words of the phrase provide more food for thought: the pronoun *you* is the fourteenth frequent word in COCA (3,081,151 occurrences), and the verb lemma *know* takes the 47th position (892,535 occurrences). If we consider their individual frequencies, it is obvious that the pronoun often occurs in contexts other than those with the verb *know*, but over 41% of occurrences of the verb form *know* are in the phrase *you know*. Undoubtedly, the phrase *you know* must be very important in the overall use of the verb *know*. To compare, *I know* returns only 100,298 matches in COCA, which implies that the verb has more uses in English in combination with the second rather than first person pronoun. Admittedly, superficial frequency data does not offer any explanations why the distribution is what it is but it serves as a good starting point for further analyses. So even if the frequency data does not yet push the word as the central unit of language from its pedestal, it certainly provides a legitimate ground for treating certain phrases as equals to words.

1.1 Phraseological research in linguistics

Frequency data of individual phrases has several implications. To start with, it justifies the decision of contemporary lexicographers to give the status of headwords to certain multi-word units rather than merely enter them as illustrative examples in the entry of the lexical headword. In other words, frequency raises the status of certain phrases to that of single words. One of the first attempts to do it was undertaken by a team of linguists lead by John Sinclair in the Collins COBUILD project which, among many other contributions to linguistics, started a new family of English dictionaries based on an electronic corpus of English. In the *Collins COBUILD English Dictionary for Advanced Learners* (2001, 3rd edition or any later edition), many phrases are entered as headwords, for instance, *of course* (adverb), *all clear* (noun), not to mention phrasal verbs and compound nouns. For many such fixed phrases, the compilers of the COBUILD dictionaries used specific labels to mark both their morphological properties and phrasal nature, for example, PHRASAL VERB (*look after*, *give in*), PHRASAL CO-ORDINATING CONJUNCTION (*or else*), PHRASAL SUBORDINATING CONJUNCTION (*just because*, *as if*), PHRASAL MODAL (*going to*, *is supposed to*), PHRASAL PREPOSITION (*due to*, *as well as*). Compilers of this dictionary also identified a group of phrases defined as 'groups of words which are used together with little variation and which have a meaning of their own' (Collins COBUILD, p. xxix), for instance, *you know*, *and the like*, *keep going* etc.

Apart from lexicography, the discovery of word co-occurrences has numerous implications for the development of automated translation tools. We might hypothesise that in the eyes of laymen machine translation is one of the most obvious applications of linguistic research. Its importance can hardly be overestimated as the society at large more often than not questions the aims and goals of linguistic research. The development of automated translation tools convincingly demonstrates how insights from corpus-driven and corpus-based phraseology may be put to very practical uses. To obtain lexically and grammatically acceptable translation, the artificial computer-based brain needs to be taught (or rather programmed appropriately) which word sequences should be processed and translated word by word and which should be processed holistically as multi-word units. Clearly, only after linguists have provided a comprehensive list of phrases, this information can be used by developers of translation tools. A simple experiment on Google Translator shows what happens when the existence of phrases is not recognized by the translation tool.

The collocation *do research* is translated into Lithuanian by Google Translator as *daryti mokslinius tyrimus* (verbatim: 'do scientific research'), which suggests that the automated translating tool recognizes the phrase and gives an appropriate Lithuanian compound noun for the English *research*. An attempt to translate *to make noise*, in contrast, gives a less successful rendering. It is translated into Lithuanian as *triukšmo* (verbatim: *noise-Gen-Sg*), so the program fails to recognize that it is a phrase and leaves out the verb altogether. Similarly, *to cast a ballot* 'to vote in an ellection' is translated as *mesti burtus* 'to toss up'; the idiomatic sequence *raining cats and dogs* is translated literally as *lyja katėmis ir šunimis* rather than by the equivalent Lithuanian idiom *pila kaip iš kibiro* (verbatim: 'pouring like out of a bucket'). Further experimenting with Google Translator shows that *pila kaip iš kibiro* is rendered back into Lithuanian as *as the rain pours*, so the Lithuanian idiom seems to be recognized yet its English equivalent is not provided. Clearly, this tool remains limited when it comes to the translation of phrases yet its accuracy is improving. A handful of examples cannot lead to any generalizations, but it could be hypothesized that a comprehensive inventory of phrases, on condition that it is at all possible, could certainly increase the level of translation equivalence above the single word level.

Another vast area in which insights from phraseological studies of vocabulary have many applications is second language acquisition (SLA) and teaching. It has long been recognized that the lexical approach (Lewis, 1993; 2000) is more

advantageous than instruction in grammar (morphology and syntax) separated from the teaching/learning of vocabulary. The practice of teaching word lists has long given way to phrase lists because memorizing individual words does not teach the learner anything about combinability of words. Expansion of research into phraseologies of non-native English was further made possible due to the development of learner corpora which represent English (or any other language acquired as foreign/second) produced by non-native speakers, or, in the case of English, EFL learners. It will suffice for now to refer to a study by Pawley and Syder (1983), which is considered by many scholars to be one of the first studies of learner English in terms of lexical phrases. Pawley and Syder raised what today might appear a simple question: why does it happen that grammatically correct L2 English still sounds unnatural? The answer, as the reader should be able to guess by now, is fairly straightforward—apart from grammatical choices, there are also lexical ones to be made. L2 learners are often unaware of the fact that, for example, a fixed expression in one language (e. g. Swedish *tusen tack*) does not have a word-for-word equivalent in English, even if its verbatim rendering would most probably make sense to any English-speaking person (verbatim from Swedish: *thousand thanks*). As a consequence, foreign learners might produce grammatically correct sentences yet they fail to make the right lexical choices, which results in unnaturally sounding language.

It is thus little surprise that numerous insights from corpus studies gave rise to a new understanding of phraseology, a branch of linguistics which deals with all kinds of multi-word units, irrespectively of their contiguity, formulaicity or semantic compositionality. Phraseology in contemporary linguistics covers a broad range of phrases while the term itself refers to either (1) a field of study dealing with all types of more or less fixed multi-word expressions or (2) the use of these expressions in natural language.

1.2 Terminology

Phraseology has 'fuzzy borders' (Granger and Paquot, 2008, p. 29) and is thus often regarded as an interdisciplinary field because its research questions and approaches draw upon principles and categories from many other branches of linguistics. While the most obvious links point to the relationship of phraseology with semantics, morphology, syntax and discourse (for a detailed discussion, see Granger and Paquot, 2008), it also has areas of overlap with psycholinguistics,

phonology and many other fields. As a consequence, phraseological research may resort to terminology originating in diverse linguistic disciplines. Therefore while it is easy to agree that phraseology is a science about phrases, what a phrase is remains a disputable issue—every scholar feels it right to come up with a definition that is most suitable to his/her research design and tradition. To an uninitiated layman, a phrase is first of all a fixed sequence of several words. A linguist, however, may argue that certain sequences of words re-occur in our language in a more or less similar shape whereas others are not fixed at all and can be freely modified by inserting other words or change their grammatical form. Here are several examples from the BNC:

(1) *I'm a fairly tidy sort of person, so I do make the bed, and sort of tidy up in the bedroom (…)*
(2) *As Beck recalled, 'He came, he slept, he left without making the bed and I never saw him.'*
(3) *Clothes lay on the floor and the bed hadn't been made.*

The examples illustrate different grammatical transformations that the phrase *to make the bed* 'neatly arrange the sheets and covers of a bed' (Collins COBUILD) may undergo in naturally used language. In this respect, it is different from, for instance, *be rolling in it* 'be very rich' (Collins COBUILD), where the Continuous form is the predominant one:

(4) *With five top 40 hits under their belt and a top ten album, you might think Take That are rolling in it.* (BNC)[2]

What is important to note at this point is the fact that changes in the form of a particular phrase might imply changes in its meaning. Hence, even though the sequence *rolled in it* is perfectly possible in English, it usually has a different meaning from the one illustrated in example (4). Although searches in the BNC give no matches for *rolls in it* and *rolled in it*, but such sequences with non-idiomatic meanings are attested in COCA, for example:

[2] Interestingly, a search in COCA gave one instance where *rolling* seems to be used in the Present Simple to express the meaning of 'being rich'. The excerpt comes from a novel published in 2003: *I can't save everybody, I'm asthmatic, and I don't want a dog, especially not one who acts like he snorts coke and looks like he rolls in it.* It is an example of change in language, in this case, an idiom changing its allegedly fixed form.

(5) *It was full of Gold 100-there must have been seventy rolls in it altogether, and she loaded one into the camera (…).*

(6) *The sheets were stripped, but I had torn down the canopy and rolled in it until I was trussed and bound.*

What we observe here suggests that certain phrases have a fixed form, and changes of their form might incur loss of the phraseological meaning. Others, in contrast, are flexible and their grammatical transformations do not cause any change in the meaning. So what is a phrase and how fixed it should be to be recognized as a phrase?

One of the most frequently cited definitions of phraseological unit was proposed by Alison Wray (2000). In 2002, she published a book titled 'Formulaic Language and the Lexicon' which was concerned with exploration of what was often termed before Wray's publications as *lexical phrases* (Nattinger and DeCarrico, 1992). Wray proposed new terminology, namely, *formulaic sequences* and *formulaicity*, which was intentionally rather broad in its scope. According to Wray, a formulaic sequence is

> *a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar.* (Wray, 2000, p. 465; 2002, p. 9)

The definition encompasses four major criteria of the phrase and each of them, albeit in varying degree, provides a basis for identification of (semi-)fixed multi-word unit, be it a prototypical idiom (*to pull a person's leg* 'to tease a person'[3]), a phrasal verb (*to give up* 'to resign, surrender'), or a collocation (*stale bread* 'bread that has lost its freshness'). Let us take a closer look at the four criteria:

1. The statement "a sequence (…) of words or other elements" implies that a formulaic sequence consists of more than one word. The "other elements" may refer to non-verbal sounds (*erm*, *mhm*) characteristic of spoken language as in *well erm I dunno*.
2. A formulaic sequence has a fixed form. We have already seen that the form may sometimes vary, but variability is limited for some phrases. For example, *raining cats and dogs* would sound unnatural if it were said as **raining dogs and cats* or **pouring kittens and puppies*. Such deviations from the established form would lead to a specific stylistic—mostly, humorous—effect

[3] If not stated otherwise, all definitions are taken from the Oxford English Dictionary Online (www.oed.com).

and would be done on purpose in order to evoke unexpected images and implications.

3. A formulaic sequence is retrieved from memory at the moment of production as a whole and not processed by combining individual words. This criterion deals with the idiom principle described by Sinclair (1991) and it will discussed in more detail in Chapter 3 (page 35).
4. A formulaic sequence has a distinct meaning in discourse. It can be fully or partially non-compositional, e.g. *take a look* 'to look' vs. *kick the bucket* '(in slang) to die'. In the latter case the meaning of the sequence cannot be understood from the individual meanings of its constituents.

The four criteria albeit in varying degrees form the basis for the definition of any phraseological unit. The definition of formulaic sequence proposed by Wray is very broad but, according to the linguist, it was intended as such in order to cover a multitude of terms that had been used by other scholars.

Terminological variation in phraseological studies is indeed impressive. Consider the following collection of terms reprinted from Wray (2000, p. 465):

> *Amalgams, automatic chunks, clichés, co-ordinate constructions, collocations, complex lexemes, composites, conventionalized forms, fixed expressions, idioms, formulaic language, formulae, fossilized forms, frozen metaphors, frozen phrases, gambits, gestalt, holistic phrases, lexical simplex, lexical phrases, lexicalized sentence stems, multiword units/items, multiword lexical phenomena, noncompositional, noncomputational, nonproductive, nonpropositional, petrifications, phrasemes, preassembled speech, precoded conventional routines, prefabs, prefabricated routines and patterns, ready-made expressions, ready-made utterances, recurring utterances, routine formulae, schemata, semipreconstructed phrases that constitute single choices, sentence builders, set phrases, stable and familiar expressions with specialized subsenses, stereotyped phrases, unanalyzed chunks of speech, units*

A brief overview of more recent literature on phraseology proves that even this list is not exhaustive. It could be further supplemented with such terms as, for instance, *catchphrases*, *lexical bundles*, *phrasal verbs*, *proverbs*, *sayings*, and most probably many others.

Whatever the term, the scope of phraseology covers such phrases which meet the four criteria discussed above. Scholars pursuing specific research designs may apply additional criteria for phrases they choose to analyse. It could be the minimal frequency in a corpus, dispersion across a specific number of texts that are part of

a corpus, a predetermined syntactic construction (e.g. 'verb + noun') and others. So any comparison of findings should be preceded by a cautious consideration of how the operational definition of item in question has been formulated by the researcher. As we know, every scholar is entitled to modify his/her approach and definitions in order to give a better account of the research, which, luckily or unfortunately, results in this terminological jungle.

1.3 Classification of phrases

Current literature describes two distinct approaches to identification of phrases, namely, phraseological and frequency-based (Nesselhauf, 2005, p. 12; Granger and Paquot, 2008, p. 28). The two approaches provide two bases for classification of phrases which will be overviewed in this section.

The development of the phraseological approach goes back to the middle of the 20th century and is inseparable from the works of the Russian linguists Vinogradov and Amosova (Granger and Paquot, 2008, p. 28), who were concerned with identification of phraseological units in Russian and English. The underlying principle in the phraseological approach is based on the degree of idiomaticity that a phrase possesses, or its (non)-compositionality. Phrases with varying idiomaticity are placed on a continuum from free combinations, whose meanings can be understood from individual meanings of their constituents, to idioms which are non-compositional and cannot be understood literally.

One of the most often cited sources for such classification is Cowie (1981), who described the continuum of idiomaticity as a cline from free combinations to pure idioms. Two middle categories, namely, restricted collocations and figurative idioms, are partly idiomatic. Table 1.3 shows how Cowie's original classification was further elaborated by Howarth (1998), who divided phrases (termed *composites*) into lexical and grammatical. Lexical and grammatical composites differ in their structure as the former consist of two open-class words and the latter include a closed-class word (preposition). The degree of semantic compositionality across the types of phrases gradually increases from left to right.

