PHRASE FRAMES IN ENGLISH PHARMACEUTICAL DISCOURSE: A CORPUS-DRIVEN STUDY OF INTRA-DISCIPLINARY REGISTER VARIATION

ŁUKASZ GRABOWSKI
Opole University
lukasz@uni.opole.pl

Abstract
Focusing on the exploration of intra-disciplinary register variation in the pharmaceutical domain, this corpus-driven study attempts to describe the use, composition and discourse functions of phrase frames, that is, contiguous sequences of words identical except for one (Fletcher, 2002-2007), found in samples of four English pharmaceutical text types, such as patient information leaflets, summaries of product characteristics, clinical trial protocols and chapters/sections from academic textbooks on pharmacology. The study deals with a specific sub-type of phrase frames, that is, 4-word units with a variable slot in the medial position, e.g. *be * with caution, to take * medicine*. The results showed, among others, that the use and discourse functions of phrase frames vary across pharmaceutical text types, that the correlation between the frequency of phrase frames and their pattern variability may depend on a register or genre, and that it is justified to treat the discourse functions of phrase frames as distinct from those of their textual variants.

Key words: corpus linguistics, phraseology, phrase frames, register variation, pharmaceutical discourse

1. Introduction

In recent years, the research methods offered by corpus linguistics have opened up a wide range of perspectives on how to explore recurrent multi-word units used in the whole variety of text types or genres. Typically identified on the basis of frequency and distribution in texts, recurrent interrupted and/or uninterrupted sequences of words have been operationalized by researchers in different ways and hence variously referred to in specialist literature, e.g. "n-grams" (Stubbs, 2007), "chunks" (O'Keefe, McCarthy, & Carter, 2007), "phrase frames" (Fletcher, 2002-2007), "formulaic frames" (Biber, 2009), "formulas" (Simpson-Vlach & Ellis, 2010), "phrasal expressions" (Martinez & Schmitt, 2012), or "lexical frames" (Gray & Biber 2013), to name but a few recently proposed labels. Thanks to these theoretical constructs grounded in corpus linguistics it is now easier for researchers to obtain new descriptive data useful to explore phraseological variation or formulaicity in texts.

Developed only recently, phrase frames (short PFs) constitute a theoretical concept specifically designed to facilitate the description of phraseological patterns in texts.

PFs were introduced by Fletcher (2002-2007) who defined them as sets of variants of n-grams (of any length) identical except for one word, e.g. *if you * any*, *in the * of, for the sake of **. In other studies, PFs are also referred to as "formulaic frames" (Biber, 2009), typically consisting of invariable function words with an intervening variable slot for content words, or discontinuous "lexical frames" (Gray & Biber, 2013). The PFs are therefore generalizations of recurrent sequences of word forms, such as n-grams, clusters or lexical bundles, as the latter ones can be textual actualizations of the PFs. According to Römer (2009, p. 91), PFs may be used as a means of comparing pattern variability across different text types or registers as they provide an insight into how fixed multi-word units are in a given register and what degree of variation they exhibit. It is believed that a high number of variants of PFs attests to the higher degree of phraseological variation in texts, which can further translate into more pronounced register variation (Römer, 2009, p. 91). Furthermore, Forsyth and Grabowski (2014) showed that PFs may be used not only for generalizing phraseologies in texts, but also for measuring the degree of formulaic language and for ranking texts or corpora from the most to the least formulaic and, by implication, from the least to the most phraseologically varied.

In many studies conducted so far (e.g. Forchini & Murphy, 2008; Römer, 2010; Gray & Biber, 2013; Fuster-Marquez, 2014) PFs have been explored in terms of their use and discourse functions in a wide variety of registers and specialist domains, including financial (financial articles in English and Italian), academic (academic book reviews, academic writing, conversations in academic contexts, all produced in English), or tourist (English hotel websites), to name but a few examples. In this paper, four different English pharmaceutical text types will be explored, namely patient information leaflets, summaries of product characteristics, clinical trial protocols and chapters/sections from academic textbooks on pharmacology.[1] The assumption at the heart of such a choice is that those four text types differ with respect to their target users, production circumstances and communicative purposes and functions, among other factors. For example, patient information leaflets are found in sales packages of medicines and their main communicative function is to provide specific information on proper use of drugs or medicines by patients. Primarily targeted at a general public, this text type should be produced by pharmaceutical companies in a plain and reader-friendly style, for example with technical terms frequently accompanied by or substituted with explanations (Montalt Resurreccio & Gonzalez Davies, 2007, pp. 68-72). Attached to the application for marketing authorization submitted to the European Medicines Agency (short EMA), or to a national competent authority in the European Union member states, the summaries of product characteristics convey detailed descriptions of medicines in terms of their pharmacological, chemical, pharmaceutical and toxicological properties (Montalt Resurreccio & Gonzalez Davies, 2007, p. 73). This text variety is highly conventionalized in that it follows a standard form for every medicinal product and provides the same types of information in a fixed order, as specified in the guidelines

[1] This study extends earlier research on keywords and lexical bundles across English pharmaceutical text types (Grabowski, 2015a). Also, by focusing only on phrase-frames with a variable slot in the medial position and by treating the discourse functions of phrase frames as distinct from their textual variants, this study is also an extension of the author's earlier research on phrase frames (Grabowski, 2015b).

issued by EMA (Montalt Resurreccio & Gonzalez Davies, 2007, p. 73). Clinical trial protocols describe objectives, design and methodology of a clinical trial; they are used as reference documents by a number of different specialists involved in the clinical trial, such as investigators, study site coordinators, pharmacists, laboratory staff, statisticians, to name but a few (Fitzpatrick, 2005, p. 2; Wang & Bakhai, 2006, p. xii). Finally, academic textbooks introduce novices to a particular field of study (in this case–pharmacology) and help explain specialist concepts to readers who are new to the field (Biber & Conrad, 2009, p. 113). The content of academic textbooks is typically factual, with information presented in a maximally objective way (Biber & Conrad, 2009, p. 113).

The main hypothesis put forward in this paper is that the four pharmaceutical text types under scrutiny, found in different pharmaceutical contexts, will prioritize different PFs and thus reveal a high degree of intra-disciplinary register variation. Furthermore, it is expected that those PFs that are found to be the most frequent will exhibit differences with respect to their composition as well as the functions in the creation of pharmaceutical discourse.

Hence, the specific aims of this study encompass the description of the frequency distributions of PFs, identification of the most frequent PFs in each pharmaceutical text type, as well as the description of the composition and discourse functions of the PFs. Also, a number of research problems pertaining specifically to the analysis of PFs will be also broached upon, for example whether there is a correlation between the frequency of occurrence of PFs and the degree of their pattern variability, or whether it is legitimate to treat the discourse functions of PFs as distinct from the discourse functions of their textual variants.

It is hoped that the results of a descriptive and exploratory study like this one will yield an insight into the specificity of a particular pharmaceutical text type relative to other pharmaceutical text types and, consequently, provide empirical evidence of considerable register variation within pharmaceutical discourse. It is also hoped that the results of this study may be useful for teaching English to those professionals in the medical and pharmaceutical who are non-native speakers of English. Since pharmacists, pharmacy technicians, clinicians or researchers in the world over have no choice nowadays but to use English in their professional or research work, the knowledge of recurrent phraseological items used to refer to drugs or medicines in various pharmaceutical contexts may facilitate their day-to-day professional communication. And last but not least, it is believed that the insights from the study will provide valuable contribution to phraseological research focused on PFs.

2. Research material, methodology and stages of the study

In this study, a purpose-designed collection of samples of English pharmaceutical texts with circa 1.82 million words is used, split into four subcorpora corresponding to four specialist text types, such as patient information leaflets (PILs), summaries of product characteristics (SPCs), clinical trial protocols (CTPs) and chapters/sections from academic

textbooks on pharmacology (CATs).[2] This accords with the claims made by Koester (2006, p. 67) and Pęzik (2013, p. 58) who argue that smaller corpora representative of a given language variety are more suitable than large multi-million-word corpora to identify the connections between linguistic patterning and specialized contexts of language use. The computer programs custom-designed for text analysis, such as kfNgram (Fletcher, 2002-2007) and WordSmith Tools 5.0 (Scott, 2008), are used in order to obtain and process the research data for the different types of linguistic analyses.

