Learner Corpus Research: LCR2013 Conference Proceedings

BeLLS 2015 Vol. 6, ISBN 978-82-998587-7-9, s. 87–106 BeLLS.uib.no

Learners' and native speakers' use of recurrent word-combinations across disciplines[1]

*Signe Oksefjell Ebeling and Hilde Hasselgård**
University of Oslo

Abstract

This paper compares the use of recurrent word-combinations (n-grams) in texts produced by Norwegian learners of English and native speakers of English in two academic disciplines, namely linguistics and business. The study explores the extent to which the same n-grams are used by learners and native speakers in the two disciplines. Using an adapted version of Moon's (1998) functional framework, we map the functions of the n-grams, distinguishing between three major functions: ideational/informational, interpersonal and textual. The n-grams are extracted from the VESPA and BAWE corpora, representing learner and native language, respectively.

The data reveal a complex picture. Informational n-grams are by far the most frequent type and they seem to be not only discipline-specific, but also topic-specific. There are more n-grams with an interpersonal function (evaluative and modalizing) in the linguistics than in the business discipline. Frequencies of n-grams with a textual/organizational function are more similar across the material. However, there is relatively little overlap in the use of individual n-grams with interpersonal and textual functions across the L1 groups. There is a higher degree of similarity between learners and native speakers in the linguistics discipline than in the business discipline. On the other hand, there is some similarity across disciplines within L1 groups as regards interpersonal and textual n-grams.

Keywords: n-grams, recurrent word combinations, academic writing, disciplinary variation, functional analysis, Norwegian learners of English

* *Principle contact:*
Signe Oksefjell Ebeling, Professor
Faculty of Humanities, University of Oslo, Norway
Tel.: (+47) 22 85 71 16
E-mail: s.o.ebeling@ilos.uio.no

[1] The authors are grateful to two anonymous reviewers for constructive comments and to Bjørn-Helge Mevik in the Department for Research Computing at the University of Oslo for helpful advice on statistics.

Bergen Language and Linguistic Studies (BeLLS), May 30th 2015. © Ebeling and Hasselgård
DOI: 10.15845/bells.v6i0.810

1. Introduction

This study investigates the use of recurrent word-combinations in texts produced by novice writers – both learners and native speakers – across disciplines. These word-combinations are defined as uninterrupted multi-word strings and are also known as lexical bundles or n-grams. More specifically we investigate how salient such n-grams really are in two different academic disciplines and to what extent the same patterns and functions are used by learners and native speakers of English.

The study of recurrent word-combinations, or n-grams, is rewarding "because they give insights into important aspects of the phraseology used by writers in different contexts" (Scott and Tribble 2006, 132). Although not all such combinations are of phraseological interest (cf. Altenberg 1998) or constitute "psycholinguistically salient sequences" (O'Donnell, Römer, and Ellis 2013, 89), they serve as a useful starting point for an investigation of patterns of lexis in student writing across disciplines. Interesting to note in this context is Hyland's (2008, 20) observation that n-grams "occur and behave in dissimilar ways in different disciplinary environments".

Our focus will be on 3- and 4-grams produced by two L1 groups – Norwegian learners of English and native speakers of English – and across two broadly defined disciplines, viz. linguistics and business. The n-grams will be functionally classified following an adapted version of Moon's (1998, 217-218) functional classification framework for "Fixed Expressions and Idioms" (FEIs), including the three main functional categories: ideational, interpersonal, and textual.

It is expected that the types of n-grams may differ between the disciplines and between learners and native speakers (cf. Hyland 2008, 7f, 20). However, due to the frequent claim that language learners are often unaware of genre conventions (e.g. Gilquin and Paquot 2008), there may be less of a disciplinary difference in the learner writing. It is also uncertain whether there will be greater differences between learner and native writing than between the linguistics and the business writing. Drawing on data from two corpora of novice academic writing – VESPA and BAWE (see Section 3) – we seek answers to the following 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?

As the Norwegian learners in question are relatively advanced in their English proficiency, we do not expect to find frequent n-grams that represent lexical errors.

We start by presenting some previous research on the potential of recurrent word-combinations as discipline discriminators and on the use of recurrent word-combinations in texts produced by learners vs. native speakers. The corpora on which the study is based are introduced in Section 3 along with a description of the n-gram extraction method. An overview of the functional classification framework is given in Section 4. The n-gram analysis proper is divided into two parts; first we compare the distribution of n-grams across the two L1 groups (Section 5.2) before we move on to a comparison across the two disciplines (Section 5.3). Section 6 provides a summary and a discussion of the findings, while Section 7 discusses further work and possible applications.

2. Background

In recent years, we have witnessed a steady increase in studies concerned with the use of recurrent word-combinations, n-grams, chains, or lexical bundles, to mention but a few of the terms that have been used. Both genre-related studies and contrastive interlanguage studies have investigated different aspects of such combinations, related to frequency, form, function and phraseological status. We will not attempt a full overview of previous studies dealing with these topics; however, a brief discussion of some of the most relevant studies is in order.

One important source of inspiration is Stubbs and Barth's (2003) study on the use of recurrent phrases as text-type discriminators. Analysing uninterrupted chains from three text-types in the Brown family of corpora: FICTION, BELLES, and LEARNED,[2] Stubbs and Barth show that "different text-types are repetitive in different ways and to different extents" (2003, 62). For example, the LEARNED texts in their material are found to be much more repetitive than the other two text categories. Moreover, although the text-types may be characterized by similar chains, e.g. DET N *of*, such chains are found to contain different nouns depending on text-type (ibid., 78).

A number of other studies have pointed to similar findings,[3] also in terms of n-gram function. In a study comparing the use and functions of n-grams in UK English student (literature) essays and academic prose, Ebeling (2011) concluded that both are academic text-types in the sense of being highly informational in nature. An additional trait of the student essays, however, is that they are typically evaluative (i.e. interpersonal) as well, e.g. including sequences such as *the importance of*, *due to the* and *a sense of*.