Free combinations are fully compositional and pure idioms are absolutely opaque, or non-compositional; restricted collocations have one constituent that is used "in a specialized, often figurative sense only found in the context of a limited number of collocates" (Howarth, 1998, p. 28), for instance, *make a suggestion* where the verb cannot be substituted by the synonymous *do*. To judge from later interpretations of this classification (cf. Nesselhauf, 2005, pp. 14–15), the phrase *blow a*

Table 1.3 A continuum of idiomaticity (reprinted from Howarth, 1998)

	Free combinations	**Restricted collocations**	**Figurative idioms**	**Pure idioms**
Lexical composites (verb + noun)	blow a trumpet	blow a fuse	blow your own trumpet	blow the gaff
Grammatical composites (preposition + noun)	under the table	under attack	under the microscope	under the weather

fuse, as it happens, is perhaps not the best example as it has two readings in contemporary English: literal and idiomatic and thus perfectly matches the definition of figurative idioms, understood by Howarth as phrases that "have metaphorical meanings in terms of the whole and have a current literal interpretation" (Howarth, 1998, p. 28). For example, *under the microscope* can be understood as 'placed on the focal plane of the microscope to see an enlarged image' and figuratively as 'being studied very closely usually because it is believed that something is wrong with it' (Collins COBUILD). The distinction of the two types in the middle (restricted collocations and figurative idioms), in particular, is far from being simple. Nesselhauf (2005, pp. 14–18) gives the most exhaustive account of varying interpretations of two major criteria of collocations (opacity and commutability of constituents) and resulting diversity of operational definitions that could be found in linguistic literature. Some of these issues will be revisited later in this course book when discussing collocations.

The frequency-based approach to the identification and classification of phrases goes back to lexical studies of languages, English among them, which became possible after the emergence of corpora. In contrast to the phraseological approach whose practical application largely rested upon the intuitive judgment of the researcher who was viewed as the ultimate authority in the identification of phrases, corpus search tools provide an objective and inherently different procedure. It is a computer software rather than the human brain which computes and processes frequencies of co-occurrences of words and determines which word combinations 'deserve' to be included in the program output because their frequency proves that they are not random co-occurrences but statistically significant linguistic events. Admittedly, or rather luckily, the human researcher cannot be totally eliminated

from the process because the computer needs to be given specific search parameters. More importantly, only the human analyst can interpret automatically retrieved corpus data.

There are two major types of phrases that can be extracted from a corpus by computerized search tools: (1) pairs of co-occurring words understood as "discontinuous combinations of two words"; (2) recurrent continuous sequences of n words known as *n-grams* (Granger and Paquot, 2008). A detailed discussion of these types of phrases will be provided in chapters on collocations and lexical bundles (Chapters 3 and 4) whereas it will suffice to note at this point that identification of frequency-based phrases is fully automated. The computer software does not evaluate any semantic relations between constituents of phrases but merely computes their frequencies or probability of co-occurrence and on this basis identifies a sequence of two or more words as something that has currency in language.

An inquisitive mind might wonder to what extent the phraseological and frequency-based approaches correlate and how much overlap is to be expected between the two. As it turns out, the frequency-based analysis offers many valuable insights into what has been pre-determined subjectively. Firstly, corpus research shows that what is conspicuous and visible to the human analyst, e.g. stylistically marked non-compositional idioms, expressions of folk wisdom, proverbs etc., are rather infrequent phenomena in natural language. Phrases which contain partial compositionality and are not necessarily viewed as fixed phrases, for example, restricted collocations (cf. Table 1.3), were found to be very frequent and very numerous. Moreover, research in applied linguistics showed that it is those 'fuzzy' types of phrases that often differentiate beginners and proficient language users as they cause many difficulties to EFL learners.

According to Granger and Paquot (2008), the two approaches to phraseological units cannot be easily reconciled; hence, they propose to make "a clear distinction between two typologies: one for automated extraction and one for linguistic analysis" (ibid., p. 41). As a result, the scholars suggest that researchers working with automated extraction of phrases could stick to terminology pertaining to computerized research tools and employ such terms as *lexical bundles, n-grams, clusters* to refer to continuous sequences of words meeting a certain frequency threshold and *co-occurrences* or *collocations* to name combinations of words identified on the basis of specific statistical measures, for example, mutual information or t-score (see Section 3.3 on association measures, p. 42). Terms used in linguistic analyses, however, where the focus is not on the extraction method but rather on functions and uses of phrases in discourse, could employ terms from a functional classifica-

tion (Figure 1.2). A detailed description of each subtype in this classification given by Granger and Paquot looks very comprehensive. Yet it remains to see whether linguists will adhere to this classification and systematically use the proposed terminology.

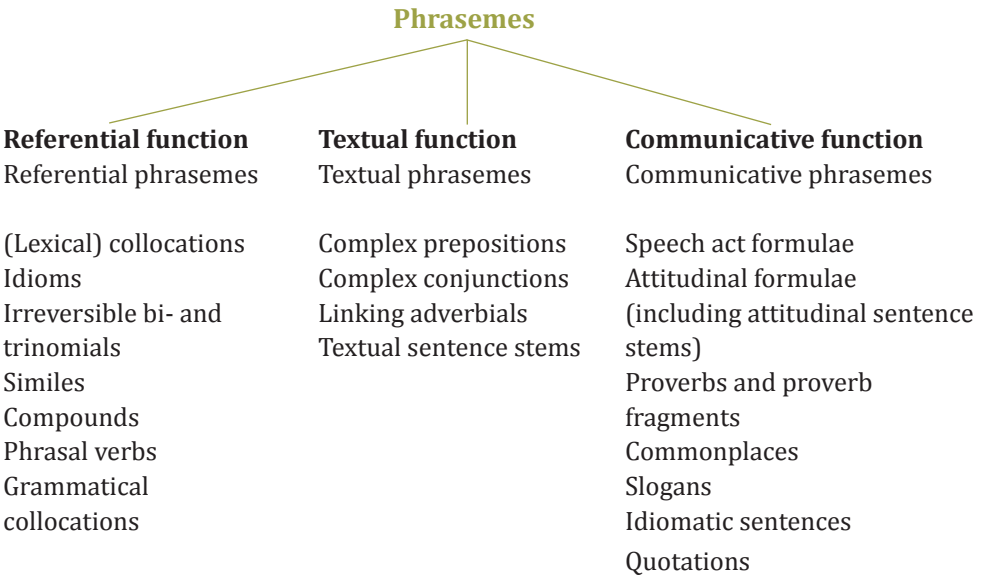


Figure 1.2 The phraseological spectrum
(reprinted from Granger and Paquot, 2008, p. 42)

As it happens, researchers do not always follow their colleagues. It may sometimes be explained by a specific design of individual research projects, traditions of linguistic schools in different institutions and many other circumstances. It is important to note, however, that whatever approach to the classification is chosen, be it phraseological, frequency-based or functional, it is always necessary to provide accurate operational definitions of phrases under study and in this way minimize any possible terminological confusion.

Finally, a very broad distinction of phraseological units could be related to two approaches to corpus analysis that appear in the title of this course book, namely, *corpus-based* and *corpus-driven* (see Tognini-Bonelli (2001) for a detailed discussion). When a corpus is merely used to exemplify uses of phrases or check their morphological and syntactic transformations, the researcher is most probably carrying out a corpus-based study and has a list of phraseological units chosen for the analysis. The corpus-driven approach, in contrast, begins from very different

premises: the researcher does not have a primary data set and therefore resorts to a corpus to find out what the linguistic reality is. Moreover, the absence of initial hypotheses means that the researcher is not certain what a corpus might yield and thus the whole research is *driven* by whatever comes out of the corpus data. So while phrases falling under the phraseological definitions are more closely related to the corpus-based approaches (e.g idioms), frequency-based identification of phrases is usually part of a corpus-driven research project.

Further reading

- Phraseology as research field; challenges and scope: Ellis (2008); Gries (2008).
- The word vs. lexical item vs. phraseological unit: chapter "The lexical item" in Sinclair (2004).
- Corpus-driven phraseology: Stubbs (2007); Biber (2009).

Study tasks

1. Consider methodologies used to calculate the degree of formulaicity in studies referred to in Table 1.1. What are their advantages and disadvantages? What exactly was being counted in each study?

2. An interesting comment on estimates of formulaicity was made by Wray (2000, p. 485), who questioned their validity on the grounds that corpora as such represent a community of speakers. In her view, corpus-based frequency counts are insensitive to idiolects and "agreed preferences of a speech community", both of which might inflate or, on the contrary, minimize the proportions of what she calls "formulaic material" (ibid., p. 466). Bearing in mind principles of corpus-based or corpus-driven research, how would you respond to Wray's remark? In your opinion, which extra-linguistic factors (e. g. age, gender, education, or occupation) may have an impact on the degree of formulaicity of a discourse community?

3. Analyse the dispersion of the phrase *you know* in different subcorpora of the BNC. Then compare co-occurrence frequency of the lemmas *you* and *know* in British English with the COCA data given on page 11. Write up your observations in a coherent paragraph (ca. 200 words).

4. Compare several dictionaries of English for foreign learners, e.g. Oxford Advanced Learners' Dictionary, Cambridge English dictionary, Longman Dictionary of

Contemporary English, for their approach to phraseology. How do they differ from the practice applied in the Collins COBUILD English Dictionary (see p. 11)? Do they enter any multi-word expressions as headwords in the body of the dictionaries, or are phrases always given as run-on entries? Write a paragraph (ca. 150-200 words) describing lexicographers' strategies for inclusion of phrases in a dictionary of your choice.

5. Reflect on your experience of learning English. Did you start with a list of individual words or rather tried to memorise complete phrases? In your opinion, how should young learners (aged 6–9 years) be taught a foreign language—by asking them to memorise individual words or rather providing them with lists of phrases? Explain your answer.

6. Apart from SLA, lexicography and translation, phraseological expressions are also being investigated in psycholinguistics, neurolinguistics, sociolinguistics and many other branches of linguistics. Find a recent research paper concerned with the analysis of any phraseological expressions and write its summary (ca. 300 words). Focus on research aims and the type of phraseological expression under study.

7. Marcinkevičienė (2010, p. 159) found that collocational pairs of words accounted for 68.1 per cent of the Corpus of Contemporary Lithuanian (the size of the corpus is 100 million words). How do you think this data from a corpus of Lithuanian, which is a morphologically inflected language, might correlate with estimates of formulaicity of English?

8. Phrases are ubiquitous. To do a simple estimate of the degree of formulaicity, read the following message from the Director of the Centre for Languages, Linguistics and Area Studies in the UK (https://www.llas.ac.uk//about.html). Can you identify which words are constituents of any type of multi-word unit? Try to count their proportion to all the running words in the text (the total number of words is 297). Compare your findings with evidence from other sources given in Chapter 1.

> *Dear colleague,*
> *As we sail into the lee of Christmas, most of us are looking forward to a break and some time out. It continues to be a difficult time for languages with a great deal of uncertainty about the future and a lot of pressures on teaching, research, management/admin, outreach, enterprise and*

all of the other activities we undertake. But the uncertainty has also brought opportunities, even if grasping them takes time and energy to think up new ideas, make business cases and submit bids.

In fact this term has shone a good deal of public attention on languages, with continued media coverage of major projects, like Language Rich Europe, and a concerted week of language-related events organized by the British Academy. There have also been significant government consultations on languages in schools and major implications flowing from proposed changes in school education and teacher training more broadly. There can be no doubting the increased concern nationally, and this is also echoed across Europe where many countries are being shaken in their complacency and beginning to worry about their capacity in languages.

One area of languages that continues to hold out hope for innovation is online education. This is an area in which languages have traditionally been very active. Technological changes are constantly challenging old approaches and offering new possibilities. LLAS at Southampton is about to host its 8th annual e-learning symposium on 24-25 January and already it is clear that the event will be bigger than ever this year. We have received twice as many session proposals as in previous years, and demand for places is correspondingly high. Readers thinking of attending should note that the early bird rate is closing on 20 Dec. It may be a reflection of where language learning is going.

Mike Kelly (follow Mike on Twitter at @ProfMikeKelly)

Director, LLAS Centre for languages, linguistics and area studies

9. Study the distribution and variability of phrases given below. Run different searches on the BNC and COCA and establish which grammatical forms of lexical verb(s) are typically used in these phrases. Apart from morphological changes, is it possible to establish a specific syntactic pattern in which these phrases usually occur?

to keep/get/stay in touch
to bring to light
to feel under the weather
to ring a bell
to get the sack
to spill the beans

10. Check the translation of different phrases on Google Translator. Try different language pairs. Do you think it is possible to establish a tendency that would account for inaccuracies in translation?

11. Consider phrases given below and choose a matching term for each of them from the list given in this section (page 15). When in doubt, check the internet for clarification of individual terms in the list.

    To whom it may concern
    To have influence
    To put up with
    On the other hand
    Bureaucratic red tape could be reduced.
    The black and white approach turned out unfruitful.
    He knew the poem by heart.
    I know I know, but I don't quite agree with you here
    He delivered the speech but he didn't tell the truth.
    Rolling stones are back.
    Bond. James Bond.
    Have you done your homework?
    What about that guy over there?

12. Misuses of phraseological units may be deliberate. 'Breaking' a fixed expression is one of the ways to create humour. Listen to a stand-up comedian's performance and find several examples of purposefully misused phrases.

13. Explain what the joke is based upon.

    An instructor is talking to a group of tourists about safety measures. At the end, he asks: "So what steps do you take if you see a grizzly bear approaching?"
    One of the tourists says: "Big ones. In the opposite direction".

CHAPTER 2.
IDIOMS

Although idioms are hardly ever associated with the development of corpus-based or corpus-driven approaches to lexicon, they have been chosen as a point of departure for the discussion of individual types of phraseological units. There are several reasons for such a decision. As the reader might have understood by now, *idiom* is often the first term that is evoked by any reference to phraseology. Apparently, to both linguists and laymen, the fact that phraseology has already gone a long way from the study of figurative expressions is not necessarily known. So it is reasonable to start an overview of phrases from idioms which by many are seen as the prototypical unit of phraseology.

