Lexical Bundles in Learner Language: Lithuanian Learners *vs.* Native Speakers

Rita Juknevičienė
Department of English Philology
Vilnius University
Universiteto St 5
LT-01513 Vilnius, Lithuania
Tel./fax: +370 5 2687 228
E-mail: rita.jukneviciene@takas.lt

Abstract. Corpus studies of recurrent word sequences, also known as lexical bundles, have outlined new directions in ELT/EFL research. The fact that naturally produced English consists of prefabricated multi-word units gave rise to the question of chunkiness in learner language. This study was thus designed to compare and contrast language produced by learners of three different levels of proficiency in terms of the use of lexical bundles. The analysis of lexical bundles retrieved from three corpora of written learner English showed that many more different bundles were established in the corpora representing less proficient learners, which could be interpreted as an indication of a more limited lexical repertoire that leads to more repetition. Furthermore, structural and functional analysis of lexical bundles showed that the language of non-native learners bears more resemblance to spoken English than the language of native speakers. These findings may be particularly useful to EFL practitioners as they seem to give new insights into the development of learner language.

Keywords: Contrastive interlanguage analysis, learner language, lexical bundles, Lithuanian EFL learners

1 Research question

The study reported in this article was designed to establish prevailing structural and functional types of four-word lexical bundles in three corpora of English representing written language of learners at different levels of achievement. Previous corpus studies of lexical bundles, which are defined here as continuous recurrent sequences of four words, have shown that different discourse types are characterised by different types of lexical bundles. It is thus possible to hypothesize that lexical bundles retrieved from learner language will provide evidence to the claim that development of written language skills progresses from spoken to written language. Furthermore, lexical bundles may also be seen as text building blocks which are processed and produced as solid lexical units. Therefore a contrastive analysis of lexical bundles in written language produced by native speakers (NS) and non-native speakers (NNS) of English may point out certain factors which determine

different quality of their written language. Owing to different reasons, the study of English produced by Lithuanian learners has been unduly neglected, so this study is expected to fill the gap by applying a modern corpus linguistics approach to the analysis of interlanguage of Lithuanian learners.

2 Review of previous research

Research of automatically retrieved recurrent lexical sequences, also known as lexical bundles (Biber *et al.* 1999, Hyland 2008), clusters (Scott 1999), recurrent sequences (De Cock 2004), or chunks (O'Keeffe *et al.* 2007) should be distinguished from the analysis of other multi-word units which perform a definite discourse function in the text. While the former are identified in a corpus on the basis of statistical recurrence, the latter, also referred to as lexical phrases (Nattinger & DeCarrico 1992) and formulaic sequences (Wray 2002), are selected intuitively as phrases serving a specific function, e.g. exemplification 'for example', discourse markers 'I think' *etc.* The present study deals with automatically retrieved lexical bundles identified on the basis of frequency regardless of their structural or semantic wholeness.

One of the earliest studies of lexical bundles was reported by Altenberg (1998) who analysed "recurrent word-combinations" in spoken English. His research showed that such combinations are "evident at all levels of linguistic organization" (Altenberg 1998, 120) while the fact that they do recur as lexical units even though their grammatical structure is often incomplete obscures the distinction between lexicon and grammar. Differences in the use of lexical bundles across the four major registers of language, namely, conversation, fiction writing, news reporting and academic prose, were first described in Biber *et al.* (1999). This study set a standard for many subsequent analyses of lexical bundles, of which only the most relevant to our study will be mentioned here. While Biber *et al.* (1999) identified the prevailing structural types of lexical bundles, Biber *et al.* (2004) and Biber (2006) proposed more elaborate taxonomies of lexical bundles in terms of structure and function. These studies clearly showed that structural types of lexical bundles have their functional correlates, which once again confirmed a close link rather than division between structure and meaning. The use of lexical bundles in articles written in different research fields was investigated by Cortes (2004) and Hyland (2008), while Cortes (2008) presented her findings from a contrastive analysis of lexical bundles in Spanish and English history writing. These studies provided evidence to the claim that different academic disciplines make use of area-specific lexical bundles which, consequently, should encourage further research in this direction and perhaps give new stimuli to ESP courses.

