



Phraseological Sequences Ending in *of* in L2 Novice Academic Writing

Kateřina Vašků¹ , Gabriela Brůhová² ,
and Denisa Šebestová² 

¹ Department of English Language and ELT Methodology, Charles University, nám. J. Palacha 2, Prague, Czech Republic

katerina.vasku@ff.cuni.cz

² Charles University, Prague, Czech Republic

Abstract. This study explores the use of recurrent four-word sequences ending in *of* in English L2 novice academic writing. Our aim is to identify to what extent and in what ways the use of phraseological patterns differs in academic texts written by L2 novice academic writers from texts authored by professional L1 academic writers. Both actual four-word sequences and structural patterns are investigated. The results show that Czech novice academic writers in the field of English literature are able to use a wide range of multi-word sequences and patterns. The frequency of the main structural patterns is very similar in the two corpora, with the prepositional sequence [prep det N *of*] representing by far the most frequent type, followed by nominal sequences [det adj/num N *of*], and verbal sequences [V det N *of*]. However, it is the prepositional type that displays most differences between the learners and native speakers, especially the use of complex prepositions. The functional analysis of the sequences has shown that the discourse functions of sequences are similar in both languages; nevertheless, Czech L2 novice academic writers tend to overuse sequences containing less advanced lexical items with transparent meaning. Pedagogical applications of the results should include improvements of pedagogical tools by increasing emphasis on advanced and semantically more complex phraseological sequences.

Keywords: Learner corpus · Phraseology · Multi-word sequences · *Of* · Structural patterns · Functions

1 Introduction

This study explores the use of multi-word sequences in L2 novice academic writing. Although recurrent multi-word sequences have been referred to by many different terms in literature, e.g. clusters (Scott 1996), recurrent word combinations (Altenberg 1998), lexical bundles (Biber et al. 1999; Cortes 2004) and n-grams (Granger and Bestgen 2014; Rayson 2015), all these approaches share the common idea that the use of multi-word sequences is crucial in language production. In other words, language users rely relatively heavily on “combinations of words that customarily occur” (Kjellmer 1991: 112). It has been demonstrated that the usage of multi-word sequences “unmistakably distinguishes native speakers of a language from L2 learners” (Granger and Bestgen 2014; cf. also

Pawley and Syder 1983; Ebeling and Hasselgård 2015). Similarly, Hyland (2008: 4) points out that “[m]ulti-word expressions are an important component of fluent linguistic production and a key factor in successful language learning”.

The focus of this study is on four-word sequences ending in *of* in L2 novice academic writing. Our aim is to identify to what extent and in what ways the use of phraseological patterns differs in academic texts written by L2 novice academic writers from texts authored by professional L1 academic writers. Previous research has suggested that language produced by advanced L2 speakers can be influenced by their limited lexical and phraseological choices (Granger 2017: 9). Hence, the present study intends to contribute towards developing a phraseology-informed approach to language instruction.

Generally, L2 language users show the tendency to use a less varied repertoire of multi-word sequences in comparison with L1 speakers, employing the same sequences more frequently (Garner 2016: 33), and in contexts where native speakers would favour a different expression. This may be explained by the fact that L2 speakers tend to feel less certain when using a foreign language, and therefore “regularly clutch for the words [they] feel safe with” (Hasselgren 1994: 237). Hasselgren describes the words favoured by L2 speakers as “lexical teddy bears”, while Ellis introduces the term “phrasal teddy bears” for “[h]ighly frequent and prototypically functional phrases like *put it on the table, how are you?, it’s lunch time*” (Ellis 2012: 29). Hasselgård (forthcoming) proposes the term “phraseological teddy bears” for multi-word units “used more frequently and in more contexts” in the language of L2 speakers when compared with those used by native speakers.

2 Material and Method

In our analysis we employ a custom-made corpus of essays written by students of English at Charles University whose L1 is Czech (here referred to as the L2 corpus).¹ These essays are credit assignments written in a literary studies seminar. As a reference corpus, we have compiled a corpus of papers published in academic journals written by professional literary critics who were native English speakers (the L1 corpus). The size of the English corpus is 234 877 tokens; the Czech corpus contains 106 668 tokens, being approximately half the size of the L1 corpus (see Table 1). While it may be argued that the two corpora under examination differ markedly in the authors’ language proficiency as well as the amount of professional experience and training that they have likely received, we base our approach on the assumption that the university students are in fact aspiring to become proficient users of academic English in the field of literary studies. Our study seeks to identify differences between L1 and L2 writers’ writing, aiming to use the results to inform future language instruction. Therefore, our L1 corpus represents the students’ target register, providing a reasonable *tertium comparationis* for our analysis.

