To appear in M. Mahlberg and V. Wiegand (eds), *Corpus Linguistics, Context and Culture*. Berlin: De Gruyter Mouton.

Phraseological teddy bears: frequent lexical bundles in academic writing by Norwegian learners and native speakers of English

Hilde Hasselgård
University of Oslo

Abstract
This paper compares frequent four-word lexical bundles in a learner corpus (VESPA) and a native speaker corpus (BAWE), both representing novice academic writing. The frequencies and dispersion of bundles in the two corpora reveal patterns of both over- and underuse among the learners. The learners are shown to use some bundles very frequently, but frequencies drop more sharply than in the native corpus. The dispersion of the frequent bundles tends to be broader in the native speaker corpus. In a closer scrutiny of four selected bundles the novice-expert dimension is addressed by consulting a corpus of published research articles. Contrasts between English and Norwegian are also considered in order to explain the learners' apparently non-native usage. Some of the most overused bundles seem to have been generalized by the learners to fit into contexts where native speakers rarely use them; these can be described as 'phraseological teddy bears'. Pedagogical applications of the results should start from the underused items in order to broaden the phraseological repertoire of the learners.

1 Introduction

It is well established that learners as well as native speakers use pre-fabricated multi-word units in their language production (see e.g. Granger 1998). Yet, "phraseology is one of the aspects that unmistakably distinguishes native speakers of a language from L2 learners" (Granger & Bestgen 2014: 229), and the phraseology of non-native users of English continues to inspire investigations into the puzzle of nativelike co-selection (Pawley & Syder 1983).

The present investigation concerns the most frequent four-word lexical bundles in two corpora of novice academic English representing advanced learners (with Norwegian as their L1) and native speakers of English. The bundles most frequently used by the two writer groups will be compared with the aim of finding out how they are used as regards their distribution, meanings and functions. I will also take a closer look at some selected bundles whose frequencies and distributions differ markedly between the corpora.

Ringbom (1998) shows that the frequencies of individual word forms tend to differ between learners and native speakers of English, with learners having a tendency to overuse vocabulary items that have high frequencies in general corpora of English. The overuse can be related to a core vocabulary that the learners have acquired early and know well. Hasselgren (1994) compares such familiar lexical favourites to children's toys: "Stripped of the confidence and ease we take for granted in our first language flow, we regularly clutch for the words we feel safe with: our 'lexical teddy bears'" (Hasselgren 1994: 237). A hypothesis of the present study is that the same tendency will be visible in the use of lexical bundles: some bundles will seem familiar and unobjectionable to learners, who will resort to them frequently as their "phraseological teddy bears". This idea is not novel; Nesselhauf (2005: 247) suggests that learners' occasional overuse of "certain native-speaker-like chunks" may partly result “from learners using some of them as lexical teddy bears”.[1] Other bundles, however, will be underused by learners, for example because most learners simply do not know them, or because they belong to a style level that the learners are not fully familiar with. At the advanced level of proficiency represented in the learner corpus used (see section 3 for details), the differences between native and non-native usage of bundles are not expected to consist in errors as much as in diverging frequencies of use.

Hasselgren (1994: 237 f) seems to imply that words characterized as lexical teddy bears are not only more frequent in learner language than in native language, but also that they are “systemically overgeneralized by advanced learners”, which leads to their being used in contexts where native speakers would choose a (near) synonym (see also Levenston & Blum 1977). Thus, a phraseological teddy bear will be a multi-word unit that learners use more frequently and in more contexts than native speakers do.

The paper is structured as follows: after a review of relevant previous research and a presentation of material and method, the most frequent lexical bundles in both corpora will be identified and discussed. Then follow four case studies of selected bundles that are either overused or underused by the learners before some concluding remarks are offered.

2 Some previous studies of lexical bundles and formulaic language in learner English

Recurrent strings of words have been studied under a number of different headings, for example ‘recurrent word combinations’ (e.g. Altenberg 1998), ‘n-grams’ (e.g. Granger & Bestgen 2014; Ebeling & Hasselgård 2015a), and ‘lexical bundles’ (e.g. Biber et al. 1999; Cortes 2004; Ädel & Erman 2012; Paquot 2013). For general overviews of phraseology in learner corpus research, see Paquot & Granger (2012) and Ebeling & Hasselgård (2015b).

Lexical bundles are defined as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (Biber et al. 1999: 990). The operationalization of the definition limits lexical bundles to “uninterrupted combinations of words” (ibid.: 993) that occur above a set frequency threshold and across a minimum number of corpus texts “to exclude individual speaker/writer idiosyncrasies” (ibid.). Biber et al. show that conversation and academic prose differ as to their use of bundles as regards lexicogrammatical structure as well as frequency (e.g. p. 997).

Using a similar method, but the term ‘chains’, Stubbs & Barth (2003) show that recurrent phrases can be used as text type discriminators. That is, they identify differences between the text types by applying a number of measures, one of which is “recurrent word-chains and/or their comparative frequency” (2003: 79). Cortes (2004) discusses lexical bundles in the academic disciplines history and biology. She classifies the lexical bundles functionally into referential bundles, text organizers and stance bundles (ibid.: 409), and shows that the academic disciplines history and biology vary in their use of lexical bundles, in terms of both structural and functional features. She also finds “that the use of target bundles by students in biology and history courses at different university levels very far from resembles the use of these bundles by published authors in these disciplines” (2004: 421).

[1] Ellis (2012: 29) uses the term *phrasal teddy bear* to refer to “highly frequent and prototypically functional phrases like *put it on the table, how are you?, it’s lunch time*”, or “formulaic phrases with routine functional purposes” (ibid.: 37). Since lexical bundles, unlike Ellis’s formulaic sequences, do not require word strings to be idiomatic or complete functional units, I have opted for the related term *phraseological teddy bear*.

Hyland (2008) similarly shows that there is variation between published academic writing and postgraduate student writing, and furthermore, that bundle usage differs across academic disciplines.

