SAL 29/2016

Received 06/2016
Accepted 11/2016

Towards a Refined Inventory of Lexical Bundles: an Experiment in the *Formulex* Method

Leksinių samplaikų sąrašo tikslinimas: bandymas taikyti „Formulex“ metodą

COMPUTATIONAL LINGUISTICS / KOMPIUTERINĖ LINGVISTIKA

Łukasz Grabowski
Dr, Associate Professor at the Institute of English, Opole University, Poland.

Rita Juknevičienė
Dr, lecturer at the Department of English Philology, Vilnius University.

http://dx.doi.org/10.5755/j01.sal.0.29.15327

Abstract

A number of corpus studies focusing on the description of the use and functions of lexical bundles have been conducted recently in order to explore the phraseology of learner language. As with any studies of lexical bundles, the problem of overlapping or structurally incomplete items poses a particular challenge. In practice, it is often difficult to align such units with specific discourse functions. The fact that lexical bundles do not constitute neat form-and-meaning mappings results from, among other reasons, their being grounded in language use rather than language system. In this pilot study we attempt to test a new method called *Formulex* (Forsyth, 2015a; 2015b) to verify whether an application of the criterion of coverage – in addition to the conventional criteria of orthographic length, minimum frequency and distribution range (Biber et al., 1999) – may help obtain a more refined inventory of lexical bundles and hence facilitate further qualitative analyses. To that end, we use Polish and Lithuanian components of the International Corpus of Learner English (ICLE, Granger et al., 2009), as well as the LOCNESS corpus (CECL), representing academic essays written by British and American students. The results revealed that many lexical bundles of fixed length identified in a conventional way are fragments of longer chunks of text and hence they should not be treated as complete or standalone 4-word lexical items. It was also revealed that the application of the *Formulex* method, where the word sequences are mutually exclusive, helps a researcher filter out overlapping or non-perceptually salient lexical bundles and, ultimately, specify more precise boundaries of lexical bundles of fixed length.

KEYWORDS: corpus linguistics, learner language, formulaicity, lexical bundles, Lithuanian and Polish learners

ktu 1922
Research Journal
Studies about Languages
No. 29/2016
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
pp. 58-73
DOI 10.5755/j01.sal.0.29.15327

Introduction

In recent years, recurrent sequences of words, known as lexical bundles, which have been shown to account for a significant part of both spoken and written English (Biber et al., 1999; Biber, Conrad, and Cortes, 2004; Cortes, 2008; Hyland, 2008a), have been frequently used as the unit of analysis in many phraseologically-oriented studies, notably the ones employing corpus-driven methodology. Lexical bundles are sequences of three or more words that occur frequently in natural discourse and constitute lexical building blocks used frequently by language users in different situational and communicative contexts (Biber et al., 1999, pp.990–991), e.g. *I don't think*, *as a result, the nature of the*. More often than not, lexical bundles are not idiomatic in meaning; on the contrary, the meaning of a lexical bundle is transparent (Biber, Conrad, and Cortes, 2003, p.134). As a result, the studies of lexical bundles position at the forefront inconspicuous and not perceptually salient multi-word sequences which acquire high frequencies in corpora. In this vein, Kopaczyk (2012, p.5; 2013, p.54, p.63) notes that lexical bundles are often either smaller than a phrase (notably, short bundles consisting of three or four words) or larger than a phrase (indicating complementation patterns of phrases).

A number of studies of lexical bundles have been conducted recently to explore the lexical characteristics of learner language, for example, to measure the extent to which these multi-word units are typical of spoken and written language produced by learners of English as a foreign language (EFL) or investigate in what ways they can shed light on the process of foreign language acquisition. Such studies fall into three major groups. One research direction involves contrastive analyses of EFL learner language representing different first language (L1) backgrounds on the one hand and a comparable corpus of native speaker data on the other (De Cock, 2004; Juknevičienė, 2009; Chen and Baker, 2010; Ädel and Erman, 2012; Baumgarten, 2014; Kizil & Kilimci, 2014). The other research strand deals with studies of lexical bundles across different proficiency levels of the learners (Hyland, 2008b; Römer, 2009; Vidakovic and Barker, 2010; Juknevičienė, 2013; Leńko-Szymańska, 2014) which gives researchers a pseudo-longitudinal perspective allowing to reveal changes in the use of lexical bundles alongside the increasing proficiency of learners. Finally, the third research direction, which up till now remains less exploited, is a comparison of lexical bundles in corpora representing learners whose mother tongues are different, e.g. Paquot (2013; 2014). However, no studies conducted so far focused on the use of recurrent lexical bundles by Polish and Lithuanian learners of English.

Despite growing popularity, the research on lexical bundles has not been devoid of methodological challenges. More specifically, the problems are directly related to the selection of salient lexical bundles from an automatically generated list which in most published research largely relies on the researcher's manual data analysis and subjective judgment. In particular, difficulties concern the methods used to deal with structurally incomplete bundles, filter out overlapping bundles, or select a representative sample of bundles other than focusing on the most frequent items (Grabowski, *in preparation*), to name but a few. In practice, many lexical bundles overlap with each other or constitute fragments of longer contiguous sequences of words, the problem that has been already identified in literature (cf. Appel & Trofimovich, 2015; Pęzik, 2015). For example, an initial list of lexical bundles might include such items as *the fact that it*, *the fact that we*, *in the fact that* etc. In order to identify the salient unit *the fact that* a researcher would need to go through the lists manually and at some point to decide that a certain recurrent sequence is more salient than others. Hence, the boundaries between many overlapping lexical bundles are not established objectively, let alone any subsequent alignment of the items with specific meanings or discourse functions (Appel & Trofimovich, 2015). That is why claims about all lexical bundles of fixed length being complete or distinct multi-word units often raise doubts and make it difficult to accept that lexical bundles may

be stored as single wholes in the mental lexicon of language users. In a similar vein, Simpson-Vlach and Ellis (2010, p.490) argue that "the fact that a formula is above a certain frequency threshold and distributional range does not necessarily imply either psycholinguistic salience or pedagogical relevance". It is the process of selecting the most salient lexical bundles that is at the focus of this article.

