2 The Functions of N-grams in Bilingual and Learner Corpora: An Integrated Contrastive Approach

Signe Oksefjell Ebeling and Hilde Hasselgård

1 Introduction

In a previous article we studied phraseological sequences (n-grams) in texts by Norwegian learners of English and native speakers of English in two academic disciplines: linguistics and business (Ebeling & Hasselgård, 2015a). The 100 most frequent n-gram types in the Varieties of English for Specific Purposes dAtabase learner corpus and the British Academic Written English corpus were functionally classified according to an adapted version of Moon's (1998) framework for analysing fixed expressions and idioms, distinguishing informational, situational, evaluative, modalising and organisational n-grams (see Section 2.4). This approach revealed differences between disciplines (e.g. significantly more modalising n-gram types in linguistics) and between L1 groups (e.g. significantly more evaluative n-gram types in native-speaker (NS) texts). This chapter adds a cross-linguistic dimension to the original study by analysing data from the Cultural Identity in Academic Prose (KIAP) corpus,[1] which is a multilingual, comparable corpus of research articles (Fløttum *et al.*, 2006).

After presenting the results from the previous contrastive interlanguage study, we will perform a contrastive analysis of n-grams in English and Norwegian in order to diagnose, in line with the Integrated Contrastive Model (Granger 1996), the extent to which the phraseology of the Norwegian learners' interlanguage may be influenced by their native language. The 100 most frequent 3- and 4-gram types for the contrastive analysis will be extracted from the English and Norwegian linguistics sections of the KIAP corpus, and the functional classification will be carried out as in Ebeling and Hasselgård (2015a). The focus is restricted to one discipline (linguistics) in order to narrow the scope slightly and to avoid problems of corpus comparability within the business material (Ebeling & Hasselgård, 2015a: 91). Such an analysis will also enable a comparison of n-gram use between student and 'expert' academic writing, and in this way throw light on the extent to which the L1 and L2 apprentice academics match what may arguably be called the target usage of their discipline (albeit not necessarily the learning target of each student). The proposed combination of corpora is hoped to differentiate between features of novice writing and L1 influence in the learners' interlanguage. More precisely, our research questions are the following:

i. What discourse functions do recurrent 3- and 4-grams have in English and Norwegian published linguistics articles? What are the cross-linguistic similarities and differences?
ii. To what extent can the cross-linguistic analysis explain the usage of n-grams by Norwegian learners of English?
iii. To what extent are the same patterns and functions used across the dimensions of L1 and writer expertise (novice/expert)?

Based on previous research, we assume that the novice writers share some characteristics regardless of L1: e.g. the students may be expected to use organisational n-grams more often than the professional academics, and the Norwegian learners even more so than the English-speaking novice writers (Ebeling & Hasselgård, 2015a; Hasselgård, 2009; Leedham, 2015). Ebeling and Ebeling (2017) found significant differences in the distribution of functional types of 3-grams between English and Norwegian fiction; we may expect to find similar cross-linguistic differences in academic writing too.

This chapter is structured as follows. Section 2 gives an account of the material and method used, including an outline of the Integrated Contrastive Model (2.1), the corpora (2.2), the n-gram extraction method (2.3), and a brief description of the functional classification procedure (2.4). In Section 3 we present the previous interlanguage study in more detail, including important observations and results, before moving on to the contrastive analysis of n-grams in English and Norwegian linguistics research articles (Section 4). Section 5 is concerned with the novice (learners and native speakers) vs. expert comparison in the English data, while the discussion in Section 6 brings together the findings from the two types of contrastive analysis – the previous Contrastive Interlanguage Analysis and the new Contrastive Analysis – in accordance with the Integrated Contrastive Model. Some concluding remarks are offered in Section 7.

2 Method, Material and Classificatory Framework

2.1 The Integrated Contrastive Model

The overall methodological framework of this study is the Integrated Contrastive Model (ICM)(Granger, 1996; Gilquin, 2000/2001). The model combines two types of analysis: Contrastive Interlanguage Analysis (CIA) (Granger, 1996, 2015), in which a comparison is typically made between an interlanguage variety and reference language variety (Granger, 2015: 17), and Contrastive Analysis (CA), in which a comparison is made between two or more different languages (Johansson, 2007: 1). The underlying assumption is that a contrastive analysis can, at least partly, either predict or diagnose transfer-related interlanguage phenomena. As Granger emphasises: 'it is important to note that the terms "predictive" and "diagnostic" refer to mere hypotheses, which can be confirmed or refuted by corpus investigation' (Granger, 1996: 46). An ICM-based study may start from a cross-linguistic analysis to make predictions about interlanguage performance, or, as in our case, start from a contrastive interlanguage analysis and form hypotheses about (i) discrepancies between the interlanguage and the reference language variety and (ii) cross-linguistic differences between the learners' first and second language. These hypotheses then form the basis for the contrastive analysis based on a bilingual corpus (cf. Granger, 2018: 189). We concur with Gilquin (2000/2001: 101) that 'this presupposes a constant movement between the two disciplines [CA and CIA], but also and above all the availability of reliable CA and CIA data in the form of well-designed and representative bilingual and learner corpora'.

2.2 The corpora

The material for the original interlanguage study was extracted from the British Academic Written English (BAWE) corpus and the Varieties of English for Specific Purposes dAtabase (VESPA) corpus. The former contains proficient student writing from UK universities in a number of academic disciplines (Alsop & Nesi, 2009; Heuboeck *et al.*, 2008), while the latter contains student L2 English writing also from several academic disciplines. Both corpora thus include course work texts written by novice writers within their respective disciplines. In the 2015a study, texts from two disciplines were investigated: linguistics and business. Those culled from the BAWE corpus were all written by students whose L1 was

English, while those from the VESPA corpus were written in English by students whose L1 was Norwegian (VESPA-NO).

The data for the contrastive analysis part of this study are culled from the KIAP (Cultural Identity in Academic Prose) corpus. KIAP is a comparable corpus of research articles in three languages (English, French and Norwegian) and three academic disciplines (economics, linguistics and medicine) (Fløttum *et al*., 2006). For the purpose of the comparison with BAWE and VESPA we use the English and Norwegian linguistics sub-corpora. In other words, the contrastive analysis will draw on published texts written by professional linguists in their native language.

Table 2.1 gives a description of the corpora used in terms of number of texts and number of running words.[2] As can be seen, the corpora differ substantially in size, a fact that will need to be borne in mind when discussing the findings. Nevertheless, since we focus almost exclusively on the most frequent combinations of words, and on *types* rather than *tokens*, this difference in size should not influence the results too much.

Table 2.1 Breakdown of data in terms of number of texts and words

Corpora	Texts	Words
VESPA-NO (L2-EN)	239	267,855
BAWE (L1-EN)	76	167,437
KIAP-NO (L1-NO)	50	269,913
KIAP-EN (L1-EN)	50	437,798

The comparability of the corpora can be described in terms of Halliday's notions of field ('what is happening'), tenor ('who is taking part') and mode ('what part is the language playing') (Halliday, 1985: 12). The corpora are comparable along the dimension of field: they all come from the discipline of linguistics. However, VESPA and BAWE differ in tenor from KIAP regarding writer expertise (novice vs. expert) and readership (unpublished, implying a limited number of addressees vs. published, implying a greater number of readers). KIAP-NO differs from the others in mode by being written in Norwegian,[3] while VESPA-NO stands out by representing second-language writing.