The functional classification in Cortes (2004) is also used by Biber et al. (2004), in a paper much referred to in subsequent research on lexical bundles in academic language. It shows clear differences as regards structural and functional categories of lexical bundles across four ‘university registers’: conversation, classroom teaching, textbooks and academic prose. Importantly the authors conclude that “lexical bundles should be regarded as a basic linguistic construct with important functions for the construction of discourse” (2004: 398).

The study of lexical bundles is extended to a comparison between learners and native speakers in Chen & Baker (2010), who compare the writing of Chinese learners of English to native speaker student and expert writing.[2] They find “that the use of lexical bundles in non-native and native student essays is surprisingly similar” while professional writing shows a wider repertoire of certain types of bundles (2010: 44). However, this applies mainly to the quantitative analysis; the qualitative analysis reveals some differences between native and non-native writing. An interesting observation for the present study is that non-native writing tends to show features of “over-generalizing and favoring certain idiomatic expressions and connectors” (ibid.). However, Ädel & Erman (2012), in a study that to some extent replicates Chen & Baker (2010) with data from Swedish learners of English, find more substantial differences between native and non-native writing. They conclude that “non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers” (2012: 90). The qualitative analysis of context is singled out as a future direction in the study of lexical bundles in learner language, since the fact that a bundle may be used by learners to the same extent as by native speakers does not necessarily entail “that it is used in the same way” by both groups (ibid.: 91).

Pérez-Llantada (2014) similarly focuses on 4-word bundles in L1 and L2 expert academic writing. Importantly she correlates her findings of L2 English with bundles in the writers’ first language, Spanish. A central conclusion is that “the L2 English variable reflects a ‘hybrid’ formulaic language” (2014: 92); i.e. it is not fully native-like and shows traces of transfer from L1 Spanish phraseology. Bundles in L1 and L2 expert academic writing is also the topic of Salazar (2014) who studies the frequency, structure and function of 3-6-word bundles extracted from corpora of biomedical research writing. As Salazar’s purpose is partly to explore pedagogical applications, bundles are selected according to a combination of frequency and MI to produce a list of pedagogically relevant bundles (2014: 46, 153).

Ebeling & Hasselgård (2015a) compare the use of n-grams across academic disciplines and L1 backgrounds in VESPA and BAWE, concluding that both factors have an impact on n-gram use, although discipline seems to be the stronger cause of differences. This study concerns functional types of n-grams, classified in line with Moon (1998) as ideational, interpersonal and textual (a framework similar to the one found in Cortes 2004 and Biber et al. 2004). The study does not focus on the frequencies of individual n-grams in the corpora, but as in Ädel & Erman (2012) one of the envisaged avenues of further research is a more

[2] Chen & Baker (2010) used subsets of the British Academic Written English corpus (BAWE) for the learner and native speaker student comparison. BAWE comprises English texts from a number of L1 backgrounds besides English; see Nesi & Gardner (2012: 268), with Chinese being the most frequent non-English L1.

qualitatively oriented study which would take token frequency and context into account. The present study can be seen as a step in that direction and an attempt to fill a gap in present research.

3 Material and method

Two corpora form the core material for the present investigation: the Norwegian component of the ***V****arieties of* ***E****nglish for* ***S****pecific* ***P****urposes d****A****tabase* (VESPA-NO) and the ***B****ritish* ***A****cademic* ***W****ritten* ***E****nglish corpus* (BAWE). Both corpora contain student writing within a variety of academic disciplines. For the present purposes only the linguistics discipline has been investigated, and only texts written by students whose L1 is Norwegian and English, respectively. Table 1 shows the size and composition of the corpora used.[3]

Table 1 The two main corpora for the study

Discipline = linguistics	Texts	Words
VESPA-NO (L2)	239	267,855
BAWE (L1, BrE)	76	167,437

Both corpora have been annotated in order for searches to ignore material not produced by the student, such as linguistic examples, quotations and bibliographies. The word counts given in Table 1 are exclusive of the ignored material. See Ebeling and Heuboeck (2007) and the respective corpus manuals (Paquot et al. 2010, Heuboeck, Holmes, and Nesi 2008) for more information regarding the annotation.

In the discussion of bundles selected for the case studies in section 5, I also draw on other corpora. The *Corpus of Research Articles* (CRA) held at Hong Kong Polytechnic University will be used to check whether there are differences between novice and expert writers within the same discipline.[4] To investigate potential influence from Norwegian, I will consult the *English-Norwegian Parallel Corpus* (ENPC) and the KIAP corpus (Cultural Identity in Academic Prose), which contains published research articles in English, Norwegian and French. From KIAP I use only the section containing linguistics articles in Norwegian, comprising 269,913 words (Fløttum et al. 2006: 7; Fløttum et al. 2013). Other corpora used for occasional reference are the *Michigan Corpus of Upper-Level Student Papers* (MICUSP), the *British National Corpus* (BNC), and the *Corpus of Contemporary American English* (COCA).[5]

The study takes a lexical-bundle approach (Biber et al. 1999, Biber et al. 2004). Lexical bundles are recurrent uninterrupted sequences of word forms that occur above a certain frequency threshold and with a certain dispersion across texts, operationalized in Biber et al. (1999: 992) as at least ten timer per million words and in at least five texts. Although lexical bundles may in principle consist of any number of words (above one), the present study is limited to four-word bundles. This decision is much in line with comparable previous studies (e.g. Hyland 2008, Ädel & Erman 2012, Pérez-Llantada 2014), and as Hyland (2008:

[3] VESPA-NO was recently updated and slightly enlarged. This study is based on the 2012 version.

[4] The corpus contains published research articles in a variety of disciplines. This study uses Applied Linguistics (170,653 words), which was considered closer than Linguistics to the topics contained in VESPA and BAWE.

[5] For further information about the corpora, see the websites listed at the end of this paper.

