COMPARING FORMULAICITY OF LEARNER WRITING THROUGH PHRASE-FRAMES: A CORPUS-DRIVEN STUDY OF LITHUANIAN AND POLISH EFL STUDENT WRITING

RITA JUKNEVIČIENĖ
Vilnius University, Lithuania
rita.jukneviciene@flf.vu.lt

ŁUKASZ GRABOWSKI
University of Opole, Poland
lukasz@uni.opole.pl

Abstract
Learner corpus research continues to provide evidence of how formulaic language is (mis)used by learners of English as a foreign language (EFL). This paper deals with less investigated multi-word units in EFL contexts, namely, *phrase-frames* (Fletcher 2002–2007), i.e. sets of n-grams identical except for one word (*it is * to*, *in the * of*). The study compares Lithuanian and Polish learner writing in English in terms of phrase-frames and contrasts them with native speakers. The analysis shows that certain differences between Lithuanian and Polish learners result from transfer from their native languages, yet both groups of learners share many common features. Most importantly, the phrase-frame approach highlights structural peculiarities of learner writing which are otherwise difficult to capture.

Keywords: EFL writing, learner corpus, Lithuanian EFL learners, phrase-frame, Polish EFL learners

1. Introduction

The rapid development of learner corpora continues to give impetus to lexical studies of learner language. Insights from lexical grammar on the one hand and the possibility of automated data extraction from corpora on the other have given rise to a number of studies of L2 learners' phraseological competence, which is broadly understood as their ability to use different formulaic sequences (Wray 2000: 465; Wray 2002: 9). Following the first publications of phraseological evidence in L2 language use (Pawley and Syder 1983; Kjellmer 1991), many studies have been undertaken to investigate the use of diverse multi-word combinations in learner corpora, for example, collocations (Nesselhauf 2005; Martelli 2006; Fan 2009), phrasal verbs (Waibel 2007), lexical bundles, also termed n-grams or recurrent word sequences (De Cock 2004; Chen and Baker

2010). This article deals with one of the least investigated multi-word unit in learner English so far, namely, a phrase-frame, first described by Fletcher (2002–2007). Identified using a bottom-up corpus-driven methodology, a phrase-frame is a set of variants of n-grams of any length identical except for one word, for example, *is the * of, it is * to*, *a part of **.[1] Phrase-frames (henceforth – PFs) constitute a theoretical concept designed to capture phraseological patterns in texts and in this respect they may be particularly interesting in learner language studies. Similarly to lexical bundles, PFs are automatically extracted from a corpus. Yet while lexical bundles offer a rather diverse lexical profile of recurrent word combinations which can be submitted to structural and functional analyses, the latter involving quite a few subjective and arguable choices for the researcher (cf. Ädel and Erman 2012: 89–90), PFs reveal a generalised picture of patterns in a corpus, which is especially valuable for a more holistic approach to the structural analysis of different language varieties, learner languages in particular.

In learner corpus research, the study of recurrent lexical combinations, PFs being one of them, usually follows one of the three research designs aimed at contrastive analyses of learner language varieties. First, such studies may be focused on one chosen EFL learner group vs. data from a comparable corpus of native speaker English (e.g. Ädel and Erman 2012; Baumgarten 2014; Chen and Baker 2010; De Cock 2004; Jalali 2013; Juknevičienė 2009; Kizil and Kilimci 2014). The second group of studies involves investigation of longitudinal or pseudo-longitudinal data representing learners at different proficiency levels (Hyland 2008a; Römer 2009; Vidakovic and Barker 2010; Juknevičienė 2013; Leńko-Szymańska 2014). Finally, the third research design is a contrastive analysis of data representing learners whose mother tongues are different (e.g. Paquot 2013; Paquot 2014; Wang 2016). Such studies usually offer an opportunity to highlight L1-specific features of the learner language varieties under study. In this respect, studies by Paquot (2013; 2014) present a significant contribution to the investigation of L1 transfer using learner corpora, most of all, owing to their methodology. It is this last research strand that the present study belongs to.

It has been only recently that PFs have become a unit of analysis in phraseological research. More specifically, PFs were explored in terms of their use and discourse functions in different registers and specialist domains (e.g. Stubbs 2007; Römer 2010; Gray and Biber 2013; Fuster-Marquez 2014; Grabowski 2015). These studies have shown that PFs may provide valuable insights into how fixed multi-word units are used in a given register and what degree of variation they exhibit (Römer 2009; 2010). Forsyth and Grabowski (2015) showed that PFs may be used not only for generalizing phraseologies in texts, but also for measuring the degree of formulaicity in language which allows

[1] On the surface, PFs bear resemblence to collocational frameworks described by Renouf and Sinclair (1991). However, the latter multi-word items are identified in a top-down corpus-based way, which means in practice that they are pre-selected by the researchers rather than automatically extracted from a corpus.

researchers to rank texts or corpora from the most to the least formulaic and, by implication, from the least to the most phraseologically varied.

PFs have been also explored in the context of English as a foreign language (EFL). For example, Römer (2009) found, first, that native and non-native students (whose L1 was German) of English often use the same PFs (with three or four words with a variable slot in the initial, medial and final position) yet with varying frequencies; second, that the students to a large extent share the slot-fillers used in the PFs; and, third, that much variation across PFs is content-related. Also, Römer (2009) found a number of PFs that occur in academic papers and yet they do not occur in native and non-native student writing, the finding that underscores the differences between expert and novice/learner language. In another study, using Michigan Corpus of Upper-level Student Papers (MICUSP), Römer and O'Donnell (2009) focused on positional variation of PFs (with 3–5 words with a variable slot in the medial position only) in native and non-native proficient academic writing, and they found that certain PFs have a strong preference for specific positions within sentences, paragraphs and texts as a whole (e.g. *it is * that* typically occurs in sentence-initial position as well as in text-final position); also, Römer and O'Donnell (2009) suggest that more research be conducted in the future on comparing student writing with expert academic writing (e.g. published research articles representing various disciplines). PFs have been also used as a unit of analysis in research on development of formulaic sequences in L1 and L2 student academic writing. For example, O'Donnell, Römer and Ellis (2013) compared the use of PFs (consisting of 3–5 words) in undergraduate native students essays collected in the LOCNESS corpus, undergraduate student writing produced by learners with eleven different L1s (sub-corpora extracted from the ICLE corpus), more advanced native and non-native student writing representing a variety of academic disciplines and collected in the MICUSP corpus as well as a corpus of expert academic writing (Hyland 1998). The said study revealed that although more advanced writers used more PFs than lower-proficiency writers (LOCNESS and ICLE), no significant effects were found of the level of language competence or native vs. non-native speaker status (O'Donnell, Römer and Ellis 2013). More importantly, the results of this study suggested that the variants of PFs should be analysed manually as otherwise no insights into their semantics or discourse functions are to be gained, and it is those functions which may help one distinguish between less and more advanced writers. In a more recent study (Garner 2016), focused on the exploration of PFs in learner language (L1 German learners of English) across five proficiency levels. The study revealed that PFs used by more proficient students exhibit a higher degree of variability and are more complex in terms of their discourse functions. An overview of PFs in EFL contexts shows that no research has been conducted so far on the comparison of the use of PFs by L2 learners with different L1 backgrounds, in particular with a focus on L1-transfer effects.

