Recurrent word-combinations in English student essays[1]

Signe Oksefjell Ebeling, University of Oslo

Abstract

The point of departure for this study is Stubbs and Barth's article from 2003 where they describe how recurrent phrases can be used as text-type discriminators. A subset of the *British Academic Written English* (BAWE) corpus—a corpus of student writing—consisting of English Studies essays is compared with two text-types taken from BNC-Baby in order to test and validate Stubbs and Barth's claims about recurrent word-combinations. Then the most salient combinations occurring in the BAWE student essays are subjected to a functional analysis, based on Moon (1998), to better be able to say something about their main functions in this text-type. Finally, the functions of recurrent phrases of the BAWE essays will be compared to the functional characteristics of the academic prose part of BNC-Baby.

Keywords: recurrent word-combinations, ngrams, lexical bundles, functional analysis, text-type discriminator, corpus-based analysis

> In a corpus of texts from different speakers and writers, they [recurrent word-chains] can be studied (a) as a predictable characteristic of different text-types, and (b) as evidence of units of routine language (Stubbs and Barth 2003:62)

1. Introduction and aims

Recurrent word-combinations have in recent years been used to characterise genre and text-type; they have even been described as "text-type discriminators" (Stubbs and Barth 2003). This relationship between recurrence and text-type will be explored in the present article. I will follow Altenberg (1998:101) in defining recurrent word-combinations as "any continuous string of words occurring more than once in identical form".

This paper sets out to give an overview of which word-combinations English Studies students tend to resort to in their essays, what these combinations typically express in functional terms, and how they compare with word-combinations typical of other text-types.

[1] I am grateful to Hilde Hasselgård and two anonymous reviewers for comments on an earlier version of this paper.

The first part of the study focuses on recurrent patterns found in essays written by UK undergraduate students in English Studies, with the ultimate aim of establishing their functional characteristics. In order to establish which patterns typically recur in English essays, quantitative corpus methods will be applied. The recurrent patterns identified for English essays will be compared with recurrent word-combinations in other text-types. For this purpose, the text categories "academic prose" and "written fiction" taken from BNC-Baby were chosen. By comparing three potentially very different text categories, Stubbs and Barth's (2003: 62) claim that "text-types are repetitive in different ways and to different extents" will be examined and validated.

The second part, and the bulk of this study, takes the most salient recurrent word combinations in two of the text types, viz. academic prose and the English essay, as its starting point, with the aim of establishing and comparing the functional qualities of these text types. This will ensure a more balanced approach towards the data, complementing quantitative findings with a qualitative, functional analysis. The functional analysis follows Moon's (1998) model of classifying expressions into organizational, informational, situational, evaluative and modalizing (cf. section 6.1).

Where other studies have focused on how novice writers compare with professional writers of the same domain (e.g. Shaw 2009 on student vs. professional writing in literary studies) or how they try to come to terms with the epistemology of their discipline (e.g. Charles 2006, Ebeling & Wickens Forthcoming), this study is more concerned with the functional roles the recurrent word-combinations play in the student essays and how these compare with academic prose in general. These functional roles may in turn point to the epistemological properties within different discourse communities, as "phraseology and epistemology are indissolubly interlinked" (Groom 2009: 125).

2. Material and method

The main data for this study are taken from the *British Academic Written English* (BAWE) corpus.[2] It is a corpus of proficient student writing

[2] The British Academic Written English (BAWE) corpus was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of

covering a wide variety of student assignment types across four disciplinary groupings, viz. Arts and Humanities, Life Sciences, Physical Sciences and Social Sciences. (See Nesi et al. (2005) and Heuboeck et al. (2007) for a more detailed account of the corpus.)

One of the largest homogeneous groups of assignments included in the corpus is the English essay, and the sample examined in this study comprises 89 essays from English studies, amounting to approximately 205,000 running words.[3]

Although homogenous in the sense that they are all classified as essays,[4] they differ in other respects; they are written by different students of different genders, different ages and different language backgrounds and at different levels of study. Nevertheless, the essays will be studied here as a single unit making up the data termed BAWE essays (i.e. English Studies essays in UK higher education), mostly concerned with English literature.

Much the same variation with respect to authors, authors' gender and written style also applies to the data taken from BNC-Baby, which will be used for comparison with the BAWE material. BNC-Baby is a subset of the *British National Corpus*, containing "four one-million word samples, each compiled as an example of a particular genre: fiction, newspapers, academic writing and spoken conversation."[5] Only two of these—academic prose and fiction—will be used for comparison with the BAWE data.

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, http://www.coventry.ac.uk/researchnet/d/505, accessed 28 Oct. 2009.

[3] Footnotes and endnotes were ignored for the purpose of this study. The main reason for this was that most notes in the English essays were references and not running text.

[4] "Essay" is one of thirteen genre families that have been identified in the BAWE material according to the learning function they serve (see Nesi & Gardner Unpublished MS). Incidentally, the essays studied here were also labelled essays by the students themselves.

[5] http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products#baby, accessed 28 Oct. 2009.

Of the two text-types chosen from BNC-Baby, fiction is perhaps the more clearly defined. Academic prose, on the other hand, is a more loosely defined category and, in addition, it represents a number of different scientific disciplines. Although both academic prose and fiction are potentially very different from the English essay in terms of recurrent word-combinations used, the academic setting for the BAWE material will probably have an impact on the results. It can be expected that there will be more overlap between the BAWE essays and academic prose than between the essays and fiction.

The method applied in this study follows Stubbs and Barth (2003), where the most frequent recurrent word-combinations are compared across the selected text-types, in order to investigate the degree to which the word combinations can be said to work as text-type discriminators.