The other argument for the inclusion of idioms in this course book is the amount of ongoing research into idiomatic treasures of languages, including diachronic and synchronic, contrastive and corpus-based studies, undertaken to inform the practice of translation, SLA, stylistic analysis and many others. Although a simple Google search is not the most reliable way of checking the currency of one or another notion, it does offer a rather crude yet convincing picture of what is happening in the virtual world and, obviously, beyond. The frequency of different phraseological terms on the internet (Table 2.1) is noteworthy. While the dominance of *idiom* could be explained by the long history of this term in language study, it is also remarkable that the term *n-gram(s)* is clearly having many more references in online resources than *collocation(s)*, which at least chronologically could have been expected to rank higher in terms of internet hits than the automatically extracted multi-word sequences. Obviously, *idiom* is the most frequent of all phrases occurring in internet sources, even if the word itself may certainly be used in other senses than the one related to phrases. Yet its currency most probably indicates that it continues to interest linguists.

In natural language, at least as it is represented in corpora, idioms are relatively infrequent. The only register where they have been found to occur in higher frequencies is fiction but even here the number is less than five per million words

Table 2.1 Results of Google search for different types of phraseological terms (searched on 18 January 2017)

Search string	Number of hits
idiom	21,100,000
idioms	25,400,000
n-gram	13,200,000
n-grams	5,190,000
collocation	5,290,000
collocations	6,420,000
"lexical+bundle"	16,700
"lexical+bundles"	29,200
"phrase+frame"	9,430
"phrase+frames"	16,700

(Biber et al., 1999, p. 1025). Another estimate is even less impressive—Moon (1998, in Vilkaitė, 2016, p. 34) argues that the frequency of idioms is less than one occurrence per million words. Yet idioms are quite conspicuous phrases in a text. The reason is that idioms are unexpected and often illogical word combinations. For example, the classic example *it's raining cats and dogs* 'raining heavily' certainly catches the eye because interpreted literally—and this is how it most probably is processed when encountered in the text for the first time—it brings to mind an unrealistic picture. To people who are less familiar with the history of British civilization, deriving the motivation behind this idiom is perhaps impossible. Its origin, as it turns out, goes back to English of the seventeenth century when

> *(…) there weren't any drains to take away heavy rain, so many pets and stray animals drowned during heavy downpours. So it was not unusual to see their bodies floating down the streets* (Watcyn-Jones 2002, p. 61).

So the meaning of this idiom is not instantly obvious because it cannot be understood from the individual meanings of constituent words. In linguistic terminology, idioms are said to be *non-compositional*, which means the meaning of an idiom is not *composed* as a sum of individual meanings of the constituents. Instead, the whole idiom has its own distinct meaning which is often impossible to explain by logical inference. Since idioms are striking and unusual, they often evoke a number of unexpected images and contribute specific emotive undertones to language. To put it differently, jumps and twists in the meanings of constituent words add colours to language.

It is thus natural that early linguists' interest in idioms was mostly related to stylistics and eloquence of speech. Hence, such research often involved analyses of literary texts. The tendency continues to persist in contemporary linguistics. For instance, the stylistic potential of idiomaticity is at the focus of Gläser (1998) and Naciscione (2010). Increasingly, however, studies of idiomaticity establish links with other realms of language study. Many valuable insights are derived from cognitive linguistics which, among other things, deals with metaphors. Cognitive linguists are trying to delimit boundaries between idioms and metaphors. While some argue that idioms are dead or frozen metaphors, others maintain that idioms offer new insights into language processing:

> *Recent research in psycholinguistics shows that the meanings of many idioms are motivated by people's conceptual knowledge, which includes metaphorical and metonymic schemes of thought. In this way, the study of idioms reveals significant aspects of how people ordinarily think.* (Gibbs, 1994, p. 277)

Advances in cognitive research provide linguists not only with methodology to analyse idioms but also enhances our understanding of natural language processing (cf. Gibbs, 1994, pp. 278–288). Finally, partly owing to the rise of automated language processing tools, linguists also engage in contrastive analyses of idioms across different languages, which has numerous applications in the fields of translation and language teaching/learning.

2.1. Definitions

According to the Oxford English Dictionary Online, the word *idiom* has several senses which are related to language, of which the one relevant in this context reads as follows:

> *a group of words established by usage as having a meaning not deducible from the meanings of the individual words.*

Notably, the word *idiom* has many other non-linguistic uses; furthermore, even as a linguistic term, it may have different interpretations. It is thus not surprising to see linguists suggesting that in order to minimize confusion the term should be avoided altogether (Naciscione 2010, p. 18) because it has been used in linguistic literature to refer to many different types of phrases which differ in the degree of compositionality. In this course book, the term *idiom* is used in its narrow linguistic sense as it is defined in the OED definition given above. Idioms are different from

proverbs and sayings which are typically full-sentence compact expressions of folk wisdom (e.g. *East or West home is best*). It should be noted, however, that scholars often modify the scope of this notion so that it suits their research designs. It is thus always necessary to consider carefully specific approaches used by every individual author and find out what exactly the term covers in a specific study.

Table 2.2 contains a selection of definitions of idioms and synonymous phraseological units from linguistic literature. The citations are taken from a number of studies and grammars and offer a number of different approaches to the understanding of this type of phrases. Yet even this small selection demonstrates terminological variation in approaches to what may appear a fairly clear-cut phraseological category. Since the approaches in some of the studies referred to in

Table 2.2 Definitions of *idiom* and similar phraseological units

Source	Definition
Gibbs (1994, p. 91)	"Idioms have traditionally been defined as expressions whose meanings are noncompositional or not functions of the meanings of their individual parts (…)."
Sinclair (1991, p. 172)	"An idiom is a group of two or more words which are chosen together in order to produce a specific meaning or effect in speech or writing. (…) The individual words which constitute idioms are not reliably meaningful in themselves, because the whole idiom is required to produce the meaning."
Biber et al. (1999, p. 1024)	"Idiomatic phrases—expressions with a meaning not entirely derivable from the meaning of their parts (…)."
Fernando (1996, pp. 35–36)	"A working definition of a pure idiom which is adequate for the present is 'a type of conventionalized, non-literal multiword expression'. *Spill the beans*, for example, has nothing to do with beans. In contrast to its literal counterpart meaning 'letting fall leguminous seeds', a non-literal meaning is imposed on the idiom as a whole: 'commit an indiscretion'."
Ishida (2008, p. 276)	"Idioms are multi-word expressions with the following three properties: i. formal frozenness (…) ii. syntactic frozenness (…) iii. semantic frozenness (…)"
Naciscione (2010, p. 32)	"[T]he phraseological unit is a stable, cohesive combination of words with a fully or partially figurative meaning."

Table 2.2 are broader or narrower in scope than the prototypical understanding of the idiom, the scholars use other terms than *idiom*. As a consequence, terminological variation in literature continues to flourish but it is most probably common practice in any research field that only time and widespread recognition of individual terms resolve the terminological confusion. For instance, Naciscione (2010) argues for the advantages of the term *phraseological unit* which, in her approach, also covers proverbs. In the opinion of Granger and Paquot (2008, pp. 42–43), *phraseological unit* is a cover term that includes all types of phrasemes (cf. Figure 1.2) and is not singled out in the classification which, interestingly, lists the term *idiom.* In order to avoid misunderstandings, it is important to find out what exactly is meant by each term. It may turn out that different terms refer to the same type of phrase, which is obvious even from the quotes in Table 2.2 where definitions of the prototypical idiom differ in scope.

2.2 Classification of idioms

The most straightforward way of grouping idioms could be derived from their surface structure. Hence, idiomatic phrases in English can be

1 – noun phrases, e. g. *pros and cons*, *chapter and verse* 'exact location or place', *a drop in the ocean*;
2 – verb phrases, e. g. *to ring the bell*, *to make mountain out of a molehill*, *to come clean*, *to go Dutch* 'pay your own bill';
3 – prepositional phrases, e. g. *in a nutshell*, *in the long run*, *on the spur of the moment*.

Biber et al. (1999, p. 1024) also mention *wh*-questions as another possible structural pattern of idiomatic phrases, for example, *what's up?*; *what on earth …?* etc. Clearly, if the operational definition of idioms covers full-sentence items, for instance, proverbs, the formal structural classification can be accordingly modified to account for any relevant syntactic feature.

Differences in the degree of non-compositionality of idioms may serve as another possible base for classification. For instance, Biber et al. (1999, p. 1025) illustrate how two idioms differ in the extent to which their meanings "can be derived from the components parts" by considering the following two expressions: *change one's mind* 'rethink a decision' and *kick the bucket* 'die'. Clearly, the meaning of the first expression is much closer to its literal meaning than is the case in the

second expression and on this basis it is logical to speak about idioms with different degree of non-compositionality.

A more thought-provoking approach to the classification of idioms was applied in a recent project, 'Widespread idioms in Europe and Beyond' (see the website at http://www.widespread-idioms.uni-trier.de/; also cf. Piirainen, 2008), where idioms were classified into five groups according to their cultural foundations. For instance, idioms with textual dependence are such phraseological units whose interpretation and origin draw upon textual sources (the Bible, fairy tales or classical literature). The other groups in this classification cover idioms originating in pre-scientific conceptual domains, cultural symbols, aspects of material and social culture. Apparently the aims of the project, formulated as identification of "the core set of idioms that actually exist in many languages, Europe-wide and beyond" (quoted from the project website) predetermined the basis for the classification and allowed researchers to draw relevant comparisons and analyse the obtained data. Whatever the approach to classification, it should be meaningful to the aims of the study in question.

2.3 Designing a study on idioms

Research projects of idioms in one language or in a contrastive perspective across several languages are popular among students of linguistics. A few important steps should be considered in order to design a valid study. The following checklist is meant to provide basic guidance to those interested in the study of idioms, especially involving corpus evidence. For a more detailed description of methodology for contrastive analysis of idioms, the reader is directed to Ishida (2008) which could be considered an exemplary paper demonstrating the contrastive idiom analysis.

2.3.1. Choosing your data

Compiling a list of idioms that describe, for example, human traits of character is not difficult. There are many dictionaries of idioms; plenty internet sites offer specific compilations of phrases in one or more languages. So collecting primary data from many sources is not particularly difficult, but it becomes clear very soon that something should be done with the long lists of idioms. It is therefore advisable from the very beginning to target a limited number of idioms related to one specific issue. It could be, for instance, idioms describing one particular trait of character or physical feature (brightness, stupidity, stubbornness, height etc.). It will certainly

yield a smaller data set but will offer more opportunity for an in-depth analysis, which is always more informative and, as it happens, more interesting than a superficial overview of a long list of items.

2.3.2. Objective judgement

In the studies of idioms, interpretation of idiomatic meanings, their implications or sometimes connotations should be made as objective as possible. The point of departure is always a lexicographic publication. Consulting several dictionaries in order to obtain authoritative definitions is essential. Apart from it, one could also engage human informants as an additional source of information. It is common practice in linguistic research to use native speaker informants (or, on the contrary, involve non-native speakers if the aims of the study in question deal with non-native interpretations of idioms). A small experiment when several informants are asked to comment on sample sentences containing idioms or identify idioms whose meanings are (non-)equivalent, or provide translational equivalents etc. allows the researcher to minimize subjectivity and, as a consequence, increase the validity of the study. An extensive survey of informants is also possible, especially with plenty of internet tools freely available, but it will inevitably increase the scope of study and require basic statistical skills to process the data.

2.3.3. Co-occurrence analysis

In order to capture the meaning of an idiom, its use in natural language should be carefully analysed. Here come corpora which represent different language varieties and offer plenty of ways to reveal how a particular idiom, let us say, *on and off*, is different from its non-idiomatic synonym *occasionally*, and whether it is equivalent to its renderings into other languages (Lith. *priešokiais*, *retsykiais*, *kartais*). It is essential to take into account naturally co-occurring data and examine lexical and grammatical patterns in which the idiom is used. Also, grammatical forms or constraints in transformation might prove that the idiomatic meaning is predominantly realised in one particular form of the idiom. For instance, the idiom *soaked to the skin* 'very wet' to judge from the BNC output seems to show preference for the verb in the Past Simple tense whereas when the prepositional phrase *to the skin* occurs with verbs in other tense forms, it is usually meant literally as in the following example:

(7) *Exercise stimulates blood flow to the skin and so gives rise to a healthy appearance* (...) (BNC)

Corpus-based analyses of idioms involve a close examination of concordances. In a contrastive study where more than one language is considered, a researcher might find himself/herself going through hundreds of instances, especially if the item in question allows for the literal and idiomatic interpretations. Hence the decision to set out on a corpus-based study of idioms inevitably means that it is important from the very beginning to consider carefully how the scope of the study could be delimited, as it was argued in section 2.2.1.

Further reading

- Idioms in the functional language perspective: Fernando (1996);
- Contrastive idiom analysis: Ishida (2008);
- Idioms in European languages: Piirainen (2008).

Study tasks

1. As you know, idioms are not necessarily reproduced in their prototypical form. Quite often they are used with certain elements omitted and/or transformed. Using the BNC or COCA, check variability of the idioms given below and decide what is their most typical form in English:

   Keep a straight face;
   Make a mountain out of a molehill;
   Pull someone's leg;
   To ring a bell;
   To take pot luck;
   To take something with a pinch of salt.

2. Gibbs (1994, pp. 92–93) gives an account of several experiments on how people process idioms which have both a literal and idiomatic reading. For instance, *he kicked the bucket* can be understood in two ways: literally as 'he pushed the bucket with his foot' and figuratively as 'he died'. The insights from such experiments are twofold. It was found that people tend to go for the literal interpretation first; yet once they know that the context supports idiomatic interpretation, they choose to process phrases as idioms. In your opinion, which of the readings of the following sentences given below (all of them taken from the BNC) is faster or primary? In your opinion, what does it depend on? Try to think about how you mentally process the sentence. How would you design an experiment to verify your answer?