In this pilot study, we test a recently proposed method, called *Formulex* (Forsyth, 2015b), in an attempt to deal with overlapping, non-perceptually salient or structurally incomplete lexical bundles. The aim of this study was to fine tune the methodology used to identify lexical bundles in texts which is expected to provide a refined – and more useful pedagogically – list of these multi-word units. We will therefore try to answer the following research question:

> Can the *Formulex* method (Forsyth, 2015b) produce a refined list of non-overlapping and more perceptually salient lexical bundles as compared with the conventional approach (Biber et al., 1999)?

In view of the above, we hypothesize that, first, the criterion of coverage, implemented in the *Formulex* method (Forsyth, 2015b), can be used in addition to the conventional criteria developed to extract lexical bundles from texts, i.e. orthographic length, minimum frequency and distribution range, see Biber et al. (1999). It is also hypothesized that by using the criterion of coverage, where the sequences of words are mutually exclusive, it will be possible to produce a more refined inventory of non-overlapping, more perceptually salient and structurally-complete lexical bundles, which can be treated as distinct multi-word units. Although it is not an explicit goal of this study, the results may also reveal similarities or differences, in terms of the use of particular lexical bundles, across essays written by Polish and Lithuanian students as compared with the bundles most frequently employed by those students who are native speakers of English.

Methodology

This pilot study adopts a corpus-driven approach (Tognini-Bonelli, 2001, p.65), which means that the empirical corpus data is used to formulate hypotheses about linguistic features of written English produced by non-native (Polish and Lithuanian EFL learners) as well as native speakers. Two computer programs designed for text analysis were used in the study to obtain and process the research material. *Formulib* (Forsyth, 2015b), written in *Python 3.4*, was used to identify contiguous n-grams with the largest coverage in the corpora under scrutiny, and the software *WordSmith Tools 5.0* (Scott, 2008) was used to extract lexical bundles using three conventional criteria of orthographic length, minimum frequency and distribution range (Biber et al., 1999). Finally, the output of both programs was compared in order to filter out the results and identify those lexical bundles that meet all the four criteria employed in the study.

Learner corpora

Two learner groups, viz. Lithuanian and Polish EFL learners, are at the focus of the present study. To analyze their written English, we used two components of the ICLE corpus (Granger et al., 2009): a corpus of Polish learner English (PICLE) from the 1st version of the ICLE and a corpus of Lithuanian learner English (LICLE), a recent contribution to the ICLE project. Both corpora represent advanced EFL learners, senior undergraduate students majoring in linguistics-based study programmes in Poland and Lithuania. For reference purposes, we used the LOCNESS corpus, representing academic essays written by British and American students (CECL). Table 1 presents more detailed information on the composition of the corpora under scrutiny.

The corpora are different in size, hence all frequencies reported in this paper have been normalized per 100,000 words to allow for comparisons across the corpora.

Table 1
Corpora used in the study

	Number of essays	Size (words)
LICLE	335	191,570
PICLE	365	234,702
LOCNESS BR and AM	298	265,229

Stages of the Study

First, we established formulas, that is, contiguous sequences of four words, in PICLE, LICLE and LOCNESS that have the highest coverage or, in other words, the greatest currency in the corpora. It allowed us to test a new method, called *Formulex* (Forsyth, 2015a, 2015b), custom-designed to specify more precise boundaries of formulaic sequences, the problem that still remains unresolved in corpus-driven research on recurrent multi-word units in texts. This should enable us to identify those sequences that account for the greater proportion of the corpora under scrutiny.

Second, using *WordSmith Tools* 5.0 (Scott, 2008), we automatically generated lists of 4-word lexical bundles from PICLE, LICLE and LOCNESS using the conventional extraction criteria with the following parameters: minimum frequency = 5 occurrences per 100,000 (or 50 per million words), distribution range = 3 % of texts. The decision to focus on four-word lexical bundles has to do with the fact that a number of previous studies of English dealt with four-word sequences which have been shown to be more semantically and pragmatically salient (Biber et al., 1999; Biber, Conrad, Cortes, 2004; Hyland, 2008a; Hyland, 2008b). Moreover, it is the four-word lexical bundles that are traditionally investigated in learner corpora of L2 English (e.g. Chen and Baker, 2010; Ädel and Erman, 2012), so in order obtain comparable data, we also focused on four-word sequences.

Finally, we compared the output of both programs (*Formulib* and *WordSmith Tools* 5.0) in order to identify the so-called 'proper' or 'refined' lexical bundles. By 'proper' we mean such bundles that meet the traditional extraction criteria, as well as the coverage criterion (0.01 % or higher) established using *Formulib* software (Forsyth, 2015a). The study is expected to yield a refined list of lexical bundles, that is, the ones that do not overlap with each other and constitute more distinct multi-word units.

Results

N-grams with the greatest currency in the corpora

Using the *Formulib* software (Forsyth, 2015a) supported by *Python 3.4*, we identified 400 contiguous n-grams, built of four words or longer, with the highest coverage of texts in each corpus under study. Importantly, *Formulib* treats coverage as a binary category, that is, a number of n-grams matching a given text sequence is irrelevant; in other words, the program only takes into account the fact whether the text sequence is covered or not (Forsyth, 2015b, pp.13–14). For example, if n-grams such as *higher education in Lithuania* and *the quality of higher education* (and *the quality of* etc.) covers a certain part of the sequence *the quality of higher education in Lithuania*, each of those seven words is marked as covered once. Based on that, the proportion of covered to uncovered characters for each text sample is calculated and, next, the character coverage for each text category is aggregated (Forsyth, 2015b, pp.13–14).[1] For the sake of illustration, the results, that is, ten n-grams with the largest coverage in PICLE are presented in Table 2.

1 Although similar to one of the algorithms (*Serial Cascading Algorithm*) proposed by O'Donnell (2011, pp.149–153) to generate adjusted frequency lists of n-grams, Forsyth (2015b, p.25) notes that his "formulex" method "is simpler and has no fixed upper limit on the length of the sequences produced".