2.3 Extraction of n-grams

To ensure comparability between the CIA and CA studies, we follow the procedure of the original study, using WordSmith Tools (Scott, 2016) to extract the top100 3- and 4-grams with a frequency threshold of 5 and a range of 3. That is, all the extracted 100 3- and 4-grams are uninterrupted sequences of three and four words that occur at least five times in identical form in at least three different corpus texts (dispersion across individual writers was not checked).[4] Note that our operationalisation of n-grams is the same as Biber *et al*.'s for lexical bundles, i.e. 'recurrent expressions, regardless of their idiomaticity, and regardless of their structural status' (Biber *et al*., 1999: 990), which occur above a set frequency threshold and across a minimum number of corpus texts (Biber *et al*., 1999: 993). See also Ebeling and Hasselgård (2015b: 209) for a survey of studies of lexical bundles in Learner Corpus Research. However, we have chosen to retain the term n-gram, as in our previous study (Ebeling & Hasselgård, 2015a).

Although the study focuses on n-gram types rather than tokens, it is useful to get a sense of token proportion. As shown in Table 2.2, the (token) frequency span of the top 100 n-gram types varies across the corpora, e.g. between 46 and 376 occurrences for English 3-grams produced by the Norwegian learners (VESPA) and between 14 and 117 for English 4-grams produced by professionals whose native language is English (KIAP-EN). (See also Appendix A, Tables A.1 and A.2 for lists of the most frequent 3- and 4-grams in the material.)

Table 2.2 Token frequency span of top 100 3-gram and 4-gram types in VESPA, BAWE and KIAP

	Frequency span 3-grams		**Frequency span 4-grams**	
	N (raw)	per 100k words	N (raw)	per 100k words
VESPA	46–376	17–140	16–102	6–38
BAWE	20–165	12–99	7–32	4–19
KIAP-NO	23–166	9–62	7–62	3–23
KIAP-EN	50–238	11–54	14–117	3–27

These discrepancies in number of occurrences demonstrate not only frequency differences relating to n-gram length but, potentially, also differences in corpus size and differences between the languages. In all the corpora the recurrence of 3-grams is generally higher than that of 4-grams, which is as expected: the shorter the n-gram the greater its chance of recurring in identical form. The largest corpus, KIAP-EN, to some extent shows that size matters, in that it has the highest frequency of 4-gram tokens as well as the most frequent 3-gram at rank 100 (with 50 occurrences) in terms of raw numbers. However, when tokens are normalised per 100,000 words, it can be seen that, relatively speaking, it is in fact the learners in VESPA who produce the most frequently recurring n-grams ranked 1-100. We can only speculate as to the reason for this, but it could be that Norwegian learners of English, as indicated in previous research (Hasselgård, 2019), draw on a smaller number of chunks that they use more frequently in the same way as they over-use high-frequency core vocabulary (Hasselgren, 1994; Ringbom, 1998). In other words, learners may be more repetitive than native speakers.[5]

Finally, we would have expected the token recurrence in KIAP-NO to be markedly lower than in the English language corpora because of what has previously been found in contrastive studies of fiction texts (Ebeling & Ebeling, 2017; Hasselgård, 2017): Norwegian is generally less recurrent than English, i.e. fewer sequences recur frequently in identical form. This may be due to several factors, including a relatively large number of accepted spelling/inflectional variants in Norwegian (e.g. *på den eine|ene sida|siden* ‘on the one hand’) and morphological and syntactic differences between English and Norwegian (for instance, Norwegian compound nouns are usually spelt as one word, and definiteness is marked by a word-final morpheme instead of a definite article, as in *the word order* = *ordstilling*<u>*en*</u>, or the Norwegian verb-second constraint which makes the English 4-gram *we have seen* correspond to both *vi har sett* and *har vi sett*, depending on the context). These differences notwithstanding, the token counts for the top 100 3- and 4-grams in Norwegian linguistics articles do not stand out in comparison with English.

2.4 Functional classification of n-grams

Moon’s (1998) taxonomy for the classification of fixed expressions and idioms is central to our functional classification of n-grams. Our adapted version of Moon’s original model (Ebeling & Hasselgård, 2015a), given in Figure 2.1, contains five broad categories: informational, situational, evaluative, modalising and organisational. Each category is exemplified in Figure 2.1 by a 3- or 4-gram. As seen to the left of the figure, the model is grounded in Halliday’s three metafunctions of language (e.g. Halliday, 1994: 36). It may be noted that the categories correspond roughly to those found in, for example, Biber *et al.* (2004), i.e. referential, stance and discourse organisers (ibid.: 384). The model we have applied here is a bit more finegrained, with three interpersonal categories (Moon, 1998: 218). These are distinguished as follows. Evaluative n-grams convey evaluations and attitudes apart from those that are modalising, i.e. that contain a modal expression. Situational n-grams 'relate to extralinguistic context' (Moon, 1998: 217). In Moon's study this category included, for example, greetings and other references to the speakers' surroundings. In our case it mostly consists of references to other texts, such as *in fletcher and garman* in Figure 2.1. Although it is challenging to apply the taxonomy to sequences that do not necessarily constitute 'complete structural units' (cf. Biber & Conrad. 1999: 183), previous research has shown that the application of this model to the functional analysis of n-grams is both possible and fruitful (e.g. Ebeling, 2011; Ebeling & Ebeling 2017; Ebeling & Hasselgård, 2015a). Indeed, and as pointed out by Conrad and Biber (2005: 58-59), n-grams (or lexical bundles, to use their term) that are 'identified purely on frequency criteria do have strong functional correlates, indicating that speakers and writers regularly use them as basic building blocks of discourse'.

	Category	**Function**	**Example**
Ideational	informational	stating proposition, conveying information	*of the brain*
Interpersonal	situational	relating to extralinguistic context, responding to situation	*in fletcher and garman*
	evaluative	conveying speaker's evaluation and attitude	*is important to*
	modalising	conveying truth values, advice, requests, etc.	*we can see*
Textual	organisational	organising text, signalling discourse structure	*in this paper*

Figure 2.1 The functional classification model (adapted from Moon, 1998: 217)

In our classification, we do not allow dual membership of an n-gram. In other words, each potentially functionally ambiguous n-gram has been assigned to one functional class only according to its most frequent use in the relevant corpus. For example, the n-gram *at the same time* was classified as organisational (see example (1)) since this function was more frequent in the material than the informational (temporal) use seen in example (2).

(1) *At the same time*, this grammatical feature is not treated very thoroughly by Tottie or Algeo… (VESPA-NO)

(2) In the example above, it seems that the process of treading water is happening *at the same time* as Bernard says he is sorry. (VESPA-NO)

3 Contrastive Interlanguage Analysis: Previous Study

Ebeling and Hasselgård (2015a) compared the use of recurrent word-combinations in texts written in English by L1 Norwegian (VESPA) and L1 English (BAWE) university students of linguistics and business. We investigated 3- and 4-grams extracted from the BAWE and VESPA corpora, classified functionally according to the model presented in Figure 2.1, to answer the following research questions:

i. What discourse functions do the recurrent word-combinations have?
ii. To what extent are the same patterns and functions used by learners and native speakers?
iii. To what extent are the same patterns and functions used in both disciplines?

iv. Is L1 background or discipline more decisive for the use of recurrent word-combinations and their functions? (Ebeling and Hasselgård 2015a: 88)

The study uncovered a somewhat complex picture. The distribution of some of the functional categories of n-grams was shown to distinguish learners from native speakers in both linguistics and business. For example, in linguistics, the learners were found to use fewer evaluative and more organisational n-grams than the native speakers (see Table 2.3). In the business material (not included in Table 2.3) the Norwegian learners were found to use more informational and fewer modalising n-grams than their native peers.