This research adopts a corpus-driven approach, which means that the empirical corpus analysis of frequency distributions of recurrent PFs will enable one to determine whether particular linguistic features are more frequent in one pharmaceutical register than another in a more objective way as compared with the intuition-based approach prioritizing unusual and rare linguistic patterns. More specifically, the study will be conducted in a number of stages, including the analyses of frequency distributions of PFs across pharmaceutical text types, followed by a more detailed exploration of the PFs' structure and discourse functions. Hence, both quantitative and qualitative methods are used in the course of the study.

It is important to emphasize that this study deals with a specific sub-type of PFs, that is, sequences of four words with a variable slot in the medial position (e.g. *you may * to, must be * by*). The rationale behind this decision is that most corpus linguistic studies conducted on lexical bundles focus on 4-word units (Hyland, 2008, p. 8, Chen & Baker, 2010, p. 32). It is therefore believed that this study may provide complementary results with respect to research on 4-word lexical bundles in pharmaceutical text types (Grabowski 2015a).

In this paper, PFs are identified directly – using the computer program kfNgram (Fletcher, 2002-2007) – based on the full list of four-word grams in each subcorpus under scrutiny (i.e. the ones with frequencies equal or higher than 1). This means that an inventory of phraseological items is meant to include also those PFs whose slot-fillers are highly variable and occur with low frequencies, i.e. the PFs that occur at least twice with at least two slot fillers. In this approach, PFs are therefore not generated based on lexical bundles (Biber et al., 2004) that are subject to somewhat stricter criteria regarding their frequency (e.g. 40 occurrences per million words) and distribution (e.g. occurrences in at

[2] The corpus was compiled for personal non-commercial research and is therefore not available to the public. PILs (463 complete texts) were extracted from the Patient Information Leaflet (PIL) Corpus 2.0 (Buoayad-Agha, 2006) compiled at the Natural Language Technology Group at the University of Brighton (The PIL Corpus 2.0 is readily available at: http:// mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL. I. Next, SPCs (136 complete texts) were downloaded from the Open Source Parallel Corpus (OPUS) Project website (Tiedemann, 2009) while CTPs (240 complete texts) were retrieved from the Clinical Trials Register (CTR) database of the European Union, hosted by the European Medicines Agency (EMA) and readily available at https://www.clinicaltrialsregister.eu/index.html. Finally, fragments of 25 book chapters were extracted from the following textbooks: Bauer, L. (2008). *Applied Clinical Pharmacokinetics*. 2nd Edition. New York: McGraw-Hill Medical (5 chapters from Part I and Part II); Hollinger, M. (2003). *Introduction to Pharmacology*. 2nd Edition. London/New York: Taylor & Francis (13 chapters); Craig, Ch. & Stitzel, R. (2004). *Modern Pharmacology with Clinical Applications*. 6th Edition. Lippincott: Williams & Wilkins (7 chapters). The size (in word tokens) of each subcorpus is as follows; PILs – 474,458; SPCs – 670,907; CTPs – 468,957; CATs – 213,159. The total size of the study corpus is 1,827,481 word tokens.

least 5 texts representing a given text variety), a procedure that is bound to yield a less comprehensive inventory of PFs. This problem is also discussed by Gray and Biber (2013, pp. 111-115) who argue that not all frequent and recurrent PFs are associated with highly frequent contiguous sequences of words; that is why in this study the PFs are identified based on the full list of n-grams identified in each pharmaceutical text type. Furthermore, only those PFs based on contiguous sequences of four words that are neither divided by sentence or clause boundaries (i.e. by full-stops, semi-colons or commas etc.) nor contain numbers will be analyzed in the study. Appendix A presents the 50 most frequent PFs in the pharmaceutical text types under scrutiny.[3]

Next, the 50 most frequent PFs with a variable slot in the medial position (A* CD and AB*D) will be explored in terms of their structure and discourse functions. Capitalizing on the observation made by Römer (2010), it was decided that those PFs with variable slots in either the initial or final position (*BCD and ABC*) will not be analyzed since they are often fragments of longer PFs and/or contain empty slots filled with function words that hardly lend themselves to qualitative analyses, as demonstrated by a number of recent studies (e.g., Römer, 2010; Gray & Biber, 2013; Fuster-Marquez, 2014). In fact, these studies showed that the most interesting insights for functional or register analyses can be gained from explorations of PFs with empty slots in medial positions, notably by means of semantic analyses of lexical slot fillers.

As for the structural analysis, the PFs will be divided – using the classification proposed by Gray and Biber (2013, p. 122) – into three groups, namely verb-based PFs (V-based) with one or more modal, auxiliary or lexical verbs (e.g. *the * should be*, *is * in patients*), PFs with content words (C-based) other than verbs (e.g. *with other * products*, *once every * weeks*) and PFs with function words only (F-based), e.g. *the * of the*, *in * the of*. Such a coarse-grained taxonomy has been chosen primarily because PFs rarely constitute complete grammatical or syntactic units, e.g. *the * of the* is a noun phrase with a post-modifier fragment. In short, the aim of the structural analysis is to determine whether different registers exhibit any similarities or differences with respect to the composition of the most frequent PFs. For example, in one of the studies, Gray and Biber (2013, p. 128) found that both F-based and V-based PFs prevail in academic writing while V-based PFs are the most frequent in conversations. This means that different registers may prioritize different structural types of PFs, the hypothesis that will be further verified in this study on the example of four distinct pharmaceutical registers.

In the next stage, the 50 most frequent PFs found in each pharmaceutical text type will be explored qualitatively in terms of their discourse functions.[4] Such studies have been already conducted by scholars, e.g. by Forchini and Murphy (2008), Römer (2010) or Fuster-Marquez (2014). For example, researching phraseological items found in academic book reviews, Römer (2010) divided PFs into functional categories such as ‘evaluation

[3] For comparative purposes, Appendix A also contains the list of the 50 most frequent PFs in the British National Corpus (Fletcher, 2010).

[4] Wray and Perkins (2000, p. 8) argue that on the whole any functional typology for recurrent phraseologies is bound to suffer from proliferation of types and subtypes, often domain-specific. That is why the functional labels assigned by researchers to multi-word units are not absolute but typically represent only tendencies or approximations foregrounding the primary rather than secondary or peripheral functions fulfilled by the multi-word units in the majority of their contexts of use. (Grabowski, 2015a, p. 26)

PFs', 'structure PF's, 'content PFs' and 'discourse PFs'; Fuster-Marquez (2014), who looked into multi-word patterns in English hotel websites, described PFs using the typology proposed by Biber, Conrad and Cortes (2004), originally applied to lexical bundles. More specifically, Fuster-Marquez (2014) divided PFs into 'referential' ones, 'discourse organizers' and 'expressing stance'.

However, the approach used in the aforementioned studies is not devoid of problems. The functional labels applied to PFs are contingent on the discourse functions of their variants or textual realizations, called n-grams, lexical bundles, clusters or otherwise. In the case when a PF has a high number of slot-fillers, its functional label is typically contingent on the semantic category of the most frequent slot-filler/s, or on the semantics of PFs' textual variants (n-grams) determined by additional analysis of their left-hand and/or right-hand context. This means that the functional labels are generalizations of the semantics of the most frequent slot-fillers or of the semantics of longer stretches of discourse where one may find a given PF. On the one hand, this is perfectly understandable considering the fact that PFs themselves are generalizations of lexical patterns in texts. On the other hand, since – to the knowledge of the author – there is no empirical evidence available showing that the variants of PFs are normally distributed, it is difficult to put forward any assumptions concerning dominant discourse functions of PFs. Also, the same slot filler may vary in terms of semantics depending on the larger discourse context. For example, the noun *end* found in the PF *the * of the* may function as a temporal or location marker, among other possibilities, depending on the right-hand context. That is why in a case like this one researchers typically assign general rather than specific functional labels; in the aforementioned example, *the * of the* would be labelled a 'referential PF' or 'content PF' since the slot-filler *end* conveys domain-specific information when followed by such post-modifiers as *of the trial*, *of the study*, *of the menstrual cycle* etc. However, when looking again at this example one may note that the referential function is inherent in either the slot-filler or longer text chunk rather than in the PF itself.