The two BNC-Baby sub-corpora contain 5 times as many words as the sub-corpus of BAWE essays (ca. 1 mill. vs. ca. 200,000 words) and it may be argued that such a discrepancy in size may distort the findings as a more varied set of recurrent combinations can be expected in a larger material. This made me want to find out if a direct comparison of frequent word combinations in the different-sized corpora would distort my findings significantly, or whether I could use the 1 mill. word corpora for comparison. I therefore took a random sample of 200,000 words from academic prose and fiction, and compared the top 30 three-word combinations in that sample with the top 30 three-word combinations in the 1 mill. word original corpora. In both fiction and academic prose there was only 63% overlap between the shorter sample and the full corpus. As the reason for this might have been my selection process of picking 6 BNC-Baby files at random, I tried a different sampling method of about 6,500 words from each of the 30 academic prose files and about 8,000 words from each of the 25 fiction files in order to reach 200,000 words. This proved to be more successful in terms of stability in the recurrent combinations, with 80% overlap for academic prose and 73% for fiction. The latter 200,000 word samples quite clearly are more representative of the text-types as a whole than the 6-file sample. The fact that the difference between the 200,000 word samples (when using the second sampling method) in both fiction and academic prose had a fairly high rate of overlap with the full million word sub-corpora, I decided, for convenience, to use the full-size ones for further comparison with the BAWE data.

3. Background

For the purpose of this study, the terms *text-type* and *genre* have been conflated. Both are hard to pin down, and as pointed out by Swales "[t]he word [genre] is highly attractive—even to the Parisian timbre of its normal pronunciation—but extremely slippery" (Swales 1990: 33). The BAWE corpus manual identifies a range of genres, which in turn have been grouped into 13 genre families, or classes of genres, sharing functional and structural properties (cf. Heuboeck et al. 2007: 7 & 45ff). Following the BAWE classification scheme,[6] influenced by Swales (1990) and Martin (1992, 1997), genre is defined as "a staged, goal-oriented social process realised through register" (Martin 1992: 505), and members of a genre family "may share a central function, or they may have evolved in the same disciplinary context" (Nesi & Gardner Unpublished MS). In the following, I will use the term *text-type*, since my point of departure is Stubbs and Barth's investigation of *text-type* discriminators.

The data for this investigation are from three text-types identified by the BNC-Baby team as academic prose and fiction and by the BAWE team as essay, one of the "making sense" genre families (cf. Nesi & Gardner Unpublished MS). The essay in the BAWE corpus is characterised by writing where the evidence to support an argument has to be sought widely and is open to debate; essays "expect students to express a viewpoint, e.g. 'Is it worthwhile to test intelligence?', which suggests students should be gathering evidence and forming their own thesis in response" (Nesi & Gardner Unpublished MS).

Previous research has shown that the use of ngrams, or clusters, as text-type discriminators is rewarding "because they give insights into important aspects of the phraseology used by writers in different contexts" (Scott & Tribble 2006: 132). This is not to say that clusters alone can identify a text-type, but they can point to typical formulations used in particular text-types, and thus be used for instructive purposes for novice writers. Also, for the purpose of the present study, they can be used to point to differences between student writers in one discipline and how they compare to other writers with "the potential to enhance our appreciation (and that of learners) of what works in particular kinds of text" (ibid.).

[6] See Nesi & Gardner (Unpublished MS).

This technique has not been applied to the BAWE material before, and it will be interesting to see what ngram statistics will tell us about the English essay as a text-type compared to the other text-types investigated here.

4. Ngrams

Although this research focuses mainly on clusters of words, a brief look at the 10 most frequent word forms in the BAWE essays show that English essays do not deviate significantly from English in general as regards the most frequent single-word frequencies (cf. Sinclair's overview of frequent word forms in the *Bank of English* (Sinclair 1999). This underlines Sinclair's point about frequent words forming a large proportion of any text (ibid.:157). Typically these frequent words are function words, which do not easily distinguish text-types, at least not on the same level as content words may do. However, in combination with other function words or content words, function words are important building blocks in the phraseology of a language. Phraseology, as pointed out by Altenberg (1998:101), involves "various kinds of composite units and 'pre-patterned' expressions such as idioms, fixed phrases, and collocations". After examining the phraseology of spoken English on the basis of recurrent word-combinations, Altenberg admits that, although not all such combinations are of phraseological interest, they are "a useful starting point for an examination of the phraseology of spoken English" (ibid.:102). In the current setting recurrent word-combinations will provide a good starting point for investigating in what way word patterns in English essays reflect the text-type in which they are used.

In this section we will identify ngrams, i.e. sequences of types, in the BAWE essays; we will only be concerned with those that occur frequently.[7] Such recurrent word-combinations differ from idioms and collocations, since they are not necessarily complete syntactic constituents or phrases. They are similar to what Biber and Conrad have termed *lexical bundles*, viz. "the most frequent recurring lexical sequences; however, they are usually not complete structural units, and

[7] To extract the ngrams, or clusters, for this study WordSmith Tools (WS 5.0) were used (Scott 2008).

usually not fixed expressions" (1999:183).[8] They view lexical bundles as extended collocations typically occurring in sequences of three or more words. In this paper we will be mainly concerned with three- and four-word combinations. The formal and functional distinction that may exist between idioms, collocations and recurrent word-combinations will not be taken into account here, since all word-combinations will be identified on the basis of frequency alone. Thus, some of the recurrent word-combinations associated with English essays could also be termed collocations. It is less likely that there will be idioms among the recurrent word-combinations; these are comparatively rare in actual writing (cf. Sinclair 1999, Biber et al. 1999).

