Russian Journal of Linguistics

2019 Vol. 23 No. 3 659—680
http://journals.rudn.ru/linguistics

DOI: 10.22363/2312-9182-2019-23-3-659-680

Research Article

Distinctive Lexical Patterns in Russian Patient Information Leaflets: A Corpus-Driven Study

Łukasz Grabowski

University of Opole
pl. Kopernika 11a, Opole, 45-040, Poland

University of Ostrava
Dvořákova 7, Ostrava 1, 701 03, Czechia

Abstract

This methodologically-oriented corpus-driven study focuses on distinctive patterns of language use in a specialized text type, namely Russian patient information leaflets. The study's main goal is to identify keywords and recurrent sequences of words that account for the leaflets' formulaicity, and — as a secondary goal — to describe their discoursal functions. The keywords were identified using three methods (G2, Hedges' g and Neozeta) and the overlap between the three metrics was explored. The overlapping keywords were qualitatively analyzed in terms of discoursal functions. As for the distinctive multi-word patterns, we focused on recurrent n-grams with the largest coverage in the corpus: these were identified using the Formulex method (Forsyth, 2015b), which provides complementary data with respect to more conservative n-gram and lexical bundles approaches. The results revealed that the most distinctive keywords were identified using Hedges' g metric, that the largest overlap occurred between G2 and Neozeta metrics, and that the frequent use and discoursal functions of the identified lexical patterns correspond with situational contexts and communicative purposes of patient information leaflets. It is hoped that this study will provide an opportunity for a methodological reflection and inspire further corpus-driven research on distinctive recurrent lexical patterns (e.g., keywords, n-grams, lexical bundles) or — more generally — on formulaic language in texts originally written in Russian.

Keywords*: keywords, n-grams, formulaic language, phraseology, patient information leaflets, Russian language*

For citation:
Grabowski, Łukasz (2019). Distinctive Lexical Patterns in Russian Patient Information Leaflets: A Corpus-Driven Study. *Russian Journal of Linguistics,* 23 (3), 659—680. doi: 10.22363/2312-9182-2019-23-3-659-680.

Научная статья

Типизированные лексические паттерны в русских инструкциях по применению лекарственных препаратов: корпусное исследование

Лукаш Грабовский

Опольский университет
pl. Kopernika 11a, Ополе, 45-040, Польша

Остравский университет
Dvořákova 7, Острава 1, 701 03, Чехия

Аннотация

Данное методологически ориентированное исследование, проведенное с использованием корпусного метода, посвящено анализу наиболее отчетливо выраженных паттернов использования языка в русских инструкциях по применению лекарственных препаратов. Цель исследования двуплановая и заключается, во-первых, в установлении и эксцерпции ключевых слов и повторяющихся словосочетаний, которые вносят вклад в формулярность данного типа текста, и, во-вторых, в последовательном описании их дискурсивных функций. Для эксцерпции ключевых слов использовались три метода: логарифмическая функция правдоподобия (ЛФП), g-Хеджеса и Неодзета. Для дальнейшего качественного анализа были выбраны только ключевые слова, совпадающие во всех трех процедурах. Рекуррентные N-граммы с самым большим лексическим охватом в корпусе извлекались с использованием метода Формулекс (Forsyth 2015b), который предоставляет взаимодополняющие данные относительно более консервативных N-грамм и лексических связок. Результаты показали, что: 1) наиболее ярко выраженные ключевые слова были выявлены с использованием формулы g-Хеджеса; 2) самое большое совпадение ключевых слов было выявлено для формул ЛФП и Неодзета; 3) частотность и дискурсивные функции отобранных слов и словосочетаний обусловлены как ситуативным контекстом, так и коммуникативными функциями инструкций по применению лекарственных препаратов. Исход проведенного анализа позволяет надеяться, что полученные результаты станут толчком для методологических размышлений, а также дальнейших корпусных исследований типичных, часто употребляемых лексических паттернов (например, ключевых слов, N-грамм, лексических связок) и в целом — формулярности русскоязычных текстов.

Ключевые слова: *ключевые слова, N-граммы, шаблонный язык, фразеология, инструкции по применению лекарственных препаратов, русский язык*

Для цитирования:
Grabowski, Łukasz (2019). Distinctive Lexical Patterns in Russian Patient Information Leaflets: A Corpus-Driven Study. *Russian Journal of Linguistics,* 23 (3), 659—680. doi: 10.22363/2312-9182-2019-23-3-659-680.

1. Introduction

Formulaicity has been a rather nebulous term encountered in a broad variety of disciplines of arts and humanities, including painting, sculpture or visual arts, among others, where an object of enquiry (a work of art, industrial design, text, etc.) has the properties that may be readily described as patterned-like, template-like, stencilled, trite, clichéd, to name but a few epithets. When studying a natural language, with various purposes in mind, linguists of various schools often refer to it as formulaic. In recent years, the linguistically-oriented research on formulaicity has been flourishing (e.g., Wray, 2002, 2008, 2009; Schmitt & Carter, 2004; Wood, 2010a, 2010b, 2015; Kecskes, 2016; Myles & Cordier, 2017; Nelson, 2018; Pęzik, 2018).

In view of the fact that every sub-discipline of linguistics approaches formulaicity from a different perspective, it is difficult to precisely determine what is meant by formulaic language. Norbert Schmitt and Ronald Carter (2004: 3) argue that this phenomenon should be treated in an inclusive way so that multiple types of linguistic units fall under an umbrella term *formulaicity*. In a similar vein, Piotr Pęzik (2018: 241), who uses the term formulaicity interchangeably with phraseological prefabrication, argues that it "is a ubiquitous, complex and multifaceted linguistic phenomenon". An attempt at a conceptual clarification as to how linguists of various schools approach formulaicity was made by Błażej Gałkowski (2006: 163—164), who singles out three major approaches, that is, a purely linguistic one, where formulaicity is explored using various lexical and grammatical categories identified primarily on the basis of formal or lexical criteria; a psycholinguistic approach, which focuses on storage and processing of linguistic data in the mental lexicon of language users; finally, a sociolinguistic approach, which explores situational and cultural underpinnings of formulaicity. To these, one may also add a corpus linguistic approach, which explores formulaicity by focusing on the frequency and distribution of various types of recurrent sequences of words in texts[1]. According to Rosamund Moon (2007: 1046), corpus linguists, in particular those conducting research on phraseology, are primarily interested in frequent and statistically significant multi-word patterns in which particular words occur. This approach contrasts with more traditional analytical methods of phraseological research, e.g., the ones proposed and developed by Viktor Vinogradov (1947/1977), Natalia Amosova (1963), which focused on the analysis of systemic invariant forms of phraseologisms (as they are or rather should be recorded in a dictionary) abstracted from situational contexts of their use. However, corpus approach is closer in spirit to more synthetic approaches to phraseology (e.g., Ivanov 1957; Bogusławski 1976, 1989; Anic'kov 1992; Mel'cuk 1995, 1998; Chlebda 1991, 2009) focusing on the identification of links between situational contexts of language use and recurrence of linguistic forms. Consequently, the frequency-driven approach is particularly attractive for the analyses of routinized or clichéd texts since they rely more on limited stocks of prefabricated chunks of text or boilerplate formulas, notably when compared with more creative literary texts.

In terms of theoretical underpinnings, this descriptive and methodologically-oriented study also draws inspiration from research conducted by Wojciech Chlebda (1991, 2009, 2010), who looks at phraseology from a perspective of a producer of an utterance in a specific social context. More precisely, Chlebda (1991) proposes that a 'phraseme' (*frazem*), in later works (Chlebda 2009, 2010) referred to as a 'reproduct' (*reprodukt*), be treated as a central unit of analysis; it is defined as "a linguistic unit (a component of the language system of a given ethnic language) isolated from texts based on the verification of its regular and repeated occurrence, functioning as a verbalizer of specific content, e.g., a notion, proposition, intention, emotion" (Chlebda 2010: 15—16 and 140)[2].

[1] More detailed discussion on different approaches to study formulaic language can be found in Forsyth & Grabowski (2015) and Grabowski (2015c, 2018).