That is why there arises the question concerning the discourse functions, if any, of the PFs as such, e.g. *the * of the*. As a syntagmatic association of function words, *the * of the* performs a textual, discourse-organizing function of framing the propositional content of the slot-fillers. This means that this PF, originally a discourse-organizing syntactic frame, may become a referential lexical bundle or n-gram (e.g. a temporal or location marker or otherwise) depending on the semantics of slot-fillers and/or larger discourse context. To avoid this problem, in this study the functional labels are assigned to PFs based on the nature of their fixed components rather than the semantics of slot-fillers and/or longer chunks of texts with a given PF.[5] Consequently, the general functional labels assigned to the 50 most frequent PFs in each text variety include the following categories:

a) 'topic PFs' related to the specialist field by referring to various aspects of the use and administration of pharmaceutical products; typically C-based or V-based PFs consisting of lexical and function words (e.g. *the * of distribution*, *the * rate constant*, *drug * the blood*, *for * treatment of*, *the * nervous system*);

[5] The latter approach will be used only when comparing textual realizations of the same PFs found in multiple pharmaceutical text types, as described later in this section.

b) 'generic PFs' that are typically C-based PFs not semantically restricted to the specialist pharmaceutical field (e.g. *as * as possible*, *in a * place*, *at the * time, last * of the*, *main * of the*, *is * the last*);
c) 'discourse-organizing PFs' that are typically F-based PFs providing syntactic frames for information conveyed in pharmaceutical texts and consisting solely of function words (e.g. *in the * of*, *the * of the*, *of the * is*, *and * is the*, *if the * is*, *have been * in*);
d) 'action-oriented PFs' that are typically V-based PFs composed of lexical and function words and used to frame or convey stance in terms of recommendations, directives or desires targeted at readers with respect to actions to be undertaken in connection with proper use or administration of medicines (e.g. *can be * to*, *should be * to*, *the * should be*, *should be * in, it is * recommended, must be * by, must be * with*);
e) 'reading-oriented PFs' that can be C-based or V-based PFs with lexical and function words, facilitating navigation through pharmaceutical texts and referring to their micro- or macro-structures (e.g. *of * see section, of * see chapter, later in * chapter*).

As can be seen, likewise the general functional typologies applied to PFs by Römer (2010) or Fuster-Marquez (2014), the taxonomy proposed in this study is a domain-specific one, corresponding to the specificity of both the pharmaceutical text types and the language used to speak or write about drugs and medicines in various pharmaceutical contexts. However, unlike other typologies, the one proposed in this paper is derived from the semantics and functions of the fixed components of PFs rather than from the meanings carried by either their slot-fillers or longer chunks of discourse. In short, this means that at that stage of the study the discourse functions of PFs are treated as distinct from the functions of their textual variants.

Finally, in order to provide a more comprehensive description of pharmaceutical phraseologies, in the last stage of the study the specific discourse functions of the variants (n-grams) of those of the 20 most frequent PFs that overlap across multiple pharmaceutical texts will be explored qualitatively based on the functional labels proposed by Biber et al. (2004) and Hyland (2008) and originally applied to lexical bundles. This is done to ensure that the description of register variation is not limited to abstract generalizations of phraseological patterns captured in the PFs, but includes also the textual realizations of PFs, that is, the most frequent contiguous sequences of four words.

3. Results

3.1 Distribution and pattern variability of phrase frames

Table 1 presents a general numerical description of frequencies and distributions of PFs with a variable slot in the medial position. More specifically, the data present the numbers of PFs divided into two frequency bands (top and medium) with corresponding values of the type token percentage index (short TTPC), a productivity measure applied to PFs, an equivalent of type/token ratio yet expressed in per cent (Forsyth & Grabowski,

2014).[6] In short, the higher the value of the TTPC, the higher the phraseological productivity of PFs.

The results show that the CATs have the lowest number of the most frequent PFs (49 in the top frequency band) yet at the same time these PFs are more phraseologically varied (TTPC = 50.16%) as compared with the remaining text types. For example, the F-based PF *of the * of*, a prepositional phrase with a post-modifier *of*-phrase fragment, occurs in the CATs 106 times (497 times per million words) with as many as 38 different variants (TTPC = 69.81%). Needless to say, in the top and medium frequency bands one may find the PFs with TTPC scores higher than 90%, e.g. *of the * to* (92.68%) or *in the * the* (96.42%). Next, the writers of PILs make frequent use of the highest number of PFs (198 with frequency of more than 200 per million words) yet these PFs are considerably less phraseologically varied (6.09%) than the ones in the CATs (50.16%). On the other end of the spectrum, in the SPCs and CTPs the most frequent PFs are simultaneously the most fixed ones. Interestingly, in the CTPs the pattern variability of PFs from the medium frequency band is higher than the one in the PILs and SPCs. After consulting the list of PFs in the CTPs, one may note that among 39 phraseologies in the medium frequency band there are F-based (*for the * of*, *as * by the*, *with a * of*, *with the * of*) and V-based PFs with auxiliary verbs (*will be * to*, *of the * is*, *is to * the*) filled by a variety of content words; these PFs exhibit high degree of pattern variability, e.g. the TTPC of the V-based PF *will be * to* found in the CTPs 61 times (130 pmw) is 57.37%.

Frequency bands (pmw)	**PILs**		**SPCs**		**CTPs**		**CATs**	
	No. of PFs	**Mean TTPC**	**No. of PFs**	**Mean TTPC**	**No. of PFs**	**Mean TTPC**	**No. of PFs**	**Mean TTPC**
More than 200 hits (Top)	198	6.09%	91	4.19%	96	2.16%	49	50.16%
200 – 101 hits (Medium)	314	10.03%	195	6.60%	39	12.01%	83	46.58%
Total no. of PFs	19,112		9,412		6,194		11,645	

Table 1. Distribution of PFs based on 4-word grams in high and medium frequency bands

The results show that the CATs have the lowest number of the most frequent PFs (49 in the top frequency band) yet at the same time these PFs are more phraseologically varied (TTPC = 50.16%) as compared with the remaining text types. For example, the F-based PF *of the * of*, a prepositional phrase with a post-modifier *of*-phrase fragment, occurs in the CATs 106 times (497 times per million words) with as many as 38 different variants (TTPC = 69.81%). Needless to say, in the top and medium frequency bands one may find the PFs with TTPC scores higher than 90%, e.g. *of the * to* (92.68%) or *in the * the* (96.42%). Next, the writers of PILs make frequent use of the highest number of PFs

[6] This metric is similar to variant-to-phrase frame ratio (VPR) originally introduced by Römer (2010, p. 105).

(198 with frequency of more than 200 per million words) yet these PFs are considerably less phraseologically varied (6.09%) than the ones in the CATs (50.16%). On the other end of the spectrum, in the SPCs and CTPs the most frequent PFs are simultaneously the most fixed ones. Interestingly, in the CTPs the pattern variability of PFs from the medium frequency band is higher than the one in the PILs and SPCs. After consulting the list of PFs in the CTPs, one may note that among 39 phraseologies in the medium frequency band there are F-based (*for the * of*, *as * by the*, *with a * of*, *with the * of*) and V-based PFs with auxiliary verbs (*will be * to*, *of the * is*, *is to * the*) filled by a variety of content words; these PFs exhibit high degree of pattern variability, e.g. the TTPC of the V-based PF *will be * to* found in the CTPs 61 times (130 pmw) is 57.37%.

In the next stage, the aim was to verify whether there exists a correlation (monotonic or linear association) between the frequency of PFs with a variable slot in the medial position and their pattern variability. In other words, the aim was to verify whether there exists the association whereby the pattern variability of PFs increases or decreases with their frequency of occurrence. In general, this finding may be pedagogically useful since positive association suggests that the most frequent PFs (and, in such a case, the most productive ones) should be emphasized in ESP teaching.

As no assumptions have been made with respect to the normal distribution of PFs, a non-parametric Spearman's rank correlation bivariate test (using paired data representing frequencies and numbers of variants of PFs) was applied to the 300 most frequent PFs in each pharmaceutical text variety (Table 2).[7] According to Rowntree (2000, p. 163) and Stangroom (2014), the rank correlation coefficient may vary from r=1 indicating strong positive association (pattern variability of PFs increases as their frequency increases) to r=-1 indicating strong negative association (pattern variability of PFs increases as their frequency decreases); r=0 indicates no correlation. As usual in social sciences, the association is considered to be statistically significant if the p value is lower than 0.05.