Earlier studies have shown that bigrams have little impact as text-type discriminators, partly due to the fact that the most frequent ones rarely form full structural units, e.g. *of the*, *is the*, *to a* (cf. Stubbs & Barth 2003, Altenberg 1998, Biber and Conrad 1999). This is also the case in the BAWE essays. The complete phrases that were identified among the top 50 bigrams typically reflect the content of the essays, e.g. *the reader*, *the poem*, *the novel, the text*. The relative overuse of content bigrams such as *the poem* in BAWE vs. academic prose (1,356 occurrences per million words vs. 26 occurrences per million words) can be easily explained by the fact that the BAWE essays are more specialized (in terms of both topic and text-type) than the academic prose corpus from BNC-Baby. A similar observation is pointed out by Pecorari (2008: 16), in her article on repeated language in academic discourse, where the corpus she used was much more specialized than the one she compared it with, viz. the academic corpus in Biber et al. (1999). In other words, it is important to keep in mind the narrowness of the BAWE material, i.e. the essays are all written within the framework of a limited number of modules by fewer authors than is the case in BNC-Baby. The four two-word noun combinations that were found among the 50 most frequent two-word combinations in the academic prose material from BNC-Baby are of a much more general kind than the ones recorded for the BAWE essays: *the same*, *per cent*, *for example* and *the other*. These

[8] In WordSmith Tools these are called *clusters*, i.e. "a *group of words which follow each other in a text*. The term *phrase* is not used here because it has technical senses in linguistics which would imply a grammatical relation between the words in it" (http://www.lexically.net/downloads/version5/HTML/?proc_definitions.htm, accessed 28 Oct. 2009).

noun combinations do not have the same ability to identify subject matter as the ones mentioned for the BAWE essays. The fact that *per cent* and *for example* are high up on the list of bigrams in academic prose may indeed point to a characteristic trait of academic writing as regards lexical choice. Still, there seems to be too few defining features, including content words, present in the most frequent bigrams for either text sample. Thus, bigrams will not be discussed any further in the present paper. Let us instead turn to trigrams and see whether our findings match the findings of Stubbs and Barth (2003:71): "With three-word chains, constituent content words appear earlier in the lists and are more frequent".

4.1 Trigrams

Using a similar approach to that of Stubbs and Barth (2003), I will compare the most frequent trigrams across three sub-corpora, listing the ones that are found in all three, followed by those that occur in two only. Finally, I will proceed to the trigrams that are most frequently found in only one of the sub-corpora.

Of the 50 most frequent trigrams, there are only four that occur in all three sub-corpora: *the end of, to be a, one of the,* and *end of the* (see Appendix I for an overview of the top 50 trigrams in the three text-types). As a three-word chain containing a content word, *the end of* also occurred among Stubbs and Barth's frequent trigrams featured in their FICTION and LEARNED categories. Their source material was taken from the Brown family of corpora, viz. Brown, LOB, Frown, and FLOB.[9] In addition, *the end of* was also recorded among the top 30 in their third genre BELLES (including belles lettres, biography, memoirs).[10] This trigram, then, seems to be a fairly general one in English overall.

In our context of looking at recurrent word-combinations that characterise specific text-types, it will therefore be potentially more rewarding to look at trigrams occurring in two, or only one, of our sub-corpora. First, let us move to trigrams that occur in two of our categories.

[9] Cf. ICAME corpus manuals, available at http://khnt.aksis.uib.no/icame/manuals/index.htm.

[10] Cf. the Brown Corpus Manual, available at http://khnt.aksis.uib.no/icame/manuals/brown/INDEX.HTM.

Table 1 Trigrams occurring among the 50 most frequent trigrams of two sub-corpora

Trigrams occurring in academic prose and BAWE essays	**Trigrams occurring in fiction and BAWE essays**	**Trigrams occurring in academic prose and fiction**
as well as	at the end	part of the
in order to		
in the first		
in which the		
it is a		
it is not		
it is possible		
on the other		
that it is		
the fact that		
the other hand		
the use of		
there is a		
there is no		

Table 1 shows that there is only one trigram that occurs in both academic prose and fiction and one that occurs in both fiction and the BAWE essays, while there are 14 that occur in both academic prose and English essays. This supports the view that English essays are closer to academic prose than to fiction. Indeed, if we look at the trigrams that are found both in academic prose and essays, quite a few of these are structuring devices that reflect a particular style, e.g. *the fact that*, *in order to, it is possible*. If they occur with a certain frequency in both academic prose and essays, we must assume that these are important building blocks of the text-types in question and in that sense point to a similarity between academic prose and essays.

Another observation that can be made on the basis of the trigrams is that the fiction ngrams often include a past tense form of a verb while both academic prose and the BAWE essays include a present tense form (cf. column 1 in Table 1). This tendency is pointed out by Biber et al. (1999: 456), where the corpus findings show that academic prose shows “a strong preference for present tense forms. Fiction shows the opposite pattern, with a strong preference for past tense verbs”; among the top 50 trigrams for BNC-Baby fiction, we find: *there was a*, *it was a*, *it was the*, *that it was*, etc. (cf. Appendix I).

The trigrams that are found among the 30 most frequent ones in one of the sub-corpora only are listed in Table 2.

Table 2 Trigrams unique among the 30 most frequent ones in one sub-corpus[11]

BAWE essays	**Academic prose**	**Fiction**
a sense of	a number of	a lot of
can be seen	and so on	back to the
due to the	in relation to	for a moment
heart of darkness	in terms of	going to be
in the novel	in this case	had been a
in the poem	it has been	he had been
it is the	it may be	he was a
of the novel	likely to be	in front of
of the poem	per cent of	it had been
the good soldier	some of the	it was a
the idea of	terms of the	it was the
the importance of	that there is	it would be
the reader is	the effect of	on to the
the reader to	the number of	out of the
the way in	the basis of	she had been
to the reader		shook his head
way in which		side of the
		that he had
		that he was
		the back of
		the rest of
		there was a
		there was no
		was going to
		would have been

Table 2 reveals quite a few things about the three sub-corpora under discussion. In fiction, as we would expect, personal pronouns are commonly found as part of trigrams; both *she* and *he* are used as part of the 30 most frequent trigrams. Another aspect worth noting is that

[11] Capital letters have not been included in the tables throughout, although some ngrams are capitalised in the original texts, e.g. *Heart of Darkness*. However, if part of the current discussion, these ngrams will be capitalised.

trigrams containing verbs are much more frequent in fiction than in the other two categories, and as mentioned above, the verbs are commonly found in their past tense form. Fiction also has the largest number of trigrams not found in the other two categories, 27 vs. 14 for academic prose and 17 for essays. This suggests that of the three categories under discussion, fiction is the one that least resembles either of the other two.

