PHRASEOLOGY IN LEARNER ACADEMIC ENGLISH: CORPUS-DRIVEN APPROACHES[1]

Markéta Malá

Abstract

The paper combines learner corpus research with contrastive analysis to test the applicability of corpus-driven methods to the study of phraseology in learner academic English. It explores phraseological patterns in English L2 academic texts written by Czech university students in comparison with English L1 novice and expert writing. Three corpus-driven approaches are employed: frequency lists, keywords and lexical bundles. The results indicate that a combination of corpus-driven methods can indeed serve as an effective starting point for the contrastive study of phraseology, highlighting potential areas of under- and overuse of multi-word patterns in English L2 novice academic texts. However, in order to give a more comprehensive picture of learner academic English, quantitative methods have to be combined with qualitative contrastive analysis.

Keywords

learner corpus research, academic English, corpus-driven methods, contrastive analysis, phraseology

1 Phraseology, learner academic English, and corpora

The present paper tests the applicability of three corpus-driven approaches to the study of phraseology in academic essays written by Czech advanced students of English. It combines learner corpus research with contrastive analysis. Two dimensions of contrast are explored: novice-expert (English texts written by novice academic writers, L1 and L2, are compared with those published in academic journals), and native-learner English.

All the key words which appear in the title of this paper and summarize its main focus and methodology have been used widely in numerous studies – their delimitation, however, seems to vary, depending on the approach applied. Let me, therefore, start by explaining how they will be used in this paper.

Language communication has been shown to rely, to a large extent, on "combinations of words that customarily co-occur" (Kjellmer 1991: 112), such as *make a decision*, *in the middle of*, or *see what I mean*. The recurrence of such multi-word linguistic units suggests that "the language we use every day is composed of prefabricated expressions, rather than being strictly compositional" (Gray & Biber 2015: 125). These expressions may then be seen as constituting

Discourse and Interaction 13/2/2020, pp. 75-88
ISSN 1802-9930
https://doi.org/10.5817/DI2020-2-75

the phraseology of a language. Various approaches to phraseology, however, differ in their views on what types of units phraseology deals with, and how these units are to be identified. The approaches which focus on formal taxonomies of phraseological units, such as *kick the bucket, under the weather* (see, e.g. Cowie 1998, Čermák 2008), tend to concentrate on the degree of non-compositionality, or idiom status, of multi-word units (cf. Ebeling & Hasselgård 2015b: 207). For the frequency-based, probabilistic approaches, on the other hand, phraseology is a characteristic feature of language, due to "the tendency of words to occur, not randomly, or even in accordance with grammatical rules only, but in preferred sequences" (Hunston 2002: 137, see also Groom 2005). As pointed out by Sinclair (1966: 411), "[there] are virtually no impossible collocations, but some are much more likely than others".[2]

Phraseology, drawing on the frequency-based approach, can therefore be understood as "… the preferred way of saying things in a particular discourse" (Gledhill 2000: 1):

> We should expect different written and spoken genres and different discourse communities to select or prioritise different phraseological patterns; the former on the grounds that they serve different communicative and institutional purposes and thus prioritise different rhetorical strategies … and the latter on the grounds that they are characterised by different ideational interests and interpersonal practices. (Groom 2005: 258)

For academic English learners, phraseological units peculiar to the academic genres, both in terms of their structure and their functional load, are the key to comprehension and fluency, as they reduce the processing effort (Nesselhauf 2005). For novice academics, the appropriate choice of phraseological patterns may also serve as an indicator of the degree to which students belong to the particular discourse community (Hyland 2008). In both respects, phraseology has been shown to be "one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners" (Granger & Bestgen 2014: 229).