Some differences between the learners and native speakers were also observed regarding the form of the n-grams used. N-grams involving first-person pronouns were more frequent among the learners, a finding that substantiates previous research reporting that (Scandinavian) learners of English tend to be visible authors (e.g. Petch-Tyson, 1998; Paquot *et al*., 2013). A prominent feature among the native speakers, by contrast, was the relatively frequent use of n-grams with non-personal projection (extraposition, e.g. *it is evident that; it is important to*) as well as n-grams including complex noun phrases (e.g. *of the language; the extent to which*); a similar trend was noted by Paquot (2013: 292). Finally, passive verb phrases, such as *been found to, has been suggested that*, were also more frequently used by the native speakers.
Regarding research questions (iii) and (iv), we found that there were more (statistically significant) differences across disciplines than across L1 groups, as attested by, for example, more overlapping n-grams between the corpora in linguistics than in business and by more evaluative and modalising n-grams in linguistics compared to business across L1 backgrounds. It was concluded that, despite the differences noted across L1 groups, 'the Norwegian learners – particularly the linguistics students – are in fact advanced users of English who are to a great extent able to adapt to disciplinary conventions' (Ebeling & Hasselgård, 2015a: 102).

The final section of the previous study suggested some avenues for further research, one of which was to compare the output of the apprentice academics represented in BAWE (native speakers of English) and VESPA (learners of English) to published academic writing in order to examine the extent to which they match the usage of experts in the field. This is, to a large degree, what the present study aims to do. Within the framework of the Integrated Contrastive Model, we follow the same research structure as the 2015a study to perform a contrastive analysis of functional types of n-grams in English and Norwegian research articles in linguistics. The results from the previous CIA of the BAWE and VESPA linguistics n-grams will then be reassessed in the light of the fresh CA based on the KIAP corpus, representing professional writing in linguistics by native speakers of English and Norwegian. As mentioned above, there are two motivations for keeping to one academic discipline: to limit the scope (and complexity) of the comparison so that a clearer picture may emerge and because there are greater problems of corpus comparability in the business/economics sections of BAWE, VESPA and KIAP than in the linguistics sections.

We thus seek to establish to what extent the Norwegian learners of English may be influenced by their L1 and to shed some light on how apprentice academics compare with professionals with regard to n-gram use. Table 2.3 (Ebeling & Hasselgård, 2015a: 95) and the observations below are repeated here to provide a starting-point and basis for the discussion in Section 5.

Table 2.3 Learners' (VESPA) and native speakers' (BAWE) use of n-gram types according to function

	3-grams			**4-grams**		
	BAWE	VESPA	p-value	BAWE	VESPA	p-value
Informational	46	57	0.1571 (p > 0.05)	42	49	0.3942 (p > 0.05)
Situational	1	0		4	0	0.1297 (p > 0.05)
Evaluative	24	8	0.003814 (p < 0.01)	29	15	0.02648 (p < 0.05)
Modalising	16	9	0.1995 p > 0.05	11	14	0.6689 (p > 0.05)
Organisational	13	26	0.03222 (p < 0.05)	14	22	0.1976 (p > 0.05)
	100	100		100	100	

Table 2.3 gives an overview of the distribution of the top 100 3- and 4-gram types according to their function in the BAWE and VESPA linguistics assignments. A test of equal proportions was carried out pairwise for each of the functional classes (`prop.test` in R), producing a p-value in each case. Cells with statistically significant results are shaded in grey. These show that the native speakers use more evaluative 3- and 4-gram types and fewer organisational 3-gram types than the Norwegian students. It is also worth mentioning that both the BAWE and VESPA n-grams are most typically informational, accounting for +/-50% of all n-gram types. Moreover, in BAWE the second-most frequent functional type is evaluative, while in VESPA it is organisational. A relatively similar distribution is found for modalising n-grams, while situational ones are marginal in both L1 groups.

To investigate whether these differences can be attributed to the influence of Norwegian, we now turn to a similar analysis of n-grams in published academic writing in L1 English and L1 Norwegian (Section 4). This analysis will also enable us to compare novice and professional writing to assess the extent to which students have acquired the functional phraseology of the field in terms of n-gram use (Section 5).

4 Contrastive Analysis

4.1 Comparing n-grams across L1s: English vs. Norwegian

In the following we present the discourse functions of the top 100 3- and 4-gram types in the English and Norwegian linguistics articles from KIAP before discussing some of the salient word-combinations included in some of these functional classes.

4.1.1 The functions of the n-grams

Table 2.4 shows the distribution of 3- and 4-grams according to function in texts produced by linguists whose native language is English (KIAP-EN) and Norwegian (KIAP-NO), respectively. To enable a direct comparison with the previous study, a test of equal proportions was again carried out pairwise for each of the functional classes, with statistically significant results highlighted in grey.

Table 2.4 English and Norwegian n-gram types according to function in published articles

	3-grams			4-grams		
	KIAP-EN	KIAP-NO	p-value	KIAP-EN	KIAP-NO	p-value
Informational	75	64	0.1246 (p>0.05)	57	39	0.01612 (p<0.05)
Situational	0	0		2	1	
Evaluative	2	19	0.0002237 (p<0.001)	15	29	0.02648 (p<0.05)
Modalising	2	4	0.6785 (p>0.05)	8	13	0.3562 (p>0.05)
Organisational	21	13	0.1876 (p>0.05)	18	18	1 (p>0.05)
	100	100		100	100	

Table 2.4 reveals that n-grams of the informational type are the most salient ones across the board. In particular, informational 3-grams constitute a very high proportion of the total 100 3-gram types. Although the proportion of informational 4-grams is lower, they are still the single-most frequent functional category among the 4-grams in both languages. Thus, not unexpectedly, linguistics as a research register is primarily informational. There is, however, a difference between English and Norwegian in the proportion of informational 4-gram types, which are significantly more frequent in the English data. One potential reason for this will be discussed below (Section 4.1.2).

Regarding the other functions, English 4-gram types show a varied distribution across organisational, evaluative, and to some extent modalising, while the only frequent functional 3-gram type, in addition to informational, is organisational. The Norwegian 'non-informational' 3-grams, on the other hand, are typically either evaluative or organisational, while the 4-grams show the same tendency as the English 4-grams. Situational n-grams are virtually non-existent among the top 100 in both English and Norwegian. This lack of situational n-gram types in the material may be due to the extraction method, which requires recurrence and dispersion of identical sequences. Although such sequences may be frequent in individual texts (e.g. the 4-gram *in hopper and traugott*), they do not often meet the recurrence/dispersion thresholds set. Moreover, Moon (1998: 225) notes that instances of the situational category 'are typically found in spoken discourse as they are responses to or occasioned by the extralinguistic context'.