On the basis of such previous findings, we infer that n-grams can also be studied as a predictable characteristic of different *disciplines*, as has indeed been shown to be the case in studies by e.g. Cortes (2004), Groom (2005), Hyland (2008), and Ädel and Römer (2012). Although Groom (2005, 272), in his study of two patterns in two genres and two disciplines, concludes that "the present study cannot claim to have proved […] that academic genres and disciplinary discourses can be described and differentiated in terms of their preferred phraseologies", he adds that such a hypothesis is well worth pursuing in the future.

Hyland (2008), for instance, sets out to explore the extent to which 4-word bundles differ by discipline. His material comprises research articles, PhD dissertations and MA/MSc theses in electrical engineering, microbiology, business studies and applied linguistics. Although the authors of the texts in his dataset represent different L1 backgrounds, this is not a factor that is discussed by Hyland as a possible influence on the use of lexical bundles. He does add, however, that he would be surprised if first language played a crucial role at "this level of proficiency" (ibid., 20), though he does not elaborate his reasons for this view.

Also important in the present context is the framework used for the functional analysis performed by Hyland. Drawing on Biber, Conrad, and Cortes (2004) and Biber (2006), Hyland operates with three functionally defined categories: research-oriented, text-oriented, and participant-oriented. He finds substantial differences in "bundle functions" by discipline. As will become evident, his functional taxonomy is very much in line with the one chosen for the present study (see Section 4).

Moreover, Hyland surveys to what extent actual 4-word bundles overlap across the disciplines, and he comments that it "may make depressing reading for commercial materials

[2] The Brown family of corpora includes the Brown Corpus, the Lancaster-Oslo/Bergen Corpus (LOB) with texts from 1961 and their 1991 counterparts, the Freiburg Brown Corpus (Frown) and the Freiburg LOB Corpus (FLOB). These corpora operate with the text categories "imaginative prose" (FICTION), "belles lettres, biography, memoirs, etc." (BELLES), and "learned", including social sciences, humanities, etc. (LEARNED).

[3] See e.g. Biber and Conrad (1999), Biber (2006), Scott and Tribble (2006).

writers seeking to identify universals of academic writing" (2008, 11). In other words, very few bundles are common across the board. His study thus supports the findings reported by e.g. Cortes (2004, 410-411), who found disciplinary variation in the use of lexical bundles in history vs. biology, both in terms of structural and functional features. The analysis of the bundles in Hyland's study indicates that "writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers" (2008, 20).

As far as the other main focus of our investigation is concerned, that of learners vs. native speakers, several researchers have addressed the puzzles of "nativelike selection and nativelike fluency",[4] by studying multi-word sequences and phraseology in L2 vs. L1 English. Publications include Granger (1998), Meunier and Granger (eds) (2008), Chen and Baker (2010), Ädel and Erman (2012), Hasselgård (2012), Paquot, Hasselgård, and Ebeling (2013).

Discussing lexical bundles that do not necessarily "represent complete structural units", Ädel and Erman (2012, 82) found that 4-word bundles produced by Swedish advanced learners of English in linguistics use "fewer and far less varied lexical bundles than native speakers". In this respect, their results resemble those of Chen and Baker (2010) who studied Chinese learners of English in several disciplines. However, the discipline-specific samples studied by Ädel and Erman showed a greater discrepancy in the use of bundles between native and non-native speakers than was the case in Chen and Baker's dataset. Since we will specifically compare learners and native speakers of English in linguistics vs. business, it will be interesting to see how our data match those of Ädel and Erman in terms of n-gram overlap within and across the disciplines.

Both Ädel and Erman (2012) and Chen and Baker (2010) base their functional classification of n-grams, or lexical bundles, on that of Biber, Conrad, and Cortes (2004). As noted above, this classification scheme is compatible with the framework adopted in the present study in the sense that both are equally inspired by Halliday's (2004) metafunctions of language. However, Ädel and Erman seem to have some reservations about the framework they (try to) adopt, due to unclear criteria for the subcategories. Thus, their final classification only takes the three main categories into account, viz. referential bundles, stance bundles, and discourse organisers. Since these do not completely match our categories, a direct comparison with our findings will be problematic. Nonetheless, there is enough overlap to draw on their insights, also in terms of the functional analysis. Ädel and Erman (2012) conclude that

> What we find are rather similar proportions of referential expressions in the two groups, but a greater proportion of stance bundles and a smaller proportion of discourse organisers among the native speakers. This confirms a pattern already spotted, the native speakers' greater reliance on, and greater variation in, stance bundles. (Ädel and Erman 2012, 90)

In the light of these findings, we can expect that the native linguistics students in our data will behave in a similar manner, producing more interpersonal n-grams than the Norwegian learners and fewer textual ones (cf. research questions (i) and (ii)). As regards the form of patterns, Hasselgård (2012) found that n-grams indicating complex phrases were more typical of native speakers; a similar tendency may be expected in the present material (cf. research question (ii)). It is also hypothesised that linguistics students, regardless of L1, will use more organizational (i.e. textual) n-grams than business students, in accordance with Hasselgård's

[4] Cf. Pawley and Syder (1983).

(in press) findings that linguistics students make use of more metadiscourse than business students do (cf. research question (iii)). Furthermore, Paquot, Hasselgård, and Ebeling (2013) found that learners are more visible authors in their texts, a feature which may also show up in their recurrent word-combinations (research question (iii)).

3. Material and method

3.1 The corpora

For the present study, native speaker data have been culled from the British Academic Written English (BAWE) corpus and (Norwegian) learner data from the Varieties of English for Specific Purposes dAtabase (VESPA-NO). At present VESPA-NO contains sufficient material in two disciplines, namely linguistics and business, to make a comparison with the corresponding native-speaker disciplines worthwhile. For this purpose, we make use of a subset of the BAWE corpus, and only include native speakers of English in the two disciplines.

Table 1 gives an overview of the material from the two corpora in terms of number of texts and number of words. As can be seen, the business part of the VESPA-NO is still relatively small, a fact that will have to be borne in mind when discussing the findings.