That is why the present research project was conceived as an exploratory corpus-driven analysis (Sinclair 2004; Tognini-Bonelli 2001: 65) to investigate the phraseological competence of Lithuanian and Polish learners of English, speakers of two different first languages hardly ever contrasted before for their EFL competence. More specifically, we will try to answer the following research questions: (1) Are Lithuanian and Polish learners similar or rather different in terms of the use of frequent PFs and how similar are they to native speakers?; (2) What are the structural properties of PFs extracted from L2 English?

It should be noted that both Lithuanian and Polish are morphologically inflected languages with free word order in the sentence and hence typologically very different from English. As for the genetic typology, Lithuanian is a Baltic language and Polish a West-Slavic one, which makes the acquisition of L2 English, a West-Germanic language, obviously challenging. The two countries, Lithuania and Poland, exist in geographical proximity, and historically they have had periods of common history. The languages, however, apart from individual lexical cognates and rich albeit different morphology and syntax, have little in common and are mutually incomprehensible. Due to historical circumstances, the traditions of teaching EFL in both countries are fairly similar, which is another reason to compare learner English coming from the same geographical region. Since PFs have been rather underexplored in EFL research, our study is also an opportunity to test suitability of PFs as a unit of analysis in research on recurrent patterns in learner English, notably targeted at identification of L1-transfer effects.

2. Materials and methods

This study was designed as a contrastive interlanguage analysis (Granger 1996) aimed at highlighting L1 specific features characteristic of Lithuanian and Polish learners. To analyse their written English, we used two components of the ICLE corpus (International Corpus of Learner English): a subcorpus of Polish learner English (henceforth – PICLE) from the 2nd version of ICLE (Granger et al. 2009) and a corpus of Lithuanian learner English (henceforth – LICLE, Grigaliūnienė and Juknevičienė 2012), which is a new addition to the currently developed version of ICLE. LICLE and PICLE represent written English of advanced EFL learners who are senior undergraduate students at universities in Lithuania and Poland majoring in linguistics-based study programmes and whose first languages are Lithuanian and Polish, respectively. As a reference corpus, we used the Louvain Corpus of Native English Essays (LOCNESS, CECL 1998) consisting of argumentative and literary essays written by British and American students (excluding A-levels examination essays). Table 1 describes the corpora under scrutiny.

Table 1. Corpora used in the study

Corpus	Number of essays	Size (in words)
LICLE	335	191,570
PICLE	365	234,702
LOCNESS	298	262,339

The study was conducted in several stages. Firstly, frequency lists of PFs were generated using the kfNgram software (Fletcher 2002–2007) which retrieves four-word PFs with one variable slot, for example, *the * of the*. The variable slot may be realized in the corpora as *the* <u>beginning</u> *of the*, *the* <u>end</u> *of the*, *the* <u>importance</u> *of the* etc. To be included in the analysis, a PF had to have at least three realisations in the corpus, each with the minimum absolute frequency of 3. This decision was taken after we observed that the kfNgram software is not sensitive to capitalized letters and returned, for instance, ** of the most* (23 occurrences, 2 variants, LOCNESS) with the two realizations *one of the most* (12 occurrences) and *One of the most* (11 occurrences), which is of little value to the study. Furthermore, although the kfNgram program can generate PFs of varying lengths, in this study we decided to focus on four-word items which in the studies of recurrent sequences have been shown to be of the optimal length (cf. Hyland 2008b: 8; Chen and Baker 2010: 32) as they have a more readily recognizable range of structures and functions than the shorter sequences and are more frequent than the longer ones.

The next stage was related to the selection of PFs with respect to the position of the variable slot. In earlier studies, e.g. Römer and O'Donnell (2009) or Römer (2010), the decision was made to leave out PFs with variable slots in either the initial or final position (*BCD and ABC*) as they are often fragments of longer PFs and/or contain empty slots filled with function words. Function words, however, could be particularly important for this study because both Lithuanian and Polish learners have many difficulties in mastering the English articles and prepositions. Hence, we decided to include PFs with variable slots in any position.

Lastly, the frequency cut-off point was set at nine occurrences in LICLE and PICLE and ten in LOCNESS, which roughly corresponds to normalized frequency of 40–45 occurrences per million words. To avoid idiosyncratic effects, we also checked the dispersion of the least frequent PFs which was done by generating concordances of individual items using WordSmith Tools (Scott 2008, version 5). On average, a PF with the absolute frequency of nine occurrences has its textual variants in at least four different texts, which we considered to be an acceptable dispersion level. The statistical analysis program R was used to run statistical tests (R Core Team 2015).

The final stage of data selection involved manual revision of PFs in order to remove topic-specific items which, as demonstrated in earlier studies (e.g. Paquot 2013; 2014), can considerably distort the results, especially when the data is retrieved from a specialised learner corpus. For this purpose, all PFs which could be linked to essay prompts were carefully checked. We considered a PF to be topic-specific if it included a lexical word from the essay prompt. Moreover, if a PF was realized as a sequence identical to a particular segment of the essay prompt on which the essay in question was written, it was excluded from further analysis. The resulting datasets are presented in Table 2 whereas a complete list of topic-specific PFs excluded from the analysis is given in Appendix 1.