Perhaps the most striking observation to be made on the basis of Table 2.4 is the Norwegian partiality to evaluative sequences: the numbers of evaluative 3- and 4-gram types are significantly higher in the Norwegian linguistics articles.

It can be concluded that 3- and 4-grams in English and Norwegian linguistics articles are similarly distributed across the functional classes, suggesting that there is some consensus among professionals, regardless of language, on the writing style of a research article at the functional level of n-grams within this discipline. The exceptions are the prominent use of evaluative 3- and 4-gram types in Norwegian and of informational 4-gram types in English.

4.1.2 The form of the n-grams

A direct comparison of the form of n-grams across languages is challenging, bordering on the impossible, for several reasons.[6] First, there is the general challenge in contrastive analysis of how to make sure to compare like with like. Although our n-grams have been classified within the same functional categories, it may not be fair to juxtapose seemingly similar n-grams, e.g. the evaluative 3-grams *the fact that* and *det faktum at* 'that fact that' or the modalising 4-grams *to be able to* and *er i stand til* 'is in condition to'. In both cases, the English and Norwegian n-grams are intuitively good correspondences of each other; however, their equivalence has not been established on the basis of any objective *tertium comparationis*. In fact, the researchers' own bilingual knowledge is arguably given too much weight, together with the formal similarity attested for the 3-grams in particular. We do not know, for instance, whether the formally similar n-grams have the same semantic and syntactic preferences. Moreover, when carving up languages into recurrent strings of words, systematic morphosyntactic differences between the languages show themselves to have a bearing on the length and internal structure of a recurrent sequence (see, for example, Ebeling & Ebeling 2017; Granger, 2014; Hasselgård, 2017). Nevertheless, some insight may be gained by examining the actual realisation of the n-grams in English and Norwegian.

For example, we notice a marked difference in the syntactic structure/realisation of English and Norwegian informational 4-grams. While the Norwegian 4-grams are mainly VP- (i.e. clausal) or PP-based, the English 4-grams are mainly NP- (i.e. nominal) or PP-based, to use Chen and Baker's (2010) terms. Typical Norwegian examples are the VP-based *at det ikke er* 'that it/there is not' and *å ta utgangspunkt i* 'to take startingpoint in' and the PP-based *i form av en* 'in form of a', and *i den forstand at* 'in the sense that'. Further, it can be noted that the VP-based Norwegian 4-grams often include the versatile pronoun *det*, which may correspond to either *it* or *there* in English, both of which are in evidence in the English VP-based 4-grams, albeit not as prominently as in Norwegian.[7] Typical examples of the more nominal English informational 4-grams, include *the referent of the* and *the extent to which* and PP-based ones: *on the basis of* and *in a number of*. Both types often include (fragments of) complex noun phrases with a determiner and the preposition *of*. In sum, the two languages clearly differ in their recurrent 4-word sequences, to the extent that English informational 4-grams (typically nominal) significantly outnumber the Norwegian informational 4-grams (typically clausal). The observations regarding English reflect a general tendency, noted by Biber *et al.* (1999: 992), for 'bundles' in academic prose to be nominal rather than clausal.

The number of evaluative n-grams also differs significantly between English and Norwegian (Table 2.4). The evaluative 3- and 4-grams are, with the exception of the two English 3-grams, mainly VP-based in both languages. The 4-grams also have in common the fact that many of them form part of non-personal (self) projection expressions, i.e. anticipatory-*it* stance constructions, such as *it is important to, it is clear that* and *det er interessant å* 'it is interesting to'. In addition, Norwegian has a productive and variable sequence that contributes to boosting the frequency of 3- and 4-grams, namely *ut til å|at* 'out to to|that', with or without the verb *se* 'look', nesting within the longer sequence *det ser ut til å|at* 'it looks out to to|that' ~ 'it seems to|that'. These n-grams are borderline cases between evaluative and modalising. Two closely related n-grams in our English material – *seems to be* and *appears to be* – have rather been classified as modalising, in line with Quirk *et al.* (1985: 146), who suggest that *seem to* and *appear to* are catenatives with 'meanings related to aspect and modality'. We consider the (*det ser*) *ut til å|at* sequences to have a stronger evaluative than modalising content. Similar sequences, such as *kan se ut til* 'can look out to' ~ 'can/may seem to', are classified as modalising due to the presence of the modal auxiliary *kan*.

To speculate further as to why Norwegian has significantly more evaluative 3- and 4-grams, we refer to the morphosyntactic differences between the languages, discussed above (Section

2.3). Could it be that 3- and 4-word sequences in Norwegian typically correspond to, for example, one or two words in English and would therefore not figure on our lists? While this may apply to some of the n-grams (e.g. *i det hele tatt* 'at all'/'overall'), it does not seem to be a major contributing factor in the material at hand. It also begs the question of why this should be a factor for the evaluative n-grams only.

5 Novice vs. Expert Use of N-grams

We have now established similarities and differences in the functions of 3- and 4-grams used by novice learners vs. native speakers of English and by academic professionals in English vs. Norwegian. The final part of the puzzle, addressing our third research question, is a comparison between novice and expert writing in English. First, Table 2.5 compares the functional distribution of 3- and 4-gram types in the English native-speaker data: novices in BAWE and professionals in KIAP-EN.

Table 2.5 English n-gram types according to function in native novice (BAWE) vs. native professional writing (KIAP-EN)

	3-grams			**4-grams**		
	BAWE	KIAP-EN	p-value	BAWE	KIAP-EN	p-value
Informational	46	75	5.119e-05 (p<0.001)	42	57	0.0477 (p<0.05)
Situational	1	0		4	2	
Evaluative	24	2	1.008e-05 (p<0.001)	29	15	0.02648 (p<0.05)
Modalising	16	2	0.001318 (p<0.01)	11	8	0.6296 (p>0.05)
Organisational	13	21	0.1876 (p>0.05)	14	18	0.5628 (p>0.05)
	100	100		100	100	

It is clear from Table 2.5 that the British linguistics students do not match the usage of the professionals, as statistically significant differences are found in three of the functional categories of 3-grams and in two categories of 4-grams. Informational 3- and 4-grams are underrepresented in BAWE compared to KIAP-EN (e.g. *account of the*, *denoted by the*, *on the basis of*), whereas evaluative 4-grams (e.g. *is important to note*, *it is clear that*) and modalising 3-grams are overrepresented (e.g. *more likely to*, *the ability to*).

Table 2.6 English n-gram types according to function in non-native novice (VESPA) vs. native professional writing (KIAP-EN)

	3-grams			**4-grams**		
	VESPA	KIAP-EN	p-value	VESPA	KIAP-EN	p-value
Informational	57	75	0.01116 (p<0.05)	49	57	0.3213 (p>0.05)
Situational	0	0		0	2	
Evaluative	8	2	0.1048 (p>0.05)	15	15	1 (p>0.05)
Modalising	9	2	0.06275 (p>0.05)	14	8	0.2585 (p>0.05)
Organisational	26	21	0.5047 (p>0.05)	22	18	0.5959 (p>0.05)
	100	100		100	100	

Second, Table 2.6 shows how the learners in VESPA compare with the native-speaker professionals in KIAP-EN. Surprisingly, the learners differ much less from the native expert writers than their native student peers (Table 2.5) in terms of n-gram functions. The only significant difference concerns the use of informational 3-grams. It is hard to interpret Tables 2.5 and 2.6 in a meaningful way as it is somewhat counter-intuitive that the learners should have a better grasp of the functional conventions of the discipline than the native-speaker students. We will supplement this analysis with more qualitative considerations in Section 6.