Table 2

PICLE coverage by frequent n-grams (by coverage)

	Coverage	Freq. (raw)	Characters	Words	Examples
1.	0.0648	49	17	4	*on the other hand*
2.	0.0405	29	18	4	*all over the world*
3.	0.0375	30	16	4	*at the same time*
4.	0.0268	26	13	4	*is one of the*
5.	0.0265	19	18	4	*it is obvious that*
6.	0.0251	11	30	5	***affect our approach to reality***
7.	0.0247	14	23	4	***our approach to reality***
8.	0.0235	16	19	4	*it is impossible to*
9.	0.0220	20	14	4	*as well as the*
10.	0.0220	20	14	4	*as a result of*

Apart from recurrent n-grams with the highest coverage, the data reveal that many of the formulaic sequences are in fact fragments of topics of students' essays. For example, the high coverage of the sequence *mass media affect our approach to reality* in PICLE, as well as of such sequences as *money is the root of all evil* (LICLE), *perception of the world* (LICLE), *the question of philosophical optimism* (LOCNESS), *in le mythe de Sisyphe* (LOCNESS), among others, shows that students, both native and non-native speakers, tend to frequently repeat the topic of the essay in their writing assignments. In fact, some LICLE essays were written as responses to long prompts (ca. 100-120 words), which were creatively used by the students in the essays. In contrast, PICLE and, to some extent, LOCNESS essays had considerably shorter topic formulations. This peculiar feature of the research material may inflate frequencies of certain n-grams that consist of lexical items found in the titles of student essays. That is why the decision has been made in this study to weed out those n-grams that are fragments of essay titles. Next, as in our study we ultimately aim to obtain a refined list of 4-word lexical bundles, it has been decided to remove the n-grams built of more than four words. Finally, as we aim to identify those n-grams that contribute the most to the formulaicity of student essays (or, in other words, have the highest currency or account for the greater proportion of the corpora under scrutiny), the decision has been made to focus only on those n-grams with the coverage of 0.01 % or higher.

Using the filtering procedures described above, we eventually obtained a list of 58 4-grams in PICLE, 75 4-grams in LICLE and 25 4-grams in LOCNESS with the highest coverage (that is, more than 0.01 %) in each corpus. Top ten n-grams in each corpus are presented in Tables 3, 4 and 5.

Table 3

PICLE coverage by frequent n-grams (by coverage)

	Coverage	Freq. (raw)	Characters	Words	N-gram
1.	0.0648	49	17	4	***on the other hand***
2.	0.0405	29	18	4	*all over the world*
3.	0.0375	30	16	4	***at the same time***
4.	0.0268	26	13	4	*is one of the*
5.	0.0265	19	18	4	*it is obvious that*
6.	0.0235	16	19	4	*it is impossible to*
7.	0.0220	20	14	4	*as well as the*
8.	0.0220	20	14	4	*as a result of*
9.	0.0201	13	20	4	*there are people who*
10.	0.0200	17	15	4	*on the basis of*

Table 4
LICLE coverage by frequent n-grams (by coverage)

	Coverage	Freq. (raw)	Characters	Words	N-gram
1.	0.0727	45	17	4	***on the other hand***
2.	0.0443	29	16	4	***at the same time***
3.	0.0341	20	18	4	*all over the world*
4.	0.0321	21	16	4	*it is clear that*
5.	0.0273	16	18	4	*it is obvious that*
6.	0.0264	21	13	4	*is one of the*
7.	0.0215	15	15	4	*will be able to*
8.	0.0214	14	16	4	*what is more the*
9.	0.0198	13	16	4	*with the help of*
10.	0.0198	13	16	4	*first of all the*

Table 5
LOCNESS coverage by frequent n-grams (by coverage)

	Coverage	Freq. (raw)	Characters	Words	N-gram
1.	0.0538	45	17	4	***on the other hand***
2.	0.0237	21	16	4	***at the same time***
3.	0.0213	16	19	4	*to a certain extent*
4.	0.0189	19	14	4	*in the case of*
5.	0.0186	14	19	4	*the majority of the*
6.	0.0159	15	15	4	*a great deal of*
7.	0.0158	14	16	4	*would be able to*
8.	0.0147	13	16	4	*when it comes to*
9.	0.0139	14	14	4	*as a result of*
10.	0.0138	13	15	4	*will be able to*

The results reveal, among others, that the sequence *on the other hand* has the highest coverage in each corpus which means that 0.0714 % of the all typed characters in LICLE consist of repetitions of the sequence *on the other hand*. Also, the n-gram *at the same time* is among the top ten by coverage in each corpus. One may also notice a number of similarities between Polish and Lithuanian learners of English, namely, the frequent use of a topic-neutral location marker *all over the world,* a sequence expressing writers' stance *it is obvious that*, or the sequence *is one of the* that functions as a focusing marker. Also, one may notice, Polish students and native-speakers often use the construction *as a result of* + 'sth' which is altogether absent in the LICLE data.

Apart from providing insights into the recurrent formulas, an additional benefit of employing the *Formulex* method is that it enables one to specify more precise boundaries between recurrent n-grams, notably overlapping or structurally incomplete ones (Forsyth, 2015b). For example, in 36 instances in PICLE the sequence *at the same time* was not a fragment of a longer sequence, namely, *and at the same time* (which occurs in PICLE 11 times); in fact, the sequence *at the same time* occurs, in total, 64 times in various patterns in the PICLE corpus, yet it appears as a 4-gram only 36 times. Hence the *Formulex* method takes into account the fact that "the sequences are mutually exclusive" and that "longer prefabricated phrases [are prevented] from being swamped by the elements of which they are composed of" (Forsyth, 2015b, p.17); this way the method enables researchers to delimit the boundaries of formulaic sequences more precisely, which has been one of the main, and still unresolved, problems in research on lexical bundles.

Hence, in what follows we will attempt to use both the data and insights from employing the *Formulex* method (Forsyth, 2015a), based on the criterion of coverage, in order to refine an inventory of lexical bundles generated in a conventional manner, that is, by using such criteria as orthographic length, minimum frequency and distribution range. Afterwards, we will identify the so-called 'proper' lexical bundles in each corpus, that is, the ones that appear in the output of both the *Formulex* method and the lexical bundles approach.