There is a close methodological link between frequency-based, probabilistic approaches to phraseology adopted in this paper, and learner corpus research:

> Phraseology has established itself as an important feature of learner language research, and learner corpus research (LCR) in particular, since corpus analysis lends itself especially well to the study of recurrent multi-word units. (Ebeling & Hasselgård 2015b: 208)

Corpus-driven methods of data extraction and evaluation appear to be particularly well-suited for the identification of phraseological patterns since they "are more inductive, in that the corpus itself is the data and the patterns of language use it represents are noted as ways of expressing regularities and exceptions in language" (Callies 2015: 36). In the present paper, three corpus-driven methods will be tested as the starting points of examining the phraseology of novice academic writing: frequency lists, keywords, and lexical bundles.

When learning to write academic texts, L2 English learners have to face two kinds of challenge: the linguistic challenge of English as a foreign language, and the academic challenge of entering the discourse community of the particular discipline. I would like to investigate the impact of both on Czech novice writers of academic English. My main goal, therefore, is to explore the areas in which phraseology may distinguish between native speakers of English and advanced EFL learners on the one hand, and between novice (L1 and L2) and expert academic writers on the other.

After the data sources have been introduced in Section 2, each of the methods will be dealt with in a separate section: frequency lists (Section 3), keywords (Section 4), and lexical bundles (Section 5). The concluding section (6) compares and evaluates the three approaches.

2 The corpora

Three corpora of written academic texts were used as data sources. The corpora are comparable in terms of academic field (English literature), medium (written English) and date of origin of the majority of texts (early 2000's).[3] Two corpora comprise students' essays (L1 and L2), one papers published in academic journals. The composition of the corpora is shown in Table 1.

corpus	**VESPA-CZ**	**BAWE-EL**	**AP**
English	L2	L1	L1
academic level	novice	novice	expert
source	Charles University, Prague, Faculty of Arts, English Studies Programme, BA 2nd year students' essays	universities of Warwick, Reading and Oxford Brookes, Arts and Humanities – English, good-standard students' essays, undergraduate and taught masters level	academic journals English Literary Renaissance, Renaissance Studies, Shakespeare Quarterly
time	2016-19	2004-17	1978-2014
register	students' essays	students' essays	academic papers

corpus	VESPA-CZ	BAWE-EL	AP
academic field	English literature: Renaissance to Restauration	English literature	English literature: Renaissance
size: tokens (approx.)	106 600	226 300	235 000
number of texts	48	89	34

Table 1: The three corpora used in the present study: VESPA-CZ, BAWE-EL and AP

The corpus of academic texts written by Czech advanced learners of English (VESPA-CZ), now about half the size of the L1 corpora used, is being compiled at the Faculty of Arts, Charles University, as a part of the international project on The Varieties of English for Specific Purposes Database learner corpus (VESPA).[4] The project, initiated by the Centre for English Corpus Linguistics (CECL) at Université catholique de Louvain, aims to build a corpus of English for Specific Purposes texts written by L2 writers from various mother tongue backgrounds in a wide range of disciplines and genres. The Czech section of the corpus comprises literature essays and linguistics term papers written by students of English Studies programmes at the Faculty of Arts, Charles University. Only the English literature Bachelor's essays were used for the research presented in this study.

The L1 corpus of novice academic writing I draw on in the present study, BAWE-EL, is a sub-corpus of the British Academic Written English Corpus,[5] which comprises good-standard student assignments written at British universities. The sub-corpus comprises essays in English Literature.

The corpus of published academic papers dealing with English Renaissance literature (AP) was compiled at the Faculty of Arts, Charles University (Tomešová 2017) as a reference corpus for the two students' corpora. The corpora were analysed using AntConc, a freeware corpus analysis toolkit for concordancing and text analysis (Anthony 2019).

3 Starting from frequency lists

As pointed out by Hunston (2002: 67), "comparing the frequency lists for two corpora can give interesting information about the differences between the texts comprising each one". This part of the study takes merely a sub-section of a frequency list as its starting point: it deals with adverbs ending in *–ly* in the three corpora. The *–ly* adverbs were chosen as possible sites of difference among academic writers due to their high representation (about 55% of common adverbs) and productivity in academic prose (Biber et al. 1999, Granger & Rayson 1998). Moreover, *–ly* adverbs are functionally diverse, occurring within all syntactic classes of adverbs (Huddleston & Pullum, 2002). From the practical point of view, *–ly* adverbs are easy to search for in a corpus which has not been POS-tagged.