The trigrams found among the top 30 in English essays undoubtedly reveal more of the actual content of the essays than is the case of either fiction or academic prose. Although the academic prose trigrams also contain content words such as *terms* and *basis*, the content words in the essays have more specific content, e.g. *poem*, *reader*, *text*. In fact, from the list in Table 2 it can be deduced that the essays are concerned with literary analysis, drawing on a narrower pool of lexis than in a less specialized corpus (cf. Pecorari 2008). In such cases it may therefore be more rewarding to look at patterns as text-type discriminators rather than exact lexical matches (e.g. *of* + det. + noun).

Before we move on to examine the function of the ngrams in our sample, let us examine quadrigrams in the material.

4.2 Quadrigrams

Four-word combinations, or quadrigrams, show similar patterns to those of trigrams across the three text categories. Of the top 50 quadrigrams, only four occur in all three categories: *the end of the*, *the rest of the*, *at the end of*, *at the same time* (see Appendix II for a complete list). It should be noted here that some of the quadrigrams include some of the trigrams, e.g. *at the end of* ◂ *at the end*, or as Biber et al. (1999:990) put it "[s]horter bundles are often incorporated into more than one longer lexical bundle".

Again, a combination featuring the noun *end* is central and seems to be an item that commonly occurs across text categories.[12] Also worth noticing is the sequence "(prep. +) definite article + noun + preposition (+ def. art.)" as a common prefabricated and recurrent pattern, which has

[12] Incidentally, this has also been noted across a variety of other apprentice writer corpora (including foreign learners), cf. Brook O'Donnell and Römer's presentation at the 2009 ICAME conference on "Proficiency development and the phraseology of learner language".

previously been identified one of “the top-5-POS-grams in the whole BNC (down to a cut-off of 3 for n-grams)” (Stubbs 2004).

As with the trigrams, recurrent quadrigrams among the top 50 show most overlap between academic prose and the BAWE essays, with 11 shared quadrigrams, in addition to the four mentioned above. Three of these are of the type DET + N + PREP + REL. PRON. (e.g. *the way in which*). The remaining eight are: *on the other hand*, *it is possible to*, *the fact that the*, *that there is no*, *in the form of*, *as well as the*, *it is clear that*, *that there is a*. These reflect the same tendency as the trigrams; most are formulaic devices representing a particular writing style. This shared feature may point to the fact that the literary criticism performed in the BAWE essays is a subset of the wider academic prose genre, triggered by the academic setting in which the essays have been produced.

As regards quadrigrams among the top 30 that are unique to one sub-corpus, Table 3 shows that most are found in fiction (26), with academic prose (20) and the BAWE essays (18) not far behind.

Table 3 Quadrigrams unique among the 30 most frequent ones in one sub-corpus

BAWE essays	**Academic prose**	**Fiction**
allows the reader to	a wide range of	a cup of tea
as can be seen	as a result of	as if he was
at the beginning of	as shown in fig	at the back of
by the use of	at the time of	at the top of
can be seen in	can be used to	did not want to
could be argued that	in terms of the	for the first time
in the winter’s tale	in the case of	he was going to
in the good soldier	in the context of	he shook his head
in heart of darkness	in the united states	i don’t want to
it could be argued	in rylands v fletcher	in the middle of
of the poem the	in the form of	in front of him
the beginning of the	in relation to the	in front of the
the image of the	in the absence of	in one of the
the importance of the	it is important to	in the first place
the death of the	on the basis of	it would have been
through the use of	one of the most	on the edge of
to the fact that	per cent of the	on the other side

way in which the	the house of lords	the edge of the
	the house of commons	the top of the
	the size of the	the back of the
		the side of the
		the other side of
		the middle of the
		the centre of the
		was going to be
		what do you mean

The most specific combinations with content nouns are found in the essays, but notably there are also three very specific noun combinations in academic prose: *the House of Lords*, *in the United States*, *in Ryland v Fletcher*, and *the House of Commons*, all being more topic-specific than the other combinations (including trigrams).

Both academic prose and fiction show clear tendencies with regard to patterns that may help define them as text-types. The most common pattern among the 20 quadrigrams found in academic prose is PREP + (DET) + N + PREP + (DET), e.g. *on the basis of.* Although the fiction texts commonly make use of a similar pattern, there is a difference in the choice of noun, e.g. *the edge of the*. A similar tendency was found by Stubbs and Barth in their learned vs. fiction categories (Stubbs & Barth 2003: 72). Furthermore, the fiction patterns typically have an element (N) indicating position: *top*, *middle*, *back*, etc. (cf. Stubbs and Barth 2003:72). Another characteristic feature of the fiction quadrigrams are the combinations with personal pronouns + VP, e. g. *I don't want to*. Based on the trigrams and quadrigrams examined from BNC-Baby fiction, it can be concluded, with Stubbs and Barth, that it has a verbal style with vocabulary from the lexical fields of knowing and wanting.