As regards learner language, one of the most extensive analyses of recurrent lexical bundles was carried out by De Cock (2004). Her study showed that "advanced learners' use of frequently recurring sequences of words displays a complex picture of overuse, underuse, misuse of target language NS sequences" (De Cock 2004, 243). Among many different reasons behind it, as shown by Koprowski (2005), inadequacy of English coursebooks might be mentioned. Koprowski argues that even recent publications often fail to include the most typical multi-word items and instead present less frequent expressions. Furthermore, Cortes (2004) argues that teaching of English at tertiary level calls for field-specific vocabulary which, in its turn, should be derived from the analysis of student writing within different disciplines. The study reported here might thus be seen as an attempt to analyse the language of Lithuanian learners of English majoring in language studies and to describe changes in the development of their lexical competence as reflected by the use of lexical bundles.

3 Data and methods

Data for this study has been retrieved from three corpora of learner English. Two of them represent the language of NNS learners of English of the first year of study and the third-fourth year of study, namely, the AFK1 corpus and the LICLE corpus. The AFK1 corpus (92 050 words) consists of student essays collected in 2006-2008 as part of my ongoing PhD research. The LICLE corpus (137 004 words) was compiled as a component of the ICLE learner corpora (Grigaliūnienė *et al.* 2008). All the students, whose essays were included in the two corpora, are native speakers of Lithuanian majoring in the English Philology BA programme in Vilnius University. The corpora contain examination essays on a variety of topics related to their study programme and its syllabus. Hence some topics deal with the study of language and linguistics, while others are written on more general issues, e.g. capital punishment, euthanasia, sex education in schools *etc.* A part of the essays (*ca.* one fifth of the LICLE corpus) are literary analyses of works of fiction. The third corpus, representing NS learners, is the LOCNESS corpus compiled by S. Granger at the University of Louvain-la-Neuve and usually used as a reference corpus in contrastive studies of learner language. For the purpose of this analysis, only part of LOCNESS corpus (164 684 words) was used so that its size and essay topics would better match the composition of the two NNS corpora[1].

Lexical bundles were retrieved from the corpora using the Wordsmith Tool software (v. 5) which produces frequency lists of lexical bundles of delimited frequency and distribution (see below).

[1] The following subcorpora of LOCNESS were included in the analysis: brsur1.cor (codes 1-15); brsur2.cor; brsur3.cor; br-alevels; usind-0001.1-28.1; usscu2.cor; usscu3.cor; usscu4.cor; usprb1.cor (codes 1-25).

Then the structural and functional analysis of the bundles was carried out following taxonomies proposed by Biber *et al.* (2004) and Biber (2006). While structural types of lexical bundles were established on the basis of constituent parts-of-speech, functional types were determined through the analysis of broader contexts of the bundles which were retrieved using the concordancing tool of the WordSmith Tools program.

This analysis deals with four-word lexical sequences as they have been found to possess fuller structure and content. The minimum frequency is four times per 100,000 words (or 40 per one million words) in four different texts. To compare, Biber *et al.* (2004) analysed only those four-word lexical bundles which recur at least 10 times per million words in five different texts (2004, 992). A more conservative approach was adopted in Biber (2006, 134) with the frequency cut-off point at 40 times per million words. In her analysis, De Cock (2004, 228), who analysed 2-, 3-, 4-, 5- and 6-word sequences, applied the threshold of approximately 10-12% of the total number of recurrent sequences which in the case of 4-word sequences covered bundles that recurred at least four times per 100,000 words. Hyland (2008) applied a minimum frequency of 20 times per million words in at least 10% of texts. Table 1 summarizes frequency cut-off points used in this study, all of which correspond to the normalized frequency of four bundles per 100,000 words.

	AFK1	LICLE	LOCNESS
Cut-off points in absolute frequency	4 times in 92,050 words	6 times in 137,004 words	7 times in 164,684 words
Normalized frequency of the cut-off point	4 occurrences per 100,000 words		

Table 1. Frequency cut-off points of lexical bundles

Unless it is specified otherwise, the terms ‘lexical bundles’ or ‘recurrent lexical sequences’ shall henceforth refer to the items of delimited frequency and distribution.

4 Research findings

The largest number of different lexical bundles was established in the AFK1 corpus (see Table 2) whereas the NS corpus contained a significantly smaller number of bundles (differences among the corpora are statistically significant at $p<0.001$).

