6 Rapid #: -2219369

Ariel
IP: 134.114.228.9

Status	Rapid Code	Branch Name	Start Date
Pending	HUA	Main Library	12/14/2008 1:15:34 PM

CALL #: **410 I61 J86**
LOCATION: **HUA :: Main Library :: maise**

TYPE:	Article CC:CCL
JOURNAL TITLE:	International journal of corpus linguistics
USER JOURNAL TITLE:	International journal of corpus linguistics
HUA CATALOG TITLE:	International journal of corpus linguistics.
ARTICLE TITLE:	Investigating Language Use through Corpus-Based Analyses of Association Patterns
ARTICLE AUTHOR:	Biber, Douglas
VOLUME:	1
ISSUE:	2
MONTH:	
YEAR:	1996
PAGES:	171-197
ISSN:	1384-6655
OCLC #:	
CROSS REFERENCE ID:	250858
VERIFIED:	

BORROWER: **AZN :: Main Library**
PATRON: **Gray, Bethany**

PATRON ID:	
PATRON ADDRESS:	
PATRON PHONE:	
PATRON FAX:	
PATRON E-MAIL:	
PATRON DEPT:	
PATRON STATUS:	
PATRON NOTES:	CSA:llba-set-c

Investigating language use through corpus-based analyses of association patterns

DOUGLAS BIBER
Northern Arizona University

The present paper argues that analyses of language use provide an important complementary perspective to traditional linguistic descriptions, and that empirical approaches are required for such investigations. Corpus-based techniques are particularly well suited to these research purposes, enabling investigation of research questions that were previously disregarded. Specifically, the paper discusses the use of corpus-based techniques to identify and analyze complex "association patterns": the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features. Several illustrative analyses are discussed, investigating the use of lexical features, grammatical features, and the overall patterns of variability among texts and registers.

KEYWORDS: association patterns, collocations, grammatical analysis, language use, lexical analysis, register variation

1. Introduction

In recent years, language use has come to be recognized as an important aspect of linguistic study, with equal status to the study of language structure. This development marks a return to the priorities of certain schools of linguistics in the 1950s, most notably Firthian linguistics (see, e.g., the papers in the volume edited by Palmer 1968).

Studies of use are concerned with actual practice, and the extent to which linguistic patterns are common or rare, rather than focussing exclusively on

INTERNATIONAL JOURNAL OF CORPUS LINGUISTICS Vol. 1(2), 1996. 171-197

potential grammaticality. As such, adequate investigations of language use must be empirical, analyzing the functions and distribution of linguistic features in natural discourse contexts. In descriptive lexicography, which is concerned with the actual use of words, new meanings are discovered only by examining the use of a word in actual discourse contexts. Grammatical structures can also be compared from a use perspective, by studying the ways in which seemingly similar structures occur in different contexts and serve different functions. In addition, a use perspective is required to investigate the stylistic preferences of individuals, the differing linguistic preferences of groups of speakers, and the ways in which "registers" (or "genres") favor some words and structures over others.

Corpus-based analyses are particularly well suited to such investigations. Over the past decade, there has been a dramatic increase in corpus-based language studies. For example, the bibliography of corpus-based studies provided by Altenberg (1991a) includes well over 600 entries, and many more studies have appeared in the last five years (see, e.g., the edited collections by Aijmer and Altenberg 1991; Armstrong 1994; Johansson and Stenström 1991; Svartvik 1990, 1992). The essential characteristics shared by these corpus-based studies are:

- they are empirical, analyzing the actual patterns of use in natural texts;
- they utilize a large and principled collection of natural texts (i.e., a "corpus") as the basis for analysis;
- they make extensive use of computers for analysis, using both automatic and interactive techniques;
- they depend on both quantitative and qualitative (interpretive) analytical techniques.

One major advantage of a corpus-based approach is that it enables a scope and reliability of analysis not otherwise feasible. Corpus-based analyses can be based on an adequate representation of naturally-occurring discourse, including analysis of complete texts, multiple texts from any given variety, and inclusion of multiple spoken and written varieties for comparative purposes. Using computational techniques, it is feasible to entertain the possibility of a comprehensive linguistic characterization of a text, analyzing a wide range of linguistic features (rather than being restricted to a few selected features); further, computational techniques can be used to analyze the complex ways in which linguistic features interact within texts. For quantitative analyses, corpus-based methods result in greater reliability and accuracy: computers do not become bored or tired — they will count a linguistic feature in the same way every time it is encountered. Finally, corpus-based analyses enable the possibility of cumulative results and public accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be analyzed using the same computational techniques. Such studies can test the results of previous research, and findings can be compared across studies, building a cumulative linguistic description of the language.