8) says, “they are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles”. Some of the resultant 4-word bundles (see next section) may be said to consist of a 3-word bundle plus a common word, for example *the meaning of the*, of which the most “salient” part (cf. Simpson-Vlach & Ellis 2010) is arguably *the meaning of*. However, as argued by Hunston (2008), “small words” play an important role in the identification of grammar patterns, which in turn can form semantic sequences (2008: 271), or in the words of Pérez-Llantada (2014: 86), “[b]undles bridge structural units in the discourse, framing semantic meanings”. For example, it may be a salient feature of sequences such as *the meaning of* and *the use of* that they occur in the context of an extended noun phrase.

Recurrent four-word bundles were extracted by means of WordSmith Tools 6 (Scott 2012). The focus is on the highest frequency band, i.e. bundles that account for at least 0.01% of the corpus according to WordSmith’s Wordlist tool. All the bundles in this frequency band occurred in at least five different texts (cf. Biber et al. 1999: 993; 2004: 376). It was not considered necessary in the present context to address the issue of overlapping bundles (cf. Simpson-Vlach & Ellis 2010: 493): the only potentially overlapping bundles in Table 2 are *is an example of* and *an example of this* (both in VESPA), but the concordance lines reveal that the overlap is limited to a single occurrence of *is an example of this*.

The bundles were classified functionally along the lines of Cortes (2004) and Biber et al. (2004). Although many of the bundles are structurally incomplete, they contain enough meaning-bearing elements to make such classification possible. The categories outlined in Biber et al. (2004) are the following:

- Referential bundles (R) “make direct reference to physical or abstract entities, or to the textual context itself” (Biber et al. 2004: 384)
- Stance bundles (S) “express attitudes or assessments of certainty” (ibid.)
- Discourse organizers (D) “reflect relationships between prior and coming discourse” (ibid.)

Bundles that were considered specific to particular topics or tasks were excluded, as they would be unlikely to occur in other corpora. Examples are *Australian and New-Zealand English, Norwegian learners of English, the second text is, as in Tager and Flusberg*.

4 Corpus analysis

The results of the search for frequent four-word bundles in the two corpora are presented in Table 2 along with a functional label and their frequency per 100,000 words. The bundles that are shared between the corpora are marked in shaded cells. More task- and topic-specific bundles have been removed from the original VESPA list than from the BAWE list (8 vs. 1): Table 2 thus shows a shorter list of recurrent bundles for VESPA. Incidentally, two of the bundles that occur in both corpora, *in the case of* and *on the other hand*, are “the most common four-word lexical bundles in academic prose” according to Biber et al. (1999: 994), apparently the only ones to reach a frequency above 100 per million words. It may be noted that the BAWE list overlaps slightly more than the VESPA list with that presented by Byrd &

Coxhead (2010: 37 ff) of widely used lexical bundles in the AWL [Academic Word List] corpus, thus suggesting that BAWE bundles are more academic.

Table 2: Most frequent bundles in both corpora (frequencies per 100,000 words)

Bundles in VESPA	**Function**	**Freq.**	**Bundles in BAWE**	**Function**	**Freq**
on the other hand	D	38.1	it is important to	S	19.1
the use of the	R	32.9	in the case of	D	17.3
when it comes to	D	18.7	as a result of	D	15.5
the meaning of the	R	18.3	the use of the	R	14.9
the rest of the	R	17.2	to be able to	R	14.9
is an example of	R	16.8	the way in which	R	13.7
an example of this	R	13.8	the fact that the	D	11.9
the fact that the	D	13.4	the way we speak	R	11.3
is the use of	R	12.3	can be found in	R	10.8
as we can see	D	11.2	on the other hand	D	10.8
I have chosen to	S	10.5	it was found that	R	10.2
in the case of	D	10.5	the context of the	R	10.2
			the meaning of the	R	10.2
			to look at the	R	10.2

The distribution of functional types of bundles is fairly similar between the corpora: VESPA has six referential bundles, five discourse bundles and one stance bundle, while BAWE has nine referential bundles, four discourse bundles and one stance bundle. VESPA has one discourse bundle and one stance bundle that are personal and self-referential (*I have chosen to* and *as we can see*), while BAWE has a personal referential bundle (*the way we speak*).

Table 3 Overused and underused bundles in VESPA (raw frequencies)

	Overuse			**Underuse**		
		VESPA	BAWE		VESPA	BAWE
p≤0.0001	*on the other hand*	102	18	*as a result of*	8	26
	when it comes to	50	0	*the way we speak*	0	19
	the rest of the	46	4	*it was found that*	0	17
	is an example of	44	9			
	is the use of	33	3			
	as we can see	30	0			
p≤0.001	*the use of the*	87	25	*it is important to*	18	32
	I have chosen to	30	3	*the way in which*	9	23
p≤0.01				*to be able to*	16	25
				the context of the	7	17
				to look at the	7	17
p<0.5	*the meaning of the*	47	17			

It is striking that the highest frequencies in VESPA far exceed those in BAWE; there is also a much steeper decline of frequencies in VESPA. This suggests that learners tend to re-use a small number of bundles to a greater extent than native speakers. To investigate overuse and underuse, the frequencies of all the bundles included in Table 2 were compared between the corpora. A number of bundles had similar frequencies in VESPA and BAWE (*an example of this, the fact that the, in the case of, can be found in*), while the rest differed significantly in frequency according to a log-likelihood test.[6] Table 3 displays those bundles that are either overused or underused in VESPA compared to BAWE.

In a next step, the dispersion of the most frequent bundles was studied so as to check whether the frequency of any bundle is boosted because of popularity in certain texts. Table 4 shows the percentage of texts in which each bundle occurs. Note that the frequency order differs slightly from that in Table 2.

Table 4 Most widely used bundles in both corpora (distribution across texts)

Bundles in VESPA	**Texts**	**Bundles in BAWE**	**Texts**
on the other hand	31.0%	it is important to	23.7%
the use of the	19.0%	the fact that the	23.7%
the rest of the	14.2%	to be able to	22.4%
the fact that the	13.4%	in the case of	19.7%
is an example of	12.6%	on the other hand	18.4%
the meaning of the	11.7%	as a result of	17.1%
as we can see	11.3%	the meaning of the	17.1%
when it comes to	10.5%	the use of the	15.8%
is the use of	10.0%	the way in which	15.8%
an example of this	9.6%	the context of the	15.8%
I have chosen to	9.6%	it was found that	13.2%
in the case of	7.9%	to look at the	13.2%
		can be found in	10.5%
		the way we speak	7.9%

The most common bundle in VESPA is shown to have a much wider dispersion than any of the bundles in BAWE, but there is a sharp drop already at rank 2. In other words, most of the frequent bundles in BAWE are more widely dispersed than those in VESPA.