[2] Chlebda (2010: 16) argues that a reproduct can be a single word or a multi-word unit with a non-compositional or compositional meaning, and that the emphasis should be put on the analysis

Although recurrent use is a defining feature of reproducts, and at the same time one of the key characteristics of formulaicity, Chlebda (2009, 2010) does not operationalize any frequency or distribution threshold — unlike in the aforementioned lexical bundles approach (Biber et al. 1999) — allowing one to decide whether a given single word or a sequence of words is a reproduct. On the contrary, the search for reproducts is largely manual, based on close reading of textual material and intuition-based analysis of a number of textual features, e.g., quotation marks, structures with reporting verbs, temporal signals, location signals, community signals, authorial signals, generic and quasi-generic operators (Chlebda 2010: 19). Such signals are also referred to as *phraseology markers* (Pęzik 2018: 213), which stand for a stable yet extensible set of lexical devices that signal phraseological prefabrication in texts. Thus, the role of corpus linguistic methods in this approach is, strictly speaking, limited to consulting language corpora to perform a frequency check when intuition, notably a subjective assessment of the degree of perceptual salience, is insufficient to decide whether a given text chunk shall be treated as a reproduct (Chlebda 2009, 2010) or — from a different perspective — as proverbial, conventional, idiomatic or prefabricated (Pęzik 2018: 213). An approach like this is often referred to as a corpus-informed one.

In this corpus-driven study, however, the search for reproducts, which can be treated as markers of formulaic language in texts, is conducted in the opposite direction. More precisely, the recurrent single words and multi-word units are first identified based on their frequent occurrence in texts and then they are aligned with specific discourse functions in order to ensure that they constitute context-sensitive form-meaning mappings. This should enable one to single out those recurrent textual patterns, i.e., keywords and recurrent n-grams, that contribute to the formulaic nature of a text type under scrutiny, namely patient information leaflets written originally in Russian. A study like this one — inspired by theoretical insights from Russian, Polish and English phraseology (traditional and distributional one) — is hoped to fill in the gap in data-driven research on recurrent patterns of language use found in Russian texts.

2. Research material and methodology

In this corpus-driven study[3], we aim to identify and describe recurrent lexical items (single- and multi-word units) that contribute to the formulaic nature of patient information leaflets originally produced in Russian[4]. As mentioned earlier, the emphasis

of the latter. According to Grabowski (2015c), this observation complies with results of certain corpus studies which showed that multi-word units with compositional meaning are considerably more frequent in texts than idioms or fixed expressions with non-compositional meaning (Moon 1998; Biber et al. 1999).

[3] I would like to thank Richard Forsyth and Costas Garbielatos for some of the comments with respect to the problems addressed in this study. Also, I am grateful to an anonymous Reviewer for a number of important remarks that helped me improve the paper.

[4] In that respect, the study may be treated as an extension of the author's earlier corpus linguistic research on keywords and/or lexical bundles in Polish and English patient information leaflets (Grabowski 2014, 2015a).

will be put on keywords and recurrent n-grams with the largest coverage in the corpus, which will be described in greater detail later in the paper. Given the various constraints pertaining to the communicative function, situational contexts of use, target audience of the text type under scrutiny as well as a highly standardized macro-structure of patient information leaflets, one may expect to find there only a limited number of recurrent lexical items. Furthermore, to the knowledge of the author, corpus linguistic studies, notably corpus-driven ones, of recurrent multi-word units in Russian texts have been scarce. For example, Maria Kunilovska, Natalia Morgoun and Alexey Pariy (2018) compared learner and professional translations from English into Russian, on the one hand, with native Russian texts, on the other, by focusing on the number of indicators, such as sentence length, lexical variety (TTR and proportion of high frequency words), lexical density and word frequencies; however, recurrent sequences of words (e.g., n-grams or otherwise) have not been explored in their study. A rare exception is the study by Daehyeon Nam and Sungmin Lee (2016), who explored the use and discourse functions of lexical bundles (Biber et al. 1999) in spoken and written Russian attested in a 1-million-word sample of the Russian National Corpus; the authors revealed that referential bundles predominate in written texts while stance bundles are more frequent in spoken texts (Nam & Lee 2016). However, in the Russian National Corpus one may find a variety of text types and genres (both contemporary and older ones, e.g., produced in the 19th century), which means that the results briefly summarized by Nam and Lee (2016) constitute generalizations that may not be applicable to any specific text type or genre.

That is why in this study we focus on a quasi-specialist text type used in the health care sector, that is patient information leaflets (short PILs). We aim to identify those single- and multi-word units that justify referring to patient information leaflets as a formulaic text type. Importantly, the goal of this research is not to measure the amount of formulaic language (cf. Forsyth & Grabowski 2015; Nelson 2018), but to describe the formulaic profile of the sample of Russian PILs by identifying those linguistic items that account for its highly patterned and clichéd style.

Generally speaking, PILs are found in sales packages of medicines and they are written in the language of the country where the medicines are sold, which in this study is the Russian language. In short, PILs are produced — in accordance with relevant legal regulations in a particular country — by pharmaceutical companies for patients, pharmacists, nurses, general practitioners, etc., who are typically target readers of this text type. However, PILs also have intermediate users, such as regulatory authorities. According to Vicent Montalt Resurrecció and Maria Gonzalez Davies (2007: 68—69), the main communicative purpose of PILs is to provide specific information on proper and safe use and administration of medicines (doses, side effects, etc.).

In fact, there have been many corpus linguistic studies exploring lexis and phraseology in PILs originally written in English and other languages, conducted by Silvia Cacchiani (2006, 2016), Rosemary Clerehan, Di Hirsh and Rachelle Buchbinder (2009) or Łukasz Grabowski (2014, 2015a, 2015b), among others. However, to the knowledge of the author of this paper, there has been no study focusing on PILs written in Russian.

The research material includes a tailor-made corpus of 100 PILs (i.e., full-texts) produced by fourteen pharmaceutical companies operating (as of the year 2016) on the

Russian market: Astellas Russia (10), AstraZeneca (10), Bayer (10), BerlinChemie (10), Biochimik Saransk (1), Boehringer Ingelheim (10), Farmstandard (10), Graminex Farma Russia (1), Kirkland Rindoxil Russia (1), Lundbeck Russia (5), Novartis (10), Sanofi Russia (10), Servier Russia (10), Takeda Russia (2). All in all, the corpus size is 229,346 word tokens, and the mean block TTR (a type-token ratio in per cent calculated using text chunks of 100 words, which can be treated as a provisional measure of lexical richness) is 78.48%, which is an average of block TTR scores of 100 documents in the study corpus. Also, the linguistic data have not been subjected to annotation or lemmatization. The research questions addressed in this primarily methodologically-oriented study are as follows:

1) What are the keywords typical of Russian patient information leaflets? To what extent do the keywords differ when identified using different keyword metrics? What are the discourse functions of overlapping keywords?
2) What are the distinctive recurrent sequences of words (n-grams) in Russian patient information leaflets? What are their discourse functions?

3.1. Units of analysis: keywords and recurrent n-grams with the largest textual coverage

In the first stage of the study, we focus on the identification of keywords, that is those words that for some reason (frequency of use, symbolic value, social or cultural significance, etc.) are more important than other words in texts (Stubbs 2011: 21). It is common knowledge that a corpus linguistic approach to the identification of keywords relies primarily on statistics. According to Michael Scott (2008: 176), keywords are those words "whose frequency is unusually high in comparison with some norm", which is found in the reference corpus constituting a benchmark for comparison. More precisely, the keywords are identified through their 'keyness', an indicator whose value is contingent primarily on word frequencies and corpus size, which — in turn — depends on subjectively specified thresholds of frequency, effect-size and statistical significance, on the choice of the unit of analysis (word forms, lemmas, constructions, senses, etc.), and on the very characteristics (representativeness, balance, size, etc.) of the corpora under comparison (Gabrielatos 2018: 252). The core component of the definition of keyness and the essence of keyword analysis is therefore the comparison of frequencies of individual linguistic items (Gabrielatos 2018). However, researchers have also recently experimented with other approaches, e.g., based on comparisons of means of frequencies of individual items (Forsyth 2014b), grounded in topic modeling (Murakami et al. 2017), etc., which go beyond the original idea of keyword analysis.

Since calculating keyness is far from straightforward[5], there are methods galore that help one identify whether a word is a *key* one in a corpus. One of the most popular approaches involves, first, comparing a frequency of a word in a study corpus with a frequency of the same word in a reference corpus and, second, by cross-tabulating the results taking into consideration the size (i.e., total number of tokens) of both corpora and by applying a test of statistical significance, e.g., Ted Dunning's (1993) log-likeli-

[5] For a more detailed overview, see Gabrielatos (2018).

hood test (also known as G2 test) or Pearson's chi-square test (Scott 2008: 122). This approach is implemented (up to version 6.0) in WordSmith Tools[6] (Scott 1996—2017).