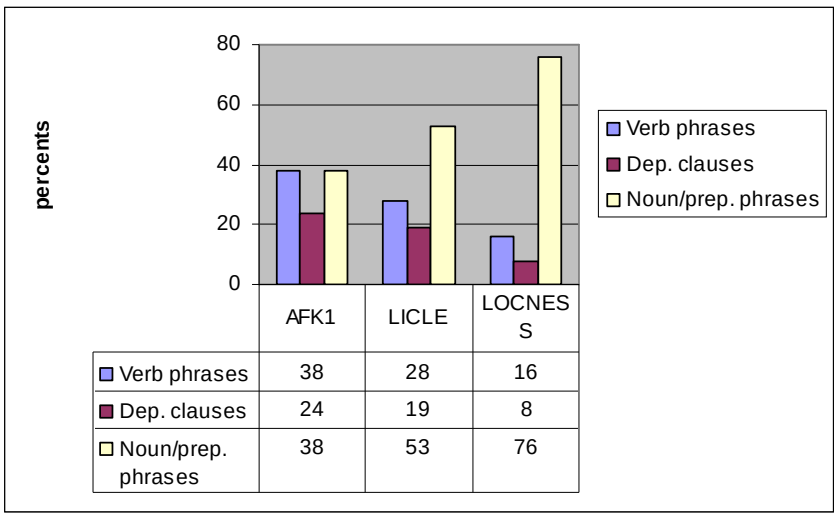
	AFK1	LICLE	LOCNESS
Number of different lexical bundles (types)	382	178	96
Total number of lexical bundles (tokens)	2316	1432	747

Table 2. Number of lexical bundles in the three corpora

Since lexical bundles are usually associated with naturalness of expression, the language of the NS learners was expected to contain many more bundles than the NNS corpora and thus possess a larger proportion of 'bundlized' vocabulary. Yet our findings gave a different picture which perhaps reflects the underlying lexical range of the learners. It is possible to assume that NNS learners rely on a limited set of lexical phrases while more proficient learners have a broader repertoire of lexical expressions. Consequently, a smaller proportion of repetitive lexis in the NS essays yields fewer lexical bundles. In contrast, repetitiveness of bundles in the language of NNS learners sometimes leads to verbosity and frequent repetition of 'safe' phrases lifted from the essay title or topic statement. Structural and functional analyses of lexical bundles provided some evidence to this claim.

4.1 Structural types of lexical bundles in learner corpora

Structural classification of lexical bundles draws on the taxonomy proposed by Biber *et al.* (1999) and later elaborated in Biber *et al.* (2004) and Biber (2006). It distinguishes three major types of bundles: 1 - verb phrase fragments (VB), e. g. *it's going to be, is based on the, everything is for the*; 2 - dependent clause fragments (DepCl), e. g. *you might want to, if you have a, that there is a*; 3- noun phrase and prepositional phrase fragments (NP), e. g. *one of the things, those of you who, at the same time*. In general, studies of English corpora showed that while conversation primarily consists of clausal bundles incorporating verb phrases and dependent clause fragments, written English makes an extensive use of phrasal bundles incorporating noun/prepositional phrases (Biber *et al.* 1999, Biber *et al.* 2004, Hyland 2008). Our analysis showed that the language of learners of lower proficiency tends to contain more verb bundles while the corpus of native speaker students has yielded a bigger proportion of noun phrases (Figure 1) and thus seems to be closer to the structural patterning of academic English prose.



	AFK1	LICLE	LOCNESS
Verb phrases	38	28	16
Dep. clauses	24	19	8
Noun/prep. phrases	38	53	76

Figure 1. Distribution of structural types of lexical bundles across the analysed corpora (in relative numbers)

A gradual shift across the three corpora in the proportion of different bundle types confirms the assumption that the three corpora might be placed on a scale from the lowest to highest level of linguistic proficiency. Furthermore, prevalence of structural types typical of spoken language in the AFK1 corpus and to a lesser degree in LICLE might be seen as evidence to the fact that the development of written language progresses from spoken to written language.

The most conspicuous difference in terms of bundle structure between the NNS corpora and the NS corpus arises from several structural subtypes incorporating verb phrase fragments. Firstly, the largest groups of verb bundles in the NNS corpora are bundles of the following structural subtype: '(conjunction) + 3rd person pronouns/nouns + VP fragment', for example,

(1) *it is better to, it is the most, it is not the, capital punishment should be, the computer has become* (AFK1);

(2) *it is possible to, language is one of, however it is not, it has to be* (LICLE).

This subtype also includes bundles with the existential *there is/are* construction which is particularly popular among the Lithuanian learners and significantly less frequent in the LOCNESS corpus. The functional analysis described below will provide some explanation of this tendency.