¹ This collection is part of the VESPA corpus, currently under construction at the Department of English Language and ELT Methodology, Faculty of Arts, Charles University. For more information about the VESPA project see <https://uclouvain.be/en/research-institutes/ilc/cccl/vespa.html>.

Table 1. The two corpora used for the analysis

Corpus	L1 corpus	L2 corpus
Tokens total	234877	106668
Texts total	48	34
Four-word sequences retrieved	233	152
Four-word sequences excluded	19	41
Four-word sequences analyzed	214	111
Tokens of four-word sequences in the sample	669	295

A keyword analysis (using AntConc 3.5.8; Anthony 2019) of the L1 corpus (with the L2 corpus as a reference corpus) has revealed that one of the most underused words in the L2 corpus is the preposition *of* (the log-likelihood value is significant at the level of $p < 0.05$). As pointed out by Groom (2010: 63), “*of* constitutes an excellent test-bed for the claim that closed-class keywords are tractable to qualitative semantic analysis”. We have therefore decided to focus on four-word sequences containing *of*, limiting our research to four-word sequences having *of* as their final element². We investigate both actual four-word sequences and the structural patterns (i.e. general structures of the retrieved sequences based on a word-class analysis of their constituents). It may be expected that some sequences and structural patterns will be underused or overused in the L2 corpus.

In the present study, we focus on four-word sequences “because they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles” (Hyland 2008: 8). Recurrent four-word sequences ending in the preposition *of* were automatically extracted by means of AntConc 3.5.8 and then analyzed both quantitatively and qualitatively. Four-word sequences that were considered “specific to particular topics or tasks” (Hasselgård, forthcoming) were excluded, as it is unlikely that they would occur in other kinds of texts (e.g. *The decision of the trial in The Merchant of Venice...* the L1 corpus). 19 four-word sequences were excluded from the L1 corpus and 41 from the L2 corpus.

First, the sequences were categorized structurally, i.e. in terms of their grammatical pattern. Next, the sequences were analyzed functionally. On the basis of previous classification proposed by Biber et al. 2004 we distinguish three primary discourse functions performed by four-word sequences³: referential sequences, discourse organizers and stance sequences. Referential sequences “make direct reference to physical or abstract entities, or to the textual context itself” in order to identify the entity or describe some particular attribute of the entity as especially important (Biber et al. 2004: 384). Four subtypes of referential sequences can be distinguished: identification/focus, imprecision

² As pointed out by Hunston (2008: 272), the study of these “small words” may prove useful in that they reveal characteristic features of “specialised corpora”, while also playing an important role in structuring the text, being involved in grammar patterns.

³ Biber et al. (2004) use the term *bundle* for what we describe as sequences. The term *lexical bundles* is defined in Biber et al. (1999: ch. 13) as the most frequent recurring word combinations in a given register, their minimum frequency being 10 IPM (ibid.: 989).

indicators, specification of attributes, time/place/text reference. Discourse organizers “reflect relationships between prior and coming discourse” (ibid.). Stance bundles “express attitudes or assessments of certainty that frame some other proposition” (ibid.).

3 Analysis

Sequences obtained by the search were classified according to the formal pattern of the sequence based on word classes.

Table 2. Structural classification of four-word sequences

Formal pattern	L1 corpus					L2 corpus				
	Types		Tokens			Types		Tokens		
	Raw freq.	%	Raw freq.	Freq. per 100k	%	Raw freq.	%	Raw freq.	Freq. per 100k	%
prep det N <i>of</i>	146	68.2	499	212.5	74.6	57	51.4	174	163.1	59.0
det adj/num N <i>of</i>	36	16.8	90	38.3	13.5	22	19.8	53	49.7	18.0
V det N <i>of</i>	15	7.0	38	16.2	5.7	14	12.6	28	26.2	9.5
conj det N <i>of</i>	8	3.7	19	8.1	2.8	2	1.8	4	3.7	1.4
other	9	4.3	23	9.8	3.4	16	14.4	36	33.7	12.2
Total	214	100	669	284.8	100	111	100	295	276.6	100