Another approach is proposed by Costas Gabrielatos and Anna Marchi (2011), who argue for measuring the effect size (i.e., the extent or magnitude of the frequency difference) rather than statistical significance of the frequency difference, the latter being highly sensitive to corpus size. Consequently, Gabrielatos and Marchi (2011) propose the effect size metric %DIFF, which is independent from sample size (Rosenfeld & Penrod 2011: 84) and calculated as follows: %DIFF = (NormFreq in SC – NormFreq in RC) × 100 / NormFreq in RC[7]. It is implemented, following the procedure proposed by Hardie (2014), in the version 7.0 of WordSmith Tools (Scott 2017), where it is called *Log-ratio*. Importantly, the log-likelihood test and %DIFF result in different rankings for keywords, i.e., a high log-likelihood score does not correlate with a high %DIFF (Gabrielatos & Marchi 2011), unlike the rankings produced by two size effect metrics (e.g., %DIFF and Ratio[8]), which turned out to be identical for all keywords (Gabrielatos 2018: 232). In short, tests of statistical significance and effect size metrics "measure different aspects of a frequency difference" and hence they "are not alternative measures of keyness" (Gabrielatos 2018: 231). In practice, this means that two rankings of keywords, e.g., based on a test of statistical significance and effect size metric respectively, are hardly comparable with each other.

According to Paul Ellis (2010: 9), another useful method that can be employed to measure effect size is Hedges' g (Hedges 1981), which — as explained by Richard Forsyth (2014b: 10) — expresses — in standard deviation units — the difference in mean frequency rates between a study corpus and a reference corpus, hence producing a z-score (i.e., a standardized difference). To sum up, in contrast to non-parametric tests of statistical significance (e.g., Pearson's chi-square test, Dunning's log-likelihood test), metrics of effect size (e.g., %DIFF, Hedges' g or Cohen's d) help one avoid the problem of small yet at the same time statistically significant differences between frequencies of words in two corpora, the problem typical of comparing corpora of large size (Gabrielatos & Marchi 2011; Gabrielatos 2018). It often happens that statistical significance is not paramount to practical significance between the observed differences. At this point, however, it is worthwhile emphasizing that different size effect metrics are based on different assumptions, e.g., %DIFF compares frequencies of individual items in a corpus while Hedges' g and Cohen's d compare means of frequencies of individual items in the texts in a given corpus[9].

There are also other methods, e.g., Simple Math metric (Kilgarriff 2009) that includes a variable that allows one to focus on either lower or higher frequency words[10].

[6] WordSmith Tools ver. 1.0 was released in 1996; the most recent version of the software (7.0) was released in 2017. The program is downloadable from the following website: http://lexically.net.

[7] 'NormFreq' stands for a normalized frequency, 'SC' stands for a study corpus and 'RC' stands for a reference corpus.

[8] The metric was proposed by Adam Kilgrariff (2001) (cited in Gabrielatos 2018: 231—235).

[9] I would like to thank Costas Gabrielatos for this remark.

[10] The Simple Math method (Kilgarriff 2009) is implemented in SketchEngine (Kilgarriff et al. 2014).

In the Keysoft software package, a collection of scripts written in Python 3.4, Forsyth (2014a) implements twelve methods to identify keywords, some of them commonly used in the field of authorship attribution (e.g., Zeta or Neozeta). Originally developed by John Burrows (2007), and later modified by Hugh Craig and Arthur Kinney (2009), Neozeta enables one to identify keywords by segmenting corpora into text chunks of equal length and counting occurrences of words in those text chunks. According to Maciej Eder (2016: 35), such a method whereby the frequencies of text chunks rather than individual words are used for corpus comparison enables one to filter out the words that appear in texts with high frequencies (e.g., function words) and, consequently, to focus on content words that convey themes or topics discussed in texts, i.e., the so-called *aboutness* (Phillips 1989).

Hence, in order to capitalize on the whole variety of approaches, we will use the Keysoft package (Forsyth 2014a) to identify and compare keywords in Russian PILs using three fundamentally different metrics, that is G2 (Dunning 1993), Hedges' g (Hedges 1981) and Neozeta (Craig & Kinney 2009), which measure different aspects of a frequency difference[11]. Since we do not focus on comparisons of statistical significance metrics only (e.g., G2), we do not additionally apply — as recently recommended by Gabrielatos (2018) — the BIC score, i.e., a metric calculated by deducting the combined size (logarithmized) of the compared corpora from the G2 value of the frequency difference.

Although the three metrics provide different flavours to keyword rankings, it is believed that there is some practical value in undertaking such a comparison. Firstly, it will enable one to further verify the correlation (if any) between rankings produced by a test of statistical significance (G2) and effect size metric (Hedges' g), the latter one focusing on means of frequencies of individual words. Second, some researchers, especially those using keywords in critical discourse analysis or sociolinguistic research, still use tests of statistical significance for keyword analysis (e.g., Baker et al. 2019), which means that it may be useful for them to compare rankings of keywords obtained using different methods, irrespective of the fact that different methods may be based on different statistical assumptions, e.g., comparisons of means of word frequencies rather than frequencies of individual items in two corpora.

As a reference corpus, we will use the Russian component of the Leeds Pentaglossal Corpus[12] (Forsyth & Sharoff 2014), which includes 113 documents or fragments of documents representing 13 text types (e.g., Bible, corporate statements, fiction, news articles, ted.com transcripts, United Nations documents). Hence, the composition of the corpus is more heterogeneous as compared with the Russian PILs. The size of the Russian Pentaglossal Corpus (henceforth RPC) is 251,204 word tokens and the mean block TTR (type-token ratio) is 79.09%. Since the STTR of Russian PILs is 78.48%, the RPC can be intuitively described as similar in terms of its lexical variation.

[11] The mathematical formulae used to calculate keyness using the three methods are described and explained in greater detail by Forsyth (2014b: 9—12).

[12] Leeds Pentaglossal Corpus is downloadable from the following website: http://corpus.leeds.ac.uk/tools/5gcorpus.zip.

Next, we will compare the keywords obtained using three different methods to identify any overlapping ones, which will be subject to further qualitative analysis. The rationale behind focusing on the overlap (i.e., similarity) between the keywords is the claim made by Gabrielatos (2018: 225) who argues that "the vast majority of keyness studies focus on difference, at the expense of similarity". Also, due to their unusually high frequency the keywords may be the center of units of meaning in texts thereby performing specific discourse functions and contributing to the texts' formulaicity. According to Stanisław Goźdź-Roszkowski (2011: 35), keywords can "reveal not only a great deal about the subject matter, the 'aboutness' of a particular genre, but they can also specify the salient features which are functionally related to the genre". This observation has two important implications. Firstly, since keywords are typically studied through their typical co-occurrence patterns, it should be possible to align them with specific discourse functions. In practice, this means developing a set of provisional categories in the form of tentative labels reflecting typical characteristics of the keywords — the type of information they convey, their role in the organization of discourse, their semantic prosody and evaluative charge etc. (Goźdź-Roszkowski 2011: 65; Grabowski 2015c). Secondly, the exploration of co-occurrence patterns or wider contexts of use of keywords should also enable one to identify distinctive or salient sequences of words that perform specific discourse functions in texts. The resulting sequences may include specialist terms or text chunks contributing to the formulaic style of a given text type or genre.

According to Bestgen (2018: 206), "one of the frequently used approaches to studying formulaic language is based on the automatic identification of recurrent continuous sequences of words". With this goal in mind, in the second stage of the study we will use a recently proposed method called *Formulex* (Forsyth 2015b), which identifies properly fragmented n-grams based on the concept of 'coverage'. According to Forsyth (Forsyth 2015b: 17), the method whereby "the sequences are mutually exclusive" and that "longer prefabricated phrases [are prevented] from being swamped by the elements of which they are composed" enables one to specify more precise boundaries of recurrent strings of words. As demonstrated by Grabowski & Jukneviciene (2016)[13], Formulex method may come in useful when dealing with overlapping sequences of n-words or with those sequences of words that constitute fragments of longer n-grams (e.g., *в недоступном для детей, в недоступном для детей месте, хранить в недоступном для детей месте* 'store in a place not accessible for children'). A problem like this one is often faced by researchers using the lexical bundles methodology (Biber et al. 1999) where the recurrent sequences of words are extracted from texts using the criteria such as orthographic length, minimum frequency and distribution range. In that approach, overlapping or structurally-incomplete items are often identified when analyzing highly-patterned, formulaic text types or genres; this is precisely the scenario that we aim to avoid when using Formulex method (Forsyth 2015b) in an attempt to extract right-sized n-grams from the study corpus.