Variables	PILs	SPCs	CTPs	CATs
Frequency & number of variants	r=0.11086 (p=0.05511). Association not statistically significant.	r=0.16398 (p=0.0044). Association statistically significant.	r=-0.13004 (p=0.02422). Association statistically significant.	**r=0.50779** (p=0). Association statistically significant.

Table 2. Correlation between frequency and number of variants among the top-300 PFs by frequency (Spearman's rho test)

The results revealed that a moderate positive association is statistically significant only for the 300 most frequent PFs found in the CATs. This means that in this text variety the pattern variability of PFs moderately increases with their frequency. For the PILs and SPCs, the test revealed a very weak (negligible) positive association, yet in the former text variety it was not found to be statistically significant (p=0.05511). In the CTPs, the test

[7] The calculations were completed using an online suite of statistical calculators called Social Science Statistics available at http://www.socscistatistics.com/tests/spearman/ (Stangroom, 2014).

revealed negligible negative association whereby the pattern variability of PFs increases when their frequency decreases, the finding corroborated by the higher value of the TTPC in the medium frequency band than in the top one (Table 1). In short, one may conclude that the correlation between the frequency of PFs and their pattern variability may be contingent on particular registers or genres.

3.2 Composition and pattern variability of phrase frames

The subsequent stage of the study pertains to the analysis of the composition of the 50 most frequent PFs (Table 3). As explained in the methodological section, the PFs were divided – using the classification proposed by Gray and Biber (2013) – into three groups: namely V-based (with one or more verbs), C-based (with function words and content words other than verbs) and F-based PFs (with function words only). For each group, the mean TTPC score was computed in order to compare pattern variability of the different structural types of the PFs.

Structural categories	**PILs**		**SPCs**		**CTPs**		**CATs**	
	No.	**Mean TTPC**	**No.**	**Mean TTPC**	**No.**	**Mean TTPC**	**No.**	**Mean TTPC**
V-based	***21***	4.13%	***22***	5.37%	15	0.94%	16	48.71%
C-based	11	3.51%	19	1.92%	***27***	1.13%	11	25.53%
F-based	18	***9.12%***	9	***8.98%***	8	***3.55%***	***23***	***64.08%***

Table 3. Composition and pattern variability of PFs across pharmaceutical text types (top-50 by frequency)

The results revealed that the PILs and SPCs rely on PFs composed of verbs, as both text types describe various properties of medicines as well as their recommended use and administration, the information primarily conveyed by verbs. The CTPs were found to rely primarily on PFs built of content words (27 out of 50), the finding that may result from a highly conventionalized macro-structure of the CTPs dominated by nominalizations. In fact, V-based PFs in the CTPs are relatively fixed: 14 items contain only nouns and function words (e.g. *information * the trial*, *objectives * the trial*, *the trial * a*) while 6 items are composed of nouns, adjectives and function words (e.g. *in * clinical trial*, *secondary objectives * the*). This shows that this text type relies more than others on the frequent use of nominalizations. Finally, the CATs rely most on PFs composed of function words (23 out of 50); these provide syntactic frames for content words conveying domain-specific information.[8] In fact, high values of both the mean TTPC (50.16%) of the PFs with frequency of more than 200 per million words and the mean TTPC (64.08%) of F-based PFs from among the 50 most frequent ones in the CATs indicate that a wide variety of content words fill empty slots in the recurrent F-based PFs. This also means that

[8] For the sake of comparison (Appendix A), among the 50 top-frequency PFs in the BNC, one may find 39 F-based PFs (composed of articles, prepositions and/or conjunctions), 9 V-based PFs (with the verb *to be* typically used as an auxiliary) and 2 C-based PFs only (*at the * time*, *the first * of*).

writers of the CATs are less restricted in terms of linguistic creativity as compared with producers of PILs, SPCs and CTPs, the clichéd text types produced in accordance with strict guidelines. A more detailed analysis of V-based PFs among the 50 most frequent ones in the PILs and SPCs (21 and 22 respectively) revealed that they are much more restricted in form in the latter text type: in the SPCs 11 PFs contain the modal verb *should* (e.g. *should be * by*, *should be * in*, *should be * to*, *should be * with*, *should be * for, should be * as*) while in the PILs only one PF contains the said modal verb (*you should * your*) and another one contains the modal verb *may* signalling a strong suggestion (*you may * to*). This also shows that the SPCs are written in an impersonal style characteristic of extensive use of passive voice; on the contrary, the modal verbs in the PILs are used in active voice so that the reader is addressed in a straightforward way. This is due to the fact that these two text varieties are directed at different types of target readers (patients vs. specialists in the pharmaceutical field). The most frequent V-based PFs are, on average, slightly more phraseologically varied in the SPCs (TTPC = 5.37% vs. 4.13 % in PILs) as the empty slots following the passive construction with the modal verb (*should be **) are filled by a wide variety of action verbs in the participial form (e.g. *advised, given, reduced, paid, taken, restricted*).

To sum up, it was revealed that, on average, the F-based PFs are the most phraseologically varied, followed by V-based and C-based PFs (the most fixed ones), the finding applicable to the 50 most frequent PFs in four pharmaceutical text types under scrutiny. The results also revealed that written academic pharmaceutical discourse (the CATs) relies mostly on variable frames composed of function words (F-based PFs). On the contrary, the PILs and SPCs rely on frames made up of verbs (V-based PFs) while the CTPs rely on the PFs composed of nouns and/or adjectives and function words (C-based PFs). This means that in these three pharmaceutical text types phraseology is primarily lexical and fixed while in the CATs it is largely grammatical and highly variable.

In order to better illustrate the quantitative findings, Tables 4-7 list the 20 most frequent PFs in each pharmaceutical register, including information on the raw and normalized frequencies of PFs, the number of their variants, pattern variability (measured by the TTPC) and with specification of the 3 most frequent slot-fillers. Also, the PFs from among the 20 most frequent ones in four text types are marked in bold and underlined; those that occur in three text types are marked in bold while those that are common to two text types are marked in bold and italics. In short, the results revealed that only five PFs (*the * of the, in the * of, on the * of, of the * of, it is * to*) overlap across the 20 most frequent items, the finding that underscores intra-disciplinary register variation in the pharmaceutical domain of language use.

PILs	Frequency (raw)	Frequency (norm. pmw)	Variants (types)	TTPC (in %)	Structural type	Three most frequent slot fillers
*if you * any*	798	1682	17	2.13	F-based	*have, notice, experience*
the * of the	632	1332	133	21.04	F-based	*lining, association, end*
*you * to take*	478	1007	12	2.51	V-based	*start, forget, need*
*if you * to*	363	765	14	3.86	F-based	*forget, want, have*
it is * to	341	719	31	9.09	V-based	*important, used, best*
*the * of your*	339	714	60	17.70	F-based	*name, advice, whites*
*do not * the*	319	672	50	15.67	V-based	*take, use, put*
*if you * not*	312	658	6	1.92	V-based	*are, do, have*
*to * your medicine*	304	641	8	2.63	C-based	*take, use, store*
of the * of	293	617	45	15.36	F-based	*reach, lining, ingredients*
*to take * medicine*	287	605	5	1.74	V-based	*your, this, the*
*if you * a*	275	580	19	6.90	F-based	*miss, have, are*
*you * any of*	261	550	12	4.60	F-based	*have, experience, get*
*your * tells you*	254	535	2	0.79	V-based	*doctor, GP*
*you should * your*	247	521	27	10.93	V-based	*tell, consult, see*
*as * as possible*	237	499	15	6.33	C-based	*soon, evenly. slowly*
*in a * place*	231	487	11	4.76	C-based	*safe, dry, cool*
*if you * taking*	230	485	4	1.74	V-based	*are, stop, start*
*before you * to*	223	470	9	4.04	C-based	*start, go, decide*
*you are * to*	221	466	21	9.50	V-based	*allergic, planning, going*