A comparison of frequencies based on dispersion gives a different perspective on overuse and underuse. Relative to the number of texts each bundle occurs in, the following phrases had similar dispersions in the two corpora: *the rest of the, is an example of, an example of this, is the use of, I have chosen to, the use of the, the meaning of the, the fact that the, can be found in* (significance was tested by means of a chi square test). The bundles that occur in significantly different numbers of texts in the corpora are shown in Table 5.

When dispersion is taken into account, the number of overused bundles is greatly reduced (compare Tables 3 and 5), while the underuse is relatively unchanged. This means that some of the 'overused' bundles in Table 3 are overused only in some texts. Interestingly, all the overused bundles in Table 5 are discourse organizers, but most of the underused ones are referential.

Table 5 Overused and underused bundles in VESPA according to text dispersion

[6] Log Likelihood was calculated with Paul Rayson's calculator available at http://ucrel.lancs.ac.uk/llwizard.html (last accessed 3 November 2015).

	Overuse	Underuse
p≤0.0001		*it is important to* *as a result of* *to be able to* *the way in which* *the way we speak* *it was found that* *the context of the*
p≤0.001	-	-
p≤0.01	*when it comes to* *as we can see*	*in the case of*
p<0.5	*on the other hand*	

5 Case studies

This section presents case studies of four discourse-organizing lexical bundles whose distributions differ significantly between the corpora as regards both frequency and dispersion. Three of these are overused in VESPA compared to BAWE (*on the other hand, when it comes to, as we can see*), and one is underused (*as a result of*). I will draw up a usage profile for each bundle on the basis of VESPA concordances, and possible reasons for the attested overuse/underuse will be discussed, such as the presence of a corresponding expression in Norwegian, along the lines of Paquot's (2013) study of transfer effects. In contrast to Paquot, I have not investigated learner behaviour in EFL corpora with other L1 backgrounds due to a lack of suitable material. However, the novice writers in VESPA and BAWE will be compared with 'expert' writers, represented in the applied linguistics section of the Corpus of Research Articles (CRA) to control for any novice-expert differences.

5.1 *On the other hand*

The most frequent four-word bundle in VESPA is *on the other hand* (see Table 2); it is also the bundle with the widest dispersion, occurring in 31% of the texts. VESPA also overuses the bundle in comparison with CRA, where there are 22 occurrences (12.9 per 100,000); the overuse is significant at LL=26.15 (p<0.0001). The VESPA concordance shows that *on the other hand* has the following characteristics:

- 30 out of 102 occurrences are clause-initial (example 1); the remaining 72 are clause-medial (example 2).
- *On the other hand* co-occurs with *on the one hand* eight times in VESPA; see example (3). This pattern is not found in BAWE.[7]
- *On the other hand* sometimes functions as a general topic shifter in VESPA, not always marking contrast (Lie 2013); see example (4).
- No extended phraseological pattern can be identified for the bundle.

(1) *On the other hand*, the overuse of the progressive is intralingual in that it reflects what has been learned and has been overgeneralized. (VESPA)[8]

[7] Byrd & Coxhead (2010: 46) also note that "*on the other hand* is most often used […] without the prior use of *on the one hand*".

(2) Coherence, *on the other hand*, is in the mind of the writer and reader: it is a mental phenomenon and cannot be identified or quantified in the same way as cohesion. (VESPA)

(3) *On the one hand*, contrastive linguists are very enthusiastic about what they have to offer L2 teaching; *on the other hand*, they seem somehow depressed by the lack of positive response among teachers. (VESPA)

(4) If it works you have substitution, and if we apply this test for this line, we find that it does. "We are the *ones*.." So far, so good. *On the other hand*, what is important to notice here is that we do not really know what the word is substituting for. (VESPA)

The preference for placing the bundle in clause-medial position is found in BAWE too,[9] with the same discourse effect, namely to set off the subject (or any other clause-initial element) as contrasting with the preceding context. Example (2), for instance, follows directly after a definition of cohesion, which makes it appropriate to steer contrastive focus to the related, but different, concept. Interestingly, the preferred position of *on the other hand* in CRA is initial (18 out of the 22 occurrences). Three medially placed instances provide contrastive focus to the clause subject, while the remaining one, example (5), occurs in an elliptical clause and gives extra focus to the final constituent. This usage is not found in BAWE or VESPA.

(5) In order to summarize the data, Fig. 9 provides a schematic representation of the main confusions identified from the perceptual results, for the adults and the 8-year old, on the one hand, and *on the other hand*, for the 4-year old. (CRA)

The Norwegian expression corresponding most closely to *on the other hand* with respect to similarity of form is *på den annen side* ('on the other side'). This expression is, however, not nearly as frequent in the Norwegian corpora consulted as the frequency of *on the other hand* in VESPA might suggest: there are 6.8 occurrences per 100,000 words in ENPC non-fiction and 9.6 in KIAP (ling.).[10] The overuse in VESPA thus cannot be explained by means of direct transfer from Norwegian, but the positional preference can: all the occurrences of *på den annen side* in KIAP occur in medial position.

Studying the wider context of *on the other hand* in VESPA, we find that it often occurs in the vicinity of other markers of contrast, as exemplified in (6).

(6) The use of cataphoric reference is not very extensive, *however*. Anaphoric reference *on the other hand*, is utilized throughout the text. (VESPA)

This suggests that the learners may have a tendency to over-express contrastive relations when the discourse moves from one topic to another. As noted above, and illustrated in (4), "VESPA contributors are as likely to use the phrase as a general topic change marker, introducing a concept only tangentially related to the preceding sentence" (Lie 2013: 39). Lie also notes (ibid.) that the VESPA texts are slightly skewed towards contrastive assignments.