As far as the relative frequency of *–ly* adverbs is concerned, a scalar increase in the number of tokens per 100,000 words can be observed, with the Czech students using the adverbs least frequently (1148 tokens), followed by the L1 novice writers (1252 tokens), and the expert writers (1430 tokens). An explanation for this tendency is to be sought in the representation of the individual syntactic classes of adverbs in the three corpora. Following (Hasselgård 2015), the adverbs were divided into adjuncts (*easily*, *usually*), disjuncts (*clearly*, *possibly*), conjuncts (*finally*, *consequently*), focus adverbs (*mainly*, *merely*), approximators (*partly*, *approximately*) and modifiers (*highly*, *extremely*). Their representation in the corpora is given in Figure 1.

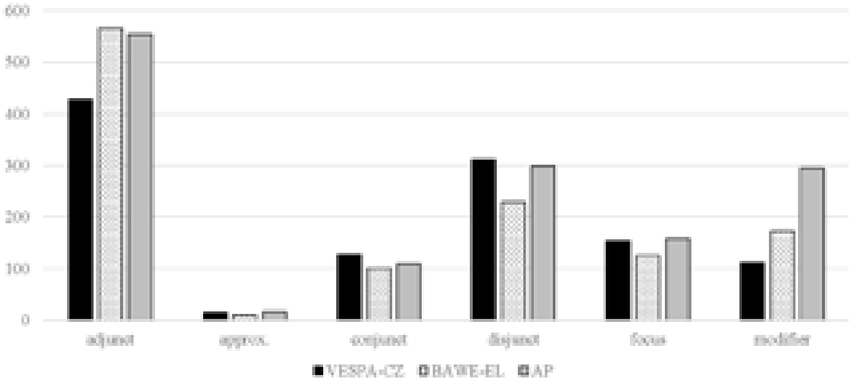

Figure 1: Syntactic classes of *–ly* adverbs

Figure 1 shows that the frequency of modifiers in the three corpora follows the same tendency as that of *–ly* adverbs in general. A closer look, however, reveals that the differences are not merely quantitative. Table 2 lists the most frequent *–ly* modifiers in the three corpora. The percentages in brackets indicate how frequent the particular adverb is within the class of modifiers in the corpus (e.g. the adverb *equally* constitutes 4.1% of modifiers in AP), suggesting that, compared to expert writers, novice writers rely on a more restricted range of adverbs, which they use frequently. The number of modifiers shared by the L1 writers, expert and novice (highlighted in bold in Table 2), is higher than that shared by the novice writers, L1 and L2 (underlined). The only modifier used by all groups is *highly*. What is most interesting is the repertoire of adverbs: L2 writers tend to use modifiers typical of spoken rather than written language (Granger & Rayson 1998) and degree adverbs not peculiar to academic discourse, e.g. *absolutely powerful*, *a completely different play*, *slightly different* (see also similar findings of Granger 1998, Biber 2006). The L1 published papers, on the other hand, display phrasal complexity typical of structurally 'compressed' academic writing (Biber & Gray 2010), e.g. *other equally important aspects of the play*, *the potentially narrow binding powers of economic obligation*.

L1 expert (AP)	***equally*** (4.1%), ***particularly*** (2.9%), ***highly*** (2.7%), ***entirely*** (2.6%), ***increasingly*** (2.3%), *potentially* (2%), *purely* (2%)
L1 novice (BAWE)	***highly*** (8.6%), ***particularly*** (7.2%), ***increasingly*** (5.9%), ***entirely*** (5.2%), ***equally*** (4.4%), *extremely* (4.2%), *completely* (4.0%)
L2 novice (VESPA)	*completely* (9%), ***highly***, *slightly* (7.4%), *purely* (6.6%), *extremely* (5.7%), *directly*, *significantly* (4.1%)

Table 2: The most frequent –*ly* modifiers

The frequency data could be used in a similar way to initiate further qualitative analyses, but I will rather proceed to another quantitative corpus-driven starting point.