6 Discussion: Linking Up the Contrastive Interlanguage Analysis and the Contrastive Analysis

6.1 Comparing functional types of n-grams across the corpora

The CIA and CA analyses presented above show that informational n-grams are the most salient ones across the board. However, while evaluative n-grams are more frequent in Norwegian than in English published articles, the novices show the opposite trend, with evaluative n-grams being more frequent with L1 English students than with Norwegian learners (although the difference in 4-grams is not statistically significant). While the Norwegian learners appear to use evaluative n-grams in a similar fashion to expert English writers (see Table 2.6) there are unexpected similarities between the L1 English novices and L1 Norwegian experts. The L1 English students in BAWE and the L1 Norwegian experts in KIAP-NO have greater proportions of evaluative n-grams than the other two corpora, apparently at the expense of (especially) informational n-grams, and thereby seem to foreground interpretations and evaluations more than the other corpora. Figures 2.2 and 2.3 visualise the distribution of functional types of n-grams across all four sub-corpora. Situational n-grams have been omitted due to their low frequencies.

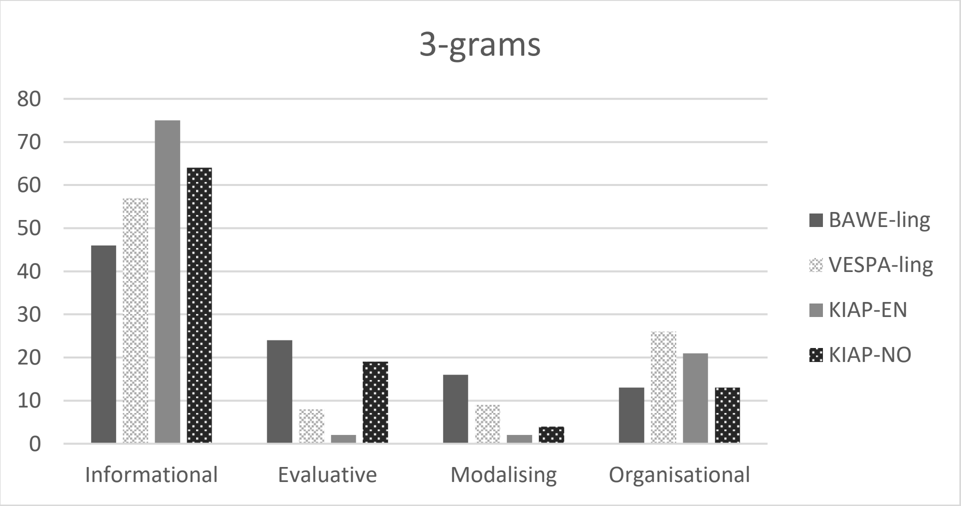

Figure 2.2 Functional types of 3-grams across the corpora

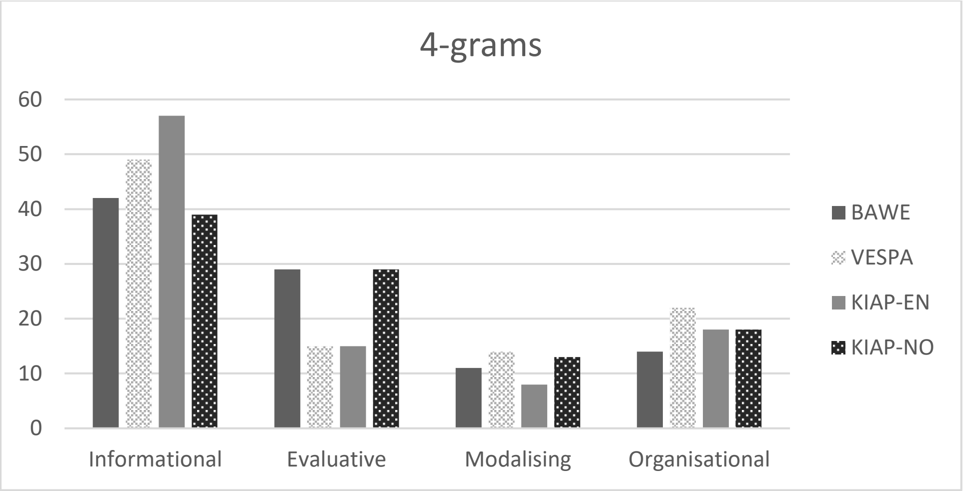

Figure 2.3 Functional types of 4-grams across the corpora

As noted above, it is difficult to see why Norwegian learners should be more similar to English than to Norwegian expert writers (particularly in their use of 3-grams). However, as the learners represented in the VESPA corpus are students of English, they are probably more used to reading about linguistics in English than in Norwegian and may thus have picked up wordings from their course reading. It is even more remarkable that the L1 English novices should resemble the Norwegian experts more than they do the L1 English experts (see also Table 2.5) in their distribution of functional types of n-grams. It is likely that our focus on *types* rather than *tokens* could be part of the explanation. For example, there are more evaluative tokens among the 20 most frequent 4-gram types in VESPA than in KIAP-EN, even if they have a similar number of types (see Figure 2.3). Furthermore, the broad classification into functional types masks both similarities and differences in the realisation and form of n-grams, to which we now turn.

6.2 The realisation and form of the n-grams

A modest number of n-gram types is shared across the corpora. Table 2.7 shows the 4-grams that occur in more than one of the English-language corpora, i.e. BAWE, VESPA and KIAP-EN. It is striking that the majority of the shared 4-grams are general in meaning and give away little about the discipline of the texts. The informational 4-grams in Table 2.7 are almost all either PP-based or NP-based, and most include the preposition *of* (see also Biber *et al*., 1999: 1014 ff; Chen & Baker, 2010: 35). However, many of the informational n-grams that are *not* shared across the corpora seem to reflect topics that are simply not present (with sufficient distribution and frequencies) in the other corpora. Examples are *the semantics of the, the argument structure of, the semantic bootstrapping, a lexical teddy bear, in the Norwegian translations*. Notably, the novice British writers in BAWE share more identical 4-grams with L1 English experts in KIAP-EN than the learners in VESPA do, which suggests that in terms of actual lexicalisation, the L1 novice writers are closer to the phrasicon of research publications within their discipline.

Table 2.7 Shared 4-gram types across the corpora

	Informational	**Evaluative**	**Modalising**	**Organisational**
BAWE + VESPA + KIAP-EN	at the end of in the case of in the use of on the basis of that there is a the meaning of the the use of the	in the same way it is important to the fact that the	it is possible to to be able to	as well as the in this case the on the other hand
BAWE + VESPA	and the use of of the use of	to the fact that	can be found in can be seen in	an example of this example of this is in this essay i is an example of
BAWE + KIAP-EN	in terms of the in the context of that there is no the nature of the the way in which the ways in which	by the fact that it is clear that	can be used to	with respect to the
VESPA + KIAP-EN	the end of the			at the same time

The shared 3-grams reveal a relatively similar pattern to the 4-grams except that only two interpersonal 3-grams occur in all three corpora (*in the same* and *the fact that*). On the other hand, the two novice corpora share five evaluative and six modalising 3-grams (e.g. *due to the, meaning of the; can also be, can be seen*). In the case of 3-grams, too, there is a greater similarity between the native speakers of English in BAWE and KIAP-EN than between the learners in VESPA and KIAP-EN, especially as regards informational n-grams, while VESPA shares a few more organisational n-grams with the English L1 experts (e.g. *in other words, in the following, in this paper*).