Table 2. The number of topic-specific phrase-frames

	LICLE	PICLE	LOCNESS
Primary list of phrase-frames	149	163	98
Topic-specific PFs	27 (18%)	24 (15%)	26 (27%)
Topic-neutral PFs	122 (82%)	139 (85%)	72 (73%)

The relative frequency of topic-specific PFs in both learner corpora is considerably lower than in the native-speaker material. The smaller density of topic-specific PFs seems to indicate that argumentative texts written by non-native learners in comparison with those in LOCNESS lack at least one textual feature, namely, density of topical lexis, which is one of the lexical means to create cohesion (Halliday and Hasan 1976). This finding confirms observations reported in Juknevičienė (2009) which dealt with lexical bundles in learner English and found that less proficient learners underuse topic-related lexical bundles in comparison to more advanced EFL learners and native speakers. Similarly, Ädel and Erman (2012: 84) reported that "topic- and discipline-specific" lexical bundles were more numerous in their native-speaker material than in the non-native data. It is also interesting to observe that although EFL students usually make their best to exploit the lexis of essay prompts which naturally lend themselves to lifting, our findings confirm the results reported in Ädel and Erman (2012). Thus, it seems valid to assume that topic-specific lexis is indeed exploited the most in LOCNESS as compared with LICLE and PICLE. Since our study is targeted at learners' vocabulary rather than their discourse competence, topic-specific PFs were eliminated from the further analysis.

Hence, PFs will henceforth refer to four-word items with a gap in any position which meet the aforementioned frequency criterion, have at least three textual realisations and do not contain topic-specific lexical words. A full list of PFs is provided in Appendix 2.

3. Results and discussion

In the following, we report our findings starting with an overview of shared and corpus-specific PFs in LICLE and PICLE and the extent of overlap between each of these two corpora with the LOCNESS data. This will enable us to check whether Lithuanian and Polish learners are similar or rather different in terms of the use of frequent PFs (research question 1). The second part of the analysis deals with the morphological properties of PFs. Firstly, we discuss PFs in terms of constituent words part-of-speech features. Secondly, we consider the gapped slots of PFs and their fillers in order to establish which lexical words are prone for frame-building in learner language. This stage of the study was undertaken to reveal the structural properties of PFs extracted from L2 English produced by Lithuanian and Polish students (research question 2).

3.1. Shared and corpus-specific phrase-frames

To answer the first research question we looked into shared and corpus-specific PFs. The analysis of the data showed that the three corpora have 33 identical PFs (Table 3). If the degree of overlap between each of the non-native speaker corpus and LOCNESS is considered, the results are similar for both groups of EFL learners: LICLE and LOCNESS share 20% of PFs whereas PICLE and LOCNESS have 19% identical PFs. In this respect, our results are similar to the ones obtained by Ädel and Erman (2012: 85) who explored lexical bundles and found that 22% of these multi-word units were shared by native speakers and advanced EFL learners (Swedish L1). The degree of overlap in our data, however, is slightly lower which might be related to the general lower level of proficiency in English of Lithuanian and Polish learners on the one hand and the type of items, viz. PFs rather than lexical bundles, on the other hand. Furthermore, the proportions of shared PFs reveal yet another interesting peculiarity of EFL learner writing. While in the case of LOCNESS the shared PFs account for the largest part of all PFs in this corpus (46 %), in the two learner corpora the 33 common PFs represent 27% of all PFs in LICLE and 24% in PICLE. Bearing in mind the fact that the primary data selection procedure involved a rather stringent removal of topic-specific items, it was not expected to find that the shared PFs account for less than one third of PFs in the learner corpora.

It was also found that LICLE and PICLE share between them quite many PFs; more specifically, 28% of PFs retrieved from LICLE and 24% from PICLE are identical. Moreover, if we add to this number PFs shared by all three corpora, the proportion of shared PFs between LICLE and PICLE is even greater and it certainly outnumbers those PFs that each of the learner corpora has in common with the native-speaker data. Bearing in mind the fact that LOCNESS represents a target language variety to advanced EFL learners, the picture is not very promising since both LICLE and PICLE seem to have less in common with LOCNESS than they have between themselves. This early observation was corroborated by further analysis.

As shown in Table 3, both LICLE and PICLE have a considerable number of corpus-specific PFs which only appear in one of the two corpora. While the greatest proportion (46%) of PFs in LOCNESS, as mentioned above, belongs to the category of items established in all corpora, in the case of EFL learners, the greatest proportion is represented by the category 'corpus-specific.' It should be noted, however, that this data refers only to those items which meet the definition of PFs applied in this study; admittedly, some PFs were not included in our dataset even though they do appear in the corpora, albeit with lower frequencies. For instance, *as well as* * has only two realizations in LOCNESS (*as well as a* and *as well as the*) and was not included in the analysis.

Table 3. Proportions of shared and corpus-specific phrase-frames

Corpora	Number of shared / corpus-specific PFs	Percentages in respective corpora
LICLE, PICLE, LOCNESS	33	27% of all PFs in LICLE 24% of all PFs in PICLE 46% of all PFs in LOCNESS
LICLE and PICLE	34	28% of all PFs in LICLE 24% of all PFs in PICLE
LICLE and LOCNESS	5	4% of all PFs in LICLE 7% of all PFs in LOCNESS
PICLE and LOCNESS	8	6% of all PFs in PICLE 11% of all PFs in LOCNESS
LICLE	50	41% of all PFs in LICLE
PICLE	64	46% of all PFs in PICLE
LOCNESS	26	36% of all PFs in LOCNESS

One of the unexpected findings is the fact that the two foreign learners' corpora have more shared PFs between them than they have in common with the native-speaker data represented by LOCNESS. Both groups of EFL learners employ a number of PFs which are considerably less frequent or do not appear even once in LOCNESS. A closer examination of the data seems to suggest several explanations for the similarities between LICLE and PICLE. Firstly, owing to geographical proximity and a common cultural and historical past, the Lithuanian and Polish languages share a number of lexical similarities which apparently provide a common linguistic background to L1 Lithuanian and L1 Polish speakers. For example, both languages have equivalents for the English phrase *in this way* * which is a common lexical calque used in both Lithuanian and Polish: Lith. *tokiu būdu* and Pol. *w ten sposób*. The existence of a close equivalent in the learners' mother tongues most probably explains why *in this way* * is significantly overused by our learners in comparison to native speakers, in whose data set this PF does not occur at all. Many more shared PFs between LICLE and PICLE, however, can be accounted for by the fact that our corpora represent inexperienced writers who are still learning to develop argumentative texts. Consequently, in comparison to native speakers they tend to overuse explicit markers of discourse organization (e.g. *as well as *, first of all *, in order to **) and stance markers. For instance, both NNS learner corpora contain such lexical boosters as *more and more *¸ * more and more*, ** the most important, a great * of* etc., which add rather categorical undertones to the texts and which could be considered characteristic of novice writers (cf. Ädel 2006; Burneikaitė 2009). Another developmental feature of learner writing, which is common both to Lithuanian and Polish EFL writers, is a frequent use of gender-neutral references to people, namely, *he or she*, and *they do not* *. Obviously, the learners are demonstrating their awareness of sexist language; in addition, it is also evident that they have not yet internalized the general reference to people, i.e. *one*, which, as a matter of fact, does not exist either in Lithuanian or Polish. Lastly, some shared PFs could be linked to the common topics of the essays in LICLE and PICLE. As explained above, the data selection procedure allowed us to weed out many topic-specific PFs except for those which are not explicitly stated in any of the essay prompts. As a consequence, the topic effect could not be completely ruled out as evidenced by, for example, *the lack of* with a gap preceding or following the sequence. It often used in the essays where the questions of fortune making and (not) having money are dealt with.