4 Starting from keywords

The term keyword will be employed here in the way it is generally used in corpus linguistics, i.e. as "a term for a word that is statistically characteristic of a text or a set of texts" (Culpeper & Demmen 2015: 90). An important advantage of working with quantitatively defined keywords is the fact that "delegating the initial task of identifying items for analysis to a computer algorithm ensures that this stage of the research is completely insulated from researcher bias" (Groom 2010: 60). Groom (ibid: 63) has shown that the typical patterns surrounding grammatical keywords can reveal both "the preferred meanings of a particular discourse community" and "the preferred stylistic features associated with this community". I will adopt a similar approach and take closed-class grammatical keywords as the basis for further qualitative analysis, looking in detail at their phraseological behaviour.

I used keywords to try to answer the question of what constructions expert academic writers employ that L2 novice writers lack in their essays. The AP corpus was therefore used as the study corpus with the learner VESPA-CZ as the reference corpus. The most salient grammatical keywords[6] were *of*, *that*, *to*, and *for*. The second step of the analysis consisted in identifying 3-4 word lexical bundles, i.e. recurrent uninterrupted multi-word strings[7] which comprise these keywords (see Table 3).

keyword	**typical lexical bundles**
of	*in terms of, a kind of, a form of, one of the*
that	*the fact that*
to	*in relation to, appear / be believed / turn out / prove / seem / be taken to be*
for	*account for*

Table 3: Grammatical keywords in expert academic papers (study corpus = AP, reference corpus = VESPA-CZ) and typical lexical bundles comprising them

A detailed look at the concordance lines of these lexical bundles suggests three areas in which expert academic writers differ from L2 novices. Academic papers employ a range of specific multi-word linking devices, such as *in terms of*, *in relation to*, which make it possible to structure the text and express connections among ideas more explicitly (Example 1). Students seldom use these lexical bundles.

(1) *The key word here,* ***in relation to*** *the exploration of female licence within intertextuality, is 'artificial'.* (AP)

The second area of difference is related to hedging strategies (cf. Hyland 2004, 2005, Hyland & Tse 2004). The lexical bundles *a kind of*, *a form of*, *one of the*, and the 'verb *to be*' constructions (e.g. *appear* / *be believed* / *seem to be*) are used frequently by expert academic writers (Example 2), while students appear to use other means to hedge their statements, such as modal verbs or the '*it can be said* / *argued that*' constructions (see Section 5).

(2) *...although in his account this* ***appears to be*** *a primarily masculine phenomenon.* (AP)

A lexical bundle frequent in academic papers, which the students seem unaware of, turned out to be *account for* (Example 3).

(3) *Mere infatuation cannot* ***account for*** *the complex, contradictory registers of weakness, resentment, and disempowerment that criss-cross this poem.* (AP)

5 Starting from lexical bundles

The last corpus-driven approach that will be tested here draws on lexical bundles which comprise no predefined lexical component. The study of lexical bundles is considered "rewarding" by Ebeling and Hasselgård (2015a: 88) since they make it possible to investigate “patterns of lexis in student writing” and their functions. A number of studies have described differences in the use of lexical bundles between native speakers and learners. Advanced learners were found to use “fewer and far less varied lexical bundles than native speakers”, rely less on stance bundles, and produce more organizational bundles than native speakers do (Ädel & Erman 2012: 82, 90).

For the analysis presented in this paper, 4-word lexical bundles with a minimum frequency of four tokens per 100,000 words and dispersion range 2 were extracted from the three corpora. Bundles comprising topic-specific words, i.e. ‘content bundles’, e.g. *the early modern period*, *of the merchant of*, were disregarded (cf. Chen & Baker 2010, Ädel & Erman 2012). The bundles which meet these criteria and their relative frequencies are presented in Table 4.