1) *No sooner was I off the train than the guard blew the whistle and the train started and I had to run for it.*
2) *It could spot border violations and blow the whistle on breaches of cease-fire agreements.*
3) *With continual borrowing over two years, the bank had blown the whistle.*
4) *We have teething problems to sort out.*
5) *'Despite various teething problems we are very happy in our new life,' said Mr Turner.*
6) *They blame budgetary pressures, persistent teething problems, and cost overruns.*

3. Translation of idioms is always challenging. On the one hand, to preserve stylistic features of a text, written or spoken, a translator should aim at phraseological equivalence and thus render source language idioms into equivalent idioms of the target language. The difficulty lies in the fact that the translator needs to find an idiom that has an equivalent meaning while the wording may be very different for the source language. How would you translate the following sentences from the BNC into another language?

   1) *They used to pull my leg, but once they'd had a go they soon changed their minds.*
   2) *They took to their heels and ran up the road.*
   3) *It was interesting first time out, but after a couple of them I realised it wasn't my cup of tea.*
   4) *Nothing's going to be hidden, no skeletons in the cupboard, no dark secrets, everything out in the open where I can deal with it.*
   5) *Pushing the wheelbarrow should have been child's play, but I still could not get the hang of it.*
   6) *Then suddenly, right out of the blue, it had gone straight down the drain.*
   7) *When positive news did come, it again seemed to arrive out of the blue.*
   8) *Bargain-hunters (…) have the right to demand their money back if what they buy turns out to be a pig in a poke.*

CHAPTER 3.
COLLOCATIONS

One of the most significant contributions of corpus linguistics to the study of lexis and, more specifically, phraseology is undoubtedly the discovery and description of collocations. In contrast to idioms, collocations are said to be ubiquitous—albeit frequent they are not easily spotted. Such phrases as *take a risk*, *make fortune*, *tender meat*, *high temperature* are seemingly simple expressions and not eye-catching vivid idioms. Moreover, collocations are often semantically transparent because they do not carry hidden figurative meanings. Yet for whatever reason the combinations **fetch/grab a risk; *do/produce fortune; *soft meat*; **tall temperature* do not sound good in English, or are not acceptable as standard ways of expressing the ideas. Apparently, there are certain restrictions in the combinability of words while transparency and fixedness of collocations is a matter of degree.

It was only after the arrival of corpora that lexical co-occurrence patterns were observed and described in a comprehensive manner. While linguistic research into collocations started gaining pace after the publication of materials from the COBUILD project which was jointly undertaken by Collins Publishers and the University of Birmingham and headed by John Sinclair in 1980s, the notion of collocation was not altogether new to linguists.

Collocation of words has long been recognized as a way to study meaning. The idea of meaning through collocation goes back to J. R. Firth (1890–1960) who provided the theoretical background to the contextual theory of meaning:

> *Meaning by collocation is a direct consequence of the fact that, for Firth, the meaning of words lies in their use, and established usage will recognize words "familiar and habitual company"* (Tognini-Bonelli, 2001).

In other words, the co-text[4] of a word gives a researcher essential evidence on which any inferences about the meaning of that word could be drawn. Corpus

[4] The term *co-text* is understood here as it was explained in Tognini-Bonelli (2001, p. 87): the term is used "to refer specifically to the verbal environment that we are aiming to formalise, and the term *context* to refer to the situational and cultural parameters involved in the interaction (…) contextual elements, such as the relevant participants or the relevant object, will often have a correlate linguistic realisation in the co-text."

evidence shows that words are selective as to their partners, and the fact that one word co-occurs more frequently with the other than we would expect by chance most probably implies that we deal with a realisation of a specific sense of that word. The classical quote of Firth "you shall know a word by the company it keeps" is often cited in linguistic literature to explain the phenomenon of collocation and its role in meaning.

As regards phraseological research, the discovery of co-occurrences of words put onto the agenda of vocabulary studies a new type of phrases, namely, *collocations*. It turns out that words do not occur one by one in naturally produced language—having said one word we tend to choose another which goes well with the previous one. For example, *environmentally* evokes *friendly* but not any other of synonyms of *friendly*, for instance, *amiable*, *amicable*, *good-natured*, *kind*, *pleasant* etc. Similarly, *mistakes* in English are *made* but not *done* whereas *research*, on the contrary, is *done*. It was John Sinclair (1991) who brought the combinability of words into mainstream linguistics and showed that words build up a network of links with other words which occur in their co-text, or, to put it differently, attract each other. In a way, he developed Firth's ideas and argued that co-occurrence of words is the way language realizes meaning, and evidence from language corpora is essential for the study of meaning (cf. Sinclair, 2004, p. 134).

To explain how meaning is created Sinclair proposed two principles which are inactivated in the process of language production: the open-choice principle and the idiom principle (Sinclair, 1991, pp. 109–112). The open choice principle implies that language is a structure consisting of slots which are filled in with grammatically acceptable words. The 'slot-and-filler' model of language is at the basis of all traditional grammars (ibid., p. 110). For example, let us consider an English sentence having the following structure: S → NP V NP, for example, *He likes* ____, where the empty slot could be realized by any noun phrase (*Julia, her, surfing, cakes, reading* etc.) that is suitable according to the rules of grammar. Such decisions are governed by the open-choice principle. The idiom principle, in contrast, works beyond the limits of one sentence and is opposed to the open-choice principle. The idiom principle requires language users to select such words which naturally combine with each other. If we want to say that coffee has a high concentration of caffeine, it is typical to describe it as *strong*, for instance: *A coffee please, not too strong* (BNC). If grammaticality were the only consideration in this utterance, it would also be possible to use synonymous words, namely, *powerful* or *forceful*, but they are clearly unacceptable to describe the drink. Hence, the combination of *coffee* and *strong* is

a realisation of the idiom principle when the words are co-selected not because of their grammatical properties but owing to certain constraints in their meaning.

Furthermore, on the surface of it, the attraction of words to each other might appear to be inexplicable as it is difficult to arrive at a logical explanation why, for instance, the English adjective *high* attracts a different set of nouns than *tall* when it is used in English (Table 3.1), because any dictionary will provide similar definitions of *high* and *tall* suggesting that the two adjectives essentially mean the same. But even a very superficial comparison of the two sets of collocates shows that *high* collocates with nouns that have more abstract meanings whereas *tall* is often used to characterise more concrete nouns.

HIGH + *court, street, level, quality, rates, proportion, buildings, standards, unemployment, degree, speed, school, ships, risk, pressure*	TALL + *man, figure, trees, woman, buildings, windows, chimneys, boy, tower, ships, girl, glass, order, grass, houses*

Figure 3.1 Noun collocates of HIGH and TALL

Such differences between the two adjectives are difficult to arrive at by intuition. Admittedly, native speakers are often capable of determining whether two words go together well but even their judgment may be limited and related to different social factors, such as the level of (linguistic) education, occupation etc. Luckily, corpus linguistics resolved the issue of subjectivity and replaced the human intuition with corpus evidence. Sinclair (1991) convincingly showed that the tendency of a particular word to appear in the context of some other words is invisible to the naked eye and can only be reliably discovered by analysing corpora. Hence, a representative corpus of a language is necessary is order to identify typical 'companions' of any word.

Before delving deeper into collocations, some clarification of terminology is due. As the reader might have noticed, the term *collocation* may be used as an uncountable noun to refer to the phenomenon of word co-occurrence. *A collocation* usually refers to one particular phrase. Collocations consist of two words which *collocate* with each other. Usually, the word which is being analysed or queried in a corpus is referred to as *the node* whereas its 'partners' are called *collocates*. It should also be noted that *to collocate* (verb) is pronounced differently from *a collocate*

(noun). When looking up for collocates of a particular word, we need to specify the distance from the node to the left and right. The distance, also known as the span of the window, is measured in words, the default in many corpus access tools being five or four words. Meant humorously, the same terminology is sometimes used by linguists to refer to people, and it is not unusual to hear them say that they often "collocate" with one or another colleague.

3.1 Definitions

Back in 1965, a simple and beautiful definition of collocations in English and French was proposed by Sir Paul McCartney in what turned out to become an award-winning song of the Beatles:

Michelle, ma belle
These are words that go together well
My Michelle
Michelle, ma belle
Sont les mots qui vont tres bien ensemble
Tres bien ensemble

As it happens, the famous songwriter and singer covered at least two essential aspects of collocation: it consists of more than one word; the words form a well-sounding combination. Linguists, as could be expected, had more to add.

Generally speaking, the definition of collocations can be formulated in two ways: it can draw on the phraseological or frequency-based approach to the understanding of collocation (cf. section 1.3; also see Nesselhauf, 2005, p. 12). The two approaches reflect not only the conceptual differences behind the notion but also pertain to the ways in which collocations are extracted from a corpus.

The phraseological approach takes into account the degree of restriction in the meaning of the constituents of a collocation and goes back to the classification of phrases proposed by Cowie (1981). The term *collocation* in this approach refers to phrases that are made up of constituents that are not freely substitutable; hence Cowie's term *restricted collocations*. For example, *to hold talks* 'negotiate' would be considered a restricted collocation because the verb is used in its non-literal meaning, the whole phrase is transparent, and it allows some substitution as in *to have talks* (cf. Nesselhauf, 2005, p. 14). The main aspect of the phraseological definition of collocations is commutability, i. e. the extent to which individual constituents of a collocation may or may not be freely substituted with synonymous words.

The phraseological approach most often points to certain peculiarities of data collection. In practice it means that a set of collocations which are defined phraseologically is inevitably a subjective selection of items performed manually by the researcher from a raw output of a corpus analysis program or manual text analysis. To put it differently, it is the researcher's responsibility to decide for each individual collocation if its constituents are freely substitutable or their substitution is restricted. Admittedly, to minimise subjectivity of judgment, it is possible to set up a transparent methodology. For instance, Nesselhauf (2005, p. 54–63) gives a detailed account of steps in the process of data selection which was used to operationalise her phraseological definition of collocations. Manually extracted collocations were carefully considered to decide how commutable they are. The linguist used information given in monolingual English dictionaries which "was then supplemented with information from corpus analysis and from native speaker tests" (ibid, p. 54). So whatever was found in the dictionaries and the BNC was put to test with native speaker informants who either confirmed or rejected certain forms of collocations as (un)acceptable in English.

The frequency-based approach, in contrast, is more inclusive as it identifies collocations on the basis of statistical co-occurrence. In principle, it means that any combination, no matter how non-compositional or idiomatic it is, may be considered to be a collocation if it meets certain statistical parameters. These parameters are understood as co-occurrence of two words, i. e. constituents of a collocation, significantly more frequently than it could be expected by chance. The statistical definition of collocations is thus based on specific statistical measures which will be overviewed in Section 3.3. In contrast to the phraseological approach, in the statistical approach collocations are extracted from a corpus automatically whereas their co-occurrence statistics is provided as a built-in function in most corpus analysis tools that are available to researchers. So automatically obtained data is objective even though it does require some interference of the human analyst. The automatically produced list of collocates contains a considerable amount of so-called 'noise', or linguistically uninteresting collocates. For instance, functional words are sometimes less interesting to scholars who focus on lexical collocations, but others, on the contrary, might be investigating collocations with prepositions. It is thus the task of the researcher to go through the list and manually select the relevant collocates.

The choice of one or the other approach to the definition of collocations always depends on the aims and scope of research. Also, it is not unusual to see studies where a combination of the two approaches is used. First, statistical collocations

are retrieved from a corpus and then a researcher manually filters all collocates to select the most salient ones. Yet the approach used is not necessarily directly obvious from the definitions found in literature. Table 3.1 offers a selection of definitions of collocations from a variety of sources. Each of them was formulated for different aims and published in books or research articles that were not necessarily concerned with phraseology.

Table 3.1 Definitions of collocations

Source	Definition
Firth (1957, p. 196)	Firth suggested that collocation is a way to analyse meaning: "Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words."
Halliday and Hasan (1976, p. 287)	Collocation is "a cover term for the cohesion that results from the co-occurrence of lexical items that are in some way or other typically associated with one another, because they tend to occur in similar environments".
Sinclair (1991, p. 170)	"Collocation is the occurrence of two or more words within a short space of each other in text. The usual measure of proximity is a maximum of four words intervening."
Kjellmer (1991, p. 116)	"[C]ollocations are defined as recurring sequences that have grammatical structure."
Biber et al. (1999, p. 988)	Collocations "are associations between lexical words, so that the words co-occur more frequently than expected by chance. (…) Unlike idioms, collocations are statistical associations rather than relatively fixed expressions. Moreover, the individual words in a collocation retain their own meaning. However, part of the extended meaning of a word is the fact that it tends to co-occur with a specific set of collocates."
Nesselhauf (2005, pp. 25–34)	"[C]ollocations are considered a type of word combination in a certain grammatical pattern" (p. 25) The definition "has only been developed for verb-noun combinations" (p. 25). "The noun can be used without arbitrary restriction in the sense in which it is being used, but the verb is, in the given sense, to some degree arbitrarily restricted to certain nouns." (p. 33).
Vilkaitė (2016, pp. 32–33)	"In this study, a corpus-based statistical approach of defining collocations as frequent co-occurrences, without any additional semantic criteria is adopted. Examples of such collocations could be *take care*, *last night*, *learning difficulties*."

Whatever the definition, it is always important to carefully consider why the linguists chose to write about collocations. Knowing the broader context, it is easier to understand why one or the other approach to the definition is preferred and appreciate why certain aspects of collocation are given more prominence than others. Clearly, a mere choice of the term *collocation* does not presume a unified approach to its operationalisation.

3.2 Classification of collocations

To classify collocations one may proceed from their formal structure and identify different syntactic patterns, for instance,

(1) verb + noun, e. g. *to shrug one's shoulders*, *to compile a corpus*
(2) adjective + noun, e. g. *rancid butter*, *low pressure*
(3) adverb + adjective, e. g. *perfectly correct*, *utterly boring* etc.