The other prominent peculiarity of the NNS corpora is frequently used bundles with non-passive verbs, for example,

(3) *becoming more and more, is the lack of, have a lot of* (AFK1);

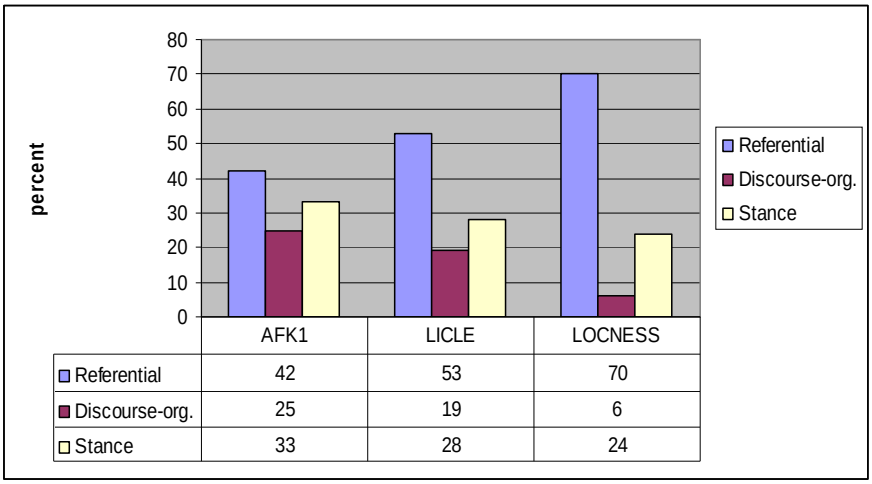
(4) *is not only the, are a lot of, is not an exception* (LICLE).

Finally, verbs in the language of NNS speakers frequently occur in the bundles with *to*- clause fragments, for example,

(5) *in order to be, in order to make, to know the English, to sum up* (AFK1);
(6) *in order to achieve, to be more precise, to pay for their, to be able to* (LICLE).
In contrast, the most widely represented structural subtypes in the NS corpus include lexical bundles incorporating noun phrases and prepositional phrase fragments, for example,
(7) *the end of the, one of the most, the rest of the, the themes of guilt* (LOCNESS).
As shown in Graph 1, these subtypes of bundles have also been established in the Lithuanian corpora yet their proportional weight is much more significant in the corpus of native speakers of English. Since the established structural types have their correlates in the functional taxonomy, the following section will show how the structural patterning affects the communicative value of learner writing.

4.2 Functional types of lexical bundles in learner language

The functional taxonomy of lexical bundles traditionally distinguishes three major functions of lexical bundles: 1 – referential (or 'research-oriented' in Hyland's terms) bundles which name physical and abstract objects, give place or time reference; 2 – discourse organizing, or text-oriented, bundles organize the text, and 3 – stance, or participant-oriented, bundles express the writer's attitudes or evaluation of the proposition (Biber *et al.* 2004, Biber 2006, Hyland 2008). The results of functional analysis in our study are summarized in Figure 2. Similarly to the structural analysis, in terms of functional types the three corpora show a gradual shift in the representation of individual function types with the LOCNESS corpus standing closest to academic English prose and the AFK1 corpus bearing more features of spoken language.



	AFK1	LICLE	LOCNESS
Referential	42	53	70
Discourse-org.	25	19	6
Stance	33	28	24

Figure 2. Distribution of functional types of lexical bundles across the analysed corpora (in relative numbers)

Previously mentioned studies of lexical bundles in larger corpora showed that written academic discourse is dominated by referential bundles while discourse-organizing and stance expressions are more characteristic of spoken language (Biber *et al.* 1999, 2004, Biber 2006, O'Keeffe *et al.* 2007). This analysis has confirmed that in terms of prevailing functional types of lexical bundles the written language of NNS speakers bears more resemblance to spoken English and in that respect it significantly differs from the written language of native speakers.

Yet all the three corpora have certain similarities. The most obvious of them is the proportion of referential bundles which account for the largest part of all the bundles in all three corpora. Examples (8) to (10) below present one subtype of this functional category, namely, bundles specifying intangible framing attributes, which deserve a special mention since they have been found to be very common in English academic prose (Biber *et al.* 2004, Biber 2006):
(8) *in the same way, the importance of English, the influence of the, that this kind of* (AFK1);
(9) *in the process of, as a means of, as a reaction to, with the help of* (LICLE);
(10) *in the case of, in such a way, the power of the, the idea of the* (LOCNESS).
Our material shows that this subtype accounts for nearly half of all referential bundles in LICLE and LOCNESS and only a quarter of referential bundles in the AFK1 corpus. It is thus possible to state that student writing makes use of at least one functional subtype of referential bundles which is particularly common in academic prose. Moreover, the frequency and proportion of this subtype of lexical bundles gradually changes across the three corpora which once again confirms that they might be placed on a scale from the lowest (AFK1 learners) to the highest (LICLE and LOCNESS learners) levels of achievement.