Lexical Bundles in Student Essays

Using *WordSmith Tools 5.0* (Scott, 2008), we generated a frequency list of lexical bundles using the traditional criteria (Biber et al., 1999) with the following parameters: orthographic length = 4; min. freq. = 5 occurrences per 100,000 words (or 50 per million words); minimum distribution range = 3 % of texts. Again, due to the specificity of the research material, the lexical bundles including words from the essay titles or prompts were excluded from further analyses. As a result, using the criteria described above, we identified 41 lexical bundles in PICLE, 40 lexical bundles in LICLE and 40 lexical bundles in LOCNESS. For the sake of illustration, top ten lexical bundles (by frequency) in each corpus are presented in Tables 6, 7 and 8. An asterisk (*) is used to mark that a given bundle meets the criterion of coverage employed in the *Formulex* method and set in this study at 0.01 % or higher. A full list of lexical bundles is presented in the Appendix to this paper.

Table 6
Top-frequency lexical bundles in PICLE

	Freq. (raw)	Norm. freq.	Range in no. of texts	Range, %	N-gram
1.	92	38.64	74	20.27	*on the other hand**
2.	64	26.88	55	15.06	*at the same time**
3.	39	16.38	34	9.31	*is one of the**
4.	37	15.54	35	9.58	*all over the world**
5.	30	12.6	26	7.12	*one of the most**
6.	24	10.08	21	5.75	*do not have to*
7.	22	9.24	19	5.20	*as a result of**
8.	21	8.82	18	4.93	*as well as the**
9.	21	8.82	20	5.47	*is the fact that**
10.	20	8.4	18	4.93	*are not able to**

Table 7
Top-frequency lexical bundles in LICLE

	Freq. (raw)	Norm. freq.	Range in no. of texts	Range, %	N-gram
1.	78	40.56	71	21.19	*on the other hand**
2.	51	26.52	43	12.83	*is one of the**
3.	42	21.84	41	12.23	*one of the most**
4.	40	20.8	31	9.25	*at the same time**
5.	25	13	20	5.97	*all over the world**
6.	25	13	21	6.26	*there are a lot*
7.	23	11.96	19	5.67	*are a lot of*
8.	23	11.96	20	5.97	*it is clear that**
9.	19	9.88	15	4.47	*it is possible to**
10.	19	9.88	18	5.37	*of the most important*

Table 8
Top-frequency lexical bundles in LOCNESS

	Freq. (raw)	Norm. freq.	Range in no. of texts	Range, %	N-gram
1.	62	23.56	37	12.41	*the end of the**
2.	48	18.24	41	13.75	*on the other hand**
3.	37	14.06	26	8.72	*at the end of*
4.	24	9.12	17	5.70	*the beginning of the*
5.	23	8.74	18	6.04	*as a result of**
6.	23	8.74	21	7.04	*one of the most**
7.	22	8.36	14	4.69	*at the beginning of*
8.	22	8.36	18	6.04	*the fact that the**
9.	21	7.98	18	6.04	*at the same time**
10.	20	7.6	17	5.70	*in the case of**

First, the data reveal that contrary to the output of *Formulex*, there are certain overlapping bundles among the ones identified in a conventional manner, e.g. *do not have to* and *they do not have* in PICLE, *of the most important* and *the most important thing* in LICLE, or *the beginning of the* and *at the beginning of* in LOCNESS, among others. This means that the 4-word bundles in question are, in fact, fragments of longer sequences of words, and that the conventional approach to the identification of lexical bundles of fixed length (e.g., 4 orthographic words) makes it difficult to identify the boundaries of such multi-word items. This problem does not apply to the list of recurrent sequences generated by the application of the *Formulex* method, where the sequences of words are mutually exclusive.

Second, the data show that not all lexical bundles identified in the traditional way meet the coverage threshold set in this study (0.01 %), which means that some of the items do not account for the greater proportion of the corpora under scrutiny. For example, the bundle *do not have to* in PICLE (ranked 6th by frequency; raw frequency of 24 occ.; normalized frequency of 10 occ. per 100,000 words; distribution range of 5.75 %, that is, 21 texts) has not been found among the 4-word grams of the greater coverage in the corpus. The reason for that is that the sequence in question, with a coverage of 0.0099 %, occurs independently, that is, not as a fragment of a longer sequence of words, only 9 times (2.4 times in terms of normalized frequency), which is even less than the normalized frequency threshold of 5 occurrences. Consequently, this and many other lexical bundles (e.g., *of the fact that, do not want to, it is better to, they do not have, the fact that the* in PICLE) are in fact fragments of longer multi-word constructions and hence they should not be treated as complete 4-word lexical items.

In view of the above, the comparison of the output of *Formulex* and the lexical bundles approach resulted in a refined inventory of 4-word lexical bundles, which meet the criteria of orthographic length, minimum frequency, distribution range and – in addition to the toolkit – coverage, the latter one applied in the *Formulex* method. The refined list (Table 9) includes 27 'proper' lexical bundles in PICLE, 26 in LICLE and 20 in LOCNESS.

The refined list of lexical bundles shows that LOCNESS which represents native speakers of English contains a smaller number of lexical bundles than the two corpora of non-native learners. This finding, in fact, confirms observations reported in Hyland (2008b), Römer (2009) and Juknevičienė (2009, 2013) that it is less proficient non-native users of language who tend to rely on repetitive sequences to a larger extent than learners of higher proficiency levels or native speakers. One way of accounting for this finding deals with the limited vocabulary range of the less proficient learners which often means that when writing they tend to

Table 9
Refined lexical bundles in PICLE, LICLE and LOCNESS in the order of decreasing frequency

PICLE	LICLE	LOCNESS
on the other hand	*on the other hand*	*the end of the*
at the same time	*is one of the*	*on the other hand*
is one of the	*one of the most*	*as a result of*
all over the world	*at the same time*	*one of the most*
one of the most	*all over the world*	*the fact that the*
as a result of	*it is clear that*	*at the same time*
as well as the	*it is possible to*	*in the case of*
is the fact that	*first of all the*	*is one of the*
are not able to	*the most important thing*	*the rest of the*
it is obvious that	*is considered to be*	*to a certain extent*
as a means of	*will be able to*	*a great deal of*
in the case of	*in order to get*	*the fact that he*
more and more people	*it is important to*	*the majority of the*
on the basis of	*one of the main*	*when it comes to*
as far as the	*the fact that the*	*would be able to*
in front of the	*a lot of people*	*will be able to*
it is impossible to	*he or she is*	*in the long run*
to the fact that	*I would like to*	*the way in which*
and that is why	*what is more the*	*it is important to*
they are able to	*in order to be*	*it is obvious that*
and what is more	*with the help of*	
become more and more	*do not want to*	
it is hard to	*for a long time*	
there are people who	*it is impossible to*	
a great deal of	*there are people who*	
it is enough to	*they do not have*	
on the one hand		

resort to repetitive lexical sequences rather than demonstrate a broader range of vocabulary as is the case in the native-speaker corpus.