[13] Using a corpus of Lithuanian and Polish students' EFL writing, Grabowski & Jukneviciene (2016) filtered out the original lists of lexical bundles, identified using three traditional criteria (Biber et al. 1999, 2003, 2004; Biber 2006), against the lists of formulas generated using the Formulex method (Forsyth 2015b).

4. Results

In the first stage of the study, we focused on exploration of the most salient words in Russian PILs. To that end, we used the Keysoft package (Forsyth 2014a) and identified keywords using three different metrics described earlier in this paper, that is G2 (Dunning 1993), Hedges' g (Hedges 1981) and Neozeta (Craig & Kinney 2009), which resulted in three lists with 52, 55 and 51 positive keywords respectively. The top-50 keywords are presented in Table 1. For the sake of clarity, all the numbers were deleted from the lists of keywords — this way three numbers were deleted from the keywords identified using G2 test, four numbers from the list identifies using Neozeta, and none from the list of keywords identified using Hedges' g metric. In brackets right next to each keyword, there is information concerning the overlap among the top-50 keywords, e.g., (3) indicates that a given keyword was identified using each method; (2) indicates an overlap between G2 and Neozeta; (2h) indicates an overlap between G2 and Hedge's g; (2n) indicates an overlap between Hedge's g and Neozeta; (1) indicates that a keyword was identified using a single method only. As mentioned earlier, this procedure should provide a preliminary insight into the similarity between the output of the three keyword metrics.

The results revealed that 22 keywords (44%) out of top-50 identified using three different metrics overlap with each other. Also, it was revealed that 20 keywords overlap in the case of using G2 test and Neozeta; 2 keywords (*реакции* 'reactions', *беременности* 'pregnancy') overlap in the case of using G2 and Hedges' g; 2 keywords overlap in the case of using Hedges' g and Neozeta (*особые* ' particular'*, взаимодействие* 'interaction'). The most distinctive keywords were identified using Hedges' g statistic: 23 keywords (46%) do not overlap with the ones identified using either G2 or Neozeta. The corresponding figure for G2 and Neozeta is 7 for both metrics. Hence, the findings confirm that the three metrics — based on different statistical assumptions — prioritize different keywords.

The provisional results were further verified using Spearman Rank Correlation (R_s)[14] applied to ranks of all positive keywords identified using each metric. In cases when a word does not occur on the list of positive keywords produced by a given metric, it was decided to assign to it a rank of N+1. For example, in comparisons of keywords identified using G2 (52 words) with the ones obtained using Hedges' g (55 words), all the words that occurred in G2 list (e.g., *после* 'after') were assigned 56th rank (55 + 1), as if they appeared on the Hedges' g list. The results confirmed our earlier observations: the highest R_s score (0.778)[15] was reported in the case of G2 vs Neozeta, which indicates rather strong positive association. The R_s scores for G2 vs Hedges' g (0.293) and Hedges' g vs Neozeta (0.189) indicate weak association between the metrics.

[14] The same approach was used by Baker (2010: 92).

[15] Ranks in G2 test: Mean: 26.5, Standard deviation: 15.15; Ranks in Neozeta: Mean: 26.5, Standard deviation: 15.1; Covariance = 90.83 / 51 = 178.1; R_s = 178.1 / (15.15 * 15/1) = 0.778.

Table 1

Keywords in Russian PILs (top-50): comparing G2, Hedges' g, Neozeta

Rank	G2	Hedges' g	Neozeta
1	*мг* (3)	*препарата* (3)	*препарата* (3)
2	*препарата* (3)	*дозы* (3)	*при* (3)
3	*пациентов* (2)	*при* (3)	*пациентов* (2)
4	*при* (3)	*следует* (3)	*тг* (3)
5	*дозы* (3)	*противопоказания*	*следует* (3)
6	*следует* (3)	*применению* (3)	*дозы* (3)
7	*крови* (3)	*особые* (2n)	*У* (2)
8	*лечения* (3)	*взаимодействие* (2n)	*применения* (3)
9	*приема* (3)	*лекарственная* (1)	*или* (1)
10	*терапии* (3)	*форма* (1)	*крови* (3)
11	*применения* (3)	*лекарственными* (3)	*лечения* (3)
12	*с* (3)	*годности* (1)	*приема* (3)
13	*мл* (2)	*инструкцией* (1)	*после* (2)
14	*применении* (3)	*препарат* (3)	*применение* (3)
15	*применение* (3)	*показания* (1)	*терапии* (3)
16	*препарат* (3)	*побочное* (1)	*применении* (3)
17	*нарушения* (2)	*тг* (3)	*препарат* (3)
18	*печени* (2)	*торговое* (1)	*с* (3)
19	*у* (2)	*выпуска* (1)	*рекомендуется* (3)
20	*сутки* (2)	*применение* (3)	*развития* (2)
21	*стороны* (1)	*фармакокинетика* (1)	*до* (1)
22	*концентрации* (2)	*хранения* (1)	*течение* (2)
23	*таблетки* (2)	*фармакотерапевтическая* (1)	*применению* (3)
24	*редко* (1)	*отпуска* (1)	*может* (1)
25	*лечение* (3)	*действие* (3)	*концентрации* (2)
26	*рекомендуется* (3)	*фармакологические* (1)	*лечение* (3)
27	*риск* (2)	*симптомы* (1)	*печени* (2)
28	*доза* (3)	*вспомогательные* (1)	*риск* (2)
29	*применению* (3)	*рекомендуется* (3)	*мл* (2)
30	*снижение* (2)	*свойства* (1)	*период* (2)
31	*течение* (2)	*реакции* (2h)	*действие* (3)
32	*прием* (2)	*регистрационный* (1)	*снижение* (2)
33	*часто* (1)	*передозировка* (1)	*нарушения* (2)
34	*дозе* (2)	*лечение* (3)	*прием* (2)
35	*почек* (2)	*приема* (3)	*необходимо* (1)
36	*после* (2)	*средствами* (3)	*другими* (1)
37	*беременности* (2h)	*беременности* (2h)	*дозе* (2)
38	*препаратов* (2)	*применении* (3)	*препаратов* (2)
39	*составляет* (3)	*состав* (1)	*средствами* (3)
40	*развития* (2)	*срока* (1)	*функции* (1)
41	*пациенты* (1)	*доза* (3)	*таблетки* (2)
42	*период* (2)	*крови* (3)	*доза* (3)
43	*плазме* (1)	*применения* (3)	*риска* (2)
44	*со* (1)	*с* (3)	*составляет* (3)
45	*лекарственными* (3)	*осторожностью* (1)	*сутки* (2)
46	*риска* (2)	*терапии* (3)	*почек* (2)
47	*действие* (3)	*лечения* (3)	*лекарственными* (3)
48	*мин* (1)	*условия* (1)	*особые* (2n)
49	*средствами* (3)	*составляет* (3)	*системы* (1)
50	*реакции* (2h)	*повышенная* (1)	*взаимодействие* (2n)

Although the application of each method results in three largely different sets of keywords, there are 22 overlapping words on three lists. These keywords may be provisionally divided into a number of functional categories, such as high-frequency function words (*при* 'at'*, с* 'with') or content words (*составляет* '(it) constitutes'), measurement keywords (*мг* 'mg'), keywords referring to administration of medicines to patients (*доза* 'dose', *дозы* 'doses', *действие* 'activity', *приема* '(drug) taking', *применение, применения, применении, применению* 'administration/application'), recommendation or advisory keywords (*рекомендуется* 'it is recommended', *следует* '(one) should'), keywords referring to human body (*крови* 'blood' gen.), procedural keywords (*лечение, лечения* 'treatment'*, терапии* 'therapy'), as well as aboutness keywords that convey a general idea about the topics raised in the Russian PILs (*препарат, препарата* 'drug', *лекарственными* 'medicinal'*, средствами* 'products').