L1 expert (AP)	**the end of the** (14.5), **at the same time** (12.8), **at the end of** (11.9), at the heart of (7.7), **as well as the** (6.8), in the context of (6.8), on the one hand (6.8), *the ways in which* (6.8), in the face of (6.4), **the fact that the** (6), as a form of (5.5), **at the beginning of** (5.5), *the way in which* (5.5), in terms of the (4.7), as a kind of (4.3), by the fact that (4.3), in a way that (4.3), *it is possible to* (4.3), **on the other hand** (4.3), **the beginning of the** (4.3), the nature of the (4.3)
L1 novice (BAWE)	*the way in which* (23.4), **the end of the** (19.4), **at the end of** (19), **on the other hand** (15.5), *it is possible to* (13.3), the use of the (11), **the fact that the** (10.2), **at the beginning of** (9.7), the ways in which (9.3), **the beginning of the** (8.8), through the use of (8.8), it could be argued (8.4), the rest of the (8), could be argued that (7.5), **at the same time** (7.1), can be seen in (6.6), that there is no (6.6), the extent to which (6.2), to the fact that (6.2), way in which the (6.2), the importance of the (5.7), allows the reader to (5.3), as can be seen (5.3), by the use of (5.3), the role of the (5.3), the structure of the (5.3), in the form of (4.9), in the light of (4.9), is an example of (4.9), as well as the (4.4), example of this is (4.4), in contrast to the (4.4), it is clear that (4.4), that there is a (4.4), *the nature of the* (4.4), with the use of (4.4)
L2 novice (VESPA)	**on the other hand** (43.1), **the end of the** (20.6), **at the same time** (16.9), **at the end of** (15), **the beginning of the** (13.1), **as well as the** (11.2), can be found in (11.2), one of the most (10.3), **at the beginning of** (8.4), can be seen as (7.5), does not have to (7.5), **the fact that the** (7.5), the role of a (7.5), as a way of (6.6), in the case of (6.6), the course of the (6.6), as a means of (5.6), be found in the (5.6), but at the same (5.6), it is clear that (5.6), the form of the (5.6), the other hand is (5.6), the role of the (5.6), the use of the (5.6), and at the same (4.7), as well as in (4.7), in the course of (4.7), in the form of (4.7), in this case the (4.7), is one of the (4.7), is presented as a (4.7), it is important to (4.7), it is obvious that (4.7), not be able to (4.7), not have to be (4.7), the contrast between the (4.7), the rest of the (4.7), the structure of the (4.7), to the fact that (4.7)

Table 4: 4-word lexical bundles in the three corpora

What is immediately apparent from Table 4 is the difference in the number of lexical bundles employed by each group of writers. This may be seen as a response to the need to strike a balance between idiomaticity and variation, which is particularly challenging for students. Writers of academic texts need to use genre-specific patterns, in accordance with the "genre-specific purposes" (Groom 2005) and stylistic requirements of the discipline and the common practices of the discourse community. At the same time, they aim at a stylistically varied (native-like, for English L2 students) expression. Novice writers, and Czech learners in particular, can be seen to employ a broader range of recurrent bundles, with the top ones occurring with high frequencies. The same observation was made by Hasselgård (2019: 347), who explored English academic texts written by Norwegian students: "learners tend to re-use a small number of bundles to a greater extent than native speakers".[8]

Highlighted in bold are those bundles which occur in all three corpora, albeit with different relative frequencies. Czech novice writers tend to overuse a relatively small number of 'academic' bundles, employing them with frequencies exceeding those attested in published papers. *On the other hand*, for instance, occurs in Czech students' texts ten times more frequently than in the experts' papers. A similar tendency can be observed in the essays written by L1 students, even though the occurrence rates are lower. In Table 4 the overlap between the more experienced and novice L1 writers is indicated in italics, the lexical bundles shared by L1 and L2 novice writers are underlined.