The analysis of corpus-specific PFs in LICLE and PICLE was expected to shed more light on L1 transfer and L1-specific patterns. A close examination of corpus-specific PFs allowed us to identify items which could be categorized as specific features of learners sharing a mother tongue. As shown in Table 3, the largest number of items retrieved from both corpora appeared to be corpus-specific PFs, namely, 50 PFs (or 41%) in LICLE and 64 (46%) in PICLE were items not attested in the data set retrieved from the other corpora used in this study. To illustrate how corpus-specific PFs may serve as evidence of L1 influence, a more detailed discussion of two characteristic cases from each NNS corpus will be provided.

LICLE data include a number of PFs with the lexical verb *say*, namely, *say that * is*, ** be said that*, *it * be said.* All of them could be linked to the Lithuanian expression *sakoma, kad* ‘it is said that’ which is a passive form of *sakyti* ‘to say’ followed by a complement *that*-clause. This expression is typical of Lithuanian argumentative discourse where it usually introduces background information or common knowledge. While sequences with the verb *say* also appear in PICLE and LOCNESS, the only one that makes it into our data set is ** said to be* (PICLE, abs. freq. 16). While this frame does appear in LICLE (its absolute frequency of 7 is below our cut-off point), Lithuanian learners, in comparison to Polish, are significantly underusing it (Log Likelihood index 46.40, p <0.0001). Instead, they are intensely exploiting such constructions which are verbatim renderings from their L1. Moreover, all PFs with *say*, with the only exception of the raising construction * *said to be*, are overused by Lithuanian learners in comparison both to PICLE and LOCNESS data. Obviously, this finding points to the inter-L1-group heterogeneity (Jarvis 2000) and could be considered a candidate for the L1-induced constructions.

Similarly, corpus-specific PFs in PICLE also indicate such ways of expression which are overused by Polish learners. An interesting case is *in front of ** which has a few instances of specific use in PICLE. Consider the following examples:

(1) *But seven years ago the brand new world opened in front of Poles.*

(2) *In front of the unifying tendencies, in Europe at least, it would be tempting to think that the cultural boundaries (...)*

The reason for the overuse is clearly L1-induced. More specifically, the Polish preposition *przed* ('in front of', 'before', 'ahead of') can be used to indicate time, place or position with respect to someone or something else or in the presence of someone else, usually important. In the examples above, the intended meaning was to signal challenges facing Poles or Europe. That is why the use of the English preposition *in front of*, typically used to indicate place, shows that Polish learners of English tend to overgeneralize its use. Once again, we have a PF significantly overused in one of the corpora in relation to the other two (PICLE vs. LICLE Log Likelihood +4.21, p<0.05; PICLE vs. LOCNESS Log Likelihood +24.88, p<0.0001), and its idiosyncratic uses in PICLE point to possible transfer from the learners' L1.

Admittedly, not all corpus-specific PFs retrieved from LICLE and PICLE can be linked to a distinct feature of the learners' L1. While a full-scale study of transfer effects, following the methodology proposed by Jarvis (2000) and applied in Paquot's study of lexical bundles (2013; 2014), was beyond the scope of this research, the phrase-frame approach is undoubtedly a promising way forward to identify features of learner language which could be linked to L1 influence.

3.2. Structural analysis of phrase-frames

In the following stage of the study, a structural analysis of PFs was conducted to explore, first, which lexical or function words are prone to appear in PFs and, second, whether the tendencies are similar or different for both learner groups as compared with native speakers. The analysis was two-fold. Firstly, the morphological structure of PFs was taken into account, and they were grouped on the basis of constituent word classes. The second part of the structural analysis dealt with the words which appear in the variable slots of PFs, or, in other words, trigger clustering and, consequently, formation of PFs.

To analyse the morphological structure of PFs, we used the classification proposed by Gray and Biber (2013: 122) who distinguish three types of PFs, namely, (1) verb-based (V-based) PFs with one or more modal, auxiliary or lexical verbs; (2) PFs with content words other than verbs (C-based), and (3) PFs with function words only (F-based). The results of the structural analysis are presented in Table 4.

Table 4. Distribution of phrase-frames across structural categories

Structural categories	LICLE		PICLE		LOCNESS	
	No	%	No	%	No	%
V-based	67	55%	73	52%	30	42%
C-based	37	30%	44	32%	19	26%
F-based	18	**15%**	22	**16%**	23	**32%**
Totals	*122*	*100*	*139*	*100*	*72*	*100*

The proportions of the structural categories in the three corpora are clearly different, although the effect size is small (Cramer's V 0.125). The χ^2 test of independence shows that differences in the frequencies of the structural categories in the three corpora are statistically significant (χ^2 10.3797, df = 4, p = 0.0345). To see which differences are the most important, we computed the residuals. It was found that it is the frequency of F-based PFs in LOCNESS that makes the statistically significant contribution to the χ^2 statistic value at the significance level of 0.05.

The underuse of F-based PFs in LICLE and PICLE in comparison to LOCNESS points to the fact that 'small' function words in the language of EFL learners do not build recurrent frames to the same extent as is the case in LOCNESS. Instead, in LICLE and PICLE patterns formed from lexical words are clearly dominating. One way of explaining it is the fact that non-native learners possess a rather limited repertoire of lexical words which inevitably leads to repetition of known words and familiar constructions and, as a result, yields a greater proportion of PFs incorporating a limited number of repeatedly used lexical words. This tendency was further confirmed by conducting a qualitative analysis of the data.

A closer examination of different structural types of PFs reveals that Lithuanian and Polish learners share a common feature which sets them apart from native speakers. As regards C-based PFs, the data sets from LICLE and PICLE include items which help express stance or act as boosters, for instance, * *more and more*, * *the most important*, *of the most* *, *it is* * *difficult*, *in my opinion* etc. In contrast, the C-based PFs in LOCNESS are mostly referential expressions (* *the end of*, *the rest of* *, * *the use of*, * *part of the*, *the use of* * etc.). So in their essays, Lithuanian and Polish learners resort to a more explicit marking of stance which, as our data shows, distinguishes them from native speakers and could perhaps be viewed as a feature of less experienced writers. The other characteristic feature of non-native learner essays is discourse-organizing frames (e.g. *first of all* *, *as a result* *, *the same time* *, *as well as* *). Interestingly, the only discourse-organizing phrase frame in LOCNESS, which is also attested in LICLE and PICLE, is *in order to* *.