The second stage of the study is aimed to identify distinctive or salient sequences of words in the sample of Russian PILs. The rationale behind our approach is that the keywords are not salient by themselves or by virtue of the communicative function of the text variety under scrutiny only. On the contrary, it is believed that the salience of keywords measured by their outstanding frequency results from the frequent use of certain text chunks and/or grammatical constructions. For example, an outstanding frequency of articles in specialist texts may result from frequent use of noun phrases or nominalizations. To that end, in the final stage of the study, we used the Formulib software (Forsyth 2015a), a collection of scripts written in Python 3.4, and attempted to identify n-grams, built of four words or longer, with the highest coverage of texts in the study corpus. More specifically, the coverage threshold was arbitrarily set at 0.02%. Since coverage is calculated in terms of the number of characters, the corresponding threshold taking into consideration the size of the study corpus is 345 or more characters. As regards the procedure of n-gram extraction, Formulib script treats coverage as a binary category, which means that the number of n-grams that match a particular text sequence is irrelevant; what the program verifies is whether the text sequence is covered or not (Forsyth 2015b: 13—14). For example, if n-grams such as *связь с белками плазмы* and *с белками плазмы крови* cover a certain part of the text sequence *связь с белками плазмы крови* 'interaction with blood plasma proteins', each of those five words in the last sequence is marked as covered once. Based on that, the proportion of covered to uncovered characters for each text is calculated and, subsequently, the character coverage for a text category, in this study — Russian PILs, is aggregated (Forsyth 2015b: 13—14)[16].

Apart from providing insights into recurrent chunks of text, the Formulex method (Forsyth 2015b) also allows one to identify boundaries between recurrent n-grams, in particular overlapping or structurally incomplete ones. To illustrate the problem, on 22 occasions in the Russian PILs a contiguous sequence of words, such as *с белками плазмы крови,* was not a fragment of a longer sequence *связь с белками плазмы крови*, which is recorded in Russian PILs 15 times; as a matter of fact, the sequence *с белками плазмы крови* occurs 46 times in total in various patterns in the Russian

[16] Forsyth (2015b: 25) notes that his method is similar to one of the algorithms ("Serial Cascading Algorithm") proposed by O'Donnell (2011: 149—153).

PILs corpus, yet it occurs by itself only 22 times. Such a solution, namely that "the sequences [of words] are mutually exclusive" and that "longer prefabricated phrases [are prevented] from being swamped by the elements of which they are composed of" (Forsyth 2015b: 17), allows one to specify more precise boundaries of formulaic sequences of words, which has been one of the challenges in research on n-grams or lexical bundles (Biber et al. 1999; Biber et al. 2004)[17].

The results in the form of n-grams with the largest coverage in the Russian PILs are presented in Table 2. For the reasons of limited space, only the 50 n-grams with the largest coverage in the study corpus are presented; in practice, this translates into coverage of more than 0.02% of the study corpus. Also, the keywords found in the n-grams are presented in bold. It is believed that the salience of the 22 keywords constituting the core vocabulary of Russian PILs and identified earlier in the study may be also contingent on the frequent occurrence of the text chunks presented in Table 2. The reason for that is that those text chunks constitute textual building blocks of the Russian PILs.

Table 2 presents 50 n-grams, i.e., contiguous sequences of 4 and 5 words, with the largest coverage in the Russian PILs under study. The results reveal that among the ten top-coverage n-grams four are in fact headings describing macro-structure of the genre (*взаимодействие с другими лекарственными средствами* 'interaction with other drugs', *инструкция по медицинскому применению препарата* 'instruction for medical use of the drug'*, способ применения и дозы* methods of administration and doses'*, взаимодействие с другими лекарственными препаратами* 'interaction with other medicinal products'), while the remaining ones are found within the PILs' contents.

Since all the n-grams presented in Table 2 are frequently used in the analyzed text type, an attempt has been made to explore their discourse functions. To that end, we capitalized on two functional typologies. The first one is largely based on the functional taxonomy originally proposed by Douglas Biber, Susan Conrad and Viviana Cortes (2004: 384—388) and Biber (2006: 139—145) and applied to lexical bundles, which are divided into three inclusive categories, namely *referential*, *discoursal* and *expressing stance*. The other one is the functional typology originally developed by Kenneth Hyland (2008: 13—14), who divided lexical bundles into three major functional categories, namely *research-oriented* (in this study called "referential" bundles), *text-oriented* and *participant-oriented bundles* (in this study called "stance/evaluation" bundles).

More specifically, in this study *referential* n-grams refer to various properties (pharmacological, pharmacokinetic etc.) of medicines or to main themes conveyed in the Russian PILs; *text-oriented* n-grams help organize and convey information presented in the analyzed text type; finally, *stance/evaluation* n-grams help express judgments or assessments of information presented in the Russian PILs. Also, more fine-grained functional subcategories are provided to account for specific functional roles of the n-grams under scrutiny. In that respect, the typology used in the present research is similar to the one applied in the study of lexical bundles in Polish patient information leaflets (Grabowski 2014).

[17] The same method was used by Grabowski & Jukneviciene (2016).

Table 2

Top-50 n-grams (by coverage) in Russian PILs

No.	Coverage	Freq.	Char.	Words	N-gram
1	0.1617	57	50	5	*взаимодействие **с** другими **лекарственными средствами***
2	0.1202	45	47	5	*инструкция по медицинскому **применению препарата***
3	0.1099	79	24	4	*способ **применения** и **дозы***
4	0.0933	43	38	5	*со стороны сердечно сосудистой системы*
5	0.0854	59	25	4	*см раздел особые указания*
6	0.0839	29	51	5	*взаимодействие **с** другими **лекарственными** препаратами*
7	0.0535	31	30	4	***при** одновременном **применении с***
8	0.0529	25	37	5	*со стороны желудочно кишечного тракта*
9	0.0514	33	27	4	*см раздел побочное **действие***
10	0.0501	30	29	4	*у пациентов пожилого возраста*
11	0.0499	28	31	5	*у пациентов **с** сахарным диабетом*
12	0.0477	22	38	5	*со стороны центральной нервной системы*
13	0.0428	22	34	4	*со стороны пищеварительной системы*
14	0.0381	18	37	5	*нарушения со стороны иммунной системы*
15	0.0379	22	30	4	*со стороны дыхательной системы*
16	0.0378	17	39	5	*у пациентов **с** почечной недостаточностью*
17	0.0369	17	38	5	*для приготовления раствора для инфузий*
18	0.0361	18	35	4	***с** другими **лекарственными средствами***
19	**0.0333**	**13**	**45**	**5**	***следует** соблюдать осторожность **при** назначении*
20	**0.0331**	**17**	**34**	**4**	***следует** соблюдать осторожность **при***
21	0.0312	16	34	4	*может потребоваться коррекция **дозы***
22	0.0304	13	41	5	*у пациентов **с** печеночной недостаточностью*
23	0.0294	12	43	4	*повышение активности печеночных трансаминаз*
24	0.0292	25	20	4	*у детей и подростков*
25	0.0285	19	26	4	*со стороны кожных покровов*
26	0.0285	16	31	5	*способ **применения** и **дозы** внутрь*
27	0.0282	22	22	4	***с** белками плазмы крови*
28	0.0275	13	37	4	*таблетки покрытые пленочной оболочкой*
29	0.0267	10	47	5	*по поводу хронической сердечной недостаточности*
30	0.0258	16	28	4	*концентрации глюкозы в крови*
31	0.0256	20	22	4	*по сравнению **с** плацебо*
32	0.0247	12	36	5	*нарушения со стороны нервной системы*
33	0.0247	12	36	4	***с** другими **лекарственными** препаратами*
34	0.0245	10	43	5	*претензии потребителей направлять по адресу*
35	0.0242	15	28	5	*связь **с** белками плазмы крови*
36	0.0234	10	41	5	*всасывается из желудочно кишечного тракта*
37	0.0233	22	18	4	*в течение 24 часов*
38	0.0231	16	25	5	*баллов по шкале чайлд пью*
39	0.0230	18	22	4	*у пациентов в возрасте*
40	0.0227	17	23	4	*не оказывает влияния на*
41	0.0223	10	39	5	*у пациентов **с** артериальной гипертензией*
42	0.0220	22	17	4	*в том случае если*
43	0.0218	14	27	5	*как и **при применении** других*
44	0.0218	14	27	4	*со стороны иммунной системы*
45	0.0217	26	14	4	*в связи **с** этим*
46	0.0217	10	38	5	*беременность и период кормления грудью*
47	0.0214	11	34	5	*со стороны костно мышечной системы*
48	0.0211	10	37	5	*у пациентов **с** фибрилляцией предсердий*
49	0.0211	10	37	4	*необходимо соблюдать осторожность **при***
50	0.0209	15	24	4	*на фоне **приема препарата***