Discussion

In comparison to the lists of lexical bundles extracted from the three corpora in the conventional manner, the refined lists present fewer overlapping n-grams that are often neighbours or near-neighbours on the frequency list. To give only a few examples, the bundles such as *but at the same* (PICLE), *there are a lot*, *are a lot of, the other hand the, is no need to, there is no need* (LICLE) or *to the fact that*, *due to the fact*, *the fact that they, the beginning of the, at the beginning of* (LOCNESS) are fragments of longer multi-word sequences. Hence, the *Formulex* method may help deal with the problem of 'syntagmatic overlap', the term proposed by Kopaczyk (2013, p.156) to refer to a situation when a given lexical bundle includes a fragment of the preceding one. Also, a number of inconspicuous multi-word units, which are syntagmatic associations hardly stored as single wholes in the mental lexicon of language users, have been removed from the refined list, e.g. *that it is not* (in PICLE, LICLE and LOCNESS), *it is not the* (LICLE), *that it is a* (LOCNESS). In that respect, one may argue that the

Formulex method helps one filter out overlapping or non-perceptually salient lexical bundles identified in the traditional way.

However, one may also notice that the refined lists of lexical bundles identified through the application of the *Formulex* method, which adopts the criterion of coverage, are not devoid of limitations. First of all, there are a number of 4-word sequences that contain more complete 3-word sequences, e.g. *first of all (the)*, *what is more (the)* in LICLE. This finding shows that it may be necessary to separately identify the n-grams shorter than or longer than four words, with the largest coverage in the corpora under scrutiny, and then filter out the results. Secondly, one may also note that a number of perceptually salient lexical bundles have been removed from the refined list, e.g. *it is better to*, *there is no doubt, seems to be the,* (expressing stance in PICLE), *it is obvious that* (expressing stance in LICLE), *is for the best*, *the only way to* (expressing stance in LOCNESS). This means that the application of the *Formulex* method may result in some information loss (as compared with the output of *WordSmith Tools 5.0*) that needs to be taken into consideration with respect to the scope and goals of a given study.

Conclusions

The aim of the study was to compare two methods to retrieve recurrent word sequences, termed lexical bundles, from a corpus and propose a more objective approach to data selection. It was found that the *Formulex* method proposed by Forsyth (2015a) allowed us to filter out a number of overlapping or not perceptually salient lexical bundles. It is thus possible to assume that application of the *Formulex* method yields a potentially more useful (pedagogically or otherwise) inventory of distinct multi-word units or – as it is the case in this study – provides a complementary insight into the recurrent multi-word units used by native and non-native English students in their essays.

A pilot study like this one may be only regarded as provisional, however. More research in the future is required to test the effectiveness of the *Formulex* method (Forsyth, 2015a) as compared with other methods or metrics developed recently to locate utterance boundaries or predict word sequence completion, e.g. a (forward and backward) transitional probability metric, which was tested by Appel and Trofimovich (2015) on a sample of 100 four-word items extracted from the BNC, or Independence-Formulaicity (IF) score (Pęzik, 2015), which gives more prominence to shorter n-grams that do not overlap with longer ones as well as to those n-grams that include multiple infrequent words. Also, the application of the *Formulex* method, in addition to the conventional criteria used to extract lexical bundles from texts, shows that the lexical bundles approach should be treated flexibly rather than strictly in order to provide a more comprehensive description of distinct and perceptually salient recurrent multi-word units in texts.

In view of the above, and irrespective of its limitations, this preliminary study is expected to be useful for corpus linguists exploring phraseological patterns and formulaicity.

References

1. Ädel, A. and Erman, B., 2012. Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: a Lexical Bundles Approach. In: English of Specific Purposes, no 31. Amsterdam: Elsevier, pp.81–92. doi: 10.1016/j.esp.2011.08.004. https://doi.org/10.1016/j.esp.2011.08.004
2. Appel, R. and Trofimovich, P., 2015. Transitional Probability Predicts Native and Non-native Use of Formulaic Sequences. In: International Journal of Applied Linguistics. [Online] Available at http://onlinelibrary.wiley.com/doi/10.1111/ijal.12100/epdf [Accessed 14 November 2016]. doi: 10.1111/ijal.12100. https://doi.org/10.1111/ijal.12100
3. Baumgarten, N. 2014. Recurrent Multiword Sequences in L2 English Spoken Academic Discourse: Developmental Perspectives on 1st

and 3rd Year Undergraduate Presentational Speech. In: Nordic Journal of English Studies, no 13 (3), pp.1–32. Available at http://ojs.ub.gu.se/ojs/index.php/njes/article/view/2889 [Accessed 10 September 2016].