[8] All corpus examples have been rendered as they occur in the corpora.

[9] In BAWE, 12 out of the 18 instances occur in medial position.

[10] This includes the variant forms *på den andre siden / på den andre sida*.

This shows up in the concordance, where about 20 of the lines reflect a comparison (e.g. between two texts) that the student has been asked make, as illustrated by (7).

(7) The second text, on the other hand, is not all that engaging, just informative. (VESPA)

5.2 *When it comes to*

When it comes to is widespread in VESPA, with 50 occurrences across 25 texts (23 writers), while it is absent from BAWE. This does not mean the expression is non-native; it occurs three times in CRA.[11] The syntactic function *when it comes to X* is respect adjunct (cf. Hasselgård 2010: 28; 244 ff), typically specifying the circumstances under which the proposition applies. The usage of *when it comes to* in VESPA has the following characteristics:

- The bundle is typically followed by an indefinite noun phrase.
- 22 (of 50) occur in sentence-initial position (six of which are preceded by a connector); these mark the sentence topic, as in (8).
- In end position the function is typically to restrict the validity of the proposition; see (9).
- Certain sentence-final occurrences are close to postmodifying function; see (10).

(8) *When it comes to collocation and sentence structure* this student has some very strange ways of saying things … (VESPA)

(9) But everything that is said in this short excerpt is in the present tense, which is quite normal *when it comes to this kind of literature*. (VESPA)

(10) …the author of this text has some problems *when it comes to word formation*. (VESPA)

Seeking to explain the massive overuse of *when it comes to* in VESPA, I searched for the following near-synonyms in the corpora: *with regard(s) to, with respect to, as regards, as to, concerning, regarding* and *in terms of.* The latter turns out to be a clear favourite, outnumbering all its synonyms across the board. Figure 1 shows *when it comes to* and *in terms of* separately while all the other near-synonyms have been lumped together. The collective frequencies of such expressions do not differ significantly across the corpora; i.e. the alternative expressions make up for the overuse of *when it comes to* in VESPA.

[11] This corresponds to 1.8 per 100,000 words. The overuse in VESPA compared to CRA is significant at $p < 0.0001$ (LL=31.90).

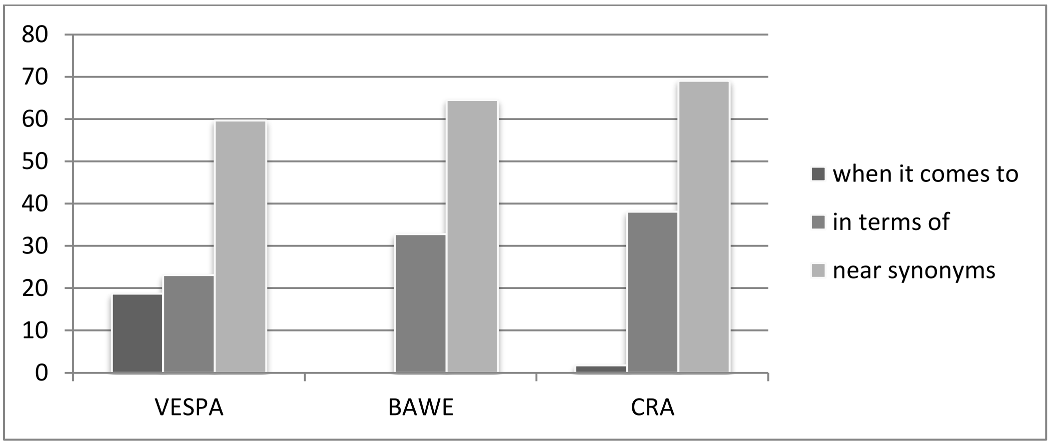

Figure 1 *When it comes to* and its near-synonyms in three corpora.

Searches in the BNC and COCA revealed that *when it comes to* is more than twice as frequent in American as in British English (12.1 per million words in the BNC vs. 26 in COCA). It is relatively common in journalistic texts in both varieties and in speech in COCA, but infrequent in academic writing. Interestingly, the bundle has almost doubled its frequency in American English within the COCA time span, with 18.36 hits per million words in 1990-94 as against 35.6 in 2010-12. A possible cause of the overuse in VESPA might thus be influence from spoken American English and journalistic texts, i.e. genres that are widespread in Norwegian society through the media. However, we may note that the linguistics part of MICUSP has only one occurrence of this bundle.

Looking to the Norwegian language for a source of the phrase *when it comes to*, we find the formally similar *når det gjelder* ('when it concerns'). This expression is fairly common in all text types, including academic prose: in the KIAP corpus, the expression was found with a frequency of 20.4 per 100,000 words in linguistics articles, i.e. very close to the frequency of *when it comes to* in VESPA. See example (11).

(11) En av de sentrale forskjellene de tar opp, er nettopp forskjellen *når det gjelder* plassering av det finitte verbet i leddsetninger. (KIAP)
"One of the central differences they take up, is indeed the difference when it comes to placement of the finite verb in subordinate clauses."

The evidence presented here suggests that the overuse of *when it comes to* can be related to the Norwegian *når det gjelder*, which is functionally and formally similar. In addition, learners may find support for their use of the expression through the media. Curiously, however, in the non-fiction part of the ENPC, *når det gjelder* is more frequent in translations (from English) than in Norwegian originals. Sometimes the source of *når det gjelder* is a preposition, as shown in (12), thus perhaps suggesting that *når det gjelder* is a relatively grammaticalized expression used for relating two concepts. Norwegian learners may have transferred this to their use of *when it comes to*, as suggested by examples such as (10) above, where a more elegant wording might have been "…has some problems with word formation".