As for *referential n-grams* (40 items), they include *topic n-grams*, referred to by Hyland (2008: 13) as topic-bundles, which identify certain themes conveyed in PILs or key aspects of medicines described therein. This group includes the following n-grams: *взаимодействие с другими лекарственными средствами* 'interaction with other drugs'*, с другими лекарственными средствами* 'with other drugs'*, с другими лекарственными препаратами* 'with other medicinal products'*, взаимодействие с другими лекарственными препаратами* 'interaction with other medicinal products', *инструкция по медицинскому применению препарата* 'instruction for medical use of the drug', *таблетки, покрытые пленочной оболочкой* ('film coated tablets', i.e., referring to a pharmaceutical form of medicines), *по поводу хронической сердечной недостаточности* 'due to chronic heart failure'*, беременность и период кормления грудью* 'pregnancy and lactation period' (i.e., referring to illnesses or physical conditions), *баллов по шкале чайлд пью* 'points on the Child-Pugh scale'. Another group in this category includes *location n-grams*, which refer to composition, parts or systems of human organism (blood plasma, central nervous system, immune system etc.) affected by illnesses or subjected to the activity of medicines, e.g., *со стороны сердечно сосудистой системы* 'of the cardiovascular system'*, со стороны желудочно кишечного тракта* 'of the gastrointestinal tract'*, со стороны центральной нервной системы* 'of the central nervous system'*, со стороны пищеварительной системы* 'of the digestive system'*, нарушения со стороны иммунной системы* 'disorders of the immune system' *, нарушения со стороны нервной системы* 'disorders of the nervous system'*, со стороны дыхательной системы* 'disorders of the respiratory system' *, со стороны кожных покровов* 'of the skin surfaces'*, со стороны иммунной системы* 'of the immunological system'*, со стороны костно-мышечной системы* 'of the osseous muscular system', *с белками плазмы крови* 'with blood plasma cells'. Next, *procedure-related n-grams* relate to various aspects of administration of medicines to patients (preparation, dose etc.), e.g., *способ применения и дозы* 'method of administration and doses'*, способ применения и дозы внутрь* 'method of administration and use inside'*, при одновременном применении с* 'when used simultaneously with', *для приготовления раствора для инфузий* 'for preparation of solution for infusion', *на фоне приема препарата* 'while taking the drug'*, по сравнению с плацебо* 'in comparison with placebo'*, как и при применении других* 'as well as in the application of'. Process-related n-grams describe chemical processes related to the activity or presence of active substances or excipients in the human body, e.g., *повышение активности печеночных трансаминаз* 'increased activity of hepatic transaminases', *концентрации глюкозы в крови* 'blood glucose concentration'*, связь с белками плазмы крови* 'interaction with blood plasma cells'*, всасывается из желудочно кишечного тракта* 'is being absorbed from the gastrointestinal tract'. Finally, one may find a single temporal n-gram (*в течение 24 часов* 'within 24 hours') related to the frequency of administration or duration of the activity of medicines.

Next, *text-oriented n-grams* include one *condition n-gram* (*в том случае если* 'in the case when'), which is used to introduce certain condition related to administration of medicines, two transition n-grams (*в связи с этим* 'in connection with this', *не оказывает влияния на* 'it does not affect'), which, according to Hyland (2008: 14), help

establish additive or contrastive links between information conveyed in PILs, and two text-deixis n-grams, e.g., *см раздел особые указания* 'see section special instruction', *см раздел побочное действие* 'see section side effects', which help readers navigate through the contents or macro-structure of PILs.

Finally, stance/evaluation n-grams include one *desire n-gram* (*может потребоваться коррекция дозы* 'dosage adjustment may be required'), which expresses a desirable course of action undertaken by patients in the event of any problems arising from the use of medicines, and four *obligation/directive n-grams*, starting with the third person present tense form of the verb *следует* used in the impersonal form and followed by the infinitive '(one) should', predicative *необходимо* '(one) needs' followed by a single action verb in its infinitive form (*необходимо/следует соблюдать осторожность при* '(one) needs/should be careful when'), or centered around the verb in the infinitive form (*направлять* 'send'), e.g., *претензии потребителей направлять по адресу* 'customer complaints should be sent to'. All in all, this last group of n-grams is used to direct patients to carry out specific actions related to the use of medicines.

5. Conclusions

Inspired by theoretical insights from Russian, Polish and English phraseology, this methodologically-oriented study falls within the scope of frequency-driven distributional phraseology (Granger & Meunier 2008; Pęzik 2013), and its main goal was to identify the keywords and distinctive recurrent sequences of words in a sample of Russian patient information leaflets, and — as a secondary goal — to describe their discoursal functions. We also compared three methods of identifying keywords in texts (G2, Hedges' g and Neozeta) and further tested — using Russian language material — a recently proposed method of identification of recurrent multi-word units called Formulex (Forsyth 2015b).

The results revealed that that 22 keywords out of top-50 identified using three different metrics overlap with each other. Those keywords refer to administration of medicines to patients or to recommendations and advice offered to patients. As for the methods of keyword identification, the biggest overlap was recorded between G2 and Neozeta methods, while the application of Hedges' g yielded the most distinctive keywords. Finally, using Formulex method (Forsyth 2015b), we identified 50 n-grams with 4 or 5 words with the largest coverage in the Russian PILs. The qualitative analysis revealed that the largest group of those recurrent sequences of words perform referential functions, that is, they refer to various aspects of the use and administration of medicines.

In general, the findings revealed that the analyzed text type relies on a limited stock of single words and prefabricated chunks of text frequently used in Russian patient information leaflets, the items that account for the formulaicity of the text type under scrutiny. Also, the comparison of the performance of the three keyword metrics — each based on different statistical assumptions — enabled one to gain an insight into both similarities and differences[18] between lexical patterns, which may come in useful for researchers using keywords in critical discourse analysis (CDA), among others.

[18] This distinction is referred to by Gabrielatos (2018: 252) as *keyness-S* vs *keyness-D*.

As regards future avenues, more research is required to compare the performance of keyword metrics other than the ones used in this study. Also, to take into account distances between keyness scores, Gabrielatos (2018) proposes that candidate key words be clustered according to an effect size score, which is another idea for the future. As for comparisons of rankings of keywords, apart from Spearman Rank Correlation test used in this study, it is possible to use Mann Whitney U test on ranks to obtain another proximity score between different keyword metrics. Furthermore, it may be useful to consider employing other approaches to extract salient vocabulary from texts and study its aboutness, e.g., unsupervised topic modeling approaches (Silge & Robinson 2017) that help cluster individual words together providing an overview of the texts' semantic content[19]. Finally, more caution is required when it comes to selection of a reference corpus, which is one of the crucial decisions in the traditional keyword analysis. Although it is now known that there is no optimum size of the corpus (Gabrielatos 2018), it often happens — usually in the case of smaller corpora with limited representativeness — that many words do not occur in them. As a result, it is necessary to carefully think about an attenuating factor (i.e., its value) assigned to zero-frequencies in a reference corpus, and — more broadly — about the very characteristics of an appropriate reference purpose given the nature of the study corpus. Crucially, those decisions have important implications on the results of any keyword analysis. Last but not least, more comprehensive research is required to further compare different keyword metrics by conducting multiple studies on corpora of different size and make-up.

As for recurrent multi-word items, the Formulex method (Forsyth 2015b) should be compared with other metrics designed to extract structurally-complete or non-overlapping sequences of words from texts, e.g., cascading serial algorithm (O'Donnell 2011), transitional probability metric (Appel & Trofimovich 2015), Independence-Formulaicity score (Pęzik 2015), frequency consolidation method (Buerki 2017), dependency-based approaches (Pęzik 2018). Furthermore, the results of a study like this one may be further verified using lemmatized Russian language data. It is also hoped that this study may be inspirational for future research on distinctive recurrent lexical patterns and on formulaicity of other genres of texts originally written in Russian. Finally, a comparison of the findings presented in this paper with the ones conducted for patient information leaflets written in other languages may yield comparable data that may be employed for developing domain-specific multilingual resources useful for translators, lexicographers or teachers of Russian for specific purposes (RSP/ПООИЯ), among others.

[19] See Murakami et al. (2017) for comparisons using topic modelling techniques, where it is not necessary to use a reference corpus, with a traditional keyword analysis involving comparisons of frequencies of individual items in a study corpus and a reference corpus.

REFERENCES

Altenberg, Bernd (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In: A. Cowie (ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. 101—122.

Amosova, Natalia (1963). Ocnovy angliiskoi frazeologii [Fundamentals of English Phraseology]. Leningrad: Izdatel'stvo Leningradskogo Universiteta (cited in Cowie 1998, 5—6, 215).

Anic'kov, Igor' (1992). Idiomatika i Semantika [Idiomatics and semantics]. *Voprosy Jazykoznanija*, 5, 136—150 (cited in Dobrovolskij & Filipenko 2007, 715).