On the basis of recurrent word-combinations, academic prose can be summarised as being nominal in nature; 17 out of the 20 quadrigrams listed in Table 3 contain a noun, most commonly an abstract noun. Our findings correspond well with what Biber et al. (1999:992) recorded for lexical bundles in academic prose: "most lexical bundles in academic prose are building blocks for extended noun phrases or prepositional phrases". Furthermore, only two lexical verbs have been recorded among the quadrigrams unique to academic prose, viz. *show* and *use*. Whether it is on the basis of facts like these that Stubbs and Barth state that their learned category is characterised by "a lack of stylistic variation" (2003:79) is hard to determine, but it is certainly one factor pointing in that direction.

The BAWE essays show less uniform patterns than either of the other two categories. They display some similarities with academic prose in that nouns are predominant; however, the noun patterns are more similar in their form and function in the academic prose texts, with more instances of preposition preceding the noun/noun phrase as instances of formulaic language. (This will be discussed in more detail in the following sections.) Furthermore, the essays show very few similarities with fiction, and I would hesitate in characterising essay as somewhere between academic prose and fiction in terms of text type. The picture seems to be more complicated than that. Nevertheless, ngrams have been shown to differ in the text-types discussed above, which suggests that they may be used for text-type discrimination, at least to some extent, and they will serve as a good starting point for the following functional analysis of the ngrams.

5. Recurrent word-combinations characterising English essays vs. academic prose

In the previous sections some similarities have been found between the BAWE essays and academic prose, particularly with regard to nominal patterns. However, differences, also in these patterns, have emerged, thus a functional analysis of the most frequent n-grams in these two text-types is called for, in order to offer a more in-depth comparison.

After introducing the model on which the functional analysis is based, I will first look at the BAWE essays. The most striking characteristic of the BAWE essays involves the relatively high number of specific proper nouns. In order to determine whether there is more uniformity in the most frequent ngrams, this section is devoted to a functional analysis of the ngrams found in the essays. I will concentrate on the recurrent word-combinations that are most commonly found in the BAWE essays. More specifically, I will take a closer look at the trigrams and quadrigrams that were found among the top 30 only in the BAWE material (cf. column 1 in Tables 2 and 3). Bigrams will not be included here, as they were seen to contain too few text defining features (cf. Section 4). A more detailed analysis, featuring a functional analysis, of trigrams and quadrigrams will hopefully tell us more about the text-type "English essay" than we have managed to reveal so far.

A similar analysis will then be carried out on the top 30 tri- and quadrigrams in academic prose before the functional characteristics of the two text-types are compared.

5.1 Classification and interpretation

In the analysis, I will adopt Moon's classification model which is based on Halliday's model of the three metafunctions: ideational, interpersonal and textual. Moon uses the model to classify fixed expressions and idioms (FEIs), and she states that "[t]he text functions of FEIs may be classified according to the way in which they contribute to the content and structure of a text. The precise contribution is instantial and bound up with context, but it is nevertheless possible to generalize and to chart typical functions" (Moon 1998:217). Figure 1 relates Halliday's ideational and interpersonal functions to Moon's FEI functions.

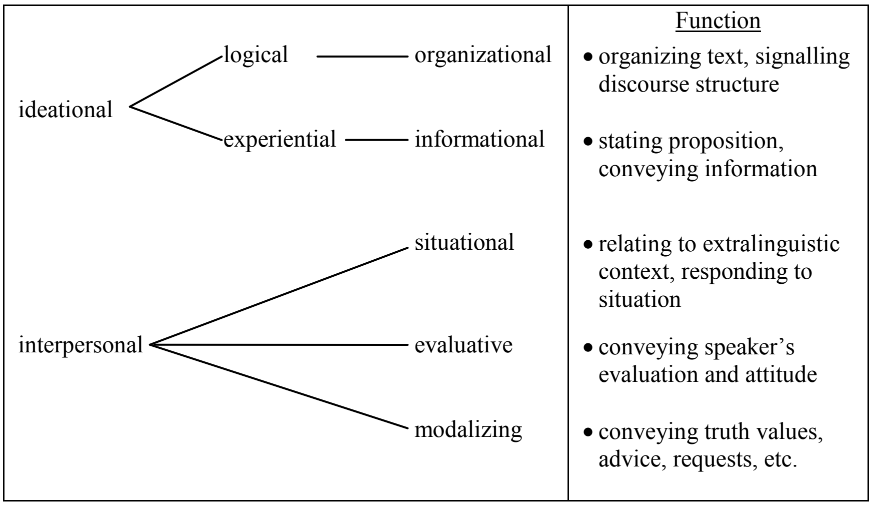
Function
ideational
logical
organizational
experiential
informational
interpersonal
situational
evaluative
modalizing
• organizing text, signalling discourse structure
• stating proposition, conveying information
• relating to extralinguistic context, responding to situation
• conveying speaker's evaluation and attitude
• conveying truth values, advice, requests, etc.

Figure 1: Breakdown of Moon's model (based on Moon 1998:217-218)

Halliday's textual function does not form part of Moon's model; "[t]he textual component [...] is best considered instantially in terms of the ways in which FEIs are placed topically and thematically, or contribute cohesion to their texts" (ibid:218-219). It should be mentioned that Moon's organizational category has been criticized for being placed within the ideational function. As Culpeper and Kytö point out "[s]ome of the organizational expressions Moon deal with seem to overlap with the textual function […]. Later in her book, Moon refers back to her organizational categories and states that they provide 'grammatical cohesion'" (Culpeper & Kytö 2002: 49). Nevertheless, I will use Moon's categories here, well aware that her organizational category is not exclusively ideational, but also textual in Hallidayan terms.