Table 4. Top-20 PFs, by frequency, with a variable slot in the middle in PILs

SPCs	Frequency (raw)	Frequency (norm. pmw)	Variants (types)	TTPC (in %)	Structural type	Three most frequent slot fillers
*the * should be*	617	919	26	4.21	V-based	*dose, haemoglobin, patient*
in the * of	614	915	52	8.47	F-based	*treatment, absence, event*
on the * of	414	617	35	8.45	F-based	*rate, surface, use*
the * of the	412	614	74	17.96	F-based	*end, participation, needs*
*once every * weeks*	326	485	3	0.92	C-based	*three, two, four*
*should be * by*	312	465	18	5.77	V-based	*reduced, administered, initiated*
*the dose * be*	311	463	5	1.60	V-based	*should, may, can*
*should be * in*	298	444	21	7.05	V-based	*considered, monitored, done*
*should be * to*	295	439	25	8.47	V-based	*advised, given, reduced*
*patients with * renal*	292	435	11	3.77	C-based	*chronic, severe, impaired*
*should be * with*	282	420	17	6.03	V-based	*used, treated, administered*
*with * medicinal products*	281	418	4	1.42	C-based	*other, antipsychotic, these*
*ability to * and*	276	411	3	1.09	C-based	*drive, concentrate, lower*
*with other * products*	268	399	2	0.75	C-based	*medicinal, intravenous*
*be * with caution*	255	380	4	1.57	V-based	*used, administered, treated*
*of * should be*	255	380	41	16.08	V-based	*administration, insulin, driving*
*to drive * use*	252	375	2	0.79	V-based	*and, or*
*drive * use machines*	251	374	2	0.80	V-based	*and, or*
*an increased * of*	240	357	6	2.5	C-based	*risk, incidence, frequency*
*dose * should be*	238	355	10	4.2	V-based	*reduction, patients, adjustment*

Table 5. Top-20 PFs, by frequency, with a variable slot in the middle in SPCs

CT	Frequency (raw)	Frequency (norm. pmw)	Variants (types)	TTPC (in %)	Structural type	Three most frequent slot fillers
*be * in the*	1143	2437	10	0.87	V-based	*used, involved, included*
*to be * in*	1133	2416	10	0.88	V-based	*used, involved, included*
the * of the	908	1936	85	9.36	F-based	*end, duration, opinion*
*be used * the*	871	1857	4	0.46	V-based	*in, for, throughout*
*Therapy * product no*	764	1629	2	0.26	C-based	*medical, medicinal*
*of the * to*	639	1362	12	1.88	F-based	*IMP, trial, investigator*
*women of * potential*	589	1256	3	0.51	C-based	*childbearing, child-bearing, reproductive*
*used in * trial*	576	1228	2	0.35	V-based	*the, this*
*of * for this*	573	1222	2	0.35	F-based	*administration, subjects*
of the * of	530	1130	32	6.04	F-based	*duration, end, effect*
*information * the trial*	496	1058	2	0.40	C-based	*on, about*
*objective of * trial*	493	1051	2	0.40	C-based	*the, this*
*in * clinical trial*	453	966	3	0.66	C-based	*the, a, another*
*the * to be*	447	953	6	1.34	F-based	*IMP, investigator, area*
*the trial * a*	447	953	8	1.79	C-based	*has, part, is*
*the * has a*	446	951	4	0.90	V-based	*trial, subject, patient*
*the * has been*	445	949	7	1.57	V-based	*IMP, study, patient*
*medical * information not*	443	945	2	0.45	C-based	*device, product*
*in the * has*	436	930	2	0.46	V-based	*trial, study*
*in this * as*	436	930	3	0.69	F-based	*indication, study, trial*

Table 6. Top-20 PFs, by frequency, with a variable slot in the middle in CTPs

ATs	Frequency (raw)	Frequency (norm. pmw)	Variants (types)	TTPC (in %)	Structural type	The most Frequent slot fillers (3)
the * of the	447	2097	225	50.33	F-based	*value, half-life, size*
in the * of	277	1299	118	42.59	F-based	*treatment, presence, case*
*the * of a*	148	694	87	58.78	F-based	*formation, presence, metabolism*
*for the * of*	139	652	61	43.88	F-based	*treatment, purposes, metabolism*
*to the * of*	134	628	78	58.20	F-based	*development, size, use*
*of * in the*	122	572	53	43.44	F-based	*drug, drugs, pharmacology*
*the * of drug*	119	558	49	41.17	C-based	*amount, rate, concentration*
*is the * of*	112	525	55	49.10	V-based	*volume, result, study*
of the * of	106	497	74	69.81	F-based	*volume, effect, development*
*can be * to*	103	483	29	28.15	V-based	*used, extrapolated, difficult*
*and the * of*	94	441	60	63.82	F-based	*volume, use, rate*
*the * of distribution*	94	441	3	3.191	C-based	*volume, volumes, processes*
*the * rate constant*	90	422	3	3.333	C-based	*elimination, absorption, distribution*
*of drug * the*	87	408	18	20.68	C-based	*in, from, action*
*in the * and*	80	375	47	58.75	F-based	*blood, body, brain*
on the * of	80	375	49	61.25	F-based	*basis, route, order*
*of the * is*	78	366	48	61.53	V-based	*drug, curve, patient*
*steady-state peak * trough*	74	347	2	2.702	C-based	*and, or*
it is * to	71	333	24	33.80	V-based	*possible, important, difficult*
*of the * and*	68	319	54	79.41	F-based	*drug, membrane, disease*

Table 7. Top-20 PFs, by frequency, with a variable slot in the middle in ATs

3.3 Discourse functions of phrase frames: quantitative and qualitative analysis

This part of the study focuses on a qualitative analysis of discourse functions of the PFs. As explained earlier, in this paper the functional labels were assigned to PFs based on the nature of their fixed components rather than the semantics of slot-fillers or longer stretches of texts containing a given PF. More specifically, the 50 most frequent PFs in each text variety were divided into the following categories (Table 8): 'topic PFs' (T) related to the specialist field (e.g. *the * rate constant*, *the * nervous system*); 'generic PFs' (G) not semantically restricted to the specialist field (e.g. *as * as possible*, *in a * place*); 'discourse-organizing PFs' (DO) performing the role of syntagmatic frames for information conveyed by pharmaceutical texts and composed of function words only (e.g. *in the * of*, *the * of the*); 'action-oriented PFs' (A) used to convey stance in terms of recommendations, directives or desires (e.g. *can be * to*, *should be * to*, *must be * by*); and 'reading-oriented PFs' (RO) facilitating navigation through pharmaceutical texts or recommending their perusal (e.g. *read * leaflet carefully, later in * chapter*).

Functional categories	PILs	SPCs	CTPs	CATs
	No.	No.	No.	No.
Topic PFs	8	19	27	11
Generic PFs	8	10	9	1
Discourse-organizing PFs	21	10	14	35
Action-oriented PFs	11	11	0	3
Reading-oriented PFs	2	0	0	0

Table 8. Discourse functions of PFs across pharmaceutical text types (top-50 by frequency)

The SPCs and CTPs have the highest number of topic PFs that are content-related. This means that these text types are not varied in terms of syntactic structures because topic PFs are typically nominalizations that refer to key properties of drugs or medicines. On the contrary, in the PILs and, in particular, in the CATs the most dominant group are discourse-organizing PFs, the finding that shows that these two text types are more syntactically and stylistically varied, notably when compared with the SPCs and CTPs. A relatively high number of action-oriented PFs in the PILs and SPCs has been expected since the majority of these items are found in sections describing recommended use and administration of medicines (e.g. in the SPCs in the section *4.2 Posology and method of administration* and in the section *4.4 Special warning and precautions for use*).

As mentioned earlier in this paper, the general discourse functions of PFs discussed above were treated in this study as distinct from the functions of their textual variants (n-grams) described using more fine-grained functional labels proposed by Biber et al. (2004), Biber (2006) or Hyland (2008). This is primarily due to the fact that PFs and n-grams (or lexical bundles, if stricter identification criteria are applied) occupy different positions on the abstract-concrete phraseological continuum. In practice, the discourse functions of PFs and their variants may either overlap or differ. For example, in the PILs the function of the variants *if you have any problems*, *if you have any question* or *if you want any* is the same as the one of the discourse-organizing PFs (*if you * a*, *if you * any*); in fact, both PFs and their variants are used to introduce conditions; on the contrary,

the most frequent variant of the discourse organizing PF *it is * to* becomes a sequence expressing stance (*it is important to*) by emphasizing the importance of the following proposition.