Appel, Randy and Trofimovich, Pavel (2015). Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics*. Article first published online: 29 Jan 2015 (accessed on 26 February 2015).

Baker, Paul (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Baker, Paul, Gavin Brookes, and Craig Evans (2019). *The Language of Patient Feedback: A Corpus Linguistic Study of Online Health Communicatio*n. London: Routledge.

Bestgen, Yves (2018). Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora*, 13(2), 205—228.

Biber, Douglas (2006). *University Language. A corpus-based study of spoken and written registers.* Amsterdam: John Benjamins.

Biber, Douglas, Susan Conrad, and Viviana Cortes (2003). Lexical bundles in speech and writing: An initial taxonomy. In Andrew Wilson, Paul Rayson, & Tony McEnery (eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Frankfurt/Main: Peter Lang, 71—92.

Biber, Douglas, Susan Conrad, and Viviana Cortes (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics,* 25 (3), 371—405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). *The Longman grammar of spoken and written English*. London: Longman.

Bogusławski, Andrzej (1976). O zasadach rejestracji jednostek języka. *Poradnik językowy 8*, 356—364.

Bogusławski, Andrzej (1978). Jednostki języka a produkty językowe. Problem tzw. orzeczeń peryfrastycznych. In: Mieczysław Szymczak (ed.), *Z zagadnień słownictwa współczesnego języka polskiego*. Wrocław: Zakład Narodowy im. Ossolińskich, 15—30.

Buerki, Andreas (2017). Frequency consolidation among word N-grams: a practical procedure. In: Ruslan Mitkov (ed.), *Computational and Corpus-Based Phraseology, Lecture notes in Computer Science*, vol. 10596. Cham: Springer, 432—446.

Burrows, John (2007). All the way through: testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22 (1), 27—48.

Cacchiani, Silvia (2006). Dis/similiarities between Patient Information Leaflets in Britain and Italy: Implications for the Translator. *New Voices in Translation Studies*, 2, 28—43.

Cacchiani, Silvia (2016). On intralinguistic translation from summaries of product characteristics to patient information leaflets. In: Giuliana Elena Garzone, Dermot Heaney & Giorgia Riboni (eds), *LSP Research and Translation across Languages and Cultures*. Newcastle upon Tyne: Cambridge Scholars Publishing, 219—251.

Chlebda, Wojciech (1991). *Elementy frazematki: wprowadzenie do frazeologii nadawcy.* Opole: Wydawnictwo WSP.

Chlebda, Wojciech (2009). Idiomatykon 4: gdzie jesteśmy, dokąd zmierzamy (i parę zdań o tym, skąd przychodzimy). In: Wojciech. Chlebda (ed.), *Podręczny idiomatykon polsko-rosyjski 4*. Opole: Wydawnictwo Uniwersytetu Opolskiego, 9—38.

Chlebda, Wojciech (2010). Nieautomatyczne drogi dochodzenia do reproduktów wielowyrazowych. In: Wojciech Chlebda (ed.), *Na tropach reproduktów: w poszukiwaniu wielowyrazowych jednostek języka*. Opole: Wydawnictwo Uniwersytetu Opolskiego, 15—35.

Clerehan, Rosemary, Di Hirs and Rachelle Buchbinder (2009). Medication information leaflets for patients: the further validation of an analytic linguistic framework. *Communication & Medicine* 6 (2), 117—128.

Cowie, Anthony (1998). *Phraseology: Theory, analysis and applications*. Oxford: Clarendon Press.

Craig, Hugh and Kinney, Arthur F. (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Dobrovolskij, Dmitri, and Tatjana Filipenko (2007). Russian phraseology. In: Harald Burger (ed.), *Phraseologie: ein internationals Handbuch zeitgenoessischer Forschung,* Vol. 2. Berlin: Walter de Gruyter, 714—727.

Dunning, Ted (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19 (1), 61—74.

Eder, Maciej (2016). Słowa znaczące, słowa kluczowe, słowozbiory — o statystycznych metodach wyszukiwania wyrazów istotnych. *Przegląd Humanistyczny*, 3, 31—44.

Ellis, Paul (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press.

Erman, Britt and Warren, Beatrice (2000). The idiom principle and the open choice principle. *Text*, 20 (1), 29—62.

Forsyth, Richard (2014a). *Keysoft*. Available at: http://www.richardsandesforsyth.net/software.html (accessed on 14 March 2017).

Forsyth, Richard (2014b). *Keysoft. User notes* http://www.richardsandesforsyth.net/docs/formulib.pdf (accessed on 14 March 2017).

Forsyth, Richard (2015a). *Formulib: Formulaic Language Software Library*. Available at: http://www.richardsandesforsyth.net/zips/formulib.zip (accessed on 30 November 2015).

Forsyth, Richard (2015b). Formulib: Formulaic Language Software Library. User notes http://www.richardsandesforsyth.net/docs/formulib.pdf (accessed on 2 November 2015).

Forsyth, Richard and Sharoff, Serge (2014). Document dissimilarity within and across languages: A benchmarking study. *Literary and Linguistic Computing*, 29 (1), 6—22.

Forsyth, Richard, and Łukasz Grabowski (2015). Is there a formula for formulaic language? *Poznań Studies in Contemporary Linguistics*, 54 (1), 511—549.

Foster, Pauline (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In Martin Bygate, Peter Skehan and Merill Swain (eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing*. Harlow: Longman, 75—93.

Gabrielatos, Costas and Marchi, Anna (2011). Keyness. Matching metrics to definitions. Paper presented at the conference Corpus Linguistics in the South: Theoretical-methodological challenges in corpus approaches to discourse studies — and some ways of addressing them. Portsmouth, United Kingdom, 5 Nov 2011. Available at: http://repository.edgehill.ac.uk/4100/7/Gabrielatos%26Marchi-Keyness-2011.pdf (accessed 15 October 2012).

Gabrielatos, Costas (2018). Keyness Analysis: nature, metrics and techniques. In Charlotte Taylor and Anna Marchi (eds.), *Corpus Approaches to Discourse: A Critical Review*. Oxford: Routledge, 225—258.

Gałkowski, Błażej (2006). Kompetencja formuliczna a problem kultury i tożsamości w nauczaniu języków obcych. *Kwartalnik Pedagogiczny,* 4, 163—180.

Goźdź-Roszkowski, Stanisław (2011). *Patterns of Linguistic Variation in American Legal English. A Corpus-Based Study.* Frankfurt am Main: Peter Lang Verlag.

Grabowski, Łukasz (2014). On Lexical Bundles in Polish Patient Information Leaflets: A Corpus-Driven Study. *Studies in Polish Linguistics*, 19 (1), 21—43.

Grabowski, Łukasz (2015a). Keywords and lexical bundles within English pharmaceutical discourse: a corpus-driven description. *English for Specific Purposes*, 38, 23—33.

Grabowski, Łukasz (2015b). Phrase frames in English pharmaceutical discourse: a corpus-driven study of intra-disciplinary register variation. *Research in Language*, 3, 266— 291.

Grabowski, Łukasz (2015c). *Phraseology in English Pharmaceutical Discourse: A Corpus-Driven Study of Register Variation.* Opole: Wydawnictwo Uniwersytetu Opolskiego.

Grabowski, Łukasz (2018). Kilka słów o formuliczności z różnych perspektyw językoznawczych. In: Alicja Pstyga, Tatiana Kananowicz and Magdalena Buchowska (eds.), *Słowo z perspektywy językoznawcy i tłumacza. Tom VII. Frazeologia z perspektywy językoznawcy i tłumacza*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego, 67—76.

Grabowski, Łukasz and Jukneviciene, Rita (2016). Towards a refined inventory of lexical bundles: an experiment in the Formulex method. *Kalbu Studijos/Studies About Languages*, 29, 58—73.

Granger, Sylviane and Meunier, Fanny (2008). Introduction: The many faces of phraseology. In: Sylviane Granger & Fanny Meunier (eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins, xix—xxx.

Hardie, Andrew (2014). Statistical identification of keywords, lockwords and collocations as a two-step procedure. Paper delivered at the ICAME 35 conference, Nottingham, UK, March 2014. Available at: http://www.nottingham.ac.uk/conference/fac-arts/english/icame-35/documents/icame35-book-of-abstracts.pdf (accessed on 15 March 2017).

Hedges, Larry (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6 (2), 107—128.

Hyland, Kenneth (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4—21.

Ivanov, Vyacheslav (1957). Lingvisticheskie vzglyady E. D. Polivanova [Linguistic views of E.D. Polivanov]. *Voprosy Jazykoznanija*, 3, 55—76. Available at: http://vja.ruslang.ru/archive/1957-3.pdf (accessed 5 August 2019).