(12) There is still much to be understood about the origin of life, including the origin of the genetic code. (ENPC, CSA1)
Det er fortsatt en hel del vi ennå ikke forstår *når det gjelder* livets opprinnelse — blant annet hvordan den genetiske koden oppstod. (CSA1T)
Lit: 'There is still a whole lot we yet not understand when it concerns life's origin…'

5.3 *As we can see*

The bundle *as we can see* occurs in 27 texts in VESPA (11.3%), but not at all in BAWE or in CRA. Its use in VESPA has a clear profile:

- It is typically used in sentence-initial position (20 of the 31 instances).
- It is typically followed by a preposition (19 of the 31 instances), of which the most frequent one is *from* (15 instances).
- 11 instances are followed by either a subject NP (7) or existential *there* (4).

In sentence-initial position *as we can see* has a connective function, as illustrated by (13). It is a metadiscursive marker, typically referring to texts, tables, figures and examples discussed. The prepositional phrase following it tells readers *where* something is to be seen.

(13) *As we can see* from the above examples, *effektiv* is not used with this reference in Norwegian, and is therefore rephrased. (VESPA)

When *as we can see* is directly followed by a subject, with no specification of where to look, the location may be well known, as in (14), where the student is referring to a text s/he has been asked to analyse.

(14) There are also examples of sentence structure that clearly derives from Norwegian influence, as in sentence (13) under 'conceptual confusion'. The spelling errors may be classified as intralingual errors, and *as we can see*, there is a lot of this type of errors as well. (VESPA)

Hyland (2008) highlights, even in the title of his paper, the frequency of the metadiscourse marker *as can be seen*, i.e. the passive counterpart of *as we can see*. Table 6 shows that *as can be seen* is recurrent in both VESPA and BAWE, and interestingly it is more frequent in VESPA. However, only five VESPA writers use it, while 27 use the active phrase. It may be noted that the active phrase occurs with about equal frequencies in spoken and academic English in both the BNC and COCA, while the passive phrase is frequent only in academic prose in both varieties. The bundle favoured in VESPA is thus the more colloquial one.

Table 6 *As we can see* and *as can be seen* across corpora. N = raw frequencies, R = relative frequencies per 100,000 words

	VESPA		BAWE		CRA	
	N	R	N	R	N	R
as we can see	31	11.6	0	0	0	0
as can be seen	14	5.2	5	3.0	6	3.5

As can be seen has the same usage profile in VESPA as its active counterpart: it is sentence-initial in 8 of the 14 instances; it is followed by a prepositional phrase in 11 instances and by a subject NP in the remaining three. There appears to be no difference in the discourse functions of the active and the passive; both are metadiscursive and guide the reader to tables, concordances, examples, and the like, as illustrated by example (15).

(15) …the phrase is expressing the simple future, *as can be seen* in one of the hits the PerlTCE produced. (VESPA)

The six examples of *as can be seen* in CRA are all followed by a prepositional phrase or the adverb *above*, four of the six are sentence-initial, and they signpost tables, figures and concordances. The VESPA writers have thus grasped the functions of *as can be seen* (and transferred them to *as we can see*), but use the phrase too often compared to native speakers.

The most closely corresponding Norwegian equivalent to *as we can see* is *som vi ser* ('as we see'). This expression is found in KIAP but is infrequent (1.1 per 100,000 words); searchers for the related *som en/man ser* ('as one sees') and *som vi/en/man kan se* ('as we/one can see') add only four hits, increasing the frequency to 2.6 per 100,000. The closest counterpart in the passive voice, *som vist* ('as shown') is more frequent at 5.9 occurrences per 100,000 words, and has much the same profile as the English *as can be seen*. In any case, direct transfer from Norwegian thus cannot be the source of the overuse of *as we can see* observed in VESPA.[12]

A possible explanation might instead be found in the general tendency of learners towards overuse of metadiscourse (Ädel 2006, Hasselgård, in press) and writer/reader visibility features (Paquot et al. 2013). *As we can see* gives the writer the opportunity to be visible, involve the reader and organize the text at the same time, thus serving three of the functions often attributed to learner discourse.

5.4 *As a result of*

The last bundle to be discussed here is one that is underused in VESPA compared to BAWE, with only eight hits (3.4 per 100,000) occurring in eight texts (3.4%) by eight different writers. The bundle is frequent in BAWE with 26 hits (13.7 per 100,000) distributed over 13 texts (17.1%) and eight writers. CRA contains eight instances of the bundle (4.7 per 100,000), which actually does not constitute a significant difference from VESPA (LL = 0.81). The use of the bundle in VESPA has the following characteristics:

- In three out of the eight instances, *as a result of* is followed by the pronoun/determiner *this* (example 16), thus marking cohesion with the preceding context.
- Three of the other instances echo a wording in the students' textbook, see (17).
- To a greater extent than the bundles discussed above, *as a result of* occurs in sentences that contain errors, see (17) and (18).

[12] Lee & Chen (2009: 289) identify 'we can see' as one of the collocations favoured by Chinese learners of English. In Lee & Chen's material too, the bundle is typically used "to refer to or explain tables or figures, and to organize the discussion" (ibid.).

(16) The sentences in text 2 are *as a result of* this shorter,… (VESPA)

(17) Johansson further explain them *as a result of* overgeneralisation from what the learner of language already has learnt… (VESPA)

(18) *As a result of* this the highly differ in style and form. (VESPA)

In BAWE *as a result of* typically precedes a complex noun phrase, as in (19), sometimes involving a nominalized process, as in (20).

(19) Once again, *as a result of* the more informal nature of the group interaction, ellipsis was commonplace. (BAWE)

(20) *As a result of* this examination of three types of instruction, each based on a different theory of language learning, I can now view my Persian learning in a more informed light. (BAWE)

The usage found in CRA closely resembles that of BAWE: the complement of *of* is most typically a complex noun phrase, as illustrated by (21).