This example shows that to further the knowledge of the scope of intra-disciplinary register variation, it is necessary to compare the functions of the variants of those PFs that overlap across multiple pharmaceutical text types. In what follows, the discourse functions of the 5 PFs (*the * of the*, *of the * of*, *it is * to*, *on the * of*, *in the * of*) from among the 20 most frequent ones in two or more pharmaceutical text types are further explored qualitatively with the help of the Concord component of WordSmith Tools 5.0 (Scott 2008). Since no other criterion than the frequency of occurrence was applied to select the variants, the specific phraseologies that were analyzed represent recurrent n-grams composed of four words (Table 9).[9] This is meant to ensure that for the purposes of phraseological description the most frequent PFs are not abstracted from their actual use across samples of pharmaceutical text types.

Phrase frame	Top-5 variants (slot-fillers) by frequency			
	PILs	**SPCs**	**CTPs**	**CATs**
*the * of the*	*lining* (65), *association* (63), *end* (58), *top* (25), *rest* (17)	*end* (35), *participation* (35), *needs* (22), *majority* (21), *course* (21)	*end* (312), *duration* (269), *opinion* (115), *course* (17), *date* (12)	*value* (14), *half-life* (13), *size* (13), *concentration* (12), *development* (10)
*of the * of*	*reach* (161), *lining* (30), *ingredients* (14), *group* (12), *tissues* (7)		*duration* (240), *end* (240), *effect* (10), *efficacy* (4), *combination* (3)	*volume* (9), *effect* (5), *development* (4), *concentration* (4), *presence* (3)
*it is * to*	*important* (170), *used* (52), *best* (32), *essential* (18), *advisable* (8)			*possible* (21), *important* (10), *difficult* (6), *necessary* (6), *easy* (4)
*on the * of*		*rate* (75), *surface* (73), *use* (45), *day* (30), *pharmacokinetics* (26)		*basis* (12), *route* (4), *order* (4), *part* (4), *development* (3)
*in the * of*		*treatment* (100), *absence* (69), *event* (55), *presence* (49), *incidence* (42)		*treatment* (26), *presence* (18), *case* (17), *value* (14), *formation* (8)

Table 9. Phraseologies based on the PF overlapping across pharmaceutical text types

The comparison revealed one PF found among the 20 most frequent items in four pharmaceutical text types, namely the discourse-organizing F-based PF *the * of the*, a noun

[9] Importantly, all five PFs in Table 9 are found in each pharmaceutical text type. However, the data focus only on the overlap among the 20 most frequent PFs, a decision made to limit the amount of data for detailed concordance analyses.

phrase with a post-modifier fragment. This item is the most phraseologically varied in the CATs (with the TTPC in the region of 50% while it is the most fixed in the CTPs with the TTPC of 9.36%). All in all, the variants of this PF generally perform referential functions in each text type. Firstly, in PILs they refer to locations by specifying parts of the human body affected by medicines, e.g. *the lining of the* (*womb/uterus/bowel/stomach*), parts of medical devices or medicines, e.g. *the end of the* (*applicator/leaflet/pen/strip/syringe/tube*), *the top of the* (*barrel of the syringe/inhaler/memo-pack/nebuliser/syringe/vial*), *the rest of the* (*medicine/pack/patch/pills/tablets*), or institutions operating in the pharmaceutical sector, e.g. *the Association of the* (*British Pharmaceutical Industry*). The n-gram *the end of the* (*course/day/menstrual cycle/month/treatment*) is also used to mark temporal references. In the SPCs, the variants of the same PF mark temporal references, e.g. *the end of the* (*dialysis session*/*first dosing interval*/*infusion*/*study*/*three weeks*), *the course of the* (*study/dosing interval*), or refer to the parties involved in the use and administration of medicines, e.g. *the participation of the* (*individual patient*), *the needs of the* (*patient*), *the majority of the* (*patients*). It is similar in the CTPs where the most frequent phraseologies mark temporal references, e.g. *the end of the* (*observation*/*study*/*trial*), *the duration of the* (*study*/*trial*), *the course of the* (*study*/*trial*), *the date of the* (*last visit*/*data capture*/*patient's treatment*). The sequence *the opinion of the* (*investigator*) refers to the formal aspect of conduct of the clinical trial. Finally, the most frequent phraseologies in the CATs refer to indicators and measurements, e.g. *the value of the* (*parameter*), *the half-life of the* (*drug*), *the size of the* (*dose*), *the concentration of the* (*drug*/*agonist*/*ion*/*substances*), or to research on medicines, e.g. *the development of the* (*drug*/*science*/*concept*). To sum up, the abstract syntactic frame *the * of the* performs the whole variety of referential functions when used across pharmaceutical text types; in most cases the specific functions differ from each other while on some occasions, notably when the same variants overlap in two or more text types, the discourse functions are similar, e.g. *the end of the* is predominantly used as a temporal marker.

Next, a single PF found in three pharmaceutical text types was revealed, namely a prepositional phrase with a post-modifier fragment *of the * of* (Table 9). As can be seen, the most frequent variants in PILs refer to locations by specifying places where medicines should be kept, e.g. (*out*) *of the reach of* (*children*), or parts of the human body affected by medicines, e.g. *of the lining of* (*the womb*), (*redness*/*swelling*/*thinning*) *of the tissues of the* (*eye/vagina*). Other frequent variants (*of the ingredients of*, *of the group of*) refer to various aspects of the use and administration as well as composition of medicines, such as allergies, ingredients, classifications of medicines, e.g. (*allergy*/*allergic to any*) *of the ingredients of* + 'medicine's trade name'; *of the group of* (*medicines + called antibiotics*/*nitrates*/*anticonvulsants*). In the CTPs, the most frequent variants are either temporal markers, e.g. *of the duration of* (*the trial*), *of the end of* (*the trial*), or refer to the activity and effectiveness of medicines, e.g. *of the effect/efficacy of* + 'chemical substance/therapy type', *of the combination of* + 'chemical substances'. Finally, in the CATs the most frequent phraseologies refer to measurements, e.g. *of the volume of* (*distribution*), *of the concentration of* (*the drug*), to the activity of medicines in the human body, e.g. *of the effect of* (*plants*/*drugs*), *of the presence of* + 'chemical substance', or to pharmaceutical research in general, e.g. *of the development of* (*the science of pharmacy*).

Finally, the analysis revealed three PFs overlapping in two pharmaceutical text types, namely *it is * to, on the * of, in the * of* (Table 9). The first of them is a construction starting with anticipatory *it* followed by either a verb phrase or adjectival phrase (*it is * to*). In both PILs and CATs, this PF is used to emphasize the importance of the contents of the following proposition. More specifically, in PILs it helps one raise awareness of patients about the importance of following doctors' instructions and manufacturers' guidelines on how and when to properly use medicines. The most frequent variant in the PILs, namely *it is important to*, is typically used in the sentence-initial position to express stance and it is frequently followed by such action verbs as *follow + your doctor's instructions*, *keep/stick + to the dose on the label*, *read + this leaflet carefully, take + your medicine*, *tell + your doctor*; the same goes for other variants of the said PFs, namely *it is best to* (*consult + your doctor*; *take + your tablet at the same time each day*), *it is essential to* (*follow + your doctor's advice/take + medical advice*), *it is advisable to* (*take + other measures/the first tablet*). The second most frequent variant, namely *it is used to* (*delay*/*help*/*prevent*/*treat* + 'medical condition') differs in that it does not express stance; on the contrary, it refers to the medicines' indications. In the CATs, the most frequent textual variants of the said PF express attitudes, opinions or evaluations concerning the whole variety of procedures associated with the use and administration of medicines to accomplish a desired pharmaceutical effect, e.g. *it is possible to* (*decrease + the dose, measure + the patient's unique pharmacokinetic parameters, determine + a pattern of drug accumulation*), *it is important to* (*understand*/*appreciate*/*remember*/*verify*), *it is difficult to* (*accomplish*/*control*/*predict)*, *it is necessary to* (*administer*/*multiply*/*know*), *it is easy to* (*compute*/*see*). In other words, these specific sequences of words signal various treatment possibilities, emphasize important characteristics of medicines and/or chemical substances, or describe recommended course of action.