4. Biber, D. et al., 1999. The Longman Grammar of Spoken and Written English. Harlow: Longman.
5. Biber, D., Conrad, S. and Cortes, V., 2003. Lexical Bundles in Speech and Writing: An Initial Taxonomy. In: Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech. Wilson, A., Rayson, P. and McEnery, T. (eds.). Frankfurt am Main: Peter Lang, pp.71–92.
6. Biber, D., Conrad, S. and Cortes, V., 2004. *If you look at…*: Lexical Bundles in University Teaching and Textbooks. In: Applied Linguistics, no 25 (3). Oxford: Oxford University Press, pp.371–405. doi: 10.1093/applin/25.3.371. https://doi.org/10.1093/applin/25.3.371
7. CECL (Centre for English Corpus Linguistics). LOCNESS. Louvain-la-Neuve: Université catholique de Louvain. Available at https://www.uclouvain.be/en-cecl-locness.html [Accessed 10 May 2016].
8. Chen, Y. H. and Baker, P., 2010. Lexical Bundles in L1 and L2 Academic Writing. In: Language Learning and Technology, no 14 (2), pp.30–49. Available at http://llt.msu.edu/vol14num2/chenbaker.pdf [Accessed 5 September 2016].
9. Cortes, V., 2008. A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish. In: Corpora, no 3 (1). Edinburgh: Edinburgh University Press, pp.43–57. https://doi.org/10.3366/e1749503208000063
10. De Cock, S. 2004. Preferred Sequences of Words in NS and NNS Speech. In: BELL – Belgian Journal of English Language and Literature, no 2, pp.225–246. Available at http://dial.academielouvain.be/vital/access/services/Download/boreal:75157/PDF_01 [Accessed 5 September 2016].
11. Forsyth, R., 2015a. *Formulib*: Formulaic Language Software Library. [Online] Available at http://www.richardsandesforsyth.net/zips/formulib.zip [Accessed 30 November 2015].
12. Forsyth, R., 2015b. *Formulib*: Formulaic Language Software Library. User Notes [Online] Available at http://www.richardsandesforsyth.net/docs/formulib.pdf [Accessed 2 November 2015].
13. Grabowski, Ł., (*in print*). Fine-tuning Lexical Bundles: A Methodological Reflection in the Context of Describing Drug-Drug Interactions. In: Patterns in Text: Corpus-driven methods and applications. Kopaczyk, J. and Tyrkkö, J. (eds.). Amsterdam: John Benjamins.
14. Granger, S. et al. (eds), 2009. International Corpus of Learner English. Version 2. Louvain-la-Neuve: Presses universitaires de Louvain.
15. Hyland, K., 2008a. As Can Be Seen: Lexical Bundles and Disciplinary Variation. In: English for Specific Purposes, no. 27. Amsterdam: Elsevier, pp.4–21. doi: 10.1016/j.esp.2007.06.001. https://doi.org/10.1016/j.esp.2007.06.001
16. Hyland, K., 2008b. Academic Clusters: Text Patterning in Published and Postgraduate Writing. In: International Journal of Applied Linguistics, no 18 (1), pp.41–62. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1473-4192.2008.00178.x/full [Accessed 1 September 2016].
17. Juknevičienė, R., 2009. Lexical Bundles in Learner Language: Lithuanian Learners vs. Native Speakers. In: Kalbotyra, no 61 (3). Vilnius: Vilnius University Publishing House, pp.61–71. https://doi.org/10.15388/klbt.2009.7638
18. Juknevičienė, R., 2013. Recurrent Word Sequences in Written Learner English. In: Anglistics in Lithuania. Cross-Linguistic and Cross-Cultural Aspects of Study. Šeškauskienė, I. and Grigaliūnienė, J. (eds.). Newcastle upon Tyne: Cambridge Scholars Publishing, pp.178–197.
19. Kizil, A. and Kilimci, A., 2014. Recurrent Phrases in Turkish EFL Learners' Spoken Interlanguage: A Corpus-driven Structural and Functional Analysis. In: Journal of Language and Linguistic Studies, no 10 (1), pp.195–210. Available at http://jlls.org/index.php/jlls/article/view/176/178 [Accessed 10 September 2016].
20. Kopaczyk, J., 2012. Long Lexical Bundles and Standardisation in Historical Legal Texts. In: Studia Anglica Posnaniensia: International Review of English Studies, no 47 (2-3). Berlin: De Gruyter, pp.3–25. doi: 10.2478/v10121-012-0001-0. https://doi.org/10.2478/v10121-012-0001-0

21. Kopaczyk, J., 2013. The Legal Language of Scottish Burghs (1380–1560). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199945153.001.0001
22. Leńko-Szymańska, A., 2014. *The Acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective*. In: International Journal of Corpus Linguistics, no 19 (2). Amsterdam: John Benjamins, pp.225–251. https://doi.org/10.1075/ijcl.19.2.04len
23. O'Donnell, M. B., 2011. The Adjusted Frequency list: A Method to Produce Cluster-sensitive Frequency Lists. In: ICAME Journal, no 35. Berlin: De Gruyter, pp.135–169.
24. Paquot, M., 2013. Lexical Bundles and L1 Transfer Effects. In: International Journal of Corpus Linguistics, no 18 (3). Amsterdam: John Benjamins, pp.391–417.
25. Paquot, M., 2014. Cross-linguistic Influence and Formulaic Language: Recurrent Word Sequences in French Learner Writing. In: EUROSLA Yearbook, no 14, Amsterdam: John Benjamins, pp.240–261. https://doi.org/10.1075/eurosla.14.10paq
26. Pęzik, P., 2015. Using n-Gram Independence to Identify Discourse-functional Lexical Units in Spoken Learner Corpus Data. In: International Journal of Learner Corpus Research, no 1 (2). Amsterdam: John Benjamins, pp.242–255. doi: 10.1075/ijlcr.1.2.03pez https://doi.org/10.1075/ijlcr.1.2.03pez
27. Römer, U. 2009. English in Academia: Does Nativeness Matter? In: Anglistik: International Journal of English Studies, no 20 (2). Heidelberg: Universitätsverlag Winter, pp.89–100.
28. Scott, M., 2008. *Wordsmith Tools. Version 5*. Oxford: Oxford University Press.
29. Simpson-Vlach, R. and Ellis, N., 2010. An Academic Formulas List: New Methods in Phraseology Research. In: Applied Linguistics, no 31 (4). Oxford: Oxford University Press, pp.487–512. doi: 10.1093/applin/amp058 https://doi.org/10.1093/applin/amp058
30. Tognini-Bonelli, E., 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6
31. Vidakovic, I. and Barker, F., 2010. Use of Words and Multi-word Units in Skills for Life Writing Examinations. In: Research Notes, no 41. Cambridge: UCLES, pp.7–14.