Table 1. Breakdown of data in terms of number of texts and words[5]

	Linguistics		Business	
	Texts	Words	Texts	Words
VESPA-NO (L2)	239	267,855	70	47,335
BAWE (L1, BrE)	76	167,437	64	141,249

A few words about the content and types of texts included in the two corpora are in order. With regard to the linguistics papers, these are fairly uniform across the two corpora. Most of the texts are essays discussing and analysing language, typically from an applied and/or functional perspective.

The business material is less uniform than the linguistics material, in that the BAWE corpus includes a varied set of texts from a number of different modules, including Introduction to Business Law, Marketing Analysis, and International Environment of Business. The business texts held in the VESPA-NO, on the other hand, are all from the same module, viz. Business Communication in English. However, both cohorts are represented by students doing different business degrees, e.g. Management Science, Economics, Business Administration. Although this is not unproblematic in the present study, we have chosen to be pragmatic and follow the BAWE team's policy on this:

> Modules are not a perfect match with disciplines – economics departments, for example, deliver modules in mathematics – but, for the purposes of this project, we treated every assignment produced for every module taught by staff belonging to the same department as belonging to the same discipline. (Alsop and Nesi 2009, 74)

The BAWE texts were produced by UK undergraduate and Master's students, "for assessment as part of taught degree programmes" (Alsop and Nesi 2009, 71); they also "meet a certain

[5] The word count excludes text in e.g. footnotes, block quotes and headlines. 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 that facilitates the automatic exclusion of text not produced by the students.

proficiency standard, as judged by the students' subject tutors", i.e. all assignments had been awarded a mark equivalent to 60 percent or more (ibid., 74). Similarly, the VESPA-NO texts are all part of the regular course work of the Norwegian students, both at the BA and MA levels; however, the papers were only assessed on a pass/fail basis. Only the ones that obtained a pass are included in the VESPA-NO corpus. Moreover, the texts in VESPA-NO are either course work essays or trial exams, while the BAWE texts include essays, case studies, critiques, etc. (see Heuboeck, Holmes, and Nesi 2008, 8ff).

The question of comparability between the corpora inevitably arises, but although there are some issues with comparability particularly in the business texts in BAWE vs. VESPA, we believe that the language produced can still tell us something about these students' ability to cope with the epistemology of the disciplines.

3.2 N-gram extraction

For the analysis, we extracted the 100 most frequent 3- and 4-grams in each subcorpus, using WordSmith Tools 6.0 (Scott 2012). Based on previous research regarding the appropriate size of n-grams to study, we chose to focus on 3- and 4-grams. Altenberg (1998), for instance, found that the majority of recurrent word-combinations cluster as 2-, 3-, or 4-grams, some as 5-grams, and very few as 6-grams,[6] while Stubbs and Barth (2003) found that three-word and four-word chains are better text-type discriminators than e.g. two-word or five-word chains.

Given the size of our corpora, we decided on a threshold of five, i.e. all 100 3- and 4-grams occur at least five times in identical form.[7] Thus, the frequency span of the top 100 grams varies across the subcorpora. An overview is given in Table 2.

Table 2. Frequency span of top 100 3- and 4-grams in VESPA-NO and BAWE

	Freq. span 3-grams	Freq. span 4-grams
VESPA-NO Ling	376-46	102-16
VESPA-NO Bus	59-11	40-5
BAWE Ling	165-20	32-7
BAWE Bus	81-15	60-12

What this table shows is that, in the case of VESPA-NO Ling, the most frequently used 3-gram occurs 376 times, while the 3-gram ranked as number 100 occurs 46 times. Not unexpectedly, there is a difference in the frequency span between 3-grams and 4-grams in all subcorpora, although the discrepancy is more marked in linguistics than in business. It is hard to assess the extent to which this may point to disciplinary differences; a qualitative analysis would be needed and will therefore have to await further research.

In order to answer our research questions and to assess the degree of overlap across the disciplines and L1 groups, the 3- and 4-grams were scrutinised both with regard to their form, i.e. the actual recurrent strings, and their function. While the form was directly accessible in the n-gram lists produced by WordSmith Tools, their function was analysed according to a functional classification framework, which is presented next.

4. Functional classification of n-grams

The classificatory framework applied in the present study draws on that of Moon (1998), though with some modifications. The model is functional, designed to classify linguistic

[6] A similar observation was made by O'Donnell, Römer, and Ellis (2013, 95).
[7] It may be noted that recurrent 5-grams did not reach the target frequency of 100 in our smallest subcorpus.

expressions "according to the way in which they contribute to the content and structure of a text" (Moon 1998, 217). Moon's model was developed for the classification of fixed expressions and idioms (FEIs). The application of the model to n-grams thus represents a challenge, since these are not necessarily "complete structural units", and "usually not fixed expressions" (cf. Biber and Conrad 1999, 183), which makes it harder to assess their meaning and function. However, Biber and Barbieri (2007, 283) point out that n-grams, though not idiomatic in meaning, "serve important discourse functions related to the expression of stance, discourse organization, and referential framing".

	Category	Function	Example
ideational	informational	stating proposition, conveying information	*of the brain*
interpersonal	situational	relating to extralinguistic context, responding to situation	*as in tager flusberg*
	evaluative	conveying speaker's evaluation and attitude	*is important to*
	modalizing	conveying truth values, advice, requests, etc.	*we can see*
textual	organizational	organizing text, signalling discourse structure	*in this paper*

Figure 1. The functional classification model (adapted from Moon 1998, 217)

Figure 1 shows our adopted taxonomy, using the categories given in Moon (1998, 217). Our modification lies simply in assigning the category of 'organizational' to the textual metafunction instead of tying it to the ideational (logical) metafunction (see Halliday 2004, 309). Thus, unlike Moon's (1998, 218), our model reflects all three of Halliday's metafunctions.[8]

The three interpersonal categories of the model may need some further comment; in particular, evaluative and modalizing may not be easy to distinguish. According to Moon (1998, 246) "the category of evaluative FEIs is, of course, especially associated with the transmission of attitude". Modalizing expressions, by comparison, are "typically epistemic or deontic in nature" (ibid., 226). However, Moon acknowledges that there are areas of overlap between the two categories, and in her analysis, which allows multifunctional expressions to have double class membership, a good number of expressions are classified as evaluative and modalizing simultaneously (ibid., 239). Moon's examples of the situational category are said to be "typically found in spoken discourse" (ibid., 225) and are mostly interactional signals. Academic written discourse naturally does not contain these types. However, references to sources, for example, point to extralinguistic context, in the sense of text-external entities. As Figure 1 shows, we have classified such references as situational, and have kept the category as interpersonal. Although it may be argued that such references are perhaps referential, this is of little practical consequence for the present study, as there were very few such references among the top 100 n-grams.