The prepositional phrase with a post-modifier fragment *on the * of* was found amid the 20 most frequent PF in the SPCs and ATs (Table 9). In the former text type, the most frequent variants perform referential functions by specifying measurements or tendencies, e.g. *on the rate of* (*increase*), locations in the human body, e.g. *on the surface of* (*cells*), or characteristics of medicines associated with their activity in the human body, e.g. (*data*/*information*) *on the use of* + 'medicine's trade name', (*effect*) *on the pharmacokinetics of* + 'chemical substance's name'. On the other hand, the most frequent n-gram in the CATs is a text-oriented discourse organizing item *on the basis of* (*risk*/*weight*/*body surface area*). More specifically, it performs the function of a framing signal that situates arguments by specifying limiting conditions (Hyland, 2008, p. 14). The remaining variants of the PF *on the * of*, such as *on the route of* (*administration*), *on the order of* + 'quantity specification', *on the part of* (*the patient*/*the pharmacist*/*the drug industry*), *on the development of* + 'chemical substance', perform referential functions by specifying locations, quantities, participants or by generally referring to pharmaceutical research.

The last overlapping PF found in two text types (the SPCs and ATs) is another prepositional phrase with a post-modifier fragment *in the * of* (Table 9). It was revealed that in the former text type the variants are primarily framing signals, in the understanding of Hyland (2008, p. 14), which specify limiting conditions imposed on the use and administration of medicines. These conditions refer to the occurrence of particular medical

conditions, characteristics of patients, access to relevant data, the presence or increase/decrease of particular chemical substances in the human body, e.g. *in the treatment of* + 'medical condition' (*arthritis*/*type 2 diabetes*) or 'patient type' (*adolescents*/*children*/*pregnant women*), *in the absence of* (*in/compatibility studies*/*histopathological changes*/*interaction studies)*, *in the event of* (*overdose*), *in the presence of* + 'medical condition', *in the incidence of* 'medical condition'. It is similar in the CATs where three variants act as framing signals specifying conditions limiting the use of particular medicines, e.g. *in the treatment* of + 'medical condition', *in the presence of* + 'chemical substance', *in the case of* + 'medical condition' / 'chemical substance' / 'medicine type'. Other frequent variants, such as (*increase*/*decrease*) *in the value of* (*the parameter*), (*result*) *in the formation of* + 'chemical substance', are primarily referential and are used to characterize the activity or specific effects of medicines in the human body.

All in all, in most cases analyzed the most frequent PFs used in multiple pharmaceutical text types vary in form and specific discourse functions. Furthermore, in most cases the discourse-organizing PFs perform referential functions when actualized in pharmaceutical texts. However, it was also reported that when actualized in texts some of the variants of discourse-organizing PFs (e.g. *on the * of*, *in the * of*) preserve the text-oriented discourse-organizing function (e.g. *on the basis of*, *in the case of*, *in the event of*).

4. Conclusions

This study was an attempt at developing a comprehensive corpus linguistic description of phraseological items, that is 4-word phrase frames (PFs), found in the pharmaceutical domain of language use. Using quantitative and qualitative methods offered by corpus linguistics, selected samples of four English pharmaceutical text types were described in terms of the use, distribution, composition and discourse functions of recurrent PFs. The paper aimed to provide answers to the following research questions:

1. Which PFs are the most frequent and pervasive in pharmaceutical text types?
2. Is there a correlation between the frequency of PFs and the degree of their pattern variability?
3. What are the differences with respect to the structure and discourse functions of the most frequent PFs across pharmaceutical text types?
4. Is it legitimate to treat the discourse functions of PFs as distinct from the ones of their textual variants?

The results showed that the patterns of use, composition and discourse functions of the 50 most frequent PFs vary across patient information leaflets, summaries of product characteristics, clinical trial protocols and chapters/sections from academic textbooks on pharmacology. This finding confirmed the existence of considerable intra-disciplinary register variation and showed that English pharmaceutical discourse is far from homogenous phraseologically, although all the text types explored in this study deal with drugs or medicines. Described in greater detail in the empirical section, these differences

are generally related to situational contexts, functions and target users of the pharmaceutical text types as well as to the varying degrees of conventionalization of their generic structure. The correlation between the frequency of occurrence of PFs and their pattern variability was found to be moderate only in the case of the 300 most frequent PFs in the CATs; as for the remaining text types, the correlation was found to be either negligible or not statistically significant. To establish this, the Spearman's rank correlation test was used. The structural analysis revealed that in PILs, SPCs and CTPs phraseology is primarily lexical and fixed while in the CATs it is largely grammatical and highly variable. The functional analysis showed that the SPCs and CTPs have the highest number of topic PFs that are content-related, while in the PILs and, in particular, in the CATs the most dominant group are discourse-organizing PFs. A high number of action-oriented PFs in the PILs and SPCs results from the fact that both text types describe methods of recommended use and administration of medicines. Finally, the results showed that by developing a domain-specific functional taxonomy of PFs, it is possible to treat the discourse functions of PFs as distinct from those of their textual variants; in fact, a number of items qualitatively examined in the paper showed that the discourse functions of PFs were found to be different from the ones of their textual variants, that is, the n-grams consisting of four words. This finding is primarily due to the fact that PFs and n-grams are two distinct conceptualizations of phraseologies in texts, the former being abstract generalizations of the latter, and that is why they may fulfil different roles in the creation of professional discourses, including the pharmaceutical one explored in this paper. All in all, the study revealed a number of findings that may be pedagogically useful for both ESP teachers and practitioners in the pharmaceutical field (notably for non-native speakers of English) as well as for researchers exploring the use, distribution and functions of PFs across various text types or genres.

The methodology used in this study could be further extended by ranking the text types from the most to the least phraseologically varied or by identifying those PFs that contribute the most to the ranking. These problems are further explored by Forsyth and Grabowski (2014). Also, it is possible to generalize the discourse functions of the PFs by applying the domain-specific functional taxonomy presented in this paper to a random sample of PFs only (e.g. by exploring 5% of the PFs in a text variety selected through systematic or random sampling) and extrapolating the results to the total population of the PFs in each text type. This way the selection of PFs would be more representative of the entire corpus of each text type rather than limited to the most frequent linguistic items.

Finally, as can be seen from this study, the identification and analysis of the most frequent PFs may also become a starting point for more detailed qualitative studies of specific linguistic patterns, also including association patterns with specialist terminology that could be regularly tied to specific PFs in a given register, the hypothesis to be tested empirically in the future. Another future challenge, notably from a methodological point, may pertain to cross-linguistic research on PFs found in text types or genres produced in typologically different languages; for example, Granger (2014), who conducted a preliminary study of lexical bundles found in English and French parliamentary debates and newspaper editorials, showed that the results can be highly informative for foreign language teachers and translators. As this study was intended as primarily descriptive and exploratory, the main challenge ahead is therefore to turn the results into useful and

actionable knowledge for researchers in phraseology, practitioners in the pharmaceutical field, teachers of ESP or translators.

Acknowledgements

This research was financed by the National Science Centre (*Narodowe Centrum Nauki*) pursuant to a decision no. DEC-2013/09/D/HS2/00543.

References

Bauer, L. (2008). *Applied Clinical Pharmacokinetics*. (2nd ed.). New York: McGraw-Hill Medical.

Biber, D. (2006). *University Language. A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

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

Biber, D., & Conrad S. (2009). *Register, genre and style*. Cambridge: Cambridge University Press.

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

Bouayad-Agha, N. (2006). The Patient Information Leaflet (PIL) corpus. <http://mcs.open.ac.uk/nlg/old_projects/pills/corpus/PIL/> (12 April 2012).

Craig, Ch., & Stitzel, R. (Eds.). (2004). *Modern Pharmacology with Clinical Applications* (6th ed.). Lippincott: Williams & Wilkins.

Fitzpatrick, S. (2005). *The Clinical Trial Protocol*. Marlow: The Institute of Clinical Research.