As follows from Table 1, the most frequent pattern in both corpora is the prepositional type [prep det N of] with 68% in the L1 corpus and 51% in the L2 corpus (ex. 1), followed by the nominal type [det adj/num N of] with 16.8% in the L1 corpus and 19.8% in the L2 corpus (ex. 2–3). The third most common is the verbal type [V det N of], which appears to be more prominent in the L2 corpus (13%) compared with 7% in the L1 corpus (ex. 4). The less common pattern [conj det N of] is illustrated by example 5.

- (1) The moral deprivation of the lyric voice **at the end of** the second group of poems is followed by three translations from Horace, Catullus and Seneca. (L1 corpus)
- (2) **A close examination of** the language of the two plays reveals that Shakespeare and Jonson frequently employ similar metaphors. (L1 corpus)
- (3) Yet **the second half of** the couplet which closes the scene registers his pride in his own skill... (L1 corpus)
- (4) That **is the case of** e.g. Lorenzo and Jessica when they confess their love for each other. (L2 corpus)
- (5) This poem portrays the change **and the end of** a relationship between two lovers. (L2 corpus)

The remaining group of patterns (marked as *other* in Table 1) contains instances of patterns which occurred only once in either corpus, for instance *point of view of* or *not a sign of*. It is noteworthy that the L2 corpus contains a significantly larger proportion of such instances (14% in L2 vs. 3% in L1), which might suggest that learners of English tend to rely on the open choice principle rather than on the idiom principle (cf. Sinclair 1991). The following sections focus on the three most frequent patterns, exploring similarities and differences between the use of phraseological sequences by L1 and L2 speakers of English.

3.1 Prepositional Type

As follows from Table 2, the prepositional type is by far the most common type of four-word sequences in the two corpora. The most frequent prepositions in the pattern are represented by *in* (38 sequences in the L1 corpus, 21 in the L2 corpus), *as* (22 sequences in L1, 10 in L2), *at* (12 sequences in L1, 5 in L2) and *of* (12 sequences in L1, 5 in L2).

Table 3 lists twenty-two most commonly used prepositional sequences in both corpora in decreasing order of frequency, with the number of tokens for each sequence

Table 3. The most frequent prepositional sequences

L1 corpus	Raw freq.	Per 100k	L2 corpus	Raw freq.	Per 100k
at the end of	24	10.2	at the end of	16	15.0
at the heart of	17	7.2	at the beginning of	9	8.4
in the context of	15	6.4	as a way of	7	6.6
in the face of	15	6.4	in the case of	7	6.6
at the beginning of	12	5.1	as a means of	5	4.7
as a form of	11	4.7	in the course of	5	4.7
as a kind of	10	4.3	in the form of	5	4.7
at the expense of	8	3.4	for the purposes of	4	3.7
at the hands of	8	3.4	in the end of	4	3.7
on the part of	8	3.4	in the eyes of	4	3.7
as a result of	7	3.0	of the idea of	4	3.7
by the end of	7	3.0	with the exception of	4	3.7
in the form of	7	3.0	as a symbol of	3	2.8
in the world of	7	3.0	as an act of	3	2.8
as a sign of	6	2.6	by the end of	3	2.8
in the history of	6	2.6	for the sake of	3	2.8
in the wake of	6	2.6	in a number of	3	2.8
as a way of	6	2.6	in the beginning of	3	2.8
as the basis of	5	2.1	in the world of	3	2.8
in the case of	5	2.1	on the role of	3	2.8
in the light of	5	2.1	with the use of	3	2.8
on the side of	5	2.1	within the structure of	3	2.8

(absolute frequency). As regards the frequencies of the most common prepositional sequences, they roughly correspond to each other (given that the L1 corpus is approximately twice as large as the L2 corpus). The items occurring in both corpora are marked in bold. The lists share only seven sequences: *at the end of*, *at the beginning of*, *by the end of*, *in the form of*, *in the world of*, *as a way of* and *in the case of*.