Santrauka

Łukasz Grabowski and Rita Juknevičienė. Leksinių samplaikų sąrašo tikslinimas: bandymas taikyti „Formulex" metodą

Svetimkalbių vartotojų produkuojamos kalbos tekstynai pastaruoju metu neretai tiriami aprašant leksines samplaikas, t. y. pasikartojančias tam tikro ilgio žodžių sekas, jų vartojimą bei funkcijas. Tokiuose tyrimuose neišvengiamai susiduriama su iš dalies sutampančiomis ar nepilnos struktūros samplaikomis. Pavyzdžiui, keturžodės anglų kalbos samplaikos *at the same time*, *the same time it* ir *but at the same* yra automatiškai generuojamos kompiuterio programa kaip tekstyne pasikartojančios sekos. Tačiau ar kiekviena jų laikytina tikrąja samplaika? Ar kiekvienai iš jų galima priskirti vienokią ar kitokią funkciją? Šiame straipsnyje siekiama parodyti, kaip „Formulex" metodas (Forsyth, 2015a, 2015b) leidžia parengti tikslesnį ir tyrimams vertingesnį leksinių samplaikų dažninį sąrašą ir tokiu būdu patikslinti programa „WordSmith Tools" ar panašiu tekstynų analizės įrankiu automatiškai parengiamą samplaikų sąrašą, kuris tradiciškai grindžiamas samplaikų ilgiu, minimaliu dažniu tekstyne bei dispersija skirtinguose tekstyno tekstuose (Biber et al. 1999), tačiau nepaiso teksto dengimo (angl. *coverage*) kriterijaus, kuriuo ir grindžiamas „Formulex" metodas.

Siekiant pademonstruoti „Formulex" metodo veikimą, straipsnyje naudojami du tarptautinio svetimkalbių produkuojamos rašytinės anglų kalbos tekstyno (ICLE, Granger et al. 2009) patekstyniai: lenkiškasis (PICLE) ir lietuviškasis (LICLE), pastarasis numatytas įtraukti į šiuo metu rengiamą atnaujintą ICLE versiją. Taip pat pateikiami duomenys iš LOCNESS tekstyno, sukaupto Liuveno anglų kalbos tekstynų lingvistikos centre (CECL). Straipsnyje aprašomas bandomasis tyrimas rodo, jog tradiciniu būdu išgaunamos leksinės samplaikos neretai yra ilgesnių pasikartojančių samplaikų dalys, tad jos turėtų būti analizuojamos ne kaip, pavyzdžiui, keturžodės, o penkiažodės samplaikos. Kitaip tariant, „Formulex" metodas leidžia tyrėjui tiksliau apibrėžti tiriamų pasikartojančių žodžių sekų ribas ir atsisakius persidengiančių ar atsitiktinių leksinių samplaikų, kokybiškiau atlikti pirminę duomenų atranką tolesniems tyrimams.

About the Authors

Łukasz Grabowsk,

Dr, Associate Professor at the Institute of English, Opole University, Poland.

Research interests

Corpus linguistics, phraseology, formulaic language, translation studies and lexicography. He is also interested in computer-assisted methods of text analysis.

Address

Institute of English
Opole University
Pl. Kopernika 11, 45-040 Opole, Poland

E-mail:

lukasz@uni.opole.pl

Rita Juknevičienė

Dr, lecturer at the Department of English Philology, Vilnius University.

Research interests

Learner corpus research, phraseology and language testing. She teaches courses on academic writing, translation, phraseology and general linguistics.

Address

Department of English Philology
Faculty of Philology
Vilnius University
Universiteto 5, LT-01122 Vilnius

E-mail:

rita.jukneviciene@flf.vu.lt

Appendix

A list of 4-word lexical bundles automatically generated by *WordSmith* Tools with 'refined' (in bold) lexical bundles, fulfilling the criterion of coverage

PICLE

No.	Lexical bundle	Freq.	Norm. freq	No of texts	Texts, %	Coverage
1.	***on the other hand***	**92**	**38.64**	**74**	**20.27**	*
2.	***at the same time***	**64**	**26.88**	**55**	**15.06**	*
3.	***is one of the***	**39**	**16.38**	**34**	**9.31**	*
4.	***all over the world***	**37**	**15.54**	**35**	**9.58**	*
5.	***one of the most***	**30**	**12.6**	**26**	**7.12**	*
6.	*do not have to*	24	10.08	21	5.75	
7.	***as a result of***	**22**	**9.24**	**19**	**5.20**	*
8.	***as well as the***	**21**	**8.82**	**18**	**4.93**	*
9.	***is the fact that***	**21**	**8.82**	**20**	**5.47**	*
10.	***are not able to***	**20**	**8.4**	**18**	**4.93**	*
11.	***it is obvious that***	**20**	**8.4**	**19**	**5.20**	*
12.	*of the fact that*	20	8.4	18	4.93	
13.	***as a means of***	**18**	**7.56**	**13**	**3.56**	*
14.	*do not want to*	17	7.14	13	3.56	
15.	***in the case of***	**17**	**7.14**	**13**	**3.56**	*
16.	*it is better to*	17	7.14	14	3.83	
17.	***more and more people***	**17**	**7.14**	**17**	**4.65**	*
18.	***on the basis of***	**17**	**7.14**	**16**	**4.38**	*
19.	***as far as the***	**16**	**6.72**	**15**	**4.10**	*
20.	***in front of the***	**16**	**6.72**	**15**	**4.10**	*
21.	***it is impossible to***	**16**	**6.72**	**15**	**4.10**	*

22.	*they do not have*	15	6.3	14	3.83	
23.	***to the fact that***	**15**	**6.3**	**14**	**3.83**	*
24.	***and that is why***	**14**	**5.88**	**13**	**3.56**	*
25.	*that we are not*	14	5.88	11	3.01	
26.	*the fact that the*	14	5.88	14	3.83	
27.	*there is no doubt*	14	5.88	14	3.83	
28.	***they are able to***	**14**	**5.88**	**11**	**3.01**	*
29.	***and what is more***	**13**	**5.46**	**12**	**3.28**	*
30.	***become more and more***	**13**	**5.46**	**11**	**3.01**	*
31.	***it is hard to***	**13**	**5.46**	**12**	**3.28**	*
32.	*seems to be the*	13	5.46	12	3.28	
33.	*the end of the*	13	5,46	13	3.56	
34.	***there are people who***	**13**	**5.46**	**13**	**3.56**	*
35.	***a great deal of***	**12**	**5.04**	**12**	**3.28**	*
36.	*but at the same*	12	5.04	11	3.01	
37.	*in such a way*	12	5.04	12	3.28	
38.	***it is enough to***	**12**	**5.04**	**12**	**3.28**	*
39.	***on the one hand***	**12**	**5.04**	**11**	**3.01**	*
40.	*that it is not*	12	5.04	12	3.28	
41.	*turn out to be*	12	5.04	11	3.01	