As to V-based PFs, the number of lexical words used in PFs is much larger in non-native English varieties than in LOCNESS. Only three forms of lexical verbs (*seen*, *say*, *continue*) appear in four PFs extracted from this corpus: *can be seen* *, * *seen to be*, * *to say that*, * *will continue to*. In contrast, the LICLE data set includes eleven PFs with the following forms of lexical verbs: *considered*, *think*, *say*, *said*, *sum up*, *want*; lexical verbs in PICLE, interestingly, are not so numerous (*want*, *said*, *take* and *realize*) yet in terms of frequencies both Lithuanian and Polish learners demonstrate a much more intense use of PFs with lexical verbs. Apparently, owing to limited vocabulary they inevitably rely on what could be seen as their 'lexical teddy bears' (Hasselgren 1994).

In an attempt to investigate which words in learner writing trigger the formation of a phrase frame, the second part of the structural analysis was focused on the variable slots. PFs retrieved from the three corpora were analysed in terms of the word class of the slot-fillers. For instance, *the * of the* may be realized by the nouns *end, beginning, majority* whereas *in order to* * is always realized by a verb. In other words, this analysis was undertaken to establish which words have the greatest potential for clustering and pattern building in the language of EFL learners and native speakers. Admittedly, some slots can be filled by different parts-of-speech, e.g. * *the fact that* is realised in LICLE by the verb *is*, preposition *to* and conjunction *and.* Such 'mixed' slots, with very few exceptions, usually occupy the initial or final position (*BCD and ABC*) of the phrase frame and they are often complete three-word formulaic sequences, e.g. * *the fact that*, * *in front of*, * *as a result* *, *in this way* *. There are PFs, however, which are formed around one particular part-of-speech. Five morphological types of slot-fillers were identified, namely, nominal (nouns and pronouns), verbal, adjectival, adverbial and functional (conjunctions, determiners and prepositions). Table 5 below presents distribution of PFs on the basis of the morphological category of the slot-filler.

Table 5. Morphological types of slot-fillers in PFs

Morphological types of slot-fillers	LICLE	PICLE	LOCNESS
Nominal	40 (32%)	45 (32%)	36 (50%)
Verbal	19 (16%)	21 (15%)	10 (14%)
Adjectival	11 (9%)	11 (8%)	2 (3%)

Morphological types of slot-fillers	LICLE	PICLE	LOCNESS
Adverbial	2 (2%)	4 (3%)	-
Functional	9 (7%)	11 (8%)	7 (10%)
Mixed types	41 (34%)	47 (34%)	17 (24%)
TOTAL	122 (100%)	139 (100%)	72 (100%)

The majority of PFs in the three corpora have a variable slot for a noun or pronoun, namely, 32% in LICLE and PICLE and 50% in LOCNESS. Since the corpora under analysis represent written English, this finding is not unexpected as noun phrases are indeed a characteristic of the written discourse (Biber et al. 1999) and thus feature prominently in written learner language. Furthermore, our findings also confirm the results of earlier studies on learner writing which showed that the proportion of noun-based recurrent sequences is directly related to the proficiency of the learners (Juknevičienė 2009; Chen and Baker 2010). Hence, while a half of PFs in the most proficient variety of English in our data, i.e. LOCNESS, contain a variable slot for a noun or pronoun, the proportions in LICLE and PICLE (32% in both) are considerably smaller.

An interesting observation of structural peculiarities of PFs in the NNS data sets refers to the use of function words. A closer examination of PFs with a nominal/pronominal slot-filler offered an explanation why PFs incorporating functional words make a significant difference between native and non-native learners in this study. It turns out that Lithuanian and Polish learners are underusing phrases with the preposition *of* in comparison to native speakers. While PFs with *of* dominate in LOCNESS (26 out of 37, or 70%), their relative frequency is significantly lower in PICLE (22, or 50%) and even more so in LICLE (17, or 42%). Among *of*-frames, those that are formulaic expressions are particularly notable in LOCNESS, e.g. *the case of* *, *the rest of* *, *in favour of* *. Although there are quite many shared PFs among the corpora, their frequencies significantly differ: the normalized frequency per 100,000 words of *the * of the* is 88 in LICLE, 77 in PICLE and 128 in LOCNESS, which shows a significant underuse of *of*-frames by EFL learners. This finding seems to be related to the fact that both Lithuanian and Polish are morphologically inflected languages, where prepositions occupy a very different place in the language system in comparison to English, while the Genitive is expressed by the case category rather than prepositional constructions equivalent to the English *of*-phrases. Undoubtedly, underuse of *of*-frames could be seen as an important feature of learner English produced by Lithuanian and Polish learners.

Another interesting finding is related to such PFs which contain a variable slot for adjective/adverb. As shown in Table 5, EFL learners significantly overuse such PFs in comparison with native speakers whereas the only ones which are found in all three corpora are *it is * to* and *it is * that* yet even those are much more frequent in non-native English varieties. Consider their normalized frequencies per 100,000 words:

	LICLE	PICLE	LOCNESS
*it is * to*	52	48	15
*it is * that*	36	19	6

The frequent use of PFs with adjectival/adverbial slots is most probably related to the overall writing competence of EFL writers rather than any other peculiarities of learner writing. As argued above, expressions of evaluation and stance, in contrast to native speakers, are overused by EFL learners (cf. PFs with such lexical slot fillers as *better, easy, difficult, possible, impossible* etc.).

4. Conclusions

The analysis of PFs in advanced Lithuanian and Polish EFL learner writing was undertaken in order to investigate whether a structural approach involving the study of recurrent PFs in learner corpora might highlight differences between the two groups of EFL learners and, consequently, reveal L1-induced features of written learner English. The answer seems to be twofold. On the one hand, it was found that the largest proportion of PFs retrieved both from LICLE and PICLE are corpus-specific items, not attested in the remaining two corpora used in the study. Yet in order to measure to what extent corpus-specific PFs indeed indicate L1 influence, a more comprehensive study should be undertaken in the future following the framework proposed by Jarvis (2000) and focusing on measuring inter-L1-group heterogeneity in language learners' performance. Such a study may help verify statistically whether PFs explored in our study come from the same or different distribution. The qualitative analysis of selected individual PFs reported in the article seems to suggest that they could serve as a starting point for further investigation of L1 influence in learner English.