Examining the intuitively similar 3- and 4-grams in Norwegian professional writing (KIAP-NO) and English learner writing (VESPA), with the reservations against direct cross-linguistic comparison of Norwegian and English n-grams expressed above, we find that the highest degree of overlap occurs in the organisational category. In fact, about half of the recurrent organisational n-grams in KIAP-NO have a counterpart in VESPA. Some examples are *i denne artikkelen* ('in this article') – *in this paper/essay, i dette tilfellet – in this case, i tillegg til – in addition to, når det gjelder* ('when it concerns') – *when it comes to, på den annen side* ('on the other side') – *on the other hand, et eksempel på en – an example of a*.[8] In addition, two of the three analogous modalising n-grams are metadiscursive, and thus also have a text-organising function, namely the pairs *jeg vil hevde at* ('I will claim that') – *I would say that* and *i denne artikkelen skal + jeg|vi* ('in this article shall + I|we') – *[in] this essay I will*. This is interesting because it suggests that Norwegian learners of English organise their texts along the lines of academic Norwegian.

The other functional types of n-grams have less 'overlap' between Norwegian and L2 English, although we may note that some evaluative n-grams are similar, e.g. *det er vanskelig å – it is hard/difficult to*. The low degree of formal similarity is presumably due to systemic differences between the languages (Ebeling & Ebeling, 2017; Hasselgård, 2017) as well as differences in topics, particularly in the case of informational n-grams. Interestingly, KIAP-NO, like VESPA, has a good number of n-grams that involve self-reference (cf. Section 3 above and Paquot *et al.*, 2013). These include, in addition to the organisational ones listed above, *det vi kan kalle* (‘what we can call’), *kan vi si at* (‘can we say that’), and *etter mitt syn* (‘in my view’). This agrees with Fløttum *et al.*’s (2006: 70) finding that first-person pronouns are more frequent in Norwegian than in English linguistics articles. The frequent use of self-reference in the English of Norwegian learners can thus potentially be linked to their L1 writing culture. However, this tendency has also been noted by, for example, Granger (2017) for learners of English more generally, as such self-referencing was shown to be typical of quite a few L1 populations, including French, Spanish, Italian, Norwegian, Swedish and German.

As noted above, there are more VP-based (clausal) n-grams in the Norwegian learner corpus than in the English L1 novice corpus, which in return contains more NP-based n-grams, suggesting a more nominal style of writing. Since a similar difference was found between English and Norwegian expert texts, it is possible that the learners’ more verbal style comes from their L1, although the developmental factor cannot be ruled out. However, a nominal style has been identified as a hallmark of English academic writing. For example, Biber and Gray (2016: 110) show that academic registers ‘have developed a distinctive grammatical style, employing a dense use of nouns and phrasal modifiers rather than verbs and clauses’. This may be illustrated by example (3), which comprises a 4-gram, *the way in which*, which is typical of English academic discourse (Groom, 2019: 303) but not shared by the Norwegian learners.

(3) Duality of language highlights *the way in which* elements and segments of language are combined to form words, expressions and phrases. (BAWE)

It was noted in Section 3 that the Norwegian learners in VESPA under-use extraposition for evaluation, as in example (4) from BAWE. The n-gram lists for KIAP-NO show, however, that evaluative extraposition is as frequent in Norwegian as in English linguistics articles; see Section 4.1.2 and example (5). The shortage of such n-grams in VESPA is thus not attributable to the learners’ L1.

(4) … and *it is clear that* these different methods of communication are learnt in different ways. (BAWE)

(5) … og *det er rimelig å* tru at det samme gjelder for norsk. (KIAP-NO)
Lit: “…and *it is reasonable to* think that the same applies to Norwegian.”

However, our finding that Norwegian learners use fewer n-grams that reflect passives and nominalisations than their peers in BAWE might be L1-related: there are no ‘passive’ n-gram types among the top 100 in KIAP-NO and very little evidence of nominalisation. VESPA, on the other hand, does contain passive n-grams, e.g. *be found in the*, *can be seen as*, which indicates that the learners have adopted wordings from their academic reading in English, albeit in smaller proportions than the novice native writers.

As noted above, Figures 2.2 and 2.3 show an unexpected similarity between BAWE (novice L1 English) and KIAP-NO (expert Norwegian) in the proportions of evaluative n-grams. A scrutiny of the 4-grams, where the pattern is most pronounced, shows that the proportional similarity is not reflected in the content of the n-grams, as there is little overlap in actual realisations of the 4-grams. The exception is the use of the evaluative frame *it is* ADJ *to/that* and its Norwegian counterpart *det er* ADJ *å/at*, as illustrated in examples (4) and (5) above. Many of the Norwegian evaluative 4-grams comprise a disjunct adverbial, e.g. *er først og fremst* (‘is first and foremost’), *i det hele tatt* (‘at all’), *til en viss grad* (‘to a certain degree’). In comparison, a large proportion of the KIAP-EN list of evaluative 4-grams consists of extraposition and sequences involving the word *fact*. The BAWE list contains more expressions denoting causes and effects, e.g. *a result of the*, *due to the fact*, *for the purposes of*, *this is due to*. The VESPA list is relatively similar to the BAWE one, but shorter and slightly more concerned with (non-causal) relations, e.g. *have to do with*, *the same meaning as*.

7 Concluding Remarks

The present study has used a contrastive analysis of English and Norwegian published academic texts to look for explanations for differences in the use of functional types of n-grams in novice writing between Norwegian learners and native speakers of English, as uncovered in Ebeling and Hasselgård (2015a).

The contrastive analysis proper revealed that the field of linguistics adopts similar writing styles in English and Norwegian in terms of functional classes of frequently occurring 3- and 4-gram types. The main difference between the languages is the markedly more frequent use of evaluative n-grams in the Norwegian research articles. At a more detailed level, regarding the form of the n-grams, it was noted that L1 English linguists prefer a nominal (NP-based) style compared to the more clausal (VP-based) style of L1 Norwegian linguists.

In compliance with the Integrated Contrastive Model (Granger, 1996), the contrastive analysis enabled us to reassess and compare the results from the previous CIA that was similarly concerned with the functions of n-grams. The quantitative analysis gave inconclusive and to some extent contradictory results, in particular the apparent similarities in the proportions of functional types of n-grams between Norwegian learners and English experts, on the one hand, and L1 English students and Norwegian experts, on the other. Moreover, the hypothesis put forward in Section 1 – that the novice writers would resort to more organisational n-grams than the experts – was not substantiated (see Tables 2.5 and 2.6).