Perhaps more interesting are instances which either occur in one of the corpora only or are significantly underused in the second corpus. The following are the most prominent sequences from the L1 corpus which are underrepresented in the L2 corpus: *at the heart of*, *in the context of*, *in the face of*, *at the expense of*, *as a sign of*. It appears that these sequences commonly used in the L1 corpus may not be stored as phraseological units in L2 speakers' mental lexicon. Interestingly, the head nouns in these expressions are all used in the abstract, often metaphorical, sense, which even advanced learners of English may find difficult to use (ex. 6–8).

- (6) But at the same time, our critical task must be to uncover the fissures, paradoxes, and contradictions that lie **at the heart of** that economy. (L1 corpus)
- (7) Authors use these surrogates, who resist the violent father/master, to solve the problem of wifely obedience **in the face of** murder. (L1 corpus)
- (8) ...when the speaker embraces present pain **as a sign of** future pleasure... (L1 corpus)

Similarly, the L2 corpus contains several sequences which are underrepresented in the L1 corpus. The most significant instances are: *in the end of*, *in the beginning of*, *in the eyes of*, *of the idea of*. The first two sequences *in the end of* or *in the beginning of* represent a common error of L2 speakers, who may often confuse the preposition *at* with *in*. Apart from such instances of language errors, the list of the most frequent sequences overused in the L2 corpus contains *in the eyes of*, which occurs four times in the L2 corpus, but only once in the L1 corpus and its overuse in L2 is thus statistically significant ($p < 0.05$). Since there is a corresponding phrase in Czech, *v očíh + N_{genitive}*, the higher frequency in the L2 corpus might reflect a possible influence of the first language⁴ (cf. Hyland 2008: 20).

All prepositional sequences were checked for their discourse function based on the classification proposed by Biber et al. (2004). Both samples contain only sequences with the referential function (see Table 4).

On the basis of previous research (Biber et al. 2004: 398; Hyland 2008: 16) the high distribution of the referential function was expected, but the fact that discourse organizers and stance sequences were not attested in our corpora at all was surprising. This may be due to the nature of sequences containing the *of*-phrase fragment, which tend to be used “to focus readers on a particular instance or to specify the conditions under which a statement can be accepted, working to elaborate, compare and emphasise aspects of an argument” (Hyland 2008: 16). Out of the four possible subtypes of referential expressions mentioned by Biber et al. (2004), our analysis revealed that prepositional sequences in academic writing display predominantly two functions,

⁴ The Czech idiom *v očíh + N_{genitive}* has a frequency of 6.76 i.p.m in the Czech corpus syn2015 as opposed to 3.76 i.p.m. of *in the eyes of* in the BNC.

Table 4. Functions of prepositional sequences

Subtypes of referential sequences	L1	%	L2	%	Examples
Identification/focus	0	0	1	2	<i>as the one of</i>
Specification of attributes	115	79	36	63	<i>as a way of, for the purposes of</i>
Time/place/text reference	31	21	20	35	<i>at the end of, at the hands of</i>
Total	146	100	57	100	

namely specification of attributes (79% in the L1 corpus and 63% in the L2 corpus) and time/place/text reference (21% in the L1 corpus and 35% in the L2 corpus).

The most common function of prepositional sequences, i.e. specification of attributes, is to “identify specific attributes of the following head noun” (Biber et al. 2004: 395). Some of the sequences specify quantity or amount, e.g. *in a range of* or *into a set of* (ex. 9), the size and form of the following head noun, abstract characteristics or logical relationships in the text, e.g. *in the case of*, *as a means of*, *as a result of*, *as a way of*, *in the form of*⁵ (ex. 10–12). The last subtype expresses abstract characteristics of the following noun or specifies logical relationships in the text, and it is by far the most common function of referential expressions in both corpora. Hyland (2008) describes sequences with this function as “framing signals”, used “to frame arguments by highlighting connections, specifying cases and pointing to limitations” (Hyland 2008: 16).