The main difference between Moon's FEIs and our ngrams is that the ngrams do not necessarily make up linguistic units. FEIs are all fixed expressions or idioms that typically form meaningful phrases or clauses. Nevertheless, we will try to apply the model to our ngrams, only excluding ngrams containing functional words only, e.g. *it is the* (where *is* is condidered to have too little content to be labelled "lexical"). Table 4 gives an overview of the 35 most common trigrams and quadrigrams (unique among the top 30) in the BAWE essays, according to functional class. Again it should be pointed out that trigrams may be subsumed under quadrigrams (see Section 5.3). However, this is not believed to distort the overall functional classification of these ngrams; we simply get some overlap.

Table 4 Functional classification of ngrams in English essays

Organizational	**Informational**	**Evaluative**	**Modalizing**
at the beginning of	heart of darkness	a sense of	as can be seen
beginning of the	in heart of darkness	allows the reader to	can be seen in
by the use of	in the good soldier	due to the	can be seen
in the first the	in the novel	the importance of the	could be argued that
through the use of	in the poem	the importance of	it could be argued
	in the winter's tale	the way in	
	of the novel	to the fact that	
	of the poem the	way in which the	
	of the poem	way in which	
	the death of the		
	the good soldier		
	the idea of		
	the image of the		
	the reader is		
	the reader to		
	to the reader		

In my analysis, most of the ngrams were seen to fall inside the text function "informational", mainly due to the fairly high number of nouns, both proper and common. In some cases it may be difficult to classify the ngrams according to the functional framework specified above, without looking at more context. Thus, an example of each of the categories might be in order.

Organizational
(1) *In the first* stanza he uses hellish imagery: …

Informational
(2) *The reader is* plunged into the mind of the character …

Evaluative
(3) 'the Winter's Tale' is an unusual play *due to the* fact that it cannot be easily categorised into any particular form of play.

Modalizing
(4) *It could be argued* that the society created, and therefore Morvern, is simply amoral, …

Examples 1-4 are all fairly clear examples of their category. However, what about the ngram *in the Winter's Tale*, which has been classified as informational rather than organizational. Typically, this particular ngram has either been used as part of a noun phrase as in (5) or as an Adverbial providing information (6).

(5) The characters *in The Winter's Tale* have a past, …
(6) Shakespeare deliberately ridicules the 'unity of time' *in The Winter's Tale …*

Only in one of the 21 occurrences of *in the Winter's Tale* has it been used in sentence initial position where it could be considered as an organizational device (7).

(7) *In The Winter's Tale* (Shakespeare, 1610), a narrator personifying time is quickly and visually used to describe a shift in time of sixteen years.

The different interpretations of *in the Winter's Tale* are based on the functional role the ngram plays in the context, and it may not always be straightforward.

Returning to Table 4, we see that, proportionally, the ngrams have been grouped as follows: organizational 13.9%, informational 44.4%, situational 0%, evaluative 27.8%, and modalizing 13.9%. This means that more than 58% of the ngrams are, in Hallidayan terms, of an ideational nature, whereas the remaining 42% or so are of an interpersonal nature. On the level of discourse, then, it could be stated that English essays are highly informational, relatively evaluative, and to some extent organizational and modalizing. Not unexpectedly no clear situational word-combinations emerged; they are "typically found in spoken discourse as they are responses to or occasioned by the extralinguistic context: they may also be illocutionary speech acts. They are therefore constrained by real-world sociocultural factors" (Moon 1998: 225). Examples include: "I beg your pardon", "go for it", "it's a small world" (ibid.).

In a study of phraseology and epistemology in academic book reviews within two humanities disciplines, Groom confirms the claim that the epistemology of the knowledge domain of the humanities is essentially reiterative in nature "(i.e. concerned with revisiting perennial questions and reinterpreting previously existing data)" (Groom 2009: 124). His study focused on recurrent combinations around key prepositions (e.g. conceptualization + *of* + phenomenon). Although the approach in the present study differs from that of Groom, I believe the functional analysis of ngrams in English studies essays (as a discipline within the humanities) equally lends support to the humanities as being reiterative, particularly shown through the high proportion of informational and evaluative ngrams.

In order to investigate how the BAWE ngrams compare in functional terms with those of academic prose, a similar analysis of the trigrams and quadrigrams in academic prose was also performed, revealing the functional distribution given in Table 5.

Table 5 Functional classification of ngrams in academic prose

Organizational	**Informational**	**Evaluative**	**Modalizing**
and so on	a number of	as a result of	can be used to
as shown in fig	a wide range of	it is important to	it may be
in relation to the	at the time of	one of the most	likely to be
in relation to	in rylands v fletcher		
in the context of	in terms of the		
	in terms of		
	in the absence of		
	in the case of		
	in the form of		
	in the united states		
	in this case		
	on the basis of		
	per cent of the		
	per cent of		
	some of the		
	terms of the		
	the basis of		
	the effect of		
	the house of lords		
	the house of commons		
	the number of		
	the size of the		

Compared to the BAWE essays, then, academic prose is even more informational in its nature, as seen in the amount of informational ngrams in Table 4 (44.4% for BAWE) vs. Table 5 (66.7% for academic prose). The other major difference is found in the evaluative function, where the ngrams studied in academic prose only reached 9.1%, while the BAWE essays were clearly evaluative with 27.8%. The BAWE material also showed a slightly more modalizing tendency with 13.9% vs. 9.1% for academic prose. The two categories were fairly similar in terms of organizational elements used (15.1% vs. 13.9%).

As was the case with the BAWE essays, most of the ngrams are informational, typically containing nouns, both proper and common. While the common nouns in the BAWE essays are mainly concrete, topic-dependent nouns (e.g. *poem* and *reader*), the nouns featuring in the informational category in academic prose are more universal, and perhaps even abstract, e.g. *terms* and *basis*.

While academic prose is found to be overwhelmingly informational, the BAWE essays are typically informational *and* evaluative in nature. In

Moon's terms, then, it appears that academic prose is typically ideational in nature, while the English essay is both ideational and interpersonal.