(21) …and decisions are made *as a result of* multiple identifications with value premises at an individual level, … (CRA)

As a result of has a literal counterpart in Norwegian: *som et resultat av* (with the variant *som resultat av*). However, translations in ENPC non-fiction show that the two expressions do not often correspond: *som (et) resultat av* occurs only once as a translation of *as a result of* and twice as its source. The related *som følge av* ('as consequence of') is a recurrent source (5 out of 21 instances), and *på grunn av* (lit. 'on reason of' = 'because of') is a recurrent translation (3 out of 8). Both *som (et) resultat av* and *som følge av* were found in KIAP. Interestingly, *som et resultat av* is also used in a more literal sense, as illustrated by (22).

(22) …at språkendringene blir forklart *som et resultat av* at en folkegruppe er blitt rammet av pest, krig eller andre sosiale tragedier… (KIAP)
"that the language changes are explained as a result of [the fact] that a population has been hit by plague, war or other social tragedies…" (My translation)

The contrastive observations may suggest that Norwegian learners of English fail to use *as a result of* idiomatically partly because of the low degree of correspondence between this bundle and its most similar Norwegian counterpart *som (et) resultat av*, and partly because the Norwegian expression seems to be less grammaticalized than the English one. But since *som resultat/følge av* is not infrequent in Norwegian, this cannot be the whole story. The VESPA contributors do write about causal relations: there is a massive overuse of the conjunction *because* compared to BAWE with 179.2 vs. 83 occurrences per 100,000 words (LL= 223.15). A more detailed investigation of causal expressions in learner language is needed to map the range of lexicogrammatical resources employed by the advanced learners. At this stage we may hypothesize that the underuse of *as a result of* has intralingual rather than interlingual causes: previous studies of Norwegian-based learner English have indicated that learners struggle with (or avoid) complex noun phrases (e.g. Hasselgård 2012); thus it is possible that *as a result of* represents a level of complexity that the learners are not prepared to handle.

5.5 Summary of case studies

The case studies presented in this section have illustrated some differences between learner and native speaker phraseology. The usage profiles worked out on the basis of concordances show that the learners do not always use bundles in the same contexts and with the same discourse functions as native speakers. *On the other hand, when it comes to* and *as we can see* are overused by the learners in terms of frequencies across the corpora as well as text dispersion. *As a result of* is underused. Explanations for both overuse and underuse can be sought in comparisons with the learners' L1 (in this case Norwegian) and in general reference corpora of spoken and written English. It has been found here that the three overused items have more similar corresponding expressions in Norwegian than the underused one. Comparison with expert writing in the same discipline is important. For example, in selecting an underused bundle for further examination I realized that most of the bundles that are underused compared to BAWE (Tables 3 and 5) are not underused if compared to CRA instead. The overuse, however, is at least as strong.[13]

The bundles *on the other hand* and *when it comes to* display two important characteristics of lexical/phraseological teddy bears: they are much more frequent in English L2 than in English L1, and they seem to have generalized their meanings and discourse functions by being used in contexts where native speakers prefer other expressions. The third overused expression, *as we can see*, does not show the same pattern of generalization, and can probably be ascribed to the learners' general leaning towards a colloquial style. The preference for a colloquial style may partly explain the underuse of *as a result of*, although limited proficiency may also cause learners to avoid this complex construction.

6 Concluding remarks

The present study has looked into four-word bundles only. Given the relatively short lists of bundles frequent enough to reveal patterns of use, it is unlikely that a study of longer bundles will be very fruitful. However, the inclusion of three-word bundles may be able to complete the picture (cf. Lie 2013; Ebeling & Hasselgård 2015a). Furthermore, as many of the studies referred to in section 2 point out, lexical bundles differ across disciplines. A natural next step would thus be to make a similar investigation of the other disciplines available, as VESPA now also contains literature and business texts.

Despite its limitations, this study has shown that frequency patterns of recurrent lexical bundles differ between learners and native speakers: the most frequent bundles are *more* frequent in learner English, but frequencies drop less sharply in L1 English. This indicates that the learners have clear favourites that they "clutch for" and "feel safe with" (Hasselgren 1994: 237), so we may justifiably speak of *phraseological teddy bears.*

[13] Unfortunately the CRA interface does not show text dispersion, so overuse and underuse can only be calculated from frequency of occurrence. Nor does the interface allow the bottom-up extraction of bundles (or downloading of raw texts). It is, however, highly likely that other underused bundles would have been identified in comparisons between VESPA and CRA.

The study of text dispersion gave a different rank frequency of the bundles than the one based on frequencies of occurrence. The most common bundles turned out to occur in a greater proportion of the texts in L1 English; learners are thus less uniform in their use of most of the frequent bundles. Some learners appear to be more fond of their phraseological teddy bears than others. For this reason it was suggested that text dispersion may be a better indicator than frequencies per 100,000 words of over- and underuse of lexical bundles.

The detailed case studies of four selected bundles corroborate Ädel & Erman's (2012) prediction that learners and native speakers may not be using bundles in the same way. The case studies were performed only on bundles with significantly different frequencies between VESPA and BAWE, but should in principle also be carried out for other bundles, as similar frequencies do not automatically mean similar usage, as also pointed out by Hasselgård & Johansson (2011: 46ff) in a case study of the patterns surrounding *quite* across a number of EFL varieties. By the same token, differences in frequency need not imply that a bundle is used incorrectly by the learners. A methodological feature of this study consists in cross-checks with parallel corpora and corpora of expert academic writing. This is believed to be indispensable in addressing potential sources of transfer as well as the issue of discrepancies between novice and expert writing. Furthermore, the qualitative study of individual bundles in terms of profiles of usage, as carried out in Section 5, seems to be a fruitful way of exploring divergences between native and non-native style.

This type of qualitative analysis is certainly required before any pedagogical recommendations can be made concerning the use of lexical bundles and phraseological teddy bears. We need to know "what they are, how and why we use them, how they affect our discourse and, hopefully, how we might be persuaded to part with them" (Hasselgren 1994: 237). Rather than just being told to use *when it comes to* less and *as a result of* more, learners should be made aware of the appropriate contexts and functions of the bundles. It is also important to compare L1 novice writing with expert writing before trying to change L2 behaviour since the professional writers are more likely to represent a learning target. Both quantitative and qualitative analyses are needed to tease out differences between native and non-native phraseology. It seems, however, that pedagogical applications should be derived from patterns of underuse: rather than depriving the learners of their phraseological teddy bears we should give them some new toys.