- (9) Amor appears **in a range of** conflicting characterizations external to the speaker... (L1 corpus)
- (10) However, the reader always gains insight into the relationship between the two and, **in the case of** Petrarch and Spenser, the main focus falls on the power dynamics therein. (L2 corpus)
- (11) **As a result of** this invention, alchemy could be studied by any literate person. (L2 corpus)
- (12) Thus he pretends to be at the point of death **as a way of** convincing himself of his own immortality. (L2 corpus)

Both corpora also contain a significant number of prepositional sequences expressing location in time, place or the text. Due to the genre of our sample texts, however, we do not attempt to distinguish temporal, locative and textual reference, as many of the prepositional sequences of this type refer to the location of some point in a play, where the distinction between the situational and textual location is irrelevant. The most common sequence in both corpora is *at the end of* (ex. 13). Surprisingly, *at the beginning of* is the second most common sequence of this type only in the L2 corpus (ex. 14), while occupying the fifth position in the L1 corpus.

⁵ Biber et al. (2004) include the sequence *in the form of* among “tangible framing attributes” (i.e. expressions specifying the size and form of the following head noun). However, in our corpus this sequence is used only in its abstract meaning; therefore, we classified *in the form of* as an “intangible framing attribute” describing the abstract characteristics of the following noun (e.g. *in the form of restrictions/a trust/retrospection/his own mortality*).

- (13) *To read Kate’s speech as an ironic performance of submission should also take into account the continued intellectual acuity and physical power Petruccio retains **at the end of** the play. (L1 corpus)*
- (14) *The poet tells his addressee **at the beginning of** the poem... (L2 corpus)*

Some authors focus on the status of prepositional sequences as formally and functionally fixed phraseological units and emphasize their grammatical function. Granger and Paquot view them as textual phrasemes, “typically used to structure and organize the content (i.e. referential information) of a text or any type of discourse” (Granger and Paquot 2008: 42) describe them as complex prepositions, “grammatized combinations of two simple prepositions with an intervening noun, adverb or adjective” (ibid.: 44).

Similarly, Klégr (1997; 2002) regards some of the prepositional sequences that “tend to be fixed in form” (Klégr 1997: 62) as complex prepositions. Based on a set of syntactic criteria (e.g. restricted variability of form, replaceability with a lexicalized primary or secondary preposition, inability to function as an independent clause element), Klégr lists over 400 complex prepositions. When comparing our samples of prepositional sequences with Klégr’s list, we identified the following complex prepositions:

Table 5. Complex prepositions in the corpora

L1 corpus (23 instances)	L2 corpus (15 instances)
<i>in the context of, at the expense of, on the part of, as a result of, in the form of, as a sign of, in the wake of, in the case of, in the light of, on the side of, in a state of, in the absence of, on the basis of, at the cost of, for the sake of, in the midst of, in the service of, on the verge of, from the point of, in the aftermath of, in the hands of, in the matter of, with the exception of</i>	<i>in the case of, in the form of, in the eyes of, with the exception of, for the sake of, during the time of, for the purpose of, from the point of, in the hands of, in the manner of, in the middle of, in the space of, on the basis of, on the side of, with the help of</i>

If we compare Table 5 with the list of most frequent prepositional sequences (Table 3), we can notice a different distribution of complex prepositions in the two corpora. While complex prepositions in the L1 corpus show a stronger tendency to occur among the most frequent prepositional sequences (10 instances of complex prepositions among the items in Table 3, cf. ex. 15), the L2 corpus contains only 5 types of complex prepositions among the most frequent sequences, each of the remaining 10 types occurring only twice. We can therefore conclude that this type of prepositional sequences is underused in the L2 corpus.

- (15) Portia’s suitors are judged not **on the basis of** their wealth or goods, but in terms of personal and moral qualities, and it must be said, racial prejudice.

The overall results of the analysis of the prepositional sequences suggest that this type is indeed underused by L2 learners of English compared with the L1 corpus, where the range of different lexemes used as nouns within the prepositional sequences is much broader, with many highly advanced vocabulary items (ex. 16–19). In addition, some of the most common prepositional sequences found in the L1 corpus, although not included in the list compiled by Klégr (1997), could be also considered instances of complex prepositions, e.g. *at the heart of* and *in the face of*.