Although we have looked in more detail at the ngrams that only figure among the top 30 of one text-type, it should also be noted that a fair amount of the trigrams and quadrigrams shared between BAWE and academic prose are structuring devices, i.e. organizational according to Moon's model, e.g. *and so on*, *at the beginning of.*

In this study, Stubbs and Barth's claim that ngrams are text-type type discriminators has been confirmed. In addition, and perhaps more important, the present study has demonstrated that, by taking ngrams as our starting point in a functional analysis of recurrent word-combinations, we are able to form better conclusions as to how text-types differ. This is true even in cases where the ngrams do not constitute full semantic units, nor full phrases in traditional terms.

6. Concluding remarks

In this study we have followed Stubbs and Barth (2003) in using recurrent word-combinations as a means of highlighting differences between groups of texts. Of the three sub-corpora we investigated, two were taken from BNC-Baby, fiction and academic prose, and one from the BAWE corpus, English Studies essay. Although the method put forward by Stubbs and Barth (2003) seems to work well in many respects, it is obvious that certain noun combinations in the BAWE essays tell us more about the subject matter of the texts than it does about the text-type in general, e.g. combinations such as *Heart of Darkness* and *The Good Soldier*. In this context it is important to stress that the BAWE material is relatively restricted compared to the two other text groups. On the other hand, combinations including words such as *poem* and *reader* are good indicators of English essays as a specific text-type. Furthermore, in the comparison of ngrams across academic prose, fiction and student essays, clear differences emerged. Academic prose and fiction showed more clear tendencies in the use of recurrent ngrams than the BAWE essays. Although it may be argued that this is due to the size of the material analysed here, since clearer tendencies might be expected in a larger material, we saw in Section 2 that size does not play a major role when investigating the top 30 ngrams.

A functional investigation of the most frequent trigrams and quadrigrams in the essays and academic prose was also carried out. It was found that most ngrams unique (among the top 30) to academic prose were overwhelmingly informational, while in BAWE they were informational or evaluative, leading to the conclusion that academic prose is highly ideational, whereas the English essay is both ideational and interpersonal. This confirms what we might have expected to be typical features of essays commenting on or analysing literature, which in many ways is a subjective activity and would call for use of interpersonal language (evaluative), such as personal pronouns, verbs of evaluation or attitude, etc. In addition, there also seems to be an urge to be informative, where the informative perhaps could be seen as a basis for the evaluative.

The overlap between academic prose and English essays is also worth noticing. The academic setting of essay writing triggers certain recurrent word-combinations commonly used in academic writing in general (e.g. *in order to*, *the fact that*, in Table 1), showing that students clearly make an effort to meet the academic writing practices that are expected, and as such, the English essay is a sub-category of academic writing.

The study reported here shows that even a small-scale investigation of recurrent word-combinations can tell us something about text-type. The method explored here paves the way for similar studies on text-types across more and different kinds of corpora. With corpora such as BAWE, and its American counterpart MICUSP,[13] more could be found out about similarities and differences across student writing. This would presumably show more clearly which combinations contribute to distinguish essays from other assignment types. Further studies could also include student vs. professional writing or native vs. non-native student writing in order to investigate how salient ngrams really are and to what extent the same patterns and functions are used by apprentice and professional writers, or by learners and native speakers.

[13] Michigan Corpus of Upper-Level Student Papers, http://micusp.elicorpora.info/, accessed 28 Oct. 2009.

References

Primary Sources—Corpora

BNC-Baby v1.0, designed and constructed by Y. Berglund and M. Wynne, Oxford Text Archive. CD edited and compiled by L. Burnard, Oxford University's Research Technologies Service, Aug. 2004, <http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf>, accessed 28 Oct. 2009.

British Academic Written English Corpus, <http://www.coventry.ac.uk/researchnet/d/911>, accessed 28 Oct. 2009.

Secondary Sources

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

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

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

Charles, M. 2006. "The construction of Stance in Reporting Clauses: A Cross-disciplinary Study of Theses." *Applied Linguistics 27/3*: 492-518.

Culpeper, J. and M. Kytö. 2002. "Lexical Bundles in Early Modern English dialogues: A window into the speech-related language of the past". In T. Fanego, B. Méndez-Naya, and E. Seoane (eds.), *Sounds, Words, Texts and Change*. Selected Papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000. Amsterdam/Philadelphia: John Benjamins, 45–63.

Ebeling, S.O. & P. Wickens. Forthcoming. "Interpersonal themes and author stance in student writing." To appear in *Proceedings from the 30th ICAME Conference*, Lancaster 2009.

Groom, N. 2009. "Phraseology and epistemology in academic book reviews: A corpus-driven analysis of two humanities disciplines." In

K. Hyland & G. Diani (eds.) *Academic Evaluation. Review Genres in University Settings*. Basingstoke: Palgrave MacMillan, 122-139.

Heuboeck, A., J. Holmes, and H. Nesi. 2007. *The BAWE Corpus Manual*. Available under 'BAWE documentation at http://www.coventry.ac.uk/researchnet/d/505/a/5160, accessed 28 Oct. 2009.

Martin. J.R. 1992. *English Text. System and Structure*. Philadelphia / Amsterdam: John Benjamins Publishing Company.

Martin. J.R. 1997. "Analysing genre: functional parameters." In Christie, F. and J.R. Martin (eds.), *Genre and Institutions: Social Processes in the Workplace and School*. London: Continuum, 3-39.

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

Nesi, H. and S. Gardner. Unpublished manuscript. "Families of genres of assessed writing."