A detailed classification based on the word classes in which constituents of collocations appear is provided in the "Oxford Collocations Dictionary for Students of English" (2002, p. ix). Benson et al. (1993), compilers of the "The BBI Combinatory Dictionary of English", also identify structural groups of collocations which they generally divide into two broad types: grammatical and lexical, as explained in the following excerpt:

> *Grammatical collocations consist of a dominant word—noun, adjective/participle, verb—and a preposition or grammatical construction. Lexical collocations, on the other hand, do not have a dominant word; they have structures such as the following: verb + noun, adjective + noun, noun + verb, noun + noun, adverb + adjective, adverb + verb.* (Benson et al., 1993, p. ix)

In many corpus-based and corpus-driven studies of collocations, linguists choose to focus on one particular syntactic type of collocations as it helps to obtain a reasonable and manageable amount of data. Coverage of a broad range of syntactic types, however, is preferred in such studies where collocation is not an object in itself but rather a means of investigating meaning of individual lexemes, synonymous pairs of words etc. Then a full picture of all types of structural categories might help a scholar to uncover specific patterns in the use of a word which, among other things, often sheds light on nuances in the meaning of that word.

Since collocations are combinations of words on a cline of non-compositionality, they can be also classified on the basis of as literalness. Hence, some collocations are

fully transparent (*give a present*) whereas others more opaque (*give a headache*). Between the two extremes, one might identify a broad category of medium non-compositionality (*give a sign*). One way of differentiating among the three examples is focusing on the verb, in this case dealing with the action of giving, and the object that is being given. Giving usually involves a transfer of an object from the agent to the recipient, which is exactly the case in *give a present*: one person transfers the object into the possession of the recipient. *Headaches* are clearly different in this respect as they cannot be transferred in the same manner as *presents*; they are sooner caused. Moreover, the doer of the action (agent) who *gives a headache*, 'shares' something he does not possess himself. *Give a sign* falls between the first two cases as the object *sign* belongs to the agent although it is not possessed in the same manner as *a present*, but neither is it so detached from the agent as in the case of *headache*. The opacity of these three collocations largely pertains to the verb, yet in other collocations, for example, *to show the way*, *to get into the way* and *way too much*, opacity stems from the noun.

Finally, collocations can be classified on the basis of commutability, understood as substitutability of their constituents. On the surface of it, commutability seems to be easier to identify as it refers to the constituent which can or cannot be substituted. For instance, *high temperature* is a collocation where none of the constituents can be replaced without a major change in meaning (*low temperature* refers to a different state of affairs), and it is impossible to say **tall/elevated temperature*. But the noun *survey* as in *to conduct a survey* collocates with several verbs of similar meaning, namely, *carry out*, *do*, *make*, *undertake*, so the two collocations, namely, *high temperature* and *to conduct a survey*, fall into two different groups on the basis of commutability: one is completely fixed whereas the other allows variability of one of its constituents.

Admittedly, a specific research design might dictate other approaches to classification, so this overview is in no way exhaustive. For instance, a study concerned with translation of collocations might propose categories related to translatability or translation strategies suitable for rendering collocations from one language into another. Hence, the choice of classification partly echoes the choice of the defining approach—it should be first of all relevant and meaningful to research questions formulated by the scholar rather than blindly follow an authoritative source.

3.3 Association measures

The discussion of ways to classify collocations so far has left out one of the most popular attributes used to describe this type of phraseological units, namely, weakness and strength, which are directly related to statistical parameters used in collocational research. More specifically, they refer to the strength of attraction between constituents of a collocation. Apparently, words, just like people, 'attract' each other, and the strength of attraction can be measured (although the measurement has not yet been shown to extend to attractions between human beings). This section will overview the basics of association statistics and will introduce several statistical measures readily available on corpus access sites and in corpus processing programs.

To understand the underlying arithmetical operations which are used to evaluate the strength of collocations in a corpus, let us take a closer look at how such measures are computed. What is measured, in fact, is the probability that one word will trigger the occurrence of the other. How certain are we that, for example, the use of *global* will necessarily trigger *warming*? Or vice versa, is it possible to claim that *warming* is a good predictor of *global*? To calculate the degree of attraction between two words, or even a word and a construction, the following frequencies should be known (cf. Levshina, 2015, pp. 223–224): *global warming* (599 occurrences in the BNC); *global* except its occurrences in *global warming* (2922 occurrences); *warming* except its occurrences in *global warming* (491). The data is usually arranged as a contingency table (see Table 3.2), which, besides the three mentioned numbers, also includes the total size of a corpus that we need not take into account to measure attraction but that is necessary for more sophisticated statistical measures.

Table 3.2 Co-occurrence frequencies of *global* and *warming* in the BNC

	global	***global*** (all other occurrences of *warming* without *global*)
warming	599	491
warming (all other occurrences of *global* without *warming*)	2922	98,308,219 (the total number of words in the corpus minus three numbers given in the other three cells)

The following computations should be performed with the data:

to measure the probability that *global* predicts *warming*: 599/(2922+599) = 0.17, or 17%;
to measure the probability that *warming* predicts *global*: 599/(491+599) = 0.55, or 55%.

So it is possible to conclude that *warming* is a much stronger predictor of *global* in the BNC because there is a probability of 55% that, in this corpus, we will find this adjective in the co-text of *warming.* In contrast, *global* has a weaker association with *warming* so there is less chance that we will see *warming* in the same sentence as *global*. Measures like these are known in corpus statistics as Attraction and Reliance (Levshina, 2015, p. 228), both being fairly simple unidirectional measures of association.

More statistical knowledge is needed in order to understand computation of bidirectional measures which take into account not only the frequency of co-occurrence of two words, but also the total corpus size. Luckily to many enthusiasts of corpus linguistics at different stages of proficiency, such tools are built-in in most corpus analysis tools and do not require a special training in statistical methods. It is, however, useful to have a basic understanding of some widely-used association measures in order to be able to interpret the automated computations. To begin with, it will suffice to know that bidirectional measures are based on statistical independence tests that evaluate how significant is the difference between the actual co-occurrence frequency of two words in a corpus and their expected frequency if no association between collocates is assumed (Levshina, 2015, p. 225). The more significant the difference, the stronger is the association in question. The ways to quantify that significance are many—Bouma (2009, p. 1) mentions over 55 measures, all involving specific statistical tests.

One of the most widely used measures is Mutual Information (MI). This measure is the default in the COCA output of collocates; it is also reported on the BNC site on the collocations output screen (one should select 'Mutual Information' in the drop-down menu for statistics). Mutual Information

> *compares the probability that the two items occur together as a joint event (…) with the probability that they occur individually and that their co-occurrences are simply a result of chance* (McEnery and Wilson, 2001, p. 86).

MI scores are fairly easy to interpret: the higher the score, the stronger the association between two words. Table 3.3 lists a few collocates of the noun *Christmas* in the

BNC output with different MI values listed. It should be remembered, however, that collocates do not necessarily represent adjacent sequences of words. The program reports any word that occurs within the window span of three words to the left/right from the node *Christmas.* To check whether it is an adjacent or rather non-adjacent sequence, one should view the actual concordance lines (easily done by going to the column 'Observed collocate frequency' in the output page and clicking on the number).

Table 3.3 A selection of collocates of *Christmas* from the BNC

Collocate	MI score
eve	8.79
decorations	8.48
carol	6.81
greetings	5.82
preparations	4.80
hope	0.11
house	0.16
parents	-0.35
services	-0.73
problems	-1.12

A high MI score proves that the two words (the node and respective collocate) are strongly connected, for instance, *Christmas + eve*, *Christmas + decorations*. If the score is negative, as in the last three rows of Table 3.3, the two words occur more often in isolation than in the same window span. Values that are close to zero (*hope* and *house*) point to such items whose occurrence in the proximity of *Christmas* may be a chance event.

Another widely-used association measure is t-score. It is, however, markedly different from MI, as t-score reflects typical associations with high-frequency words. The difference between MI and t-score is related to the type of co-occurrences that are treated by the program as significant. We have seen that MI collocates are sometimes fairly infrequent words in a language (*eve*, *carols*) but they build quite exclusive links with the node and form salient lexical collocations. When computing MI scores, the program disregards the absolute (raw) frequency of each word and gives more importance to such collocates that are normally rare words in language (cf. Gries, 2015) and their co-occurrence with the node is thus treated as a significant event. Collocates ranked by t-scores, in contrast, include high-frequency

words, functional word classes among them, therefore collocate rankings are very different from MI lists.

To illustrate differences among various association measures, Table 3.4 shows top twenty collocates of the noun *lung* in the BNC ranked by the strength of four types of measures. Besides MI and t-score, it also includes z-score whose computation takes into account not only the frequencies of two collocates but also the frequencies of all other words that occur in the specified window span (McEnery and Wilson, 2001, p. 86). The Log-likelihood measure, which is the default association measure in the BNC output, is computed with regard to absolute frequencies of each collocate and its list of collocates roughly falls in-between the MI and t-score lists.

Table 3.4 Top twenty collocates of *lung* in the BNC ranked by different association measures

rank	MI	t-score	Log-likehood	z-score
1	parenchyma	cancer	cancer	cancer
2	asbestosis	and	heart	parenchyma
3	punctured	heart	function	punctured
4	kai	his	disease	asbestosis
5	nonsmokers	function	air	heart
6	et-1	her	his	kai
7	lobes	into	and	transplant
8	cancer	from	punctured	function
9	obstructive	air	smoking	lobes
10	transplant	disease	transplant	obstructive
11	bronchitis	your	liver	et-1
12	pulmonary	my	into	nonsmokers
13	lymph	with	her	liver
14	cancers	the	deaths	smoking
15	kidneys	in	tissue	disease
16	fibrosis	smoking	collapsed	cancers
17	carcinoma	risk	chronic	pulmonary
18	bursting	patients	risk	deaths
19	respiratory	liver	parenchyma	collapsed
20	liver	deaths	lobes	chronic

As seen in Table 3.4, collocations with the highest MI are field-specific terms, medical in this case, whereas the t-score list includes many high-frequency words. Therefore, MI rankings might be relevant in studies dealing with terminology or more salient collocations (cf. Wang, 2016, p. 61). If one is concerned with grammatical

patterns of the node, collocates ranked by their t-score values are certainly more informative. While the z-score association measure is said to be better suitable for the extraction of collocates from a corpus of special text types, for example, literary, the Log-likelihood measure seems to yield a combined version of the MI/z-score and t-score list.

There is still no one clear answer in literature as to which of the available measures of association is the best. An approach quite often applied in studies of collocations is a combination of several measures when a researcher considers rankings for several different measures and produces a combined list of collocates that have high ranks by the majority of measures tested. Such approach is supported in literature:

> *One of the lessons taught by systematic evaluation of association measures against different gold standards is that there is not one association measure that is best in all situations. Rather, different target collocations may be found most effectively with different methods and measures. It is therefore useful to have access to a wide array of association measures coupled with an understanding of their behaviour if we want to do collocation extraction.* (Bouma, 2009, p. 2)

So the real collocations are most probably not to be established by application of one association measure. Interestingly, the only collocate that appears in all four columns of Table 3.4 is *cancer*; collocates that rank high for at least three measures are *punctured*, *heart* and *function.* This is hardly a list of collocates of the lexeme *lung* that a native speaker of English would produce if asked to give the most common words that *lung* combines with.

The remaining unanswered question is what kind of mental links one or another association measure captures most effectively. The author of a recent textbook on statistics in corpus linguistics very aptly observed that it is not quite clear what kind of reality is reflected in corpora:

> *It is still unclear which [association] measures represent the information that the speakers store in their minds more adequately than the others, since recent empirical studies based on corpus-based and experimental evidence have yielded divergent results.* (Levshina, 2015, p. 238)

So it is perhaps impossible to give a simple answer to the question of whether a particular phrase is a collocation because we still know very little about how associations between words are processed in the brain. As it should be clear to the reader by now, to call any phrase a collocation a linguist would first need to know

the approach to its definition (phraseological or statistical) used; the size of the corpus and the association measure used to extract collocates. If a phrase meets the phraseological definition, is extracted from a representative corpus of a language, has a high rank in the lists based on two or more association scores, we might be dealing with a real collocation. It remains to see, however, if findings from a corpus-driven analysis are confirmed by psycholinguistic experiments.

If one may think that measures of association between words are difficult to grasp, it should be remembered that they are merely capturing the linguistic competence which is possessed by any native speaker of a language. Links between words may not be visible to the naked eye yet native speakers have an intuitive awareness of combinatory properties of words and their co-selective preferences. Non-native speakers, in this respect being in a less advantageous position, need to develop their *collocational competence* in order to be able to produce native-like sounding natural language. So the association measures discussed in this section are not complex; complexity is the feature of the empirical world and one of its parts, i. e. our mental lexicon, which is so masterly and smoothly controlled by the human brain.

3.4 The extended unit of meaning

The analysis of collocates, indicating lexical and grammatical patterns of word co-occurrence, is closely related to the development of the theory of extended unit of meaning (Sinclair, 1996). It introduced the lexical item as a new and more broadly understood lexical structure which in the hierarchical structure of language deserves a place above the word. The identification of the lexical items draws on evidence from four levels of analysis required for the description of lexical items:

1. Collocation;
2. Colligation;
3. Semantic preference;
4. Semantic prosody.

Sinclair maintained that those "structural categories" had the potential of assuming "a central rather than a peripheral role in language description" (2004, p. 39) as they provide a comprehensive picture of any lexical unit, be it a word or a phrase. His analyses of the phrases *naked eye*, *true feelings* and the lexemes *brook* (verb) and *place* (noun) are convincing demonstrations of how the four levels may be instrumental in language study (2004, pp. 24–48). More importantly, a purely

corpus-driven approach was shown to provide an objective rather than intuition-based account of language use.

So what are the four levels of analysis? Collocation, as the term suggests, is the analysis of collocates of a particular word or phrase. The resulting collocational profile of a lexical unit shows significant lexical realisations of words that occur in the co-text of the item under study. Colligation refers to a set of grammatical choices, namely, articles, prepositions and specific constructions which occur when the item is used in natural language, or a corpus. Semantic preference pertains to semantic fields which are represented by the lexical collocates of the item. Finally, semantic prosody is the level of analysis that deals with pragmatic functions and involves examination of, for instance, negative, positive or neutral implications, of the analysed lexical items.