LICLE

No.	Lexical bundle	Freq.	Norm. freq	No of texts	Texts, %	Coverage
1.	***on the other hand***	**78**	**40.56**	**71**	**21.19**	*
2.	***is one of the***	**51**	**26.52**	**43**	**12.83**	*
3.	***one of the most***	**42**	**21.84**	**41**	**12.23**	*
4.	***at the same time***	**40**	**20.8**	**31**	**9.25**	*
5.	***all over the world***	**25**	**13**	**20**	**5.97**	*
6.	*there are a lot*	25	13	21	6.26	
7.	*are a lot of*	23	11.96	19	5.67	
8.	***it is clear that***	**23**	**11.96**	**20**	**5.97**	*
9.	***it is possible to***	**19**	**9.88**	**15**	**4.47**	*
10.	*of the most important*	19	9.88	18	5.37	
11.	***first of all the***	**18**	**9.36**	**18**	**5.37**	*
12.	*there is no need*	18	9.36	17	5.07	
13.	*it is obvious that*	17	8.84	15	4.47	
14.	***the most important thing***	**17**	**8.84**	**15**	**4.47**	*
15.	***is considered to be***	**16**	**8.32**	**13**	**3.88**	*
16.	*that there is no*	16	8.32	16	4.77	
17.	***will be able to***	**16**	**8.32**	**13**	**3.88**	*
18.	***in order to get***	**15**	**7.8**	**13**	**3.88**	*
19.	***it is important to***	**15**	**7.8**	**15**	**4.47**	*
20.	***one of the main***	**15**	**7.8**	**14**	**4.17**	*

21.	***the fact that the***	**15**	**7.8**	**15**	**4.47**	*
22.	***a lot of people***	**14**	**7.28**	**14**	**4.17**	*
23.	***he or she is***	**14**	**7.28**	**12**	**3.58**	*
24.	***i would like to***	**14**	**7.28**	**11**	**3.28**	*
25.	*it is not a*	14	7.28	13	3.88	
26.	*it is very hard*	14	7.28	14	4.17	
27.	***what is more the***	**14**	**7.28**	**12**	**3.58**	*
28.	*do not think that*	13	6.76	11	3.28	
29.	*in order to be*	**13**	**6.76**	**13**	**3.88**	*
30.	*it is not the*	13	6.76	11	3.28	
31.	***with the help of***	**13**	**6.76**	**13**	**3.88**	*
32.	*at the end of*	12	6.24	12	3.58	
33.	***do not want to***	**12**	**6.24**	**11**	**3.28**	*
34.	***for a long time***	**12**	**6.24**	**12**	**3.58**	*
35.	*is no need to*	12	6.24	11	3.28	
36.	***it is impossible to***	**12**	**6.24**	**12**	**3.58**	*
37.	*that it is not*	12	6.24	12	3.58	
38.	*the other hand the*	11	5.72	11	3.28	
39.	***there are people who***	**11**	**5.72**	**11**	**3.28**	*
40.	***they do not have***	**11**	**5.72**	**11**	**3.28**	*

LOCNESS

No.	Lexical bundle	Freq.	Norm. freq	No of texts	Texts, %	Coverage
1.	***the end of the***	**62**	**23.56**	**37**	**12.41**	*
2.	***on the other hand***	**48**	**18.24**	**41**	**13.75**	*
3.	*at the end of*	37	14.06	26	8.72	
4.	*the beginning of the*	24	9.12	17	5.70	
5.	***as a result of***	**23**	**8.74**	**18**	**6.04**	*
6.	***one of the most***	**23**	**8.74**	**21**	**7.04**	*
7.	*at the beginning of*	22	8.36	14	4.69	
8.	***the fact that the***	**22**	**8.36**	**18**	**6.04**	*
9.	***at the same time***	**21**	**7.98**	**18**	**6.04**	*
10.	***in the case of***	**20**	**7.6**	**17**	**5.70**	*
11.	***is one of the***	**20**	**7.6**	**17**	**5.70**	*
12.	*to the fact that*	20	7.6	17	5.70	
13.	*due to the fact*	17	6.46	14	4.69	
14.	***the rest of the***	**17**	**6.46**	**14**	**4.69**	*
15.	***to a certain extent***	**16**	**6.08**	**11**	**3.69**	*
16.	***a great deal of***	**15**	**5.7**	**15**	**5.03**	*
17.	*is for the best*	15	5.7	11	3.69	
18.	***the fact that he***	**14**	**5.32**	**10**	**3.35**	*

19.	***the majority of the***	**14**	**5.32**	**12**	**4.02**	*
20.	***when it comes to***	**14**	**5.32**	**10**	**3.35**	*
21.	***would be able to***	**14**	**5.32**	**12**	**4.02**	*
22.	*the fact that they*	13	4.94	11	3.69	
23.	*the only way to*	13	4.94	9	3.02	
24.	***will be able to***	**13**	**4.94**	**12**	**4.02**	*
25.	*for the good of*	12	4.56	10	3.35	
26.	***in the long run***	**12**	**4.56**	**10**	**3.35**	*
27.	***the way in which***	**12**	**4.56**	**10**	**3.35**	*
28.	*by the end of*	11	4.18	10	3.35	
29.	*do not want to*	11	4.18	9	3.02	
30.	***it is important to***	**11**	**4.18**	**11**	**3.69**	*
31.	***it is obvious that***	**11**	**4.18**	**10**	**3.35**	*
32.	*not be able to*	11	4.18	11	3.69	
33.	*that it is a*	11	4.18	11	3.69	
34.	*that it is not*	11	4.18	11	3.69	
35.	*a part of the*	10	3.8	9	3.02	
36.	*and the fact that*	10	3.8	10	3.35	
37.	*both sides of the*	10	3.8	9	3.02	
38.	*by the fact that*	9	3.42	9	3.02	
39.	*example of this is*	9	3.42	9	3.02	
40.	*to be able to*	9	3.42	9	3.02	