- (16) **In the wake of** extensive critical discussion revolving around the analysis of sonnet sequence personae, this, too, seems self-evident. (L1 corpus)
- (17) **In the midst of** this spatial dissonance, the presence of the island-mountain shows how the mythology of The Faerie Queene depends on examples taken from the poet's life in Ireland. (L1 corpus)
- (18) Her findings should be viewed **against the backdrop of** the rising number of women involved in litigation more generally... (L1 corpus)
- (19) In both versions of this origin myth, the island is a geographically peripheral place **on the cusp of** the known world. (L1 corpus)

As has been pointed out by Hasselgård (forthcoming), the reason for the underuse of these prepositional sequences may be that “most learners simply do not know them, or [...] they belong to a style level that the learners are not fully familiar with.”

3.2 Nominal Type

The nominal type [det adj/num N of] is the second most common type of four-word sequences in both corpora. Especially in the L1 corpus, there is a significant drop between the frequency of the prepositional and nominal type (see Table 1). The type includes sequences with a determiner followed by an adjective/numeral, a noun and the preposition *of*. In the L1 corpus, 30 sequences contain an adjective (ex. 20) and 6 sequences contain a numeral (ex. 21), while in the L2 corpus, 18 sequences contain an adjective and 4 sequences contain a numeral.

- (20) For Volumnia, martial honour is **the logical outcome of** maternal nurture. (L1 corpus)
- (21) **The first instance of** role-playing in the play is Jessica's dressing as a page. (L2 corpus)

The raw frequencies of nominal sequences are considerably lower than those of prepositional sequences, especially in the L1 corpus, where only three sequences reach the level of at least four occurrences in the corpus: *the second half of* (six occurrences), *the very act of* (five occurrences) and *a close examination of* (four occurrences). Since four occurrences in the L1 corpus should roughly correspond to two occurrences in the L2 corpus (given the size of the two corpora), all 22 nominal sequences in the L2 corpus have the corresponding frequency. This discrepancy between the L1 and L2 corpus may be caused by a limited vocabulary of English learners, or by the effect of “phraseological teddy bears” (cf. Hasselgård forthcoming). However, due to the low numbers of raw frequencies in both corpora these conclusions should be viewed as tentative.

All nominal sequences found in both corpora can be classified as referential. However, the sequence *a close examination of* (L1 corpus) could alternatively be classified as a discourse organizer, which is used to “provide overt signals [...] that a new topic is being introduced” (Biber et al. 2004: 391). As follows from Table 6, sequences in the sample either contribute to the specification of attributes or express time, place or text reference. The table also shows that the latter type is more prominent in the L2 corpus than in the L1 corpus.

Table 6. Functions of nominal sequences

Subtypes of referential sequences	L1 corpus	%	L2 corpus	%	Examples
Specification of attributes	26	72.2	13	59.1	<i>the first instance of, the individual features of, the very act of, a good deal of</i>
Time/place/text reference	10	27.8	9	40.9	<i>the second half of, the final act of, the very end of</i>
Total	36	100	22	100	

Similarly to the distribution of prepositional sequences (cf. Table 4), the most common function of specifying nominal sequences is to specify abstract characteristics or logical relationships in the text (ex. 22–23). In addition, our L1 corpus revealed 5 sequences specifying quantity (ex. 24). Two of them (*a good deal of, a great deal of*) can be seen as lexicalized phraseological units corresponding to the single-word quantifier *many*. Apart from these two sequences, the L1 corpus also contains the sequences *a wide range of* (three occurrences), *a particular set of* (two occurrences) and *the same set of* (two occurrences). By contrast, there are no sequences of this type in the L2 corpus.

- (22) Readers are advised imaginatively to invent (literally, to reinvent) the making of the poem, to reconstruct the conceptual design of its fictional landscape in order to profit in **the very act of** so doing, from its teaching. (L1 corpus)
- (23) “Sonnet 130” does the opposite: the focus is on **the individual features of** the lover, however, it does not serve to emphasize her beauty but rather to draw attention to her imperfections. (L2 corpus)
- (24) They also foster **a wide range of** quasi-religious or magical beliefs about the malevolent agency of objects that are not properly exchanged. (L1 corpus)
- (25) He is named as such **at the very end of the play** by the advocates... (L2 corpus)

It appears that L2 users tend to use more frequently sequences which refer to time, place or location in the text. Similarly to the corresponding function within the prepositional type, it is impossible to distinguish temporal, locative and textual reference, as many of the nominal sequences of this type refer to the location of some point in a play or its part (ex. 25).