The four levels proposed as a means to describe any lexical item provide a robust methodology to anyone willing to describe a particular word or phrase; they also serve as a practical step-by-step approach to the study of synonymous words.

Further reading

- Cross-linguistic and contrastive studies of collocations: Molina-Plaza & de Gregorio-Godeo (2010); Xiao & McEnery (2006);
- Collocation as a way to describe meaning: Moon (2008); Römer (2009a); Stubbs (2007);
- Classification of collocations: Stulpinaitė et al. (2016);
- Collocations in second language acquisition: Durrant (2014);
- Collocational resonance: Williams (2008);
- Ways to test word associations: Zareva & Wolter (2012);
- The extended unit of meaning in vocabulary research: Tognini-Bonelli (2001).

Study tasks

1. Collocation and collocations were first described in detail when electronic corpora came into being. Can you explain why lexical patterns were discovered only through the study of corpora while grammatical patterns had been extensively described by linguists in the pre-corpus era?
2. Extract frequency data of the words *you* and *know* from the spoken subcorpus of the BNC and compare the strength of the collocation *you know* with its strength

in the spoken part of COCA. Compare your findings with the data given on p. 11 in Chapter 1.

3. Check which word of the following phrases is a better predictor of the other one? Is it the same in the BNC and COCA?

   a splitting headache
   a sweeping statement
   tectonic plates
   to clench one's teeth
   to look back
   to tell the truth
   to nod one's head
   to shrug one's shoulders
   totally crazy

4. A list of collocates of a particular word may differ from one variety of English to the other (British vs. American, spoken vs. written). Compare collocates of the following lexemes in different registers of English represented in the BNC or COCA:

   Nouns: *point*; *sense*; *way*
   Verbs: *like*; *give*; *set*

5. Collect relevant corpus data in order to answer the following questions: What is the collocational resonance (see Williams (2008) for the explanation and methodology) of *pen friend* (British English) and *pen pal* (American English) in current language use? Has it been changing over the last decades? Run searches on COCA, COHA (Corpus of Historical American English accessible through the COCA site) and the BNC.

6. Read the following extract from the style guide of *The Economist* about the meanings of *quite* in different varieties of English. Extract adjective collocates of *quite* from the BNC and COCA to verify the claims of editors at *The Economist.*

   quite *In America,* quite *is usually an intensifying adverb similar to* altogether, entirely *or* very*; in Britain, depending on the emphasis, the tone of voice and the adjective that follows, it usually means* fairly, moderately *or* reasonably*, and often damns with faint praise.* (Economist Books, 2005, p. 119)

7. Read another extract from the style guide of *The Economist*:

> ***red and blue*** *In Britain, colours that are associated with socialism and conservatism respectively; in the United States, colours that are associated with Republicans and Democrats respectively.* (Economist Books, 2005, p. 120)

How would you check whether the claims are valid? Use different corpora to provide evidence. Pay attention to possible differences across registers: spoken, written (newspaper language vs. academic) etc.

8. To see the difference between intuitive linguistic judgment and corpus evidence, do a simple experiment. Ask your friends to write down 5-7 words that often go together with *despair*, *spoon, looming*. Then extract collocates of these words from a corpus of English. Do you observe any differences between human intuition and corpus evidence?

9. Analyse corpus evidence to capture differences between the following pairs of synonyms:

> Verbs: *attempt* vs. *try* vs. *endeavour*; *buy* vs. *purchase*
> Nouns: *building* vs. *edifice*; *research paper* vs. *research article*
> Adjectives: *big* vs. *large*, *little* vs. *small*, *beautiful* vs. *handsome; subjective* vs. *biased*; *good* vs. *kind*; *quick* vs. *fast*
> Adverbs: *soon* vs. *shortly*; *absolutely* vs. *totally*

Write up your observations for one pair of synonyms in a coherent text (ca. 1,000 words).

10. Write an essay (ca. 600 words) in response to the following quote from John Sinclair (2008, p. 409):

> *[H]owever we circumscribe the unit of meaning, there will be connections like tentacles stretching out to the surrounding cotext, supporting or modifying the selection. Then, we have to concede that the normal primary carrier of meaning is the phrase and not the word (…).*

CHAPTER 4.

LEXICAL BUNDLES, N-GRAMS, RECURRENT SEQUENCES

The two types of phrases discussed so far, i. e. idioms and collocations, share a common feature—their constituents are related through meaning. In the case of idioms, we may observe a rather unexpected semantic shift of the primary senses of the words which produces non-compositional phrases. Collocations, in contrast, are often semantically transparent even though they involve a certain restriction in meaning which explains why the constituent words cannot be freely substituted with synonymous words. The present chapter introduces an inherently different approach to phrases which, on the surface of it, has nothing to do with word meanings. Moreover, *lexical bundles*, to call them by one of the terms used to refer to the new type of phrase, are automatically extracted from a corpus on purely formal criterion of recurrence at a certain frequency in identical form consisting of n words. Therefore they are often structurally and semantically incomplete, for example, *it was a*, *in order to*, *in the case of*, *you know* and similar. Only at a later stage, lexical bundles are investigated in terms of structure and functions in discourse. In this respect, they are totally different from idioms and collocations. But let us take care of first things first.

4.1 Definitions

In linguistic literature, the term *lexical bundles* and synonymously used terms *n-grams*, *clusters, chunks, recurrent sequences* usually refer to the same type of phrase. Table 4.1 below contains a selection of definitions taken from different sources.

Different terms for automatically generated sequences of words often reflect the tradition that scholars belong to, but essentially they denote a sequence or a string of words, both lexical and functional, which appears in a corpus in the identical form at a certain frequency. In the following, I will stick to the term *lexical bundles*

Table 4.1. Definitions of lexical bundles and related terms

Term	Definition
Lexical bundle	"Lexical bundles can be regarded as extended collocations: bundles of words that show a statistical tendency to co-occur." (Biber et al., 1999, p. 989) "Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status." (Biber et al., 1999, p. 990) "Lexical bundles are recurrent expressions, regardless of their idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in natural discourse." (Biber et al., 1999, p. 991)
Recurrent sequences; recurrent word combinations	"any continuous string of words occurring more than once in identical form" (Altenberg, 1998, p. 101) "sequences of word forms of length *n* which recur in identical form with frequency greater than *m* from a corpus using specialised software" (De Cock, 2004, p. 228)
N-grams	"repeated units of four words" (Forchini and Murphy, 2008) "combinations of *n* words" (Römer, 2009b, p. 91)
Chunks	"(…) recurrent strings of words, delimited by establishing frequency cut-off points, for example, that a string must occur at least 10 times per million words of text (…) and must be distributed over a number of different texts, to qualify as a bundle." (O'Keeffe et al., 2007, p. 61) "items in the automatically extracted strings which display pragmatic integrity and meaningfulness regardless of their syntax or lack of semantic wholeness" (O'Keeffe et al., 2007, p. 64)

which is the one introduced by the authors of *The Longman Grammar of Spoken and Written English* (Biber et al., 1999) where this type of phrase was extensively described for the first time in linguistics (Cortes, 2015, p. 203).

To better understand how lexical bundles are identified, let us take a closer look at the nature of these units. Imagine that we want to extract 4-word lexical bundles from a particular corpus which is uploaded on a computer. To generate lexical bundles, one has to use a corpus analysis program, the most popular in present-day research being *WordSmith Tools* (Scott, 2005) or *AntConC* (Anthony, 2015). There is also an online tool made freely available by Tom Cobb (http://lextutor.ca/n_gram/). These programs have a special function to generate sequences of

recurrent words termed clusters in *WordSmith Tools*, n-grams in *AntConC* and on the *Lextutor* site. When performing the function, the programs break up the entire corpus into sequences of n words. The length of lexical bundles depends on the aims of the study. In most studies of English, however, scholars prefer to analyse 4-word lexical bundles as they have been found to be more salient sequences than the shorter ones. For example, the sentence *There are several limitations associated with this research* (COCA) would be broken up into the following 4-word sequences:

there are several limitations
are several limitations associated
several limitations associated with
limitations associated with this
associated with this research

Usually the programs recognise sentence breaks by default, and the procedure is repeated anew for each punctuated sentence in the corpus. As a result, in a few seconds we get an output file in which all sequences of four words that are repeated at least twice in the uploaded corpus are listed in the order of frequency. The list may be easily exported into the required format and saved, for instance, as an Excel file, for further analysis. A sample output list may look like the one given in Figure 4.1. What is instantly obvious is the fact that many lexical bundles are structurally incomplete units (*there are a lot*, *of the most important*). The other peculiarity of phrases of this type is their semantic transparency. Most of them are fully compositional; yet there are also such discourse markers as *on the other hand* and *at the same time* which are semi-compositional. Obviously, if any idiom is repeated in a particular corpus more than once, it can make its way into the list of lexical bundles extracted from that corpus.

So in order to obtain a list of lexical bundles, a researcher needs to make several important decisions and formulate the operational definition of lexical bundles for any study. The first aspect to consider, as already explained above, is the length of the items. Many studies based on the English language deal with 3- and 4-word lexical bundles, but the decision is directly related to the aims of the research. The criterion of salience in the choice of the length was already mentioned above. In practice, it means that the shorter bundles, for instance, 2-word, are usually incorporated in the longer ones so it is reasonable to focus on a unit that is not to be split up into shorter sequences. The shorter lexical bundles are naturally much more frequent than the longer ones and manual revision of the automatically generated list might lead to a situation where the researcher doesn't see the wood for the trees,

		Frequency	Dispersion
1	on the other hand	79	71
2	is one of the	51	43
3	one of the most	42	41
4	at the same time	41	32
5	all over the world	26	21
6	there are a lot	25	21
7	are a lot of	24	20
8	it is clear that	21	19
9	of the most important	19	18
10	it is possible to	19	15
11	first of all the	17	17
12	the most important thing	17	15
13	the fact that the	16	16
14	there is no need	16	15
15	is considered to be	16	13
16	will be able to	16	13
17	a lot of people	15	15
18	it is important to	15	15
19	that there is no	15	15
20	it is obvious that	15	14

FIGURE 4.1 Twenty most frequent 4-word lexical bundles in LICLE

which is another reason why the longer lexical bundles are often preferable. More importantly, 2-word lexical bundles are less interesting for the qualitative analysis (structural and/or functional) because it is not so easy to generalize about the functions of such items as *in the*, *on the*, *can be* etc. Yet the short lexical bundles may sometimes be handy in order to filter out from a corpus some specific items, e.g. complex prepositions (*as to*, *as regards*, *according to*, *out of*), complex conjunctions (*as if*, *just because*) and similar. Whatever the length, its choice is one of the first decisions to be made.

The second important decision is related to the frequency threshold, or the minimum cut-off point. In practice, it means that linguists shorten the raw list of lexical bundles to those that have the greatest currency in the corpus and are the most frequent. It is not unusual to see different cut-off points applied to lexical bundles of different lengths. For instance, Biber et al. used the minimal frequency of at least ten times per million words for four-word lexical bundles (1999, p. 990)

and five times per million words for five- and six-word lexical bundles (ibid., p. 992). In studies where the approach to data selection is more stringent, the minimal cut-off point may be set at forty times per million words, which is often the case in studies based on corpora that are smaller in size and that, as research shows, yield many more lexical bundles than large corpora (Cortes, 2015, p. 205). This decision is also linked to the scope of a study—a sample adequate for a term paper might be smaller than the one chosen for a more extensive study.

Finally, the last aspect to consider is the dispersion of a lexical bundle across different texts that make up a corpus. A lexical bundle is only considered to be such if it is spread across different texts rather than occurs in one text. This way linguists ensure that whatever they analyse in the corpora is not idiosyncratic, i. e. characteristic of only one text or person who wrote that text. This decision is fairly flexible and is freely modified by linguists, usually in combination with the frequency criterion when delimiting the list of automatically generated lexical bundles. An interesting approach is described in Biber (2006, pp. 148–149) where the researcher relaxed the parameters for retrieval of data from certain subcorpora in order to avoid artificially inflated frequencies and obtain a reasonable sample of data. While it is difficult to formulate a rule of thumb for dispersion, common sense usually helps: if a lexical bundle occurs in two texts in a corpus that consists of 500 texts, any generalizations about currency of that bundle in the corpus would hardly be valid. If it occurs in more than half of the texts, it certainly has a different status in that corpus. Any dispersion in-between the two extremes, coupled with the criterion of absolute frequency, usually leads to a more or less reasonable sample.

A lexical bundle is thus a phraseological unit defined on the basis of the following parameters: 1) length, usually set between two and six words; 2) normalised frequency per million (or any other number of) words; 3) dispersion: the minimal number of texts in which it should occur to be included in the sample. While some linguists do not fully define these criteria when they present the definition of lexical bundles (see Table 4.1), these parameters are inevitably explained very accurately when discussing research methods and operationalising the general definition of lexical bundles.

4.2 Qualitative analyses of lexical bundles

Qualitative analysis of lexical bundles was first introduced in Biber et al. (1999) where the major structural types and their distribution across two registers of English (spoken conversation and written academic) were overviewed. A refined

classification of lexical bundles into structural and, for the first time explicitly, functional types was proposed in Biber et al. (2004) and further revised, especially the functional subtypes in Biber (2006). The structural classification is based on the formal word class features and the functional classification takes into account the functions of lexical bundles in discourse. While the two bases of classification cause little argument, the ways of delimiting categories, as it will be demonstrated below, leave much space for doubt and ongoing debate among linguists. The following presentation of the two classifications largely draws on Biber et al. (2004).

Structurally lexical bundles fall into three major types:

I. Lexical bundles that incorporate fragments of verb phrases;
II. Lexical bundles that incorporate fragments of dependent clauses;
III. Lexical bundles that incorporate fragments of noun phrases and prepositional phrases.