As mentioned in Section 2, Moon's functional classification has clear parallels with the taxonomy first found in Biber, Conrad, and Cortes (2004, 384), where lexical bundles are classified as stance expressions, discourse organizers, and referential expressions (roughly corresponding to the interpersonal, textual and ideational functions outlined in Figure 1). There are also similarities with the taxonomy used by Hyland (2008, 13 f.), whose categories

[8] See Culpeper and Kytö (2002, 49) for a similar discussion.

are (i) participant-oriented (including stance and engagement); (ii) text-oriented; and (iii) research-oriented.

As noted above, the model proposed by Biber, Conrad, and Cortes (2004) is also applied by Chen and Baker (2010). While Chen and Baker use the taxonomy without reporting any difficulties of classification, Ädel and Erman (2012, 89) express "several reservations" about it. "The main problem is that no clear criteria are given for how to decide which (sub)category a given bundle should belong to" (ibid.), a problem that is compounded by the multifunctionality of several bundles.[9]

In applying the analytical framework to our material, we encountered the same kind of problems as those reported by Ädel and Erman (2012). Some of the difficulty was that the n-grams tend to be incomplete units (unlike Moon's FEIs), which entails that their meaning and function are not clear out of context. Thus it was necessary to consult concordances of the troublesome n-grams and to discuss the classification of ambiguous cases; for instance, does *different* mean 'unlike' or 'many' in *different types of*? The former meaning would indicate an evaluative function of the 3-gram and the latter an informational function. In the case of multifunctional n-grams, we chose the function that seemed to be the dominant one from an examination of the material; for instance *as well as* occurred as evaluative ('do something as well as somebody else'), but was most frequently a conjunction and therefore classified as organizational.

Note finally that Chen and Baker (2010) as well as Ädel and Erman (2012) identify a set of "content bundles", i.e. bundles that are very closely connected with the topic of a particular text or a particular field. Content bundles are removed from the referential expressions in these studies. In ours, however, no such exclusion has been made, so the category of informational n-grams includes some that are clearly topic-specific, such as *lexical teddy bears*, *in the corpus*, *in New Zealand English* and *the Norwegian original texts*.

5. Corpus investigation

5.1. Introduction

As stated in Section 1, our investigation concerns 3-grams and 4-grams. Sometimes a 3-gram is almost invariably part of a specific 4-gram: for example *the other hand* is inevitably part of the 4-gram *on the other hand* in the material for the present study, perhaps suggesting that *the other hand* may be discarded. However, as Biber et al. (1999, 990) point out "shorter bundles are often incorporated into more than one longer lexical bundle" (for example *the use of* is part of at least two 4-grams, *in the use of* and *by the use of*); thus it would be misleading to exclude all 3-grams that are also part of a 4-gram. For this reason, 3-grams and 4-grams have simply been kept separate in the analysis.

Note that the analysis presented below is concerned with *types* rather than *tokens* of n-grams. That is, once the 100 most frequent 3-grams and 4-grams have been identified in each subcorpus, they are analysed without any further regard to their frequency (hence the need to determine a single category for each n-gram), and the analysis does not take account of the actual frequencies of specific n-grams.[10]

The remainder of this section compares the use of n-grams across L1 groups (learners vs. native speakers of English), and then discusses disciplinary differences within the two writer groups.

[9] It may be noted that the functional analysis reported in Biber and Barbieri (2007, 273, 279) has an additional category of 'other', but we cannot find any reason why such a category is needed, or what it consists of.

[10] For a related investigation looking at actual frequencies of recurrent n-grams in VESPA, see Lie (2013).

5.2 Comparing L1 groups: learners vs. native speakers

Having classified the n-grams functionally, we are now in a position to present the discourse functions of the top 100 3- and 4-grams in the four subcorpora. We will first look at the functional categories of the n-grams before discussing some of the salient forms included in these categories.

5.2.1 The functions of the n-grams

Table 3 shows the distribution of 3- and 4-grams according to function in texts produced by linguistics students in the Norwegian learner corpus (VESPA) vs. the native English corpus (BAWE). A test of equal proportions was carried out pairwise for each of the functional classes, producing a p-value in each case.[11] Cells with statistically significant results are shaded in grey.

Table 3. Learners' vs. native speakers' use of n-grams according to function: Linguistics

	BAWE-ling 3-grams	VESPA-ling 3-grams	p-value	BAWE-ling 4-grams	VESPA-ling 4-grams	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)
Modalizing	16	9	0.1995 p > 0.05	11	14	0.6689 (p > 0.05)
Organizational	13	26	0.03222 (p < 0.05)	14	22	0.1976 (p > 0.05)
	100	100		100	100	

We see that informational grams constitute the largest functional type for both 3-grams and 4-grams across the L1 groups. The distribution of modalizing n-grams is also fairly similar between the learners and native speakers. There are few situational n-grams overall, and none at all recorded in the learner data. While the second most frequently used functional type in the native speaker material is that of evaluative, it is the organizational type that is more frequently used among the learners. In fact, the p-values show that native speakers use significantly more evaluative n-grams than Norwegian learners do, in the same discipline. In a similar vein, the Norwegian learners are shown to use significantly more organizational 3-grams than their native peers.