3.3 Verbal Type

The analysis of four-word sequences revealed that the verbal type is represented by 15 sequences in the L1 corpus and 14 instances in the L2 corpus, but the overall distribution is very low since the vast majority of sequences is only represented by two occurrences. What is characteristic of the verbs in this pattern is that they are semantically empty, the most frequent verb being the copular *be* (ex. 26), see Table 7.

Table 7. Verbal sequences in the corpora

L1 corpus (15 instances)	L2 corpus (14 instances)
<i>is a form of, are a number of, is a figure of, is a kind of, is a parody of, is a sign of, is the body of, is the product of, is the site of, takes the form of, emphasizes the importance of, suggests the presence of, becomes a sign of, had the effect of, occupies the place of</i>	<i>is a cycle of, is a part of, is a way of, is an example of, is the case of, is the climax of, is the subject of, is the use of, adopts the role of, assume the role of, keep a memory of, playing the role of, share a number of, uses a metaphor of</i>

- (26) For Augustine idolatry **is a form of** forgetting, a failure to notice and honor this indicative relationship: a failure to interpret properly. (L1 corpus)
- (27) While it might initially seem that time **is the subject of** many of these poems it is in fact just as much, if not more, the mutability or inherent inconstancy produced by it. (L2 corpus)
- (28) She thinks that [...], which to her **is a sign of** Antony's regression.

As for the function of the verbal sequences, almost all function as referential expressions specifying attributes of the following noun. With the exception of two sequences expressing quantity, *are a number of* (L1 corpus) and *share a number of* (L2 corpus), all these sequences specify a form (ex. 26), abstract characteristics (ex. 27) or logical relationships (ex. 28). There are two instances of referential expressions referring to time, place or location in the text (*is the site of, occupies the place of*).

4 Conclusions

The present study explored four-word sequences ending in the preposition *of* in L2 novice academic writing in order to identify to what extent and in what ways the use of these multi-word sequences differs in academic texts written by L2 novice academic writers from texts written by professional L1 academic writers. The sequences were analyzed both from the structural and functional point of view.

The study has shown that the language of novice L2 academic writers and professional L1 writers displays both similar features and differences. Generally, the L2 corpus analysis proved that Czech novice academic writers in the field of English literature are able to use a wide range of multi-word sequences and patterns. As far as the use of structural patterns is concerned, it can be concluded that the frequency of the main structural patterns is very similar in the two corpora under examination, with the prepositional sequence representing by far the most frequent type. At the same time, it is

the prepositional type that displays the most differences between learners and native speakers. Our findings have shown that the most formally fixed prepositional sequences, i.e. complex prepositions, tend to be underused by Czech learners of English. Although some complex prepositions are found in the L2 corpus, the range of different lexemes used as nouns within the prepositional sequences is much narrower, with few advanced vocabulary items and predominantly with transparent meaning. The nominal and verbal type of sequences were considerably less frequent in both corpora, but the drop between the prepositional and nominal/verbal types is considerably more striking in the L1 corpus than in the L2 corpus. Our L2 corpus contains significantly more structural patterns that we included in the category “other” (14.4% of sequences in L2 corpus vs. 4.3% in L1 corpus), which may suggest that structural patterns are less fixed in the language of learners than in the language native speakers and that learners tend to rely rather on the “open-choice principle” than on the “idiom principle” (cf. Sinclair 1991).

The functional analysis of four-word sequences ending in the preposition *of* revealed that their typical discourse function is referential, discourse organizers and stance sequences being not at all attested in our corpora. It is argued that this is due to the nature of sequences containing the *of*-phrase fragment, which describes “some attribute of the object being discussed” (Garner 2016: 40). A closer examination of the referential expressions showed that regardless of the structural type, the examined sequences display predominantly two functions, namely specification of attributes and time/place/text reference.

Our findings have confirmed that Czech learners indeed do underuse some phraseological sequences. Differences between native and non-native language production on the phraseological level tend to be rather subtle, yet they present an interesting challenge and room for development even for advanced L2 speakers, especially so for students aiming to become language professionals. As pointed out by Granger (2017: 9), general academic vocabulary (i.e. words and sequences typical of academic discourse in general, not limited to a particular discipline) tend to be difficult for learners to master, mainly because they “are not particularly salient and tend to pass unnoticed”. Pedagogical applications of the results should include improvements of pedagogical tools by increasing emphasis on advanced phraseological sequences.