References

Ädel, A. & B. Erman. 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: 81-92.

Altenberg, B. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In *Phraseology. Theory, analysis, and applications*, ed. A.P. Cowie, 101–122. Oxford: Oxford University Press.

Biber, D., Conrad, S. & Cortes, V. 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. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Byrd, P. & A. Coxhead. 2010. *On the other hand*: Lexical bundles in academic writing an din the teaching of EAP. *University of Sydney Papers in TESOL,* 5, 31-64.

Chen, Y.-H. & P. Baker. 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,* 397–323.

Ebeling, S.O. & H. Hasselgård. 2015a. Learners' and native speakers' use of recurrent word-combinations across disciplines. *Bergen Language and Linguistics Studies* (*BeLLS*) vol.6, 87-106.

Ebeling, S.O. & H. Hasselgård. 2015b. Phraseology in learner corpus research. In S. Granger, G. Gilquin, F. Meunier (eds), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 207-230.

Ebeling, S.O. & A. Heuboeck. 2007. Encoding document information in a corpus of student writing: The *British Academic Written English* corpus. *Corpora* 2 (2): 241-256.

Ellis, N.C. 2012. Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics* 32, 17-44.

Fløttum, K., T. Dahl & T. Kinn. 2006. *Academic Voices*. Amsterdam: Benjamins.

Fløttum, K., Dahl, T., Didriksen, A. A., & Gjesdal, A. M. 2013. KIAP–reflections on a complex corpus. *Bergen Language and Linguistics Studies*, *3*(1).

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

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

Hasselgård, H. 2010. *Adjunct Adverbials in English*. Cambridge: Cambridge University Press.

Hasselgård, H. 2012. *Facts, ideas, questions, problems*, and *issues* in advanced learners' English. *Nordic Journal of English Studies*, 11:1, 22-54.

Hasselgård, H. In press. Discourse-organizing metadiscourse in novice academic English To appear in María José López-Couso et al. (eds), *Corpus linguistics on the move: Exploring and understanding English through corpora*. Brill | Rodopi.

Hasselgård, H. & S. Johansson. 2011. Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (eds), *A Taste for Corpora. In honour of Sylviane Granger*. Amsterdam: Benjamins, 33-62.

Hasselgren, A. 1994. Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics* 4: 237-259.

Heuboeck, A., J. Holmes, and H. Nesi. 2008. The BAWE Corpus Manual. University of Warwick, University of Reading, Oxford Brookes University.

Hunston, S. 2008. Starting with the small words. Patterns, lexis and semantic sequences. *International Journal of Corpus Linguistics* 13:3, 271-295.

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

Lee, D.Y.W. & S.X. Chen. 2009. Making a bigger deal of the smaller words: Funtion words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18, 281-296.

Levenston, E.A. & S. Blum. 1977. Aspects of lexical simplification in the speech and writing of advanced adult learners. In S.P. Corder & E. Roulet (eds), *The notion of simplification, interlanguage and pidgins and their relations to SL Pedagogy*. Neuchâtel: Actes du 5ème colloque de Linguistique Appliquée de Neuchâtel, 51-71.

Lie, J. 2013. "The fact that the majority seems to be…" A corpus-based investigation of non-native academic English. Unpublished MA thesis, University of Oslo. [available at http://urn.nb.no/URN:NBN:no-46382]

Moon, R. 1998. *Fixed Expressions and Idioms in English. A corpus-based approach*. Oxford: Clarendon Press.

Nesselhauf, N. 2005. *Collocations in a Learner Corpus*. Amsterdam/New York: John Benjamins.

Paquot, M. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18:3, 391–417.

Paquot, M., S.O. Ebeling, A. Heuboeck, & L. Valentin. 2010. The VESPA tagging manual. CECL, Université catholique de Louvain.

Paquot, M. & S. Granger. 2012. Formulaic language in learner corpora. *Annual Review of Applied Linguistics* 32, 130-149.

Paquot, M., H. Hasselgård & S. O. Ebeling. 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 S. Granger, G. Gilquin & F. Meunier (eds) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain: Presses Universitaires de Louvain, 377-387.

Pawley, A. & F.H. Syder. 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards, S.C. and Schmidt, R.W. (eds.), *Language and Communication*. London/New York: Longman, 191–226.

Pérez-Llantada, C. 2014. Formulaic language in L1 and L2 expert academic writing: convergent and divergent usage. *Journal of English for Academic Purposes*, 14: 84-94.

Ringbom, H. 1998. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. In *Learner English on Computer*, ed. S. Granger, 41-52. London: Longman.

Salazar, D. 2014. *Lexical Bundles in Native and Non-native Scientific Writing*. Amsterdam: John Benjamins.

Scott, M., 2012, WordSmith Tools version 6. Stroud: Lexical Analysis Software.

Simpson-Vlach, R. & N.C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4), 487-512. doi:10.1093/applin/amp058.

Stubbs, M. & I. Barth. 2003. Using recurrent phrases as text-type discriminators. A quantitative method and some findings. *Functions of Language* 10 (1): 61–104.

Corpora

BAWE (British Academic Written English Corpus): http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/

VESPA (Varieties of English for Specific Purposes dAtabase): http://www.uclouvain.be/en-cecl-vespa.html, http://www.hf.uio.no/ilos/english/services/vespa/
CRA (Corpus of Research Articles): http://rcpce.engl.polyu.edu.hk/RACorpus/

BNC (British National Corpus): http://www.natcorp.ox.ac.uk/
COCA (Corpus of Contemporary American English): http://corpus.byu.edu/coca/
ENPC (English-Norwegian Parallel Corpus): http://www.hf.uio.no/ilos/english/services/omc/enpc/
KIAP (Cultural Identity in Academic Prose): http://KIAP.uib.no/KIAPCorpus.htm
MICUSP (Michigan Corpus of Upper-Level Student Papers) http://micusp.elicorpora.info/