A slightly different picture emerges when comparing learners' and native speakers' use of n-grams in business. As Table 4 reveals, both BAWE and VESPA students show a preference for informational n-grams. However, the preference among learners is even greater, and the discrepancy between the two L1 groups in their use of informational 4-grams is found to be statistically significant.

[11] These and subsequent tests were performed using the R prop.test (R-3.0.2, R Development Core Team 2013).

Table 4. Learners' vs. native speakers' use of n-grams according to function: Business

	BAWE-bus 3-grams	VESPA-bus 3-grams	p-value	BAWE-bus 4-grams	VESPA-bus 4-grams	p-value
Informational	64	73	0.2233 (p > 0.05)	65	80	0.02662 (p < 0.05)
Situational	0	0		1	1	
Evaluative	8	6	0.7817 (p > 0.05)	12	4	0.06808 (p > 0.05)
Modalizing	9	1	0.02314 (p < 0.05)	2	4	0.6785 (p > 0.05)
Organizational	19	20	1 (p > 0.05)	20	11	0.118 (p > 0.05)
	100	100		100	100	

The situational category is not popular with either group. Table 4 also shows a slightly higher number of evaluative n-grams with the native speakers, although the difference is not significant. Organizational n-grams constitute the second-most frequent type in both groups, displaying a fairly similar distribution. Finally, the native speakers tend to make more frequent use of modalizing 3-grams than the Norwegian learners do, but it is important to stress that the numbers are too low for the results to be conclusive.

5.2.2 The form of the n-grams

If we look at the actual n-grams used by Norwegian learners across the two disciplines, it is remarkable how little overlap there is. Table 5 gives the complete list of shared 3- and 4-grams in texts produced by Norwegian linguistics and business students. Only 6% of the 3-grams and 9% of the 4-grams are shared. These all belong to the interpersonal and textual categories.

Table 5. Learners: Shared n-grams across the disciplines (full list)

	3-grams:	4-grams
informational		
situational		
evaluative		*it is important to*
modalizing	*we can see*	*i would like to* *we can see that*
organizational	*in this essay* *it comes to* *on the other* *the other hand* *when it comes*	*at the same time* *in this essay i* *on the other hand* *the other hand is* *this essay i will* *when it comes to*

On the basis of what we have observed in Tables 3, 4 and 5, we can sum up the features that are typical of the Norwegian learners. In terms of function, the n-grams are typically ideational and textual. In other words, the learners generally use more informational n-grams and slightly more organizational n-grams than the native speakers do. We may note that the

Norwegian learners appear to have this in common with the Chinese learners studied by Chen and Baker (2010).

As regards form, quite a few n-grams with author presence are attested, including *i will look at*; *in this paper i*; *i would like to*; *i will discuss*; *we can see that*. There is also fairly frequent use of n-grams that are sentence stems or rhemes (Altenberg 1998), i.e. either subject + verb or verb + following constituent, as in *the* [*first*/*second*] *text is*; *is an example of*; *decisions are made, the boss has more*.

Finally, some specific n-grams are unique to the Norwegian learners, which may indicate overuse compared to native speakers. This is noted particularly for the 4-gram *when it comes to* (see also Lie 2013 for discussions of specific n-grams).

Moving on to what is common for the BAWE students of linguistics and business, we find a much greater overlap across the disciplines, both in their use of 3-grams and 4-grams. More than half of the overlapping n-grams are informational, contrary to the findings for learners (Table 5).

Table 6. Native speakers: Shared n-grams across the disciplines (frequencies and examples)

	3-grams		4-grams	
informational	18	*a number of, it is a, part of the, such as the, that it is …*	8	*at the end of, in the form of, the nature of the…*
situational	0		0	
evaluative	5	*due to the, is important to, the fact that, the importance of…*	4	*it is clear that, it is important to, the fact that the …*
modalizing	4	*be able to, can be seen, it can be, need to be*	1	*to be able to*
organizational	6	*a result of, as well as, in terms of, in this case, one of the …*	4	*a result of the, as a result of, on the other hand…*

Table 6 illustrates a general tendency in academic prose for "bundles" to be nominal rather than clausal (Biber et al. 1999, 992). In contrast, the n-grams shared between business and linguistics in the learner corpus are more clausal (see Table 5). This suggests that the learners have not emulated the epistemology of the (general) academic register to the same degree as the native novice writers have.

5.2.3 Summary of the learner vs. native speaker comparison

Summing up the features that are typical of the native speakers, we can conclude that the n-grams typically have ideational and interpersonal functions. In other words, although the native speakers use fewer informational n-grams than the learners, it is still the predominant function. Moreover, the BAWE students generally use more evaluative and modalizing n-grams than the Norwegian learners, and they appear to have a shared pool of informational n-grams across the disciplines.

Formal features characterizing the native speakers' n-grams include non-personal (self) projection,[12] e.g. *it is clear that, it is argued that*, complex noun phrases, typically represented by a noun followed by the preposition *of*, e.g. *the majority of the, the nature of the, as a result of*). Another common trait of the n-grams used by the native speakers is that many reflect the

[12] This was a feature particularly noted for English (native) student writing in the Social Sciences (represented by Anthropology and Business), but also to a certain extent for Arts and Humanities (represented by English Studies and History) (Ebeling and Wickens 2012).

passive voice, as in *it can be seen*, a 4-gram that is paralleled in the learner data by *we can see that*. As hypothesized in Section 2, this use of author reference in the n-grams is more typical of learners than it is of native speakers.

The results from linguistics (see Table 3) may be compared with Ädel and Erman's (2012, 90), whose material also consisted of linguistics papers. While their findings for native speakers were similar to ours, their non-native material had a lower proportion of referential bundles (corresponding to informational n-grams) and a higher proportion of stance bundles (corresponding to evaluative n-grams), leading to greater similarity between native and non-native speakers. Some methodological differences may explain this discrepancy: as noted above, Ädel and Erman removed content-specific bundles from the referential category, thus reducing its size. Moreover, their investigation is concerned with tokens rather than types of bundles, which is likely to affect the proportional distribution of functions. However, Chen and Baker (2010, 38) present the functional distribution of *types* in their multi-discipline EAP material. Interestingly, their Chinese L2 writers – like our Norwegian L2 writers – have a higher proportion of referential bundles and a lower proportion of stance bundles than the English L1 writers (from BAWE).