Kecskes, Istvan (2016). Deliberate Creativity and Formulaic Language Use. In: Keith Allan, Alessandro Capone amd Istvan Kecskes (eds.), *Pragmemes and Theories of Language Use*. Berlin: Springer, 3—20.

Kilgarriff, Adam (2009). "Simple maths for keywords". In: Michaela Mahlberg, Victorina González-Díaz and Catherine Smith (Eds), *Proceedings of Corpus Linguistics Conference CL2009.* University of Liverpool, UK, July 2009. Available at: https://www.sketchengine.co.uk/wpcontent/uploads/2015/04/2009-Simple-maths-for-keywords.pdf (accessed on 12 March 2017).

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Milos Jakubícek, Vojtech Kovář, Jan Michelfeit, Pavel Rychlý and Vit Suchomel (2014). The Sketch Engine: ten years on. *Lexicography* 1 (1), 7—36.

Kunilovska, Maria, Natalia Morgoun and Alexey Pariy (2018). Learner vs. professional translations into Russian: Lexical profiles. *Translation & Interpreting*, 10 (1), 33—52. Available at: https://trans-int.org/index.php/transint/article/view/585/304 (accessed on 16 December 2018).

Mel'cuk, Igor' (1995). Phrasemes in language and phraseology in linguistics. In: Martin Everaert, Erik-Jan van der Linden, Andre Schenk and Rob Schreuder (eds.), *Idioms: Structural and Psychological Perspectives*. Hillsdale: Lawrance Eribaum Associates, 167—232. Available at: http://bookre.org/reader?file=1500171&pg=175 (accessed 10 March 2014).

Mel'cuk, Igor' (1998). Collocations and Lexical Functions. In: Anthony Cowie (ed.), *Phraseology: Theory, analysis and applications*. Oxford: Clarendon Press, 21—53.

Montalt Resurrecció, Vicent and Gonzalez Davies, Maria (2007). *Medical Translation Step by Step. Translation Practices explained*. Manchester: St. Jerome Publishing.

Moon, Rosamund (2007). Corpus linguistic aspects of phraseology. In: Harald Burger (ed.), *Phraseologie: ein internationales Handbuch zeitgenoessischer Forschung* Vol. 2, Berlin: Walter de Gruyter, 1045—1059.

Murakami, Akira, Paul Thompson, Susan Hunston and Dominik Vajn (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora*, 12 (2), 243—277.

Myles, Florence and Cordier, Caroline (2017). Formulaic Sequence(fs) Cannot be an Umbrella Term in SLA: Focusing on Psycholinguistic FSs and Their Identification. *Studies in Second Language Acquisition*, 39, 3—28. Available at: https://www.cambridge.org/core/services/aop-cambridge-core/content/view/AFCD7233ACEC89C2A4314392127C5967/S027226311600036Xa.pdf/div-class-title-formulaic-sequence-fs-cannot-be-an-umbrella-term-in-sla-div.pdf (accessed on 12 December 2016).

Nam, Daehyeon and Lee, Sungmin (2016). Lexical bundles in spoken and written Russian. *Corpus Linguistics Research*, 2, 46. Available at: http://www.kacl.or.kr/read.php?pageGubun=journalsearch&pageNm=article&search=&journal=Vol.%202&code=286336&issue=21290&Page=2&year=2016&searchType=&searchValue= (accessed on 12 March 2017).

Nelson, Robert (2018). How 'chunky' is language? Some estimates based on Sinclair's Idiom Principle. *Corpora*, 13(3), 431—460.

O'Donnell, Matthew Brook (2011). The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal*, 35, 135—169.

Pęzik, Piotr (2013). Wybrane aspekty reprezentatywności małych i średnich korpusów. In: Wojciech Chlebda (ed.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów*. Opole: Wydawnictwo Uniwersytetu Opolskiego, 45—58.

Pęzik, Piotr (2015). Using n-gram independence to identify discourse-functional lexical units in spoken learner corpus data. *International Journal of Learner Corpus Research*, 1 (2), 242—255.

Pęzik, Piotr (2018). *Facets of prefabrication. Perspectives on modelling and detecting phraseological units*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Phillips, Martin (1989). *Lexical Structure of Text*. *Discourse Analysis Monographs 12*. Birmingham: University of Birmingham (cited in Scott 2001: 110).

Rosenfeld, Barry and Penrod, Steven (2011). *Research Methods in Forensic Psychology*. London: John Wiley and Sons (cited in Gabrielatos & Marchi 2011).

Schmitt, Norbert and Carter, Ronald (2004). Formulaic sequences in action: An introduction. In: Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins, 1—22.

Scott, Michael (1996—2017). *WordSmith Tools*. Liverpool: Lexical Analysis Software. Available at: http://www.lexically.net/wordsmith/ (accessed on 30 May 2017).

Scott, Michael (2001). Mapping key words to *problem* and *solution*. In Michael Hoey, Michael Scott and Geoff Thompson (eds.), *Patterns of text: In Honour of Michael Hoey*. Amsterdam: John Benjamins, 109—127.

Scott, Michael (2008). *WordSmith Tools Help*. Liverpool: Lexical Analysis Software.

Sinclair, John (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Silge, Julia and Robinson, David (2017). *Text Mining with R. A Tidy Approach*. [Section 6: Topic modelling]. Sebastopol: O'Reilly Media.

Stubbs, Michael (2011). Three concepts of keywords. In Michael Scott and Marina Bondi (eds.), *Keyness in Texts*. Amsterdam: John Benjamins, 21—42.

Vinogradov, Victor (1947/1977). O osnovnykh tipakh frazeologicheskikh edinits v russkom yazyke [About Basic Types of Phraseological Units in Russian]. In Alexey Shakhmatov (ed.), *Сборник статей и материалов* [Collection of Papers and Materials]. Moscow: Nauka, 339—364 (cited in Cowie 1998, 2—4 and Dobrovolskij & Filipenko 2007, 714). Available at: http://www.philology.ru/linguistics2/vinogradov-77d.htm (accessed on 10 August 2012).

Wood, David (2015). *Fundamentals of Formulaic Language*. London: Bloomsbury.

Wood, David (ed.) (2010a). *Perspectives on Formulaic Language: Acquisition and Communication.* London: Continuum.

Wood, David (ed.) (2010b). *Formulaic Language and Second Language Speech Fluency. Background, Evidence and Classroom Applications.* London: Continuum.

Wray, Allison and Perkins, Michael (2000). The functions of formulaic language: an integrated model. *Language & Communication*, 20, 1—28.

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

Wray, Allison (2008). *Formulaic language. Pushing the boundaries*. Oxford: Oxford University Press.

Wray, Allison (2009). Identifying formulaic language. Persistent challenges and new opportunities. In Roberta Corrigan, Edith Moravcsik, Hamid Ouali and Kathleen Wheatley (eds.). *Formulaic Language. Vol. 1. Distribution and historical change*. Amsterdam: John Benjamins. 27—51.

Article history:
Received: 29 January 2019
Revised: 02 March 2019
Accepted: 15 April 2019

История статьи:
Дата поступления в редакцию: 29 января 2019
Дата принятия к печати: 15 апреля 2019

Bionote:
ŁUKASZ GRABOWSKI — Associate Professor at the Institute of Linguistics, University of Opole (Poland), and Department of English and American Studies, University of Ostrava (Czechia). His research interests include corpus linguistics, phraseology, formulaic language, translation studies and lexicography. He is also interested in computer-assisted methods of text analysis. He has published research articles in *International Journal of Corpus Linguistics*, *English for Specific Purposes*, *International Journal of Lexicography* as well as book chapters with John Benjamins, Emerald or Springer, among others.
Contact information: e-mail: lukasz@uni.opole.pl or Lukasz.Grabowski@osu.cz
ORCID ID: 0000-0002-3968-9218

Сведения об авторе:
ЛУКАШ ГРАБОВСКИЙ — доцент Института языкознания Опольского университета (Польша) и кафедры английских и американских исследований Остравского университета (Чехия). Научные интересы включают корпусную лингвистику, фразеологию, лексикографию, теорию и практику перевода, а также методы анализа текста с привлечением компьютерных технологий. Имеет публикации в международных журналах, в частности *International Journal of Corpus Linguistics*, *English for Specific Purposes, International Journal of Lexicography* и др.
Контактная информация: e-mail: lukasz@uni.opole.pl или Lukasz.Grabowski@osu.cz
ORCID ID: 0000-0002-3968-9218