In addition, each type is subdivided further into several subtypes (Biber et al., 2004, p. 381). The analysis of different registers of English shows that while lexical bundles of type I are more frequent in spoken language, nominal bundles (type III) have more prominence in the written registers of English (Biber, 2006). Applications of the structural classification do not cause any disagreement among linguists and seem to be generally accepted. More difficulty, however, arises from the application of the functional classification.

The functional classification proposed by Biber et al. (2004) and revised in Biber (2006) reflects the three metafunctions of language described in the systemic-functional theory of grammar that was formulated by the famous British linguist Michael Halliday. The three metafunctions are ideational (we use language to refer to real and imagined world), interpersonal (we use language to communicate stance) and textual (we use language to create coherent texts in speech and writing). Accordingly, all lexical bundles fall into one of three functional types associated with the metafunctions of language, namely,

I. Referential lexical bundles, including expressions of imprecision, quantity specification, multifunctional reference;
II. Stance lexical bundles, including stance, attitudinal, modality, imperatives etc. bundles;
III. Discourse-organising lexical bundles, including lexical bundles of topic introduction/focus, topic elaboration/clarification, identification/ focus.

In Biber (2006), the fourth type, i. e. special functions lexical bundles referring to politeness and inquiries, was added to account for certain bundles established in spoken university registers.

The broad functional types are subdivided into several subtypes which may differ from one variety of language to another. The reason is that a particular function and, consequently, lexical bundles which express it, are well-attested in one register but may be absent in the other registers of English. For instance, referential bundles expressing imprecision (*and something like that*, *and stuff like that*, *or something*) are typical for speech but hardly ever occur in written academic English. Similarly, lexical bundles *in accordance with the* or *with regard to* may be characteristic expressions of written language but quite rare in conversation. Therefore, an analysis of the functions of lexical bundles usually begins with the examination of concordances of individual bundles and thus involves a considerable amount of text analysis and interpretation. Unavoidably, it leaves space for divergent interpretations and different readings.

To give an example, let us consider the function of what is a typical lexical bundle in any corpus of English, namely, *one of the most*. Structurally, this bundle is NP-based as it consists of a noun phrase with *of*-phrase fragment (cf. Biber et al., 2004, p. 381). In terms of its function in discourse, however, this lexical bundle may be attributed to different categories. Biber et al. (2004) categorized it as a referential bundle performing the function of identification/focus in most university registers, excluding textbook language. In Biber (2006, p. 159) it was classified as a discourse-organising bundle used for identification/focus in written university registers, but on p. 166, it appears among referential bundles performing the identification function in textbook English. The point is that it is not always possible to give one interpretation of an instance of use. The following three examples from the BNC illustrate challenges in the functional analysis:

(8) Bedford Park is regarded as the first Garden Suburb in England, and thus has an international importance, while Shaw is <u>one of the most</u> important architects of the Victorian period, best known as the designer of the old New Scotland Yard building on the London Embankment.

(9) Dictionaries describe a monograph as an account of a single subject; by this definition monographs make up <u>one of the most</u> common categories of art publishing.

(10) All species will do well, but C luciliae is <u>one of the most</u> widely available and dependable (…).

Apparently, the function of *one of the most* should be related to the adjective immediately following the bundle and the broader context. In all cases, it is related to identification yet what this identification is derived from is debatable. While in (8) it is easy to link it to the author's personal preferences and evaluation which is signalled by *important*, which would mean that we have a stance lexical bundle, in (9) and (10) the adjectives *common* and *widely available* imply focus on one object out of many known. Is it a focus in order to refer to the object or in order to narrow the following discussion? In the first case, we would categorize (9) and (10) as referential expressions; in the second, they could be interpreted as discourse organisers. It is therefore important to remember that a large part of interpretation remains necessarily subjective; it is usually a slow process with many returns and re-examinations.

An alternative functional classification of lexical bundles was proposed by Hyland (2008a; 2008b) who investigated Master's theses and dissertations written in English. Hyland identified the following functional types of lexical bundles:

I. Research-oriented, including lexical bundles referring to location, procedure, quantification, description, topic;
II. Text-oriented, including transition signals, resultative signals etc.;
III. Participant-oriented, including lexical bundles of stance and engagement.

Irrespectively of terminology used, Hyland's approach largely overlaps with the one proposed by Biber and his colleagues. The fact that one classification was published earlier than the other perhaps explains why Biber et al. (2004) is a much wider referenced source in literature than Hyland's works.

A new development in the analysis of lexical bundles involves research into how individual lexical bundles correlate with specific communicative purposes and rhetorical moves characteristic of academic prose (Cortes, 2015, pp. 210–212; Cortes, 2013). The focus in such studies is on lexical bundles used in different sections of academic research articles representing the four moves proposed by Swales (1990), namely, presenting the research topic, overviewing previous research, identifying a gap and introducing the current project. Hence, the qualitative analysis of lexical bundles may be based on their use to express one or the other rhetorical move of an academic text.

The structural and functional classifications of lexical bundles have been used to analyse academic English used across different language varieties (Biber and Barbieri, 2007) and disciplines (Cortes, 2004). The lexical bundle approach, as it is increasingly known among linguists, is also used to investigate non-native

varieties of English (Ädel and Erman, 2012; Chen and Baker, 2010; De Cock, 2004; Juknevičienė, 2009; 2013; Nekrasova, 2009; Paquot, 2013). Moreover, research into lexical bundles is also concerned with contrastive studies of different languages (Granger, 2014) and cross-linguistic (translation) research (Shrefler, 2010). Most probably, the relative ease of extraction of lexical bundles from a corpus continues to generate a considerable interest in the lexical bundles approach both among novice and experienced researchers. The design of a study of lexical bundles, however, requires a careful consideration of a number of aspects in order to avoid possible pitfalls in the analysis of those seemingly superficial phraseological units.

4.3 Methodological considerations

One may wonder whether a purely formal or technical approach to corpus data is at all valid. Apparently, it is. Moreover, it is even seen as advantageous:

> *The 'recurrent word combination' method is an illustration of corpus linguistics methodology at its most heuristic, i.e. as a raw discovery procedure. The method does not presuppose any linguistic category or pre-established list of sequences.* (De Cock, 2004)

Hence, it provides the researcher with objectively obtained raw data and, obviously, is a non-biased starting point. Yet it poses a number of methodological challenges in order to be fully appreciated.

The first important consideration at the initial stage of any study on lexical bundles is the overall representativeness of the corpus used, which is perhaps universal in any corpus-based and corpus-driven research. If a corpus is not representative of general language use, for instance, it is too small and biased towards one specific language variety, social group etc., it will yield highly specialised data. So caution should be taken to obtain an adequate set of lexical bundles and avoid hasty conclusions when formulating general statements about such data and contrasting it with material extracted from a different corpus. Frequent lexical bundles extracted from a corpus of research papers will most probably be different from those extracted from a corpus of newspaper articles, and it is the responsibility of the researcher to consider very carefully whether the data obtained lends itself to any contrastive comparisons or sweeping generalisations. Obviously, the smaller the corpus, the less generalizable are research findings.

The automatically generated list of lexical bundles is then submitted to qualitative (structural or functional) analysis. Yet prior to any qualitative analyses, it

is often necessary to carry out a manual revision of the list and weed out all kinds of irrelevant material. Sometimes the programs fail to recognize certain punctuation marks and treat words with the apostrophe (*don't* or *isn't*) as two words, namely, *don t* or *isn t* are counted as two words, which naturally leaves inaccuracies that have to be manually treated by the researcher. The other frequent issue has to do with capital letters which might yield two different tokens: *there are many* and *There are many*. Again, it is the researcher's responsibility to check if capitalisation could be disregarded and bundles with different capitalisation merged into one item with their frequencies reported as a sum. Lastly, a similar formal issue has to do with mistyped words all of which, if repeated, produce irrelevant items. Such technicalities have to be dealt with before any other steps of analysis are taken.

Another aspect deserving attention is lexical bundles that are made up of words from (book) titles, proper names, recurring quoted words and the like. In studies of lexical bundles such items are called topic-related or topic-specific lexical bundles. For instance, if one is analysing a corpus of film or book reviews, it is inevitable that the titles of films or books, if not eliminated when compiling the corpus or pre-treated as unanalysable material, will appear in the list of recurrent lexical bundles. It might be necessary to remove such topic-related bundles from the sample in which case it is done manually by checking the whole list and deleting the irrelevant material. In specific studies, however, the proportion of such 'quoted' words might be useful to consider as it provides evidence on the extent to which production of language involves reliance of writing input or quotes. But topic-related lexical bundles present an issue a researcher needs to take into account.

It is also important to remember that corpus linguistics requires a cautious treatment of frequency statistics which in this case refers to the frequencies of lexical bundles. In order to be able to compare frequencies of one particular lexical bundle in two corpora of different sizes, one should first normalise the absolute frequencies so that the comparison makes sense. Furthermore, some scholars include a normalised frequency parameter in the operational definition of lexical bundles which again requires that one has a clear understanding of what is involved in the normalisation counts. Let us consider the frequency data of *a lot of* in the BNC (Figure 4.2). It is difficult to make any claims about the currency of that sequence in the spoken and written subcorpora simply because the two subcorpora are different in size.

To be able to make a straightforward comparison, it is necessary to re-calculate the absolute number of hits per certain number of words. In this case, when the

Category	No. of words	No. of hits
Spoken	10,409,858	5,747
Written	87,903,571	8,885
total	**98,313,429**	**14,632**

Figure 4.2 Distribution of *a lot of* in the BNC

size of the corpus reaches millions of words, it is convenient to use a sample of one million words. So let us calculate the normalised frequency of *a lot of* in spoken and written subcorpora of the BNC per one million words (sometimes abbreviated as 'pmw'):

spoken: 5747*1000000/10409858=552.07 occurrences/pmw

written: 8885*1000000/87903571=101.08 occurrences/pmw

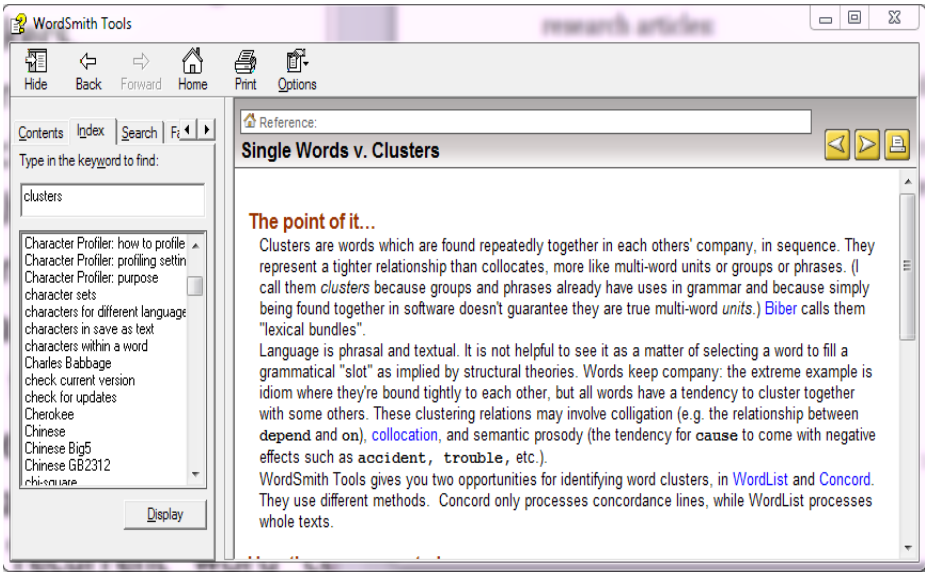
The sample for normalisation need not be one million; its size should be either meaningful or close in number to corpora under study or set to such value that would make calculations reasonably convenient, for instance, per 1,000 or 100, 000 words. Once the normalised frequencies are available, it is possible to compare the use of items in different corpora. In the case of *a lot of*, one might conclude that this phrase is 5.5 times more frequent in speech rather than in written language, which, obviously, explains why it is sometimes considered inappropriate in academic essays.

Further reading

- Lexical bundles in different registers of English: Biber (2006); Biber & Barbieri (2007);
- Lexical bundles in different disciplines: Cortes (2004); Forchini & Murphy (2008); Hyland (2008b);
- Lexical bundles in native and non-native language varieties: Ädel & Erman (2012); Chen & Baker (2010); De Cock (2004); Juknevičienė (2009; 2013); Nekrasova (2009);
- Applications of the lexical bundle approach to contrastive research: Granger (2014);
- Lexical bundles for the teaching of academic English: Martinez & Schmitt (2012).

Study tasks

1. Study the HELP window screenshot from the WordSmith tools (Scott, 2005). How does the definition of *cluster* relate to definitions given in Table 4.1 and other definitions of different types of phraseological units?



2. Study the description of the structural types of lexical bundles in Chapter 13.2 of the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). Write an essay (ca. 700 words) entitled 'Distribution of structural types of lexical bundles across two English registers: academic prose vs. conversation'.

3. Biber (2006) reports data on the distribution of lexical bundles in university registers. Study his findings and consider the use of English in the classroom. In your opinion, in terms of structural types of lexical bundles, is it closer to conversational English or written academic prose?

4. Figure 4.1 gives a list of top twenty lexical bundles extracted from a corpus of written English comprising student essays (LICLE corpus). Below (Figure 4.3) there is a list of top twenty lexical bundles from the LINDSEI-LT corpus which represents EFL learner speech. Analyse lexical bundles in terms of their structures and, in as much as it is possible, functions. Report the quantitative findings in a chart and write up the analysis in a short report (ca. 500 words).

LINDSEI			
1	I would like to	130	47
2	I don't know Ii	42	25
3	er I don't know	24	19
4	and I don't know	21	13
5	a lot of time	19	15
6	would like to visit	18	9
7	erm I don't know	17	9
8	I don't know what	17	16
9	I would love to	16	8
10	like to talk about	16	15
11	would like to talk	16	14
12	and I would like	15	12
13	I don't know it's	15	8
14	and it was very	14	6
15	but I would like	14	12
16	in the first year	14	11
17	and I think that	13	12
18	so I don't know	13	13
19	at the same time	12	9
20	I think I would	12	10

Figure 4.3 Twenty most frequent 4-word lexical bundles in LINDSEI-LT.