5.3 Comparing disciplines

We now turn to a comparison of disciplines within the L1 groups to investigate whether there are disciplinary differences in the distribution of functional categories of n-grams (5.3.1), and how the functional categories are realized in the two disciplines (5.3.2).

5.3.1 The function of the n-grams

Table 7 shows the distribution of functional types of n-grams across disciplines in the learner corpus.

Table 7. The use of n-grams across disciplines: Learners

	VESPA-ling 3-grams	VESPA-bus 3-grams	p-value	VESPA-ling 4-grams	VESPA-bus 4-grams	p-value
Informational	57	73	0.02617 ($p < 0.05$)	49	80	9.287e-06 ($p < 0.0001$)
Situational	0	0		0	1	
Evaluative	9	7	0.7944 ($p > 0.05$)	15	4	0.01588 ($p < 0.05$)
Modalizing	9	1	0.02314 ($p < 0.05$)	14	4	0.02617 ($p < 0.05$)
Organizational	25	19	0.3934 ($p > 0.05$)	22	11	0.05678 ($p > 0.05$)
	100	100		100	100	

The most striking finding is that linguistics students use significantly more modalizing n-grams (and evaluative 4-grams) than business students, while the business students show a significantly greater preference for informational n-grams. Organizational n-grams are also more frequent in the linguistics corpus, but the difference is not significant according to the test of equal proportions used here.

Table 8 shows that BAWE, like VESPA, contains significantly more informational n-grams in the business component than in the linguistics component, accompanied by a

significantly more frequent use of evaluative (3-grams) and modalizing n-grams in business. The organizational n-grams show the opposite tendency from VESPA, with higher frequencies in business than in linguistics. However, the differences are not significant. Thus, in both corpora there are clear disciplinary differences in the use of n-gram function types.

Table 8. The use of n-grams across disciplines: Native speakers

	BAWE-ling 3-grams	BAWE-bus 3-grams	p-value	BAWE-ling 4-grams	BAWE-bus 4-grams	p-value
Informational	46	64	0.01568 ($p < 0.05$)	42	65	0.001815 ($p < 0.01$)
Situational	1	0		4	1	0.365 ($p > 0.05$)
Evaluative	25	9	0.004748 ($p < 0.01$)	29	12	0.005071 ($p < 0.01$)
Modalizing	16	9	0.1995 ($p > 0.05$)	11	2	0.02175 ($p < 0.05$)
Organizational	12	18	0.3221 ($p > 0.05$)	14	20	0.3466 ($p > 0.05$)
	100	100		100	100	

The distribution of functional types shown in Tables 7 and 8 may be compared to Hyland's (2008, 14) figures for the two disciplines "applied linguistics" and "business studies": he too, finds a higher proportion of so-called "research-oriented bundles" in business studies, accompanied by slightly fewer text-oriented and (especially) participant-oriented bundles. Although the disciplinary differences in Hyland's material are not as great as in ours, the comparable results suggest that the present findings may have some general validity.

5.3.2 The form of the n-grams

There is some degree of overlap between Norwegian learners and native speakers of English as regards the form of the n-grams they use in their linguistics papers. Table 9 shows that the two linguistics corpora share a total of 36% of 3-grams and 25% of 4-grams across L1 background.

Table 9. Linguistics: shared n-grams across the L1 groups (frequencies and examples)

	3-grams		4-grams	
informational	16	*that there are, the number of, the use of, part of the…*	7	*and the use of, at the end of, by the use of…*
situational	0		0	
evaluative	6	*in the same, meaning of the, the fact that…*	6	*in the same way, it is important to, the fact that the…*
modalizing	6	*be found in, can also be, can be seen, can be used…*	5	*can be found in, can be seen in, it is possible to…*
organizational	8	*an example of, in this case, in this essay, looking at the…*	7	*an example of this, example of this is, in this case the…*

These findings are similar to Ädel and Erman's (2012, 85) comparison of 4-grams used by Swedish learners of English and native speakers. Most of the shared n-grams are rather general in meaning – there is a striking lack of linguistics terminology and specialized vocabulary. One exception is that *example* often refers to linguistic examples as in (1) and (2).

(1) *An example of* this is 'fall into' (within), and its correspondence 'falle innenfor', which often feature in texts regarding laws etc. (VESPA-NO)
(2) However, classifiers are a more closed class of adjectives and *an example of* this is shown in (9). (BAWE)

Table 9 should be compared with Table 10, which shows those n-grams that are shared between L1 groups in the business material. The table presents a full list, as only 9% of 3-grams and 3% of 4-grams are shared. As in linguistics, content-specific n-grams are completely absent from the list. Notably no modalizing n-grams are shared, and all the shared evaluative n-grams contain *important/importance*.

Table 10. Business: Shared n-grams across the L1 groups (full list)

	3-grams	4-grams
informational	*a lot of* *that they are*	
situational		*at the same time*
evaluative	*is important to* *it is important* *the importance of*	*it is important to*
modalizing		
organizational	*based on the* *in order to* *there is a* *one of the*	*on the other hand*

5.3.3 Summary of the discipline comparison

We have seen that the n-grams that seem typical of linguistics are predominantly ideational and interpersonal. Many informational n-grams are topic-specific, and these are not shared across the L1-subcorpora. The linguistics students generally use more evaluative and modalizing n-grams than the business students, a trend which is also evident in Hyland's (2008) study. There are also more overlapping n-grams between L1 backgrounds in the linguistics material than in the business material. The form of n-grams in the linguistics material indicate a higher frequency of complex noun phrases among the ideational n-grams (e.g. *at the end of, by the use of, in the case of*), while n-grams with *can* predominate in the modalizing function (e.g. *can be found in, can be seen in, can also be*).