On the other hand, the study also revealed a number of shared PFs in Lithuanian and Polish learner writing that are not found in the LOCNESS corpus. These PFs often indicate developmental issues that the two learner groups are facing. Typically, the shared PFs are expressions of stance or text-organizing devices which are often favoured by less proficient learners. In this respect, it would be particularly interesting to consider PFs which are frequent in LOCNESS but underused by EFL learners. Possibly, they might represent a number of features that should be specifically targeted in EFL classrooms for at least two learner groups, i.e. Lithuanian and Polish.

A study like this one, i.e. conducted using basic quantitative methods and involving two corpora of learner language, can only be regarded as a preliminary one. There are many possible ways in which this research may be pursued further in the future. One of the most obvious continuations would be application of the phrase-frame approach to corpora representing texts produced by learners of mother tongues other than Lithuanian and Polish. Next, if PFs indeed prove to be useful in EFL contexts, the natural line of research in the future would be to identify those PFs that carry the most salience to EFL learners of different L1 languages. In fact, similar studies have been already conducted using lexical bundles approach (e.g. Simpson-Vlach and Ellis 2010; Martinez and Schmitt 2012) even though L1 bias was beyond their focus. In this study, we focused on PFs based on contiguous sequences of four words and with a variable slot in any position. However, one may try employing longer or shorter phrase-frames in order to develop more comprehensive descriptions of phraseological patterns in learner language. Finally, bearing in mind specificity of the LOCNESS corpus, it would be possible to verify our findings by using other reference corpora representing more advanced argumentative essays, e.g. Michigan Corpus of Upper-Level Student Papers (MICUSP).

All in all, this descriptive and exploratory research may be useful for corpus linguists exploring phraseological patterns in learner language, notably when selecting phrase-frames as the unit of analysis, the concept that has been rather underexplored so far in ELF contexts.

References

Ädel, Annellie. 2006. *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Ädel, Annellie and Britt Erman. 2012. Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: a Lexical Bundles Approach. *English for Specific Purposes* 31. 81–92.

Baumgarten, Nicole. 2014. Recurrent Multiword Sequences in L2 English Spoken Academic Discourse: Developmental Perspectives on 1st and 3rd Year Undergraduate Presentational Speech. *Nordic Journal of English Studies* 13(3). 1–32.

Biber, Douglas et al. 1999. *The Longman Grammar of Spoken and Written English.* Harlow: Longman.

Burneikaitė, Nida. 2009. Metadiscoursal Connectors in Linguistics MA Theses in English. *Kalbotyra* 61(3). 36–50.

CECL (Centre for English Corpus Linguistics). 1998. LOCNESS. Louvain-la-Neuve: Universite catholique de Louvain. Available from https://www.uclouvain.be/en-cecl-locness.html. [Accessed: 12th September 2016].

Chen, Yu-Hua and Paul Baker. 2010. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology* 14(2). 30–49.

De Cock, Sylvie. 2004. Preferred Sequences of Words in NS and NNS Speech. *BELL – Belgian Journal of English Language and Literature* 2. 225–246.

Fan, May. 2009. An Exploratory Study of Collocational Use by ESL Students. A Task Based Approach. *System.* [Online] ScienceDirect 37. 110–123. Available from: www.sciencedirect.com. [Accessed: 3rd January 2017].

Fletcher, William, H. 2002–2007. KfNgram. Annapolis: USNA. [Online] Available from: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html. [Accessed: 20th November 2011].

Fletcher, William, H. 2010. Phrases in English. [Online] Available from: http://phrasesinenglish.org/. [Accessed: 20th September 2014].

Forsyth, Richard, S. and Łukasz Grabowski. 2015. Is There a Formula for Formulaic Language? *Poznań Studies in Contemporary Linguistics* 51(4). 511–549.

Fuster-Marquez, Miguel. 2014. Lexical Bundles and Phrase-frames in the Language of Hotel Websites. *English Text Construction* 7(1). 84–121.

Garner, James, R. 2016. A Phrase-frame Approach to Investigating Phraseology in Learner Writing Across Proficiency Levels. *International Journal of Learner Corpus Research* 2(1). 31–67.

Grabowski, Łukasz. 2015. Phrase-frames in English Pharmaceutical Discourse: a Corpus-Driven Study of Intra-disciplinary Register Variation. *Research in Language 13(*3). 266–291.

Granger, Sylviane. 1996. From CA to CIA and Back: An Integrated Contrastive Approach to Computerized Bilingual and Learner Corpora. In: Karin Aijmer, Bengt Altenberg & Stig Johansson (eds.), *Lund Studies in English 88: Languages in Contrast. Text-based cross-linguistic studies*, 37–51. Lund: Lund University Press.

Granger, Sylviane et al. 2009. *The International Corpus of Learner English*. *Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gray, Bethany and Douglas Biber. 2013. Lexical Frames in Academic Prose and Conversation. *International Journal of Corpus Linguistics* 18(1). 109–135.

Grigaliūnienė, Jonė and Rita Juknevičienė. 2012. Corpus-based Learner Language Research: Contrasting Speech and Writing. *Darbai ir dienos* 58. 137–152.

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

Hasselgren, Angela. 1994. Lexical Teddy Bears and Advanced Learners: A Study Into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics* 4. 237–260.

Hyland, Ken. 1998. *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, Ken. 2008a. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics* 18(1). 41–62.

Hyland, Ken. 2008b. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* 27. 4–21.

Jalali, Hassan. 2013. Lexical Bundles in Applied Linguistics: Variations Across Postgraduate Genres. *Journal of Foreign Language Teaching and Translation Studies*. [Online] Available from: http://efl.shbu.ac.ir/efl4/1.pdf. [Accessed: 3rd January 2017].

Jarvis, Scott. 2000. Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. *Language Learning* 50(2). 245–309.

Juknevičienė, Rita. 2009. Lexical Bundles in Learner Language: Lithuanian Learners vs. Native Speakers. *Kalbotyra* 61(3). 61–72.

Juknevičienė, Rita. 2013. Recurrent Word Sequences in Written Learner English. In: Inesa Šeškauskienė and Jonė Grigaliūnienė (eds.), *Anglistics in Lithuania. Cross-Linguistic and Cross-Cultural Aspects of Study*, 178–197. Newcastle upon Tyne: Cambridge Scholars Publishing.

Kizil, Aysel S. and Abdurrahman Kilimci, A. 2014. Recurrent Phrases in Turkish EFL Learners' Spoken Interlanguage: A Corpus-driven Structural and Functional Analysis. *Journal of Language and Linguistic Studies* [Online] 10(1). 195–210. Available from: http://jlls.org/index.php/jlls/article/view/176/178. [Accessed: 3rd January 2017].