The analysis of n-grams gives an indication of how similar texts are in terms of function. However, a deeper understanding is gained if we look ‘behind the scenes’ at the actual realisations of the n-grams, where we can see how the functions are lexicalised across languages and interlanguages. Comparing the lexicalisations of the n-grams more qualitatively, we found that the writing of Norwegian learners may indeed be coloured by the style of academic articles in their L1. Hence, the Norwegian learners share some lexical and discursive features with L1 expert Norwegian which distinguish their academic writing from L1 English academic writing. In particular, this concerns more clausal n-grams and fewer nominal ones and a more frequent use of self-reference. There were also important similarities in the organisational n-grams between Norwegian learners and Norwegian L1 experts, suggesting that the Norwegian learners of English bear traces of a Norwegian writing culture. However, the scarce use of the evaluative frame *it is* ADJ *that/to* among the learners cannot be attributed to L1 influence, since a formally similar pattern is frequent in KIAP-NO.

The survey of shared n-grams across the corpora showed that the L1 English novices seem closer to the L1 English experts than the learners are. However, at the same time, the Norwegian learners of English also show similarities with L1 writing in English, such as the somewhat more frequent use of passive n-grams than L1 Norwegian and the simple fact, not commented on above, that none of the recurrent n-grams seems unidiomatic. Hence, the present investigation confirms the impression formed in our 2015a study, that ‘the Norwegian learners […] are in fact advanced users of English who are to a great extent able to adapt to disciplinary conventions' (Ebeling & Hasselgård. 2015a: 102) although we can trace a slight Norwegian accent in their writing.

This study has some obvious limitations that need to be reiterated. Not unexpectedly, some of these concern comparability, both in terms of corpus size (see Table 2.1) and the challenges of comparing n-grams across languages (see Section 4.1.2). These are not trivial matters but we have tried to reduce the effect of these variables by pointing them out and, in the latter case, to mainly compare functional classes, thereby keeping the direct cross-linguistic comparison of individual n-grams to a minimum. Nevertheless, the contrastive analysis may not have given a true picture of similarities and differences between English and Norwegian academic phraseology because of the generally greater variability of Norwegian in terms of, for example, spelling and syntax. For example, while the English 3-gram *we have seen* may meet the frequency requirements, the two Norwegian variants *vi har sett|har vi sett* may not, simply because neither of them will meet the frequency threshold (cf. Section 2.3 and previous studies of n-grams in English and Norwegian: Ebeling & Ebeling, 2017; Hasselgård, 2017).

Another limitation concerns the problems related to cross-linguistic comparisons based on comparable corpora, and the absence of a completely unbiased common ground against which comparisons across languages can be made (see Section 4.1.2). However, functional classes are arguably better suited for contrastive analysis based on comparable data of this kind than, for example, lexical studies, as they are abstracted from established grammatical categories and lexicalisations.

In spite of these limitations, the study has contributed further insight into the use of phraseological sequences across several writer groups and we would strongly encourage further research to be conducted in this field. It would be of great interest to apply the same integrated contrastive approach to more disciplines, more L1 learner groups as well as more languages, in order to gain even more knowledge in this area, not least to further differentiate L1 influence from the interlanguage factor (cf. Jarvis, 2000; Paquot, 2013).

Notes

[]

[1] The acronym stems from the Norwegian name of the (corpus) project: Kulturell Identitet i Akademisk Prosa.

[2] The word counts exclude text in footnotes, block quotes and headlines. See Ebeling and Heuboeck (2007) and the corpus manuals for VESPA and BAWE (Heuboeck *et al*., 2008; Paquot *et al*., 2010) for information on the annotation that facilitates the automatic exclusion of text not produced by the students, and Fløttum *et al.* (2006: 7) on the word counts in KIAP.

[3] Halliday's definition of mode does not mention different languages but uses the phrase 'symbolic organisation of the text' (Halliday, 1985: 12), which explicitly includes the speech/writing contrast, and has been extended here to also include language code.

[4] Where the number of occurrences of n-gram number 100 was identical for several n-grams, we included the (alphabetically) first n-gram to reach the top 100, in order to get an equal number from each (sub-)corpus.

[5] It is also possible that course assignments in relatively large student groups will have prompted the use of certain expressions across corpus texts in VESPA. See, for example, Ädel (2015: 409), who notes: 'even small differences in prompts or assigned topics affect the written production'.

[6] However, measures have been proposed to counter these challenges (see, for example, Chlumská & Lukeš, 2018; Cortes, 2008; Granger, 2014; Milička *et al*., 2019).

[7] This is in line with what previous studies have reported regarding the use of dummy subject *det* vs. *it*/*there* constructions in Norwegian and English (Ebeling, 2000; Ebeling & Ebeling, 2020; Gundel, 2002).

[8] Those Norwegian n-grams that are not followed by glosses correspond word for word to their English counterparts.

References

Ädel, A. (2015) Variability in learner corpora. In S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research* (pp. 401-421). Cambridge: Cambridge University Press.

Alsop, S. and Nesi, H. (2009) Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 (1), 71-83.

Biber, D. and Conrad, S. (1999) Lexical bundles in conversation and academic prose. In H. Hasselgård and S. Oksefjell (eds) *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Biber, D. and Gray, B. (2016) *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2004) 'If you look at…': Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25 (3), 371-405.

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

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

Chlumská, L. and Lukeš, D. (2018) Comparing the incomparable? Rethinking n-grams for free word-order languages. In S. Granger, M.-A. Lefer and L. Aguiar de Souza Penha Marion (eds) *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition)*. CECL Papers 1. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain, 40-41. Available at: https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/c850523d-1953-4204-964d-c6d1ee174bfe/UCCTS2018_book_of_abstracts_with%20correction.pdf?guest=true.

Conrad, S. and Biber, D. (2005) The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* vol. 20, 56-71.

Cortes, V. (2008) A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3 (1), 43-57.

Ebeling, J. (2000) *Presentative Constructions in English and Norwegian: A Corpus-based Contrastive Study*. Oslo: Acta Humaniora.

Ebeling, S.O. (2011) Recurrent word-combinations in English student essays. *Nordic Journal of English Studies* 10 (1), 49-76.

Ebeling, S.O. and Hasselgård, H. (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. and Hasselgård, H. (2015b) Learner corpora and phraseology. In S. Granger, G. Gilquin and F. Meunier (eds) *The Cambridge Handbook of Learner Corpus Research* (pp. 207-229). Cambridge: Cambridge University Press.

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

Ebeling, S.O. and Ebeling, J. (2017) A cross-linguistic comparison of recurrent word-combinations in a comparable corpus of English and Norwegian fiction. In M. Janebová, E. Lapshinova-Koltunski and M. Martinková (eds) *Contrasting English and* *other Languages through Corpora* (pp.2-31). Newcastle: Cambridge Scholars Publishing.

Ebeling, S.O. and Ebeling, J. (2020) Dialogue vs. narrative in fiction: A cross-linguistic comparison. In S. Granger and M-A. Lefer (eds) *The Complementary Contribution of Comparable and Parallel Corpora to Crosslinguistic Studies*, special issue of *Languages in Contrast* 20 (2), 389-314.

Fløttum, K., Dahl, T. and Kinn, T. (2006) *Academic Voices*. Amsterdam: Benjamins.

Gilquin, G. (2000/2001) The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3 (1), 95-124.