The texts written by business students are characterized by n-grams that are informational and topic-specific, even more so than is the case for linguistics. There are some organizational grams, but very few interpersonal ones. Very few overlapping n-grams were found across the two L1 groups, apart from evaluative n-grams containing *importan**. This may be linked to the fact that topics differ even more between the corpora in business than in linguistics, although that ought not to have an impact on the use of interpersonal and textual expressions.

6. Summary and discussion of findings

At the end of Section 2 we hypothesized that the functional types of n-grams might differ between learners and native speakers and across disciplines. The investigation has shown this to be the case. The difference between native and non-native writers is more pronounced among business students than among linguistics students. It is, however, possible that different types of assignments and topics are responsible for some of the dissimilarities in the use of n-grams in the business corpora. Importantly, the disciplinary differences were greater among the learners than among the native speakers.

On the basis of previous studies of writer/reader visibility (e.g. Paquot, Hasselgård, and Ebeling 2013), we hypothesized that the recurrent word combinations would reveal the learners as more visible authors in their texts. While the learners used more n-grams involving a first-person pronoun than the native speakers in both disciplines, this was not such a salient feature as we had thought. Based on Hasselgård's (in press) finding that the use of metadiscourse is more widespread in linguistics than in business we also thought that the linguistics students would use more organizational n-grams than would business students. This hypothesis was not confirmed. Although there were slightly more organizational n-grams in the linguistics material in VESPA, BAWE showed the opposite tendency, and the difference was not statistically significant in either case.

It is possible that our results would have been different if we had counted tokens rather than types of n-grams (see Section 3), thus getting a more comprehensive survey of the functional classes of n-grams across disciplines and L1 backgrounds. Yet, our findings are largely in agreement with comparable previous studies, suggesting that some of these are related to general processes of L2 development/production, regardless of the learners' L1. We may thus conclude that the type-based n-gram approach and the functional classificatory framework were able to uncover differences between disciplines and between L1 groups.

As shown above, the ideational/informational grams are typical for both L1 groups and for both disciplines, with a proportion above 50% in all the subcorpora except NS linguistics. This is not surprising, since academic disciplines have been found to be highly informational (Ebeling 2011) and we are dealing with novice academic writers. Perhaps more unexpectedly, situational n-grams were rare in all the corpora; they were found mainly in NS linguistics, and were mainly references to external sources, such as *Hunston and Francis*. This should incidentally not be taken to mean that learners of English or business students do not refer to their sources in the running text, only that such references do not come up as 3-grams or 4-grams above the frequency threshold set for inclusion in the present study (see Section 3.2).

The fact that there were far fewer overlapping n-grams across the disciplines among the learners than among the native speakers may indicate that learners have a smaller stock of general-purpose academic vocabulary. However, the linguistics papers were more similar across the L1 groups than the business papers as regards both the functional distribution of n-grams and a greater percentage of shared n-grams, which suggests a more native-like phraseological repertoire among the NNS linguistics students than among the business students.

With reference to the research questions outlined in Section 1, we sum up our findings in this and subsequent paragraphs. Certain features of the distribution of *functional* types of n-grams distinguish learners from native speakers in both linguistics and business. First, the learners in both disciplines use fewer modalizing and evaluative grams than their NS peers. There is furthermore a slight tendency for the learners to use more informational n-grams,

although this is significant only with 4-grams in the business material. The use of organizational n-grams remains inconclusive: in linguistics, the learners have more organizational grams than native speakers, but in business they have slightly fewer.

There were also differences across the L1 groups as regards the *form* of the n-grams used. Most strikingly, the learners used more n-grams involving first person pronouns, thus partly reflecting the tendency for Scandinavian learners of English to be visible authors (see Petch-Tyson 1998). On the other hand, native speakers used more n-grams with non-personal projection (extraposition, e.g. *it is evident that; it is important to*). Furthermore, native speakers were found to use more n-grams reflecting complex noun phrases (e.g. *of the language; the extent to which*); a similar trend was noted by Paquot (2013, 292). Native speakers also used more verb phrases in the passive voice, such as *been found to; has been suggested that*. Both construction types are known to be salient features of academic prose (Biber and Gray 2011) and thus show that the NS students are familiar with the genre.

An important finding of our study is that the discipline comparison involved more statistically significant differences than the NS/NNS comparison.[13] As noted, there are more overlapping n-grams between the corpora in linguistics than in business; linguistics students have fewer informational n-grams than business students (across L1 backgrounds), and linguistics students have more evaluative and modalizing n-grams than business students. Thus, in spite of the differences across L1-groups, this investigation suggests that 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.

7. Further work and potential applications

The present study suggests several avenues of further research. First of all, we have paid relatively little attention to the syntactic form of n-grams, unlike Hyland (2008) and Biber, Conrad, and Cortes (2004). Those features that have been noted give indications of more formal grammar in the NS material (complex noun phrases and passive verb phrases), which may constitute a teaching point for learners of English. A follow-up of the study could take a more qualitative approach, examining the n-grams in their contexts to evaluate the extent to which they are used appropriately by the learners. A similar extension could take token frequency into account. The frequency spans of n-grams presented in Table 2 show that the most frequent n-grams have higher frequencies in the learner material than in the NS material (see also Lie 2013), thus suggesting that learners may be overusing a small number of high-frequency n-grams in the same way as they overuse high-frequency core vocabulary (Ringbom 1998) or delexical verb + noun collocations (Wang 2013). Another consideration for future studies of a similar kind is text dispersion of the n-grams. Because of the threshold set for the frequency of n-grams in the present study, the large majority of n-grams had a more or less even distribution across texts and individuals; however we did notice a few n-grams that reached the top 100 because they were very frequent in a small number of texts. Since our main focus here is on *functional* types of n-grams, we do not believe this to have impacted the results much, but in the kind of qualitative study outlined above, text dispersion needs to be taken into account.