Kjellmer, Göran. 1991. A Mint of Phrases. In: Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, 111–127. London: Longman.

Leńko-Szymańska, Agnieszka. 2014. The Acquisition of Formulaic Language by EFL Learners: A Cross-sectional and Cross-linguistic Perspective. *International Journal of Corpus Linguistics* 19(2). 225–251.

Martelli, Aurelia. 2006. A Corpus Based Description of English Lexical Collocations Used by Italian Advanced Learners. [Online] *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-*

9 settembre 2006, 1005–1011. Available from: https://dialnet.unirioja.es/servlet/articulo?codigo=4685334. [Accessed: 15th September 2016].

Martinez, Ron, and Norbert Schmitt. 2012. A Phrasal Expressions List. *Applied Linguistics* 33(3). 299–320.

MICUSP (Michigan Corpus of Upperlevel Student Papers). 2009. Ann Arbor, MI: The Regents of the University of Michigan.

Nesselhauf, Nadia. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

O'Donnell, Matthew B., Römer, Ute and Nick C. Ellis. 2013. The Development of Formulaic Sequences in First and Second Language Writing. Investigating Effects of Frequency, Association, and Native Norm. *International Journal of Corpus Linguistics* 18(1). 83–108.

Paquot, Magali. 2013. Lexical Bundles and L1 Transfer Effects. *International Journal of Corpus Linguistics* 18 (3). 391–417.

Paquot, Magali. 2014. Cross-linguistic Influence and Formulaic Language: Recurrent Word Sequences in French Learner Writing. *EUROSLA Yearbook* 14. 240–261.

Pawley, Andrew and Francis H. Syder. 1983. Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In: Jack C. Richards and Richard W. Schmidt (eds.), *Language and Communication*, 191–225. London: Longman.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Viena, Austria. [Online] Available from: http://www.R-project.org. [Accessed 15th September 2016].

Renouf, Antoinette and John Sinclair. 1991. Collocational Frameworks in English. In: Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics*, 128-143. New York: Longman.

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

Römer, Ute. 2010. Establishing the Phraseological Profile of a Text Type. The Construction of Meaning in Academic Book Reviews. *English Text Construction* 3(1). 95–119.

Römer, Ute and Matthew O'Donnell. 2009. Positional variation of phrase frames in a new corpus of proficient student writing. [Online] Paper presented at AACL conference. Edmonton, Canada, 9 Oct 2009.
Available from: http://www.ualberta.ca/~aacl2009/PDFs/RoemerODonnell2009AACL.pdf. [Accessed: 15th September 2016].

Scott, Mike. 2008. Wordsmith Tools. Version 5. Oxford: Oxford University Press.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Simpson-Vlach, Rita. and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4). 487–512.

Stubbs, Michael. 2007. Quantitative Data on Multi-word Sequences in English: the Case of the Word 'World'. In Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert (eds), *Text, Discourse and Corpora*, 163–190. London: Continuum.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Vidakovic, Ivana and Fiona Barker. 2010. Use of Words and Multi-word Units in Skills for Life Writing Examinations. *Research Notes* 41. 7–41.

Waibel, Birgit. 2007. *Phrasal Verbs in Learner English: A Corpus-based Study of German and Italian Learners*. [Online] Unpublished PhD dissertation. Freiburg: Albert-Ludwigs-Universität. Available from: https://freidok.uni-freiburg.de/dnb/download/3592. [Accessed: 15th September 2016].

Wang, Ying. 2016. *The Idiom Principle and L1 Influence*. Amsterdam: John Benjamins.

Wray, Alison. 2000. Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics* 21(4). 463–489.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Appendices

Appendix 1

Topic-specific phrase-frames

LICLE

* as I lay
* has done more
* in the society
* men and women
* money is the
* of higher education
* of the world
* same sex marriages
* the English language
* the European Union
* the higher education
* the process of
* the quality of
* the right to
language is a *
of * in Lithuania
of higher education *
of the * language
quality of studies *
that writing is *
the * of education
the * of money
the problems of *
the quality of *
the reform of *
the right to *
to study at *

PICLE

* are exposed to
* have the right
* of mass media
* the development of
* the European community
* the influence of
* the mass media
* the outside world
* the right to
* to adopt children
approach to reality *
have * right to
in the * world
of mass media *
of the * world
the * of mass
the * of money
the development of *
the influence of *
the opponents of *
the right to *
the role of *
to * a child
to bring up *

LOCNESS

* a loss of
* ethnic American literature
* for the best
* invention of the
* is for the
* Le Mythe de
* loss of sovereignty
* of the absurd
* of the play
* the ##th century
* the #th Republic
* the death penalty
* the idea of
* the right to
* the United States
for the best *
invention of the *
of the * Republic
the ##th century *
the * of optimism
the age of *
the death of *
the death penalty *
the people of *
the right to *
the United States *

Appendix 2

Lists of PFs used for the analysis (in the frequency order). The first number indicates the absolute frequency of the phrase-frame and the second number shows the number of realizations it has in the corpus.

LICLE			PICLE			LOCNESS		
the * of the	169	42	the * of the	180	42	the * of the	335	50
one of the *	104	12	it is * to	113	12	in the * of	97	17
it is * to	100	13	in the * of	93	17	* the fact that	76	11
* a lot of	83	14	* the fact that	79	9	the fact that *	71	10
* one of the	82	8	it is not *	76	12	at the * of	61	7
it is not *	74	12	as a * of	71	7	* be able to	58	6
in order to *	69	13	one of the *	70	11	* the end of	58	6