Another common feature of the three corpora is the frequency of topic vocabulary. Quite a big number of referential lexical bundles in all the analysed corpora are directly related to the topics of the essays. Such bundles often consist of words lifted from the essay titles or topic statements. For example, the bundles *educational system in Lithuania* and *the quality of education* clearly deal with the topic of education reform while such bundles as *in the case of* or *a part of the* could be considered to be topic-neutral. Furthermore, titles of literary work or quotes in English or other languages used by the students in their essays, e. g. *le Mythe de Sisyphe* (LOCNESS), *the Joy luck club* (LOCNESS), *Desire under the elms* (LICLE), *As I lay dying* (LICLE) were also treated as topic-related bundles. The analysis showed that the proportion of topic-related bundles is one of the aspects on which more similarity is to be found between the LICLE and the LOCNESS corpora while in the AFK1 corpus only 33 per cent of the bundles were categorized as topic-related (cf. 45%

in LICLE and 44% in LOCNESS). Since all referential bundles, topic-related included, mostly contribute to the propositional density of the text, underuse of referential lexical bundles in the NNS language in relation to NS language indicates that the language of NNS learners is less topic-focused than the language of native speakers. Yet this lack in propositional content is compensated by an extensive use of more formal elements of written discourse represented by stance and discourse-organizing bundles.

Stance bundles, which by definition 'express attitudes or assessments of certainty that frame some other proposition' (Biber *et al.* 2004: 384), most often are epistemic stance expressions, for example,

(11) *are more likely to, due to the fact, argue that it is* (AFK1);
(12) *the fact that the, I do not think, it is clear that* (LICLE);
(13) *the fact that he, to the fact that, that it is a* (LOCNESS).

The other numerous subtypes of stance bundles are expressions of attitudinal stance, of which ability bundles are well-represented in all the three corpora, for example, *would be able to, to be able to, it is possible to etc.* In general, attitudinal stance bundles are more common in the NNS corpora than in the LOCNESS corpus. The Lithuanian learners seem to be more categorical and uncompromising, which is evidenced by their frequent use of directives (*it should not be, that it should be*, (*death*) *penalty should be established*), or bundles expressing importance (*are the most important, is a very important, it is important to*) and evaluation (*has a great influence, it is better to, the best way to*). Apart from the effect of the essay topics, which might call for a more direct expression of personal opinion, another factor behind the frequency of these bundles could be the younger age of the Lithuanian learners, particularly the first-year students of English (the AFK1 corpus).

Finally, discourse-organizing bundles have also been found to be significantly more frequent in the AFK1 corpus while the other two corpora yielded considerably shorter lists of this type of bundles (see Figure 2). The most striking difference here is related to the use of the construction *there is/are* for topic introduction or focus. Though the bundles with this construction may be seen as multi-functional varying between stance and discourse-organizing functions, in the NNS corpora they usually serve the discourse-organizing function by introducing a new topic, for example:

(14) *<u>There are a lot</u> of ways of communication and language is one of them.* (AFK1)
(15) *<u>There are a lot</u> of similarities between modernism and post-modernism.* (LICLE)

(16) *These days there is much talking about the challenged language, especially of politicians.* (LICLE)

In contrast, the presentative *there is/are* construction does not appear at all in the bundle list of the NS corpus. Moreover, only five discourse-organizing bundles were established in this corpus, namely, *on the other hand, at the same time, one of the most, is one of the,* and *when it comes to* (LOCNESS). The overuse of such bundles as *to sum up it, on the other hand, at the same time, what is more the, in order to be, taking everything into account, all in all the etc.* in the Lithuanian student writing perhaps might be explained by the aims and teaching foci of academic writing courses where much attention is paid to issues of text organization and cohesion. Obviously, text connectives are seen by the learners as one of the main features of good writing.