5. If lexical bundles are recurrent word sequences in one language (English), arguably, they should have more or less regular equivalents in other languages. Using any parallel corpus of English and any other language, for example, the online multilingual resource GLOSBE (https://glosbe.com/), analyse translational equivalents of the following English lexical bundles in the language of your choice:

   at the rate of
   in relation to
   in the case of
   in terms of
   the amount of
   the degree of
   the number of

to a certain extent
with regard to

Write an essay (ca. 700 words) in which you describe your findings and observations about the translational equivalents of one chosen item.

6. Forensic stylometry is an area where n-grams have been found to be particularly useful. Combined with several other measures, n-grams were used to uncover the true identity of Robert Galbraith, the author of *The Cuckoo's Calling*. Read the account of this linguistic detective by Patrick Juola to find out the true identity of the writer and explain how it was revealed by linguistic analyses (available from http://languagelog.ldc.upenn.edu/nll/?p=5315).

POSTSCRIPT

This brief journey through the jungle of English phraseology was inevitably limited. Owing to the scope of the course for which it was written (one semester) and two-fold aims of the course, i. e. training both in phraseology and research writing, it would be difficult to cover all types of phraseological units. Among such regrettable omissions is the phrase-frame, a further development of the lexical bundle. This is essentially a lexical bundle with one variable slot, for example, *it is * to*, *as it could **, *is the * of*. The variable slot may be realised by a number of different lexemes while in its basic form a phrase-frame represents a recurrent lexical pattern. Phrase-frames of English extracted from the BNC are available on a website developed by Fletcher (n.d.), which the reader is encouraged to explore. Research into phrase-frames is gaining pace across different linguistic areas: specialised discourses (Fuster-Marquez, 2014; Grabowski, 2015); different registers of English (Gray and Biber, 2013); analysis of novice and proficient academic writing and learner English (Garner, 2016; Juknevičienė & Grabowski, n.d.). The phrase-frame approach reveals a construction-driven picture of language and more than any other phraseological unit links lexis and grammar.

The ongoing research into the nature of phraseological units and corpus research continues to provide evidence that language consists of chunks of words, whatever the term chosen for the chunks, be it collocations, lexical bundles or, more generally, formulaic expressions. Research in psycholinguistics demonstrates that chunkiness of language eases its processing and thus deserves a place not only in the applied avenues of language such as teaching, learning, translation etc. but also in the description of the language system. It is thus hardly surprising that phraseology as a branch of linguistics is gradually developing into a full-fledged research field. In his Preface to a volume of research papers on phraseology, Ellis aptly described the role of the field in the world of linguistics:

> *[P]hraseology pervades theoretical, empirical, and applied linguistics. Like blood in systemic circulation, it flows through heart and periphery, nourishing all*. (2008, p. 9)

As an interdisciplinary research field, phraseology attracts scholars from a variety of research strands and keeps growing. The fact that it does have a professional association, regular publications and conferences certainly testifies to a rise of its status. The European Society of Phraseology founded in 1999 and having its seat in Zurich arranges international conferences and publishes the Yearbook of Phraseology, all of which points to increasing institutionalisation of phraseology. Alongside, smaller research communities, for example, Formulaic Language Research Network, FLaRN, co-ordinated by Alison Wray (University of Cardiff), continue to provide discussion fora to both novice and seasoned researchers. Information available on their respective websites might be interesting to anyone seeking inspiration and recent updates on the latest developments.

REFERENCES

I. Corpora

BNC – British National Corpus. Available at: http://bncweb.lancs.ac.uk or http://www.natcorp.ox.ac.uk/.

COCA – The Corpus of Contemporary American English: 520 million words, 1990-present. Compiled by Mark Davies. Available at: http://corpus.byu.edu/coca/.

LICLE – Lithuanian component of the International Corpus of Leaner English (Grigaliūnienė et al., 2008).

LINDSEI-LT – Lithuanian component of the Louvain International Database of Spoken English Interlanguage (Grigaliūnienė and Juknevičienė, 2011).

LOCNESS – Louvain Corpus of Native English Essays, compiled at the Centre for English Corpus Linguistics, University of Louvain-la-Neuve, Belgium. 1998. Online description at: http://www.learnercorpusassociation.org/.

II. Dictionaries

Benson, M., Benson, E., & Ilson, R. (Eds). 1993 [1986]. *The BBI Combinatory Dictionary of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Collins COBUILD English Dictionary for Advanced Learners. 2001. Third edition. Edited by J. Sinclair. Glasgow: HarperCollins Publishers.

Oxford Collocations Dictionary for Students of English. 2002. Oxford: Oxford University Press.

Oxford English Dictionary Online. 2017. Oxford: Oxford University Press. Available at: www.oed.com, accessed on 2017-01-15.

III. Software

Anthony, L. (2015). AntConc (version 3.4.4). Available at: http://www.laurenceanthony.net/software/antconc/, accessed 2017-02-07.

Cobb, T. Compleat Lexical Tutor. Available at: http://www.lextutor.ca/, accessed 2017-01-10.

Scott, M. (2005). WordSmith Tools (Version 5). Oxford: Oxford University Press.

IV. Literature

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

Altenberg, B. 1998. On the Phraseology of Spoken English: The Evidence of Recurrent Word-Combinations, in: Cowie, A.P. (Ed.), *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press, pp. 101–122.

Biber, D. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam/Philadelphia: John Benjamins Publishing.

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

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

Biber, D., Conrad, S. & Cortes, V. 2004. 'If you look at …': Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25, 371–405.

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

Bouma, G. 2009. Normalized (pointwise) mutual information in collocation extraction, in: *Proceedings of the Biennial GSCL Conference*. Available at: https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf, accessed 2017-01-20.

Chen, Y.-H. & Baker, P. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14, 30–49.

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

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

Cortes, V. 2015. Situating lexical bundles in the formulaic language spectrum. Origins and functional analysis developments, in: Cortes, V. & Csomay, E. (Eds) *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*. Amsterdam: John Benjamins Publishing Company, pp. 197–216.

Cowie, A. P. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2, 223–235.

De Cock, S. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures* (BELL) 2, 225–46.

Durrant, P. 2014. Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics* 19, 443–477.

Economist Books (Ed.). 2005. *Style Guide*, 9th ed. London: Profile Books.

Ellis, N.C. 2008. Phraseology: The periphery and the heart of language, in: Meunier, F. & Granger, S. (Eds), *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 1–13.

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

Fernando, C. 1996. *Idioms and Idiomaticity*. Oxford: Oxford University Press.

Firth, J. R. 1957. *Papers in Linguistics*. *1934–1951.* London: Oxford University Press.

Fletcher, W. (n.d.) Phrases in English. Online resource. Available at: http://phrasesinenglish.org/, accessed 2017-01-20.

Forchini, P. & Murphy, A. 2008. N-grams in comparable specialized corpora: Perspectives on phraseology, translation, and pedagogy. *International Journal of Corpus Linguistics* 13, 351–367.

Fuster-Marquez, M. 2014. Lexical bundles and phrase-frames in the language of hotel websites. *English Text Construction* 7, 84–121.

Garner, J. 2016. A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research* 2, 31–67.

Gibbs, R. W., Jr. 1994. *The Poetics of Mind. Figurative Thought, Language, and Understanding*. Cambridge: Cambridge University Press.

Gläser, R. 1998. The stylistic potential of phraseological units in the light of genre analysis, in: Cowie, A. P. (Ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxfors University Press, pp. 125–143.

Grabowski, Ł. 2015. Phrase-frames in English pharmaceutical discourse: a corpus-driven study of intra-disciplinary register variation. *Research in Language* 3, 266–291.

Granger, S. 2014. A lexical bundle approach to comparing languages. Genre- and register-related discourse features in contrast. *Languages in Contrast* 14:1, 58–72.

Granger, S. & Paquot, M. 2008. Disentagling the phraseological web, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 27–49.

Gray, B. & Biber, D. 2013. Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics* 18, 109–135.

Gries, S.T. 2008. Phraseology and linguistic theory: A brief survey, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 2–26.

Gries, S. T. 2015. Quantitative designs and statistical techniques, in Biber, D. & Reppen, R. (Eds), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambrigde University Press, pp. 50-72.

Grigaliūnienė, J., Bikelienė, L. & Juknevičienė, R. 2008. The Lithuanian component of the International Corpus of Learner English (LICLE): a resource of English language learning, teaching and research at Lithuanian institutions of higher education. *Žmogus ir žodis* 10, 62–66.

Halliday, M. A. K. & Hasan, R. 1976. *Cohesion in English.* London: Pearson Education.

Grigaliūnienė, J. & Juknevičienė, R. 2011. Formulaic language, learner speech and the spoken corpus of learner English LINDSEI-LITH. *Žmogus ir žodis* 13, 12–18.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied linguistics* 19, 24–44.

Hyland, K. 2008a. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics* 18, 41–62.

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

Ishida, P. 2008. Contrastive idiom analysis: The case of Japanese and English idioms of anger, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 275–291.

Juknevičienė, R. 2013. Recurrent word sequences in written learner English, in: Šeškauskienė, I. & Grigaliūnienė, J. (Eds), *Anglistics in Lithuania: Cross-Linguistic and Cross-Cultural Aspects of Study*. Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 178–197.

Juknevičienė, R. 2011. *Lexical Bundles in Non-Native Speaker and Native Speaker Written English.* Summary of doctoral dissertation in English. Vilnius: Vilnius University.

Juknevičienė, R. 2009. Lexical bundles in learner language: Lithuanian learners vs. native speakers. *Kalbotyra* 61, 61–72.

Juknevičienė, R. & Grabowski, Ł. (n.d., *submitted*) Comparing formulaicity of learner writing through phrase-frames: a corpus-driven study of Lithuanian and Polish EFL student writing.

Kjellmer, G. 1991. A mint of phrases, in: Aijmer, K. & Altenberg, B. (Eds), *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London: Longman, pp. 111–127.

Levshina, N. 2015. *How to do Linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Lewis, M. 2000. *Teaching Collocation. Further Developments in the Lexical Approach*. London: Language Teaching Publications.

Lewis, M. 1993. *The Lexical Approach. The State of ELT and the Way Forward*. Hove: Language Teaching Publications.

Marcinkevičienė, R. 2010. *Lietuvių kalbos kolokacijos*. Kaunas: Vytauto Didžiojo universitetas.

Martinez, R. & Schmitt, N. 2012. A Phrasal Expressions List. *Applied Linguistics* 33, 299–320.

McEnery, T. & Wilson, A. 2001. *Corpus Linguistics*. 2nd ed. Edinburgh: Edinburgh University Press.

Molina-Plaza, S. & de Gregorio-Godeo, E. 2010. Stretched verb collocations with give: their use and translation into Spanish using the BNC and CREA corpora. *ReCALL* 22, 191–211.

Moon, R. 2008. Dictionaries and collocation, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 313–336.

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

Naciscione, A. 2010. *Stylistic Use of Phraseological Units in Discourse*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Nattinger, J. R. & DeCarrico, J. S. 1992. *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Nekrasova, T. M. 2009. English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59, 647–686.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

O'Keeffe, A., McCarthy, M. & Carter, R. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Paquot, M. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18, 391–417.

Pawley, A. & Syder, F. H. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency, in: Richards, J.C. & Schmidt, R.W. (Eds), *Language and Communication*. London & New York: Longman, pp. 191–226.

Piirainen, E. 2008. Phraseology in a European framework, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 234–258.

Römer, U. 2009a. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7, 140–162.

Römer, U. 2009b. English in academia: Does nativeness matter. *Anglistik: International Journal of English Studies* 20, 89–100.

Shrefler, N. 2011. Lexical bundles and German bibles. *Literary and Linguistic Computing* 26(1), 89–106.

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

Sinclair, J. 1996. The search for units of meaning. *Textus* 9, 75–106.

Sinclair, J. 2004. *Trust the Text*. London & New York: Routledge / Taylor & Francis Group.

Sinclair, J. 2008. The phrase, the whole phrase and nothing but the phrase, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 407–410.

Stubbs, M. 2007. Notes on the history of corpus linguistics and empirical semantics, in: Nenonen, M. & Niemi, S. (Eds), *Collocations and Idioms*. Joensuu: Joensuun Yliopisto, pp. 317–329.

Stulpinaitė, M., Horbačauskienė, J. & Kasperavičienė, R. 2016. Issues in translation of linguistic collocations. *Kalbų studijos/Studies about Languages* 29, 31–41.

Swales, J.M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge & New York: Cambridge University Press.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins Publishing.

Vilkaitė, L. 2016. Formulaic language is not all the same: comparing the frequency of idiomatic phrases, collocations, lexical bundles, and phrasal verbs. *Taikomoji kalbotyra* 8, 28–54.

Wang, Y. 2016. *The Idiom Principle and L1 Influence: A contrastive learner-corpus study of delexical verb + noun collocations*. Amsterdam & Philadelphia: John Benjamins Publishing Company.

Watcyn-Jones, P. 2002. *Test Your Idioms*. Harlow: Pearson Education Limited.

Williams, G. C. 2008. The Good Lord and his works: a corpus-driven study of collocational resonance, in: Granger, S. & Meunier, F. (Eds), *Phraseology: An Interdisciplinary Perspective*. Amsterdam & Philadelphia: John Benjamins Publishing Company, pp. 159–174.

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

Wray, A. 2000. Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21, 463–489.

Xiao, R. & McEnery, T. 2006. Collocation, semantic prosody, and near synonymy: a cross-linguistic perspective. *Applied Linguistics* 27, 103–129.

Zareva, A. & Wolter, B. 2012. The "promise"of three methods of word association analysis to L2 lexical research. *Second Language Research* 28, 41–67.

Rita Juknevičienė

ENGLISH PHRASEOLOGY AND CORPORA

An introduction to corpus-based
and corpus-driven phraseology

ISBN 978-609-459-820-3

4,00 aut. l.

Išleido Vilniaus universitetas,
Vilniaus universiteto leidykla
Universiteto g. 3, LT-01513 Vilnius