Fletcher, W. (2002–2007). *KfNgram*. Annapolis: USNA. <http://www.kwicfinder.com/kfNgram/kfNgramHelp.html> (20 November, 2011)

Fletcher, W. (2010). Phrases in English. <http://phrasesinenglish.org/> (20 September, 2014)

Forchini, P., & Murphy, A. (2008). N-grams in comparable specialized corpora. Perspectives on phraseology, translation and pedagogy. *International Journal of Corpus Linguistics*, *13*(3), 351–367.

Forsyth, R., & Grabowski, Ł. (2014). "Is there a formula for formulaic language?". Paper presented at 6th Formulaic Language Research Network Conference. Swansea, UK, 14-16 July 2014. < http://flarn.viviennerogers.info/wp-uploads/2014/02/Forsyth.pdf> (6 August 2015) [paper under review].

Fuster-Marquez, M. (2014). Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction*, *7*(1), 84–121.

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

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

Granger, S. (2014). A lexical bundle approach to comparing languages. Stems in English and French. In M-A. Lefer, & S. Vogeleer (Eds.), *Genre- and register-related discourse features in contrast*. Special issue of *Languages in Contrast*, *14*(1), 58–72.

Gray, B., & Biber D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics*, *18*(1), 109–135.

Hollinger, M. (2003). *Introduction to Pharmacology* (2nd ed.). London/New York: Taylor & Francis.

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

Koester, A. (2006). *Investigating Workplace Discourse*. London: Routledge.

Martinez, R., & Schmitt N. (2012). A Phrasal Expressions List. *Applied Linguistics*, *33*(3), 299–320.

Montalt Resurreccio, V., & Gonzalez Davies, M. (2007). *Medical Translation Step by Step. Translation Practices explained*. Manchester: St. Jerome Publishing.

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

Pęzik, P. (2013). Wybrane aspekty reprezentatywności małych i średnich korpusów. In W. Chlebda (Ed.), *Na tropach korpusów. W poszukiwaniu optymalnych zbiorów tekstów* (pp. 45–58). Opole: Wydawnictwo Uniwersytetu Opolskiego.

Römer, U. (2009). English in Academia: Does Nativeness Matter? *Anglistik: International Journal of English Studies*, *20*(2). 89–100.

Römer, U. (2010). Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews. *English Text Construction*, *3*(1), 95–119.

Rowntree, D. (2000). *Statistics Without Tears*. London: Penguin Books.

Scott, M. (2008). *WordSmith Tools 5.0*. Liverpool: Lexical Analysis Software.

Simpson-Vlach, R., & Ellis, N. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, *31*(4), 487–512.

Stangroom, J. (2014). *Social Science Statistics*. <http://www.socscistatistics.com/tests/spearman> (10 August, 2014)

Stubbs, M. (2007). Quantitative data on multi-word sequences in English: the case of the word 'world'. In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, Discourse and Corpora* (pp. 163–190). London: Continuum.

The European Clinical Trials Register. <https://www.clinicaltrialsregister.eu/index.html> (2 February, 2012)

Tiedemann, J. (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, & R. Mitkov (Eds.), *Recent Advances in Natural Language Processing, 5* (pp. 237–248). Amsterdam: John Benjamins.

Wang, D., & Bakhai, A. (Eds.). (2006). *Clinical Trials: A Practical Guide to Design, Analysis, and Reporting*. London/Chicago: Remedica.

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

Appendix A

The top-50 (by frequency) 4-word PFs with a variable slot in the middle across pharmaceutical text types, and in the British National Corpus (BNC) extracted from "Phrases in English" website (Fletcher, 2010)*

PILs	SPCs	CTPs	CATs	BNC
*if you * any*	*the * should be*	*be * in the*	*the * of the*	*the * of the*
*the * of the*	*in the * of*	*to be * in*	*in the * of*	*in the * of*
*you * to take*	*on the * of*	*the * of the*	*the * of a*	*the * of a*
*if you * to*	*the * of the*	*be used * the*	*for the * of*	*of the * of*
*it is * to*	*once every * weeks*	*therapy * product no*	*to the * of*	*to the * of*
*the * of your*	*should be * by*	*of the * to*	*of * in the*	*at the * of*
*do not * the*	*the dose * be*	*women of * potential*	*the * of drug*	*on the * of*
*if you * not*	*should be * in*	*used in * trial*	*is the * of*	*and the * of*
*to * your medicine*	*should be * to*	*of * for this*	*of the * of*	*for the * of*
*of the * of*	*patients with * renal*	*of the * of*	*can be * to*	*a * of the*
*to take * medicine*	*should be * with*	*information * the trial*	*and the * of*	*with the * of*
*if you * a*	*with * medicinal products*	*objective of * trial*	*the * of distribution*	*of the * and*
*you * any of*	*ability to * and*	*in * clinical trial*	*the * rate constant*	*by the * of*
*your * tells you*	*with other * products*	*the * to be*	*of drug * the*	*the * and the*
*you should * your*	*be * with caution*	*the trial * a*	*in the * and*	*the * in the*
*as * as possible*	*of * should be*	*the * has a*	*on the * of*	*it is * to*
*in a * place*	*to drive * use*	*the * has been*	*of the * is*	*in the * and*
*if you * taking*	*drive * use machines*	*medical * information not*	*steady-state peak * trough*	*as a * of*
*before you * to*	*date of * authorisation*	*in the * has*	*it is * to*	*that the * of*
*you are * to*	*an increase * of*	*in this * as*	*of the * and*	*and * of the*
*you * any other*	*dose * should be*	*in the * trial*	*a * of the*	*from the * of*
*please * this leaflet*	*to the * of*	*used * the trial*	*it is * that*	*the * of his*
*in the * of*	*haemoglobin * greater than*	*this imp * use*	*can be * by*	*it is * that*
*as * as you*	*four weeks * the*	*end of * trial*	*the * of drugs*	*is the * of*
*the * of this*	*for the * of*	*the last * of*	*of the * in*	*a * in the*
*group of * called*	*the haemoglobin * be*	*visit * the last*	*used to * the*	*the end * the*
*a * of medicines*	*once every * week*	*of the * and*	*the * and the*	*the * of this*
*what you * know*	*have been * in*	*last * of the*	*such as * and*	*in a * of*
*before taking * medicine*	*to * the dose*	*last visit * the*	*and * is the*	*the * to the*
*you may * to*	*has been * in*	*trial * end of*	*with the * of*	*of * in the*
*any * or are*	*should be * if*	*of the * subject*	*is * to the*	*of the * in*
*at the * time*	*the * of anaemia*	*the * visit of*	*where * is the*	*with a * of*
*on the * of*	*and * tissue disorders*	*other trial * description*	*drug * the blood*	*is * to be*
*before * your medicine*	*should * be used*	*last * undergoing the*	*is * by the*	*the * that the*
*never give * to*	*one * two weeks*	*last subject * the*	*as a * of*	*is a * of*
*this * is for*	*with * renal failure*	*the last * undergoing*	*a * in the*	*the * for the*

*read this * carefully*	*in the * group*	*subject * the trial*	*a * dose of*	*to be * in*
*to * of these*	*may be * to*	*for the * trial*	*be * to the*	*to the * and*
*do * suffer from*	*the haemoglobin * to*	*by the * for*	*is a * of*	*the * on the*
*what * in your*	*the * half life*	*investigator * to be*	*can be * in*	*the * of an*
*you * know about*	*is * in patients*	*secondary * of the*	*if the * is*	*at the * time*
*never * it to*	*used with * in*	*secondary objectives * the*	*is * in the*	*be * in the*
*forget to * a*	*of the * of*	*main * of the*	*and * of the*	*the * of their*
*if * are not*	*should be * for*	*of * potential not*	*the drug * the*	*was * by the*
*read * leaflet carefully*	*as a * injection*	*childbearing potential * using*	*that the * of*	*about the * of*
*you are * or*	*the * in haemoglobin*	*objectives of * trial*	*drug * in the*	*to * in the*
*to * of the*	*if the * in*	*subjects * be included*	*the * of an*	*and a * of*
*and * your doctor*	*at the * of*	*is * the last*	*for * treatment of*	*the first * of*
*to take * tablets*	*should be * as*	*is * of a*	*by the * of*	*it was * that*
*out * the reach*	*no * adjustment is*	*other * products*	*of the * the*	*to * to the*

***** The "Phrases in English" website (Fletcher, 2010) is available at: http://phrasesinenglish.org/explorep.html (accessed September 2014).