5 Conclusions

The study has proved that corpus linguistics approach to learner language might give new insights into the development of lexical competence and, more specifically, differences in the quality of learner language. The sheer number and frequency of lexical bundles extracted from the three corpora analysed here indicate the proportion of repetitive lexis and thus point to varying lexical range and richness of vocabulary represented in the corpora. Structural and functional analyses, on the other hand, might provide information about learners' awareness and control of the register. This study has shown that structural and functional types of lexical bundles established in the AFK1 corpus testify to its similarity to spoken English – prevalence of bundles with fragments of verb phrases and dependent clauses, a relatively small proportion of referential bundles and overuse of bundles serving stance and discourse-organizing functions. In contrast, the native speaker corpus LOCNESS bears more features typical of written academic prose – it has the largest proportion of referential bundles which are usually expressed by noun and prepositional phrases. Finally, the LICLE corpus occupies an intermediary position between the other two corpora. Yet due to the limited scope of this study, these conclusions are rather tentative. To get a more accurate picture of differences between NNS and NS language, lexical bundles of other lengths might also be analyzed. Furthermore, in studies of small corpora topic impact is very strong, which necessitates a more cautious approach to the selection of corpus material.

6 References

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

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

Biber, D., Conrad, S. & V. Cortes. 2004. If You Look at…: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics* 25, 371-405.

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

Cortes, V. 2004. Lexical Bundles in Published and Student Writing in History and Biology. *English for Specific Purposes* 23 (4), 397-423.

Cortes, V. 2008. A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish. *Corpora* 3 (1), 43-57.

De Cock, S. 2004. Preferred Sequences of Words in NS and NNS Speech. *BELL (Belgian journal of English language and literature)*, 225-246.

Grigaliūnienė, J., Bikelienė, L. & R. Juknevičienė. 2008. The Lithuanian Component of the International Corpus of Learner English (LICLE): A Resource for English Language Learning, Teaching and Research at Lithuanian Institutions of Higher Learning. *Žmogus ir žodis* 10 (III), 62-66.

Hyland, K. 2008. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* (27), 4-10.

Koprowski, M. 2005. Investigating the Usefulness of Lexical Phrases in Contemporary Coursebooks. *ELT Journal* 59 (4), 322-332.

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

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

Scott, M. 1999. *Wordsmith Tools*. Software. Oxford: Oxford University Press.

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

Leksinių samplaikų vartojimas mokinių kalboje: kontrastyvinis lietuvių ir gimtakalbių anglų kalbos vartotojų kalbos tyrimas

Rita Juknevičienė

Leksinių samplaikų arba pasikartojančių žodžių sekų tyrimai anglų lingvistikoje atvėrė naujų tyrimo krypčių anglų kaip svetimosios kalbos studijoms. Tekstynų tyrinėjimai parodė, jog autentiškoje anglų kalboje itin dažnai pasitaiko pasikartojančių žodžių sekų, kurios sakytinėje ar rašytinėje kalboje atkuriamos kaip nedalomi vienetai. Todėl tikėtina, jog kintant kalbos vartotojų pasiekimams tokių sekų kiekis ir pobūdis jų kalboje taip pat gali kisti. Straipsnyje pristatomas kontrastyvinis mokinių kalbos (angl. *learner language*) tekstynų tyrimas, analizuojantis leksinių samplaikų vartojimą skirtingų pasiekimo lygių mokinių kalboje. Keliama hipotezė, kad samplaikų vartojimo skirtumai gali lemti kokybinius įvairių lygių mokinių kalbos skirtumus. Tyrimo duomenys rinkti iš trijų tekstynų, kurių du (AFK1 ir LICLE tekstynai) reprezentuoja lietuvių gimtosios kalbos mokinių rašytinę anglų kalbą, o trečiasis (LOCNESS) – gimtakalbių anglų kalbos mokinių kalbą. Tyrimo rezultatai parodė, jog tekstynai skiriasi tiek pagal juose nustatytų samplaikų kiekį, tiek pagal vyraujančius struktūrinius ir funkcinius samplaikų tipus, apibrėžiamus pagal Biber ir kt. (2004) ir Biber (2006) taksonomijas. Žemesnio pasiekimų lygio mokiniai vartoja gerokai daugiau samplaikų, o tai rodo, jog jų kalboje yra daugiau pasikartojančios leksikos negu gimtakalbių mokinių kalboje, kuri laikytina leksiškai turtingesne. Be to, pagal struktūrinius ir funkcinius samplaikų tipus lietuvių mokinių kalba turi daugiau sakytinės anglų kalbos bruožų negu gimtakalbių mokinių kalba.

Įteikta 2009 m. lapkričio 20 d.