Acknowledgements. This study is a result of the research funded by the Czech Science Foundation as the project GA CR 19-05180S “Phraseology in English academic texts written by Czech advanced learners: a comparative study of learner and native speaker discourse”, and by the Faculty of Arts, Charles University within the project “Specifický vysokoškolský výzkum - Jazyk a nástroje pro jeho zkoumání”.

References

- Altenberg, B.: On the phraseology of spoken English: the evidence of recurrent word-combinations. In: Cowie, A.P. (ed.) *Phraseology, Theory, Analysis, and Applications*, pp. 101–124. OUP, Oxford (1998)
- Anthony, L.: *AntConc (Version 3.5.8)* (Computer Software), Waseda University, Tokyo (2019). <http://www.laurenceanthony.net/>

- Biber, D., Conrad, S., Cortes, V.: *If you look at ...* Lexical bundles in university lectures and textbooks. *Appl. Linguist.* **25**, 371–405 (2004)
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E.: *Longman Grammar of Spoken and Written English*. Longman, London (1999)
- Ebeling, S.O., Hasselgård, H.: Learner corpora and phraseology. In: Granger, S., Gilquin, G., Meunier, F. (eds.) *The Cambridge Handbook of Learner Corpus Research*, pp. 207–230. Cambridge University Press, Cambridge (2015)
- Ellis, N.C.: Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annu. Rev. Appl. Linguist.* **32**, 17–44 (2012)
- Garner, J.R.: A phrase-framed approach to investigating phraseology in learner writing across proficiency levels. *Int. J. Learn. Corpus Res.* **2**(1), 31–68 (2016)
- Granger, S.: Academic phraseology. A key ingredient in successful L2 academic literacy. In: Fjeld, H., Henriksen, J. (eds.) *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives*. *Oslo Studies in Language*, vol. 9, no. 3, pp. 9–27. Olsen & Prentice (2017)
- Granger, S., Bestgen, Y.: The use of collocations by intermediate vs. advanced nonnative writers: a bigram-based study. *Int. Rev. Appl. Linguist. Lang. Teach.* **52**(3), 229–252 (2014)
- Granger, S., Paquot, M.: Disentangling the phraseological web. In: Granger, S., Meunier, F. (eds.) *Phraseology. An Interdisciplinary Perspective*, pp. 27–50. John Benjamins, Amsterdam (2008)
- Groom, N.: Closed-class keywords and corpus-driven discourse analysis. *Keyness Texts* **41**, 59–78 (2010)
- Hasselgård, H.: Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English. In: Mahlberg, M., Wiegand, V. (eds.) *Corpus Linguistics, Context and Culture*. De Gruyter Mouton, Berlin (forthcoming)
- Hasselgren, A.: Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *Int. J. Appl. Linguist.* **4**, 237–259 (1994)
- Hunston, S.: Starting with the small words. In: Römer, U., Schulze, R. (eds.) *Patterns, Meaningful Units and Specialized Discourses* (2008). *Int. J. Corpus Linguist.* **13**(3), 271–295
- Hyland, K.: As can be seen. Lexical bundles and disciplinary variation. *English Specific Purposes* **27**(1), 4–21 (2008)
- Kjellmer, G.: A mint of phrases. In: Aijmer, K., Altenberg, B. (eds.) *English Corpus Linguistics*, pp. 111–127. Routledge, London (1991)
- Klégr, A.: English complex prepositions of the prepositional phrase type. *Acta Universitatis Carolinae–Philologica. Prague Stud. English* **22**(5), 51–78 (1997)
- Pawley, A., Syder, F.H.: Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In: Richards, S.C., Schmidt, R.W. (eds.) *Language and Communication*, pp. 191–226. Longman, London (1983)
- Rayson, P.E.: Computational tools and methods for corpus compilation and analysis. In: Biber, D., Reppen, R. (eds.) *Cambridge Handbook of English Corpus Linguistics*. CUP, Cambridge (2015)
- Scott, M.: *Wordsmith Tools 4*. Oxford University Press, Oxford (1996)
- Sinclair, J.: *Corpus, Concordance, Collocation*. OUP, Oxford (1991)