The scope of the study could also be widened by expanding the dataset. In particular, the business subcorpus of VESPA-NO should be added to, not only because it is smaller than the corresponding section of BAWE (see Table 1), but also because the discrepancy between

[13] Ädel and Römer (2012, 28) report a similar result regarding the use of n-grams and phrase frames across course levels and disciplines in MICUSP: disciplines are found to be the stronger variable.

the two corpora as regards topics and task types may be responsible for some of the differences found here between L1 and L2 business writing. Furthermore, because of the strong influence of discipline on the use of n-grams in general, there is a clear need to investigate more disciplines. It would similarly be interesting to make comparisons with other learner groups so as to make it possible to distinguish potential L1-specific features of n-gram use from more general features of non-native English, as is inherent in the Contrastive Interlanguage Analysis model (e.g. Granger 1996). And as writers in both BAWE and VESPA are apprentice academics (cf. Scott and Tribble 2006), their data could be compared to published academic writing in the relevant disciplines, to examine the extent to which they match the target usage.

The results obtained in the present study – despite its shortcomings – clearly have a number of applications. For example, insights into disciplinary and L1-specific use of n-grams could feed into EAP courses and teaching materials or inform writing instruction in other types of academic courses (see Cortes 2006 for an example). One could also picture the development of a "multi-word academic word list" along the lines of Coxhead (2000), or indeed a "phrase book with grammatical notes" as envisaged by Pawley and Syder (1983, 220), but containing discipline-specific n-grams as well as the more general, discipline-independent academic vocabulary of Coxhead's list. Finally, the differences uncovered between L1 and L2 apprentice academics indicate a need for explicit instruction in academic vocabulary and phraseology among the learners, a need which may be more urgent among the business students than among the linguistics students, according to the findings of the present investigation.

References

Ädel, A. and 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 (2): 81-92.

Ädel, A. and U. Römer. 2012. Research on advanced student writing across disciplines and levels. Introducing the *Michigan Corpus of Upper-level Student Papers*. *International Journal of Corpus Linguistics* 17 (1): 3-34.

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

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. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. and F. Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26 (3): 263–286.

Biber, D. and S. Conrad. 1999. Lexical bundles in conversation and academic prose. In *Out of corpora: Studies in honour of Stig Johansson*, eds. H. Hasselgård and S. Oksefjell, 181–190. Amsterdam: Rodopi.

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

Biber, D. and B. Gray. 2011. The historical shift of scientific academic prose in English towards less explicit styles of expression. In *Researching specialized languages*, eds. V. Bhatia, P. Sánchez Hernández, and P. Pérez-Paredes, 11–24. Amsterdam: John Benjamins Publishing Company.

Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Chen, Y.-H. and 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 (4): 397–323.

Cortes, V. 2006. Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education* 17 (4): 391-406.

Coxhead, A. 2000. A new academic wordlist. *TESOL Quarterly* 34 (2): 213-238.

Culpeper, J. and M. Kytö. 2002. Lexical bundles in Early Modern English dialogues: a window into the speech-related language of the past. In *Sounds, words, texts and change: selected papers from 11 ICEHL, Santiago de Compostela, 7-11 September 2000*, eds. T. Fanego, B. Méndez-Naya, and E. Seoane, 45–64. Amsterdam: John Benjamins.

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

Ebeling, S.O. and A. Heuboeck. 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 P. Wickens. 2012. Interpersonal themes and author stance in student writing. In *English corpus linguistics: Looking back, moving forward. Papers from the 30th international conference on English language research on computerized corpora (ICAME 30)*, eds. S. Hoffmann, P. Rayson, and G. Leech, 23–40. Amsterdam: Rodopi.

Gilquin, G. and M. Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1): 41–61.

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In *Languages in contrast. Papers from a symposium on text-based cross-linguistic studies, Lund 4-5 March 1994*, eds. K. Aijmer, B. Altenberg, and M. Johansson, 37–51. Lund Studies in English 88. Lund: Lund University Press.

Granger, S. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In *Phraseology. Theory, analysis, and applications*, ed. A.P. Cowie, 145–160. Oxford: Oxford University Press.

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

Halliday, M.A.K. 2004. *An introduction to functional grammar*. 3rd ed., revised by C.M.I.M. Matthiessen. London: Arnold.

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 *Corpus linguistics on the move: Exploring and understanding English through corpora*, eds. M.J. López-Couso, B. Méndez-Naya, P. Núñez-Pertejo, and I. Palacios. Amsterdam: Brill / Rodopi.

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

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

Lie, J. 2013. 'The fact that the majority seems to be...': A corpus-driven investigation of lexical bundles in native and non-native academic English. Master's thesis, University of Oslo. [available at www.duo.uio.no]

Meunier, F. and S. Granger (Eds.). 2008. *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins Publishing Company.

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

O'Donnell, M.B., U. Römer and N.C. Ellis. 2013. The development of formulaic sequences in first and second language writing. Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics* 18 (1): 83–108.

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, and L. Valentin. 2010. The VESPA tagging manual. CECL, Université catholique de Louvain.

Paquot, M., H. Hasselgård, and 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 *Twenty years of learner corpus research: Looking back, moving ahead*, eds. S. Granger, G. Gilquin, and F. Meunier, 377–387. Louvain-la-Neuve: Presses universitaires de Louvain.

Pawley, A. and F. H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and communication*, eds. J. C. Richards and R. W. Schmidt, 191–226. London: Longman.

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

R Development Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. (http://www.R-project.org).

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.

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

Scott, M. and C. Tribble. 2006. *Textual patterns: Key words and corpus analysis in language education*. Amsterdam / Philadelphia: John Benjamins.

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

Wang, Y. 2013. Delexical verb + noun collocations in Swedish and Chinese learner English. Doctoral dissertation, Uppsala University.

Corpora

BAWE, see http://www.coventry.ac.uk/research/research-directory/art-design/british-academic-written-english-corpus-bawe/.

BAWE was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

VESPA, see http://www.uclouvain.be/en-cecl-vespa.html and http://www.hf.uio.no/ilos/english/services/vespa/.

The Norwegian component of VESPA is being compiled by Signe Oksefjell Ebeling and Hilde Hasselgård, and has been funded by the Department of Literature, Area Studies and European Languages at the University of Oslo.