Granger, S. (1996) From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, and M. Johansson (eds) *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies* (pp. 37-51). Lund Studies in English 88. Lund: Lund University Press.

Granger, S. (2014) A lexical bundle approach to comparing languages: Stems in English and French. In M.-A. Lefer and S. Vogeleer (eds) *Genre- and Register-related Discourse Features in Contrast*, special issue of *Languages in Contrast* 14 (1), 58-72.

Granger, S. (2015) Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research* 1 (1), 7-24.

Granger, S. (2017) Academic phraseology: A key ingredient in successful L2 academic literacy. In R. V. Fjeld, K. Hagen, B. Henriksen, S. Johansson, S. Olsen and J. Prentice (eds) *Academic Language in a Nordic Setting – Linguistic and Educational Perspectives, Oslo Studies in Language* 9 (3), 9-27.

Granger, S. (2018) Tracking the third code. A cross-linguistic corpus-driven approach to metadiscursive markers. In A. Čermáková and M. Mahlberg (eds) *The Corpus Linguistics Discourse. In Honour of Wolfgang Teubert* (pp. 185-204). Amsterdam: Benjamins.

Groom, N. (2019) Construction grammar and the corpus-based analysis of discourses. The case of the WAY IN WHICH construction. *International Journal of Corpus Linguistics* 24 (3), 291-323.

Gundel, J. (2002) Information structure and the use of cleft sentences in English and Norwegian. In H. Hasselgård, S. Johansson, B. Behrens and C. Fabricius-Hansen (eds) *Information Structure in a Cross-linguistic Perspective* (pp. 113-128). Amsterdam: Rodopi.

Halliday, M. A. K. (1985) Context of situation. In M. A. K. Halliday and R. Hasan (eds) *Language, Context and Text: Aspects of Language in a Social-semiotic Perspective* (pp. 3-14). First edition. Sydney: University of New South Wales Press.

Halliday, M.A.K. (1994) *An Introduction to Functional Grammar*. London: Arnold.

Hasselgård, H. (2009) Temporal and spatial structuring in English and Norwegian student essays. In R. Bowen, M. Mobärg and S. Ohlander (eds) *Corpora and Discourse – and Stuff. Papers in Honour of Karin Aijmer* (pp. 93-104). Göteborg: Acta Universitatis Gothoburgensis.

Hasselgård, H. (2017) Temporal expressions in English and Norwegian, In M. Janebová, E. Lapshinova-Koltunski and M. Martinková (eds) *Contrasting English and Other Languages through Corpora* (pp. 75-101). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

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

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., Holmes, J. and Nesi, H. (2008) The BAWE Corpus Manual. UK: University of Warwick, University of Reading, Oxford Brookes University. http://www.reading.ac.uk/internal/appling/bawe/BAWE.documentation.pdf

Jarvis, S. (2000) Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language Learning* 50 (2), 245-309.

Johansson, S. (2007). *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies.* Amsterdam: Benjamins.

Leedham, M. (2015) *Chinese Students Writing in English: Implications from a Corpus-driven Study*. London & New York: Routledge.

Milička, J., Cvrček, V. and Lukešová, L. (2019) N-gram length correspondence in typologically different languages based on a parallel corpus. Conference presentation at CL2019, Cardiff, Wales, UK, 22-26 July 2019, http://www.cl2019.org/.

Moon, R. (1998) *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford: Clarendon Press.

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

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

Paquot, M., Ebeling, S. O., Heuboeck, A. and Valentin, L. (2010) The VESPA Tagging Manual. Centre for English Corpus Linguistics (CECL), Université catholique de Louvain.

Petch-Tyson, S. (1998) Writer/reader visibility in EFL written discourse. In S. Granger (ed.) *Learner English on Computer* (pp. 107-118). London: Longman.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

R Core Team. (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

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

Scott, M. (2016) *WordSmith Tools*. Version 7. Stroud: Lexical Analysis Software.

Corpora

BAWE – British Academic Written English corpus: http://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/

KIAP – Cultural Identity in Academic Prose: http://www.uib.no/fremmedsprak/23107/kiap-korpuset

VESPA – Varieties of English for Specific Purposes dAtabase, Norwegian component: http://www.hf.uio.no/ilos/english/services/vespa/

Appendix A

Table A.1 Top 10 3-gram types according to frequency in the four corpora

	VESPA	**BAWE**	**KIAP-EN**	**KIAP-NO**
1	THE USE OF	THE USE OF	THE FACT THAT	AT DET ER ‘that it/there is’
2	IN THE TEXT	IN ORDER TO	THERE IS NO	I FORHOLD TIL ‘in relation to’
3	OF THE TEXT	THE FACT THAT	IN TERMS OF	UT TIL Å ‘out to to’ (≈ as if)
4	AN EXAMPLE OF	AS WELL AS	THERE IS A	NÅR DET GJELDER ‘when it concerns’ (≈ when it comes to)
5	THERE IS A	DUE TO THE	IN WHICH THE	VED HJELP AV ‘with help of’ (≈ through)
6	THE TEXT IS	IN TERMS OF	THAT THERE IS	MEN DET ER ‘but it/there is’
7	USE OF THE	BE ABLE TO	THE USE OF	OG DET ER ‘and it/there is’
8	SEEMS TO BE	ONE OF THE	IT IS NOT	MED ANDRE ORD ‘with other words’ (≈ in other words)
9	PART OF THE	THERE IS A	THE CASE OF	DET ER IKKE ‘it/there is not’
10	IN ORDER TO	MEN AND WOMEN	AS WELL AS	I DENNE ARTIKKELEN ‘in this article’

Table A.2 Top 10 4-gram types according to frequency in the four corpora

	BAWE	**VESPA**	**KIAP-EN**	**KIAP-NO**
1	IT IS IMPORTANT TO	ON THE OTHER HAND	IN THE CASE OF	SER UT TIL Å ‘look out to to’ (≈ looks as if)
2	IN THE CASE OF	THE USE OF THE	ON THE BASIS OF	UT TIL Å VÆRE ‘out to to be’ (≈ (seems) to be)
3	AS A RESULT OF	WHEN IT COMES TO	ON THE OTHER HAND	I OG MED AT ‘in and with that’ (≈ because of)
4	THE USE OF THE	THE MEANING OF THE	THAT THERE IS A	AT DET IKKE ER ‘that it/there is not’
5	TO BE ABLE TO	THE REST OF THE	WITH RESPECT TO THE	I DET HELE TATT ‘in the whole taken’ (≈on the whole)
6	THE WAY IN WHICH	IS AN EXAMPLE OF	HOPPER AND TRAUGOTT 1993	I DEN FORSTAND AT ‘in the sense that’
7	THE FACT THAT THE	AN EXAMPLE OF THIS	AT THE SAME TIME	PÅ DEN ANNEN SIDE ‘on the other side’ (≈ on the other hand)
8	THE WAY WE SPEAK	THE FACT THAT THE	THE END OF THE	DET VIL SI AT ‘it will say that’ (≈ i.e.)
9	CAN BE FOUND IN	IS THE USE OF	IN TERMS OF THE	PÅ SAMME MÅTE SOM ‘on same way as’ (≈ in the same way as)
10	ON THE OTHER HAND	AS WE CAN SEE	THE FACT THAT THE	SER DET UT TIL ‘looks it out to’ (≈ it looks as if)