The lexical bundles were classified on the basis of their function into three classes, drawing on Ebeling and Hasselgård (2015a)[9]: bundles presenting and discussing content (also referred to as ideational, informational, research-oriented, or referential bundles), bundles organizing discourse (textual, or text-oriented bundles), and bundles expressing attitudes (interpersonal, participant-oriented bundles). The lexical bundles shared by the three groups of writers were found to perform text-oriented functions (*on the other hand*, *as well as the*), sometimes overlapping with referential ones (e.g. *the end of the*, *at the same time*). L1 writers, both experienced ones and students, use descriptive, research-oriented patterns *the way(s) in which*, *the nature of the*, and the bundle *it is possible to* related to expressing attitudes. There is, however, a difference in the extent to which the interpersonal bundle *it is possible* is used: L1 novice writers employ it three times more frequently than expert writers (Example 4). In both L1 corpora it tends to co-occur with other stance expressions (e.g. *I think*, *I hope* in Example 5).

(4) ***It is possible to** suggest that in fact T.S. Eliot does not invite the reader to understand the text. However, **it is possible** to 'make sense' of the poem through an understanding of how the text was written, and how it came to be written.* (BAWE-EL)

(5) ***I think it is possible to** interpret the play in such a way as to provide an affirmative answer - indeed, **I hope** my reading has done so. (*AP)

More stance-oriented patterns can be observed among the lexical bundles used by both groups of student writers. Novice writers, whether English or Czech L1, employ numerous bundles comprising modal verbs: *it could be argued*, *could be argued that*, *can be seen in*, *as can be seen*, *can be found in*, *can be seen as*, *does not have to*, *not be able to*, *not have to be* (Example 6), as well as bundles with adjectives or nouns expressing attitude: *the importance of the*, *it is important to*, *it is obvious that*, *one of the most* (Example 7).

(6) *Second possible way of dealing with mutability is seeing it as the mighty master, who measures the lovers' time. Here it **can be seen as** the long and gradual progression of life ended by the greatest change of all, death.* (VESPA-CZ)

(7) *However, **it is important to** point out that generally the rhyme scheme of the final sestet tends to be flexible…* (VESPA-CZ)

These results are in accordance with those of Paquot et al. (2013: 385), who claim that EFL learners are "generally more overtly present within their texts than native students", but the frequency of visibility markers decreases in their academic-like writing.

6 Conclusions

Even though limited in its extent and depth, the present study has pointed out several areas of difference between academic texts written by expert and by novice academic writers on the one hand, and between English L1 and advanced English L2 texts. Due to the relatively high level of proficiency of the Czech students, the novice-expert dimension of difference appears to play a more prominent role than the English L1-L2 one. The divergences between the two groups of novice writers are often merely a matter of degree when compared to the published academic papers.

Learners were found to overuse conjuncts (*finally*) and several discourse-organizing lexical bundles (*on the other hand*). They, however, underuse complex phrasal linking expressions (*in terms of*, *in relation to*). The lower complexity of learners' writing is also reflected in their underuse of modifiers; students tend to rely on a limited repertoire of modifiers which are not peculiar to academic discourse (*completely/slightly different*). Novice academic writers tend to use other forms of attitudinal expressions and hedges than expert ones: there are more stance adverbs (*interestingly*) and bundles comprising modal verbs (*can be seen*) and evaluative adjectives (*it is obvious that*) in students' essays than in published papers. The analysis based on keywords, on the other hand, revealed a stance pattern typical of expert academic papers, *appear / be believed / be taken / seem to be*. More generally, the results corroborate the findings of other studies (especially Hasselgård 2019) that learners tend to overuse a limited number of multi-word units characteristic of academic discourse.

I hope to have shown that corpus-driven methods can indeed serve as effective starting points for the study of the features of English L2 academic texts. A combination of methods, such as analysis of keywords and lexical bundles, may provide a more comprehensive view than a single approach. Nevertheless, as pointed out by Ebeling and Hasselgård (2015b: 217), "[the] relative phraseological success of learners should be studied both quantitatively and qualitatively. A multi-word unit may be grammatically correct, but inappropriate in the context in which it occurs".