Nesi, H., S. Gardner, R. Forsyth, D. Hindle, P. Wickens, S.O. Ebeling, M. Leedham, P. Thompson, and A. Heuboeck. 2005. "Towards the compilation of a corpus of assessed student writing: An account of work in progress." In Danielsson, P. and M. Wagenmakers (eds.) *Proceedings from The Corpus Linguistics Conference Series*, Vol. 1, No.1 <http://www.coventry.ac.uk/researchnet/external/content/1/c4/30/09/v1192628796/user/towards_compilation.pdf>.

Pecorari, D. 2008. "Repeated language in academic discourse: the case of biology background students." *Nordic Journal of English Studies*, vol. 7 No. 3. 9-33.

Scott, M., 2008. WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

Scott, M. and C. Tribble. 2006. *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam / Philadelphia: John Benjamins.

Shaw, P. 2009. "Linking adverbials in student and professional writing in literary studies: What makes writing mature." In M. Charles, D. Pecorari, and S. Hunston (eds.) *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum, 215-235.

Sinclair, J. 1999. "A way with common words." In H. Hasselgård & S. Oksefjell (eds.), *Out of Corpora. Studies in Honour of Stig Johansson*. Amsterdam / Philadelphia: Rodopi, 157-180.

Stubbs, M. 2004. “On Very Frequent Phrases in English: Distributions, Functions, and Structures.” Plenary lecture at ICAME 25, Verona, Italy, 2004. http://web.archive.org/web/20070828004603/http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/icame-2004.htm, accessed 20 April 2010.

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

Swales, J.M. 1990. *Genre Analysis*. Cambridge: Cambridge University Press.

Appendix 1

Top 50 trigrams in the three text-types in alphabetical order (academic prose, fiction, and BAWE essays)

BAWE essays	**Academic prose**	**Fiction**[14]
a sense of	a number of	a couple of
an example of	and it is	a kind of
as it is	and so on	a lot of
as well as	as a result	and it was
at the end	as well as	as if he
can be seen	because of the	as soon as
due to the	but it is	at the end
end of the	end of the	back to the
heart of darkness	in order to	be able to
image of the	in relation to	but it was
in order to	in terms of	do you think
in the first	in the case	end of the
in the novel	in the first	for a moment
in the poem	in this case	going to be
in which the	in which the	had been a
is able to	is likely to	he did not
it is a	is not a	he had been
it is not	it can be	he was a
it is possible	it has been	i don't know
it is the	it is a	i want to
nature of the	it is not	in front of
of the city	it is possible	it had been
of the novel	it may be	it was a
of the play	likely to be	it was the
of the poem	many of the	it would be
of the text	on the other	must have been
on the other	one of the	on to the
one of the	part of the	one of the

[14] WordSmith seems to counts e.g. inverted commas as part of ngrams. For the purpose of this study I have not included trigrams or quadrigrams of the type *he said "*, where the inverted comma is looked upon as the third word of the trigram.

such as the
that it is
the beginning of
the end of
the fact that
the good soldier
the idea of
the image of
the importance of
the other hand
the poem is
the reader is
the reader to
the use of
the way in
the winter’s tale
there is a
there is no
to be a
to the reader
use of the
way in which

per cent of
shown in fig
some of the
terms of the
that it is
that there is
the basis of
the case of
the effect of
the end of
the fact that
the form of
the house of
the number of
the other hand
the terms of
the united states
the use of
there is a
there is no
this is not
to be a

out of the
part of the
she did not
she had been
shook his head
side of the
that he had
that he was
that it was
that she had
that she was
the back of
the end of
the first time
the rest of
there was a
there was no
to be a
to do with
to have a
was going to
would have been

Appendix II

Top 50 quadrigrams in the three text-types in alphabetical order (academic prose, fiction, and BAWE essays)

BAWE essays
allows the reader to
and the good soldier
as can be seen
as well as the
at the beginning of
at the end of
at the same time
be seen in the

Academic prose
a wide range of
as a result of
as shown in fig
as we have seen
as well as the
at the end of
at the same time
at the time of

Fiction
a cup of tea
as if he was
at the back of
at the bottom of
at the end of
at the same time
at the top of
did not want to

by the use of
can be seen in
could be argued that
gender and gender roles
importance of being earnest
in contrast to the
in heart of darkness
in the form of
in the good soldier
in the winter's tale
is an example of
it could be argued
it is clear that
it is possible to
mise en scene
of the poem the
on the other hand
rest of the poem
tess of the d'urbervilles
that there is no
that there is a
the beginning of the
the death of the
the end of the
the extent to which
the fact that the
the image of the
the importance of the
the importance of being
the nature of the
the rest of the
the role of the
the structure of the
the use of the
the use of language
the way in which
the ways in which

by the fact that
can be used to
in relation to the
in rylands v fletcher
in terms of the
in the absence of
in the case of
in the context of
in the course of
in the form of
in the united states
in this case the
is likely to be
it is clear that
it is difficult to
it is important to
it is necessary to
it is possible to
more likely to be
on the basis of
on the other hand
on the part of
one of the most
per cent of the
regions ii and iii
rule in rylands v
that there is a
that there is no
the basis of the
the court of appeal
the end of the
the extent to which
the fact that the
the house of lords
the house of commons
the nature of the
the rest of the
the rule in rylands

for the first time
from time to time
got to his feet
he shook his head
he was going to
I don't want to
in front of him
in front of the
in one of the
in the direction of
in the first place
in the middle of
it had been a
it was as if
it would have been
nothing to do with
on the edge of
on the other side
other side of the
out of the window
she shook her head
she was going to
that sort of thing
the back of the
the back of his
the bottom of the
the centre of the
the edge of the
the end of the
the middle of the
the other side of
the rest of the
the side of the
the top of the
the two of them
there had been a
there was no sign

through the use of
to make sense of
to the fact that
way in which the
with the use of

the size of the
the structure of the
the way in which
the ways in which

to be able to
was going to be
was one of the
what do you mean
what do you think