LICLE			PICLE			LOCNESS		
it is * that	69	7	* one of the	66	7	the * of a	50	13
a lot of *	67	11	in order to *	64	13	to the * of	43	12
there is no *	58	9	there is no *	64	13	as a * of	43	7
is the * of	54	7	the fact that *	64	10	it is * to	38	6
in the * of	51	10	do not * to	55	4	end of the *	37	5
* there is no	51	9	* it is not	48	8	a * of the	35	5
that it is *	48	8	as well as *	47	7	in order to *	33	7
* the fact that	43	8	* more and more	45	6	* one of the	32	4
* be able to	42	5	* be able to	45	5	is * of the	32	4
do not * to	42	5	that it is *	44	8	that it is *	31	4
* it is not	40	6	it is * that	44	6	is * to be	30	5
first of all *	39	6	the other hand *	43	8	one of the *	30	5
the fact that *	38	6	they do not *	43	7	the rest of *	30	4
as well as *	37	8	in * of the	41	6	to the * that	28	3
* a part of	37	6	more and more *	41	6	for the * of	27	4
they do not *	36	7	as * as the	40	3	* that it is	24	6
* that it is	35	8	that is why *	39	7	that * is a	24	4
it is very *	33	4	* do not have	39	5	of the * of	23	6
* the most important	3	32	* not have to	39	3	is a * of	22	6
in the world *	31	8	do not have *	38	4	that * is not	20	3
* they do not	30	6	* that it is	36	8	in the * and	19	5
* in the world	29	6	* there is no	34	7	* it is not	19	3
the other hand *	29	4	they are * to	34	6	do not * to	19	3
is a * of	28	6	at the * of	34	5	the idea of *	19	3
is * to be	28	3	in front of *	34	5	the * of his	17	5
of the most *	28	3	as a result *	34	4	can be seen *	17	4
* more and more	27	6	* a lot of	33	7	it is * that	17	4
it is * a		27	for the * of	33	6	a part of *	17	3
the most important *	4	27	seems to be *	33	5	do not have *	17	3
* considered to be	27	3	* of the world	31	7	* that they are	16	4
* is one of	25	4	he or she *	31	7	* the use of	16	4
of the world *	24	5	of the * of	30	9	* they do not	16	4
he or she *	24	3	* aware of the	30	6	* have to be	16	3
* it is a	23	5	first of all *	30	5	* of the world	16	3
a part of *	23	5	a great * of	29	4	on the * of	15	4
of the * of	23	5	is not * to	28	5	* part of the	15	3
is very * to	23	3	* are able to	28	4	* seen to be	15	3
there is a *	22	7	to the * of	27	7	* not have to	14	4
that the * of	22	5	* at the same	27	3	the right to *	14	4
* it is the	22	4	is the * of	26	7	* the world and	14	3
as a * of	22	4	* they do not	26	5	* use to the	14	3
in my opinion *	22	4	to the * that	26	3	in favor of *	14	3
that * is a	22	4	* should not be	25	7	of the * and	14	3
the majority of *	22	4	is * to be	25	5	be able to *	13	4
as a * to	22	3	a lot of *	24	4	in * of the	13	4
do not think *	22	3	in such a *	24	3	* should not be	13	3
* be said that	21	3	it is very *	23	6	* there is no	13	3
* do not have	21	3	* to be a	23	3	a lot of *	13	3

LICLE			PICLE			LOCNESS		
the * that the	21	3	* out to be	22	3	the use of *	13	3
the * way to	21	3	there are * who	21	3	* to be a	12	4
* the lack of	20	5	they * not have	21	3	it is not *	12	4
* do not think	20	4	* have to be	20	4	* the rest of	12	3
do not have *	20	4	* seems to be	20	3	and the * of	12	3
it * be said	20	3	a * number of	20	3	I * that the	12	3
that * is not	20	3	does not * to	20	3	in a * of	12	3
* not have to	19	3	that there is *	20	3	is the * of	12	3
in the * world	19	3	to be the *	20	3	that this is *	12	3
the * of a	18	5	not * to be	19	4	the * of their	12	3
people do not *	18	4	* of the fact	19	3	they are * to	12	3
the * is that	18	4	we do not *	19	3	this is not *	12	3
to sum up *	18	4	the most important *18	18	5	* to say that	11	3
* at the same	18	3	the same time *	18	4	that the * of	11	3
* it does not	17	4	what is more *	18	4	the case of *	11	3
a * number of	17	4	* do not want	18	3	with the * of	11	3
in * of the	17	4	* they are not	18	3	* will continue to	10	3
of the * and	17	4	does not have *	18	3			
as * as the	17	3	the * of a	17	5			
because it is *	17	3	the majority of *	17	5			
more and more *	17	3	should not be *	17	4			
* are able to	16	4	* in front of	17	3			
a great * of	16	4	* the most important	17	3			
* the world and	16	3	in the world *	17	3			
considered to be *	16	3	is * it is	17	3			
is very important *	16	3	that * is a	17	3			
that * is the	16	3	* said to be	16	3			
is a * to	15	5	aware of the *	16	3			
it is a *	15	4	it is a *	15	4			
it is the *	15	4	* he or she	15	3			
* not able to	15	3	* it is a	15	3			
can be * that	15	3	it is * difficult	15	3			
say that * is	15	3	of the most *	15	3			
and it is *	14	4	the only * of	15	3			
in this way *	14	3	they are not *	15	3			
it can be *	14	3	we are * to	15	3			
most of the *	14	3	* the lack of	14	4			
they are * to	14	3	as long as *	14	4			
to * with the	14	3	that there are *	14	4			
would not be *	13	4	there are also *	14	4			
* do not need	13	3	* are not able	14	3			
most important thing *	3	13	* do not need	14	3			
* a number of	12	4	* people who are	14	3			
* in order to	12	4	* take into consideration	14	3			

LICLE			PICLE			LOCNESS
it is * for	12	4	are * to be	14	3	
the * and the	12	4	in this way *	14	3	
to the * of	12	4	* in the world	13	4	
* as a means	12	3	* the idea of	13	4	
* is the most	12	3	* most of the	13	3	
is * difficult to	12	3	is * difficult to	13	3	
that they * not	12	3	it is * a	13	3	
* amount of money	3	11	point of view *	13	3	
* should not be	11	3	they * to be	13	3	
* they want to	11	3	would not be *	13	3	
should not be *	11	3	* that they are	12	4	
the lack of *	11	3	fact that * are	12	4	
they want to *	11	3	* seem to be	12	3	
* is a very	10	3	a * variety of	12	3	
* that they are	10	3	from the * of	12	3	
i think that *	10	3	with the * of	12	3	
of a * language	10	3	* it possible to	11	3	
there * be no	10	3	* the number of	11	3	
there are many *	10	3	* the rest of	11	3	
* people do not	9	3	* the world and	11	3	
* to say that	9	3	* there is a	11	3	
is not * to	9	3	* we do not	11	3	
there are more *	9	3	are * likely to	11	3	
			by * of the	11	3	
			is * reason why	11	3	
			is the most *	11	3	
			that * is the	11	3	
			the idea of *	11	3	
			we must * that	11	3	
			* a kind of	10	3	
			* do not realize	10	3	
			it * not be	10	3	
			that the * of	10	3	
			the most * of	10	3	
			there are * many	10	3	
			to * about the	10	3	
			* would not be	9	3	
			in * to the	9	3	
			there are many *	9	3	
			there is * a	9	3	