Since English L2 novice academic writers have to face two types of challenge (EFL and EAP), a comparison of their texts with those produced by two types of academic authors (novice and expert) can reveal the relative weight of each of the factors. Another factor which may be worth exploring is the potential influence of the learners' L1. Longitudinal studies of learners' writing could also point out areas requiring more attention when designing teaching materials. Both these extensions, however, would require specific corpora large enough to allow the use of corpus-driven methods. The research based on corpora of learner academic English may, hopefully, help map the area and eventually lead to new pedagogical applications drawing on a phraseological approach to L2 teaching.

Notes

[1] This work was supported by the Czech Science Foundation grant 19-05180S Phraseology in English academic texts written by Czech advanced learners: a comparative study of learner and native speaker discourse.

[2] Cf. also Sinclair's "idiom principle": "The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments." (Sinclair 1991: 110)

[3] I am aware of the fact that the results may be affected by the difference in size between VESPA-CZ on the one hand, and the L1 corpora on the other. The preliminary results presented here will have to be revisited once the L2 corpus has reached the size of its L1 counterparts. The size of the L2 corpus, however, appears sufficient to draw some conclusions despite these limitations. I also believe that the composition of the published academic papers corpus did not affect the results, since most of its texts were published between 1990 and 2014.

[4] https://uclouvain.be/en/research-institutes/ilc/cecl/vespa.html

[5] https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

[6] The statistical method used was Log-likelihood, with significance the threshold set at $p < 0.05$.

[7] Only the lexical bundles with a minimum frequency of 5, occurring in at least 2 texts, and the difference between their frequency in the two corpora significant at the level of $p < 0.05$ were considered.

[8] Hasselgård (2019: 358) uses the term 'phraseological teddy bears' to refer to the bundles which "are much more frequent in English L2 than in English L1, and […] seem to have generalized their meanings and discourse functions by being used in contexts where native speakers prefer other expressions".

[9] The functional taxonomy employed in Ebeling and Hasselgård (2015a) draws on Ädel and Erman (2012), Chen and Baker (2010), Cortes (2004), and Biber et al. (2004).

References

Ädel, A. and Erman, B. (2012) 'Recurrent word combinations in academic writing by native speakers and non-native speakers of English: A lexical bundles approach.' *English for Specific Purposes 31*(2), 81-92.

Anthony, L. (2019) AntConc (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software.

Biber, D. and Gray, B. (2010) 'Challenging stereotypes about academic writing: Complexity, elaboration, explicitness.' *Journal of English for Academic Purposes 9*, 2-20.

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

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English.* Harlow: Pearson.

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

Callies, M. (2015) 'Learner corpus methodology.' In: Granger, S., Gilquin, G. and Meunier, F. (eds) *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press. 35-55.

Čermák, F. (2008) *Frazeologie a idiomatika česká a obecná.* Prague: Karolinum.

Chen, Y. H. and Baker, P. (2010) 'Lexical bundles in L1 and L2 academic writing.' *Language Learning and Technology 14*(2), 30-49.

Cortes, V. (2004) 'Lexical bundles in published and student disciplinary writing: Examples from history and biology.' *English for Specific Purposes 23*(4), 397-323.

Cowie, A. P. (ed.) (1998) *Phraseology: Theory, Analysis, and Applications.* Oxford: Clarendon Press.

Culpeper, J. and Demmen, J. (2015) 'Keywords.' In: Biber, D. and Reppen, R. (eds) *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press. 90-105.

Ebeling, S. O. and Hasselgård, H. (2015a) 'Learners' and native speakers' use of recurrent word-combinations across disciplines.' *LCR2013 Conference Proceedings. BeLLS 6*, 87-106.

Ebeling, S. O. and Hasselgård, H. (2015b) 'Learner corpora and phraseology.' In: Granger, S., Gilquin, G. and Meunier, F. (eds) *The Cambridge Handbook of Learner Corpus Research.* Cambridge: Cambridge University Press. 207-230.

Gledhill, C. (2000) *Collocations in Science Writing.* Tübingen: Gunter Narr.

Granger, S. and Rayson, P. (1998) 'Automatic profiling of learner texts.' In: Granger, S. (ed.) *Learner English on Computer.* London and New York: Longman. 119-131.

Granger, S. and Bestgen, Y. (2014) 'The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study.' *IRAL - International Review of Applied Linguistics in Language Teaching 52*(3), 229-252.

Granger, S. (1998) 'Prefabricated patterns in advanced EFL writing: Collocations and formulae.' In: Cowie, P. (ed.) *Phraseology: Theory, Analysis, and Applications.* Oxford: Oxford University Press. 145-160.

Gray, B. and Biber, D. (2015) 'Phraseology.' In: Biber, D. and Reppen, R. (eds) *The Cambridge Handbook of English Corpus Linguistics.* Cambridge: Cambridge University Press. 125-145.

Groom, N. (2005) 'Pattern and meaning across genres and disciplines: An exploratory study.' *Journal of English for Academic Purposes 4*(3), 257-277.

Groom, N. (2010) 'Closed-class keywords and corpus-driven discourse analysis.' In: Bondi, M. and Scott, M. (eds) *Keyness in Texts.* Amsterdam and Philadelphia: John Benjamins. 59-78.

Hasselgård, H. (2019) 'Phraseological teddy bears Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English.' In: Wiegand, V. and Mahlberg, M. (eds) *Corpus Linguistics, Context and Culture.* Berlin: De Gruyter. 339-362.

Hasselgård, H. (2015) 'Lexicogrammatical features of adverbs in advanced learner English.' *International Journal of Applied Linguistics 166*(1), 163-189.

Huddleston, R. and Pullum, G. K. (2002) *The Cambridge Grammar of the English Language.* Cambridge: Cambridge University Press.

Hunston, S. (2002) *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hyland, K. (2004) *Disciplinary Discourses: Social Interactions in Academic Writing.* Ann Arbor: University of Michigan Press.

Hyland, K. (2005) *Metadiscourse.* London: Continuum.

Hyland, K. (2008) 'As can be seen. Lexical bundles and disciplinary variation.' *English for Specific Purposes 27*(1), 4-21.

Hyland, K. and Tse, P. (2004) 'Metadiscourse in academic writing: A reappraisal.' *Applied Linguistics 25*(2), 156-177.

Kjellmer, G. (1991) 'A mint of phrases.' In: Aijmer, K. and Altenberg, B. (eds) *English Corpus Linguistics. Studies in Honour of Jan Svartvik.* London and New York: Longman. 111-127.

Nesselhauf, N. (2005) *Collocations in a Learner Corpus.* Amsterdam: John Benjamins.

Paquot, M., Hasselgård, H. and Ebeling, S. O. (2013) 'Writer/reader visibility in learner writing across genres. A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora.' In: Granger, S., Gilquin, G. and Meunier, F. (eds) *Twenty Years of Learner Corpus Research: Looking Back, Moving Ahead.* Louvain-la- Neuve: Presses universitaires de Louvain. 377-387.

Sinclair, J. M. (1966) 'Beginning the study of lexis.' In: Bazell, C. E., Catford J. C., Halliday, M. A. K. and Robins, R. H. (eds) *In Memory of J. R .Firth.* London: Longman. 410-430.

Sinclair, J. M. (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Tomešová (Cilcová), K. (2017) *Result/inference Discourse Connectives in Academic Texts.* Unpublished Mgr. thesis. Prague: Filozofická fakulta. Univerzita Karlova.

Markéta Malá is associate professor of English Linguistics at Charles University, Prague, Czech Republic. The main focus of her work is on using corpora of naturally occurring language data to explore patterning in the English language in contrast with Czech. She is currently carrying out research on phraseology in advanced learners' academic English and on the language of children's literature. Her publications deal mainly with contrastive linguistics, corpus linguistics, and academic English.

Address: Markéta Malá, Department of Linguistics, Faculty of Arts, Charles University. Nám. Jana Palacha 2. 116 38 Praha 1. Czech Republic. [e-mail: Marketa.Mala@ff.cuni.cz]