Even more important, corpus-based techniques enable investigation of new research questions that were previously disregarded because they were considered intractable. In particular, the corpus-based approach makes it possible to identify and analyze complex "association patterns": the systematic ways in which linguistic features are used in association with other linguistic and non-linguistic features.[1]

Association patterns can be regarded as an extension of Firth's notion of collocation (e.g., Firth 1952). Collocations are characterizations of a word in terms of the other words that it typically co-occurs with. Firth also paid attention to the relationship between collocations and the "context of situation", focussing primarily on the different purposes for which a word might be used.

The notion of association pattern extends the concept of collocation in several ways. First, association patterns are identified empirically from analysis of a representative corpus; many stereotypical collocations do not in fact represent strong association patterns, while other unanticipated collocations are identified by empirical analyses of association patterns. Second, association patterns represent continuous relationships that must be analyzed in quantitative terms. Given a large, representative corpus and the appropriate analytical tools, association patterns can be specified in precise quantitative terms, identifying the extent to which a particular type of relationship is found. Finally, association patterns are used to characterize grammatical features as well as words, with respect to systematic co-occurrence patterns with other words, other grammatical features, or non-linguistic characteristics of the context.[2]

As Table 1 shows, association patterns are used to investigate two major kinds of research question: the variablity of a linguistic feature, and the variablity among texts.

Table 1. Kinds of association patterns

A) Investigating the variablity of a linguistic feature (lexical or grammatical)
i) Non-Linguistic associations of the feature:
– distribution across registers
– distribution across dialects
– distribution across time
ii) Linguistic associations of the feature:
– co-occurrence with particular words
– co-occurrence with grammatical features
B) Investigating the variablity among texts.
– "dimensions" = co-occurrence patterns of linguistic features

Investigating the variability of a linguistic feature in terms of its association patterns has two major components: 1) non-linguistic associations, and 2) linguistic associations. Non-linguistic association patterns describe how certain linguistic features are differentially associated with registers, dialects, or historical change.

There are two main types of linguistic association patterns: lexical associations and grammatical associations. Both individual words and grammatical constructions can be studied with respect to their association patterns. For a corpus-based study of an individual word, the lexical associations are the collocations of the target word (other words that the target word frequently co-occurs with). The grammatical associations of the target word describe structural preferences, for example, whether a particular adjective typically occurs with attributive or predicative functions, or whether a particular verb typically occurs with transitive or intransitive functions.

Corpus-based studies of a grammatical construction can similarly include both lexical and grammatical associations. In this case, the lexical associations are the tendencies for the target grammatical construction to co-occur with particular words. For example, what matrix-clause verbs typically occur with a *that*-clause, and do a different set of matrix-clause verbs typically occur with *to*-clauses? Grammatical associations in this case identify contextual factors associated with structural variants. For example, are *that*-clauses used in extraposed constructions as often as *to*-clauses?

All of these linguistic association patterns interact with non-linguistic associations. In fact, corpus-based analyses show that linguistic association patterns are generally *not* valid for the language as a whole. Rather, linguistic and non-linguistic associations interact with one another, so that strong linguistic associations in one register often represent only weak associations in other registers.

One final type of association pattern is important when the research goal is to describe texts and registers rather than individual linguistic features: the ways in which groups of linguistic features commonly co-occur in texts. For example, frequent nouns, adjectives, and prepositional phrases commonly co-occur in academic prose texts, working together to provide a dense integration of information. Textual co-occurrence patterns such as these are important in identifying the salient linguistic characteristics of registers and styles.

What is just now coming to be realized is how extensive and systematic the patterns of language use are. Such association patterns are well beyond the access of intuitions, and yet these patterns are much too systematic to be disregarded as accidental. While future research is required to determine the theoretical underpinnings of these patterns (and the extent to which they can be attributed to cognitive, situational, or textual factors), we are now in a position to document the extent and nature of these patterns much more fully than has heretofore been possible.

The following sections provide example analyses of each type of association pattern: association patterns for individual words are illustrated in Section 2; association patterns for grammatical constructions are illustrated in Section 3; and register analyses with respect to textual co-occurrence patterns are illustrated in Section 4. The analyses are carried out on a 10 million-word subsample from the Longman/Lancaster Corpus (c. 5 million-word samples from fiction and academic prose), supplemented by a 5 million-word sample of conversation from the British National Corpus. In the conclusion, the paper outlines some of the future investigations needed for an integrated description of linguistic structure and use.

2. Association Patterns for Individual Words

Over the past 10 to 15 years, lexicographic researchers have been at the leading edge of work that applies corpus-based techniques to standard linguistic research questions. There are now many concordancing packages that are commercially available for doing lexicographic research, and the most important new dictionaries (e.g., by COBUILD, Longman, and Cambridge)

are all based entirely on corpus analysis. (In contrast, there are few adequate commercially available tools for doing grammatical research on a corpus, and most publishers continue to rely on traditional methods for developing new grammars.) When coupled with a concordancing program, a corpus provides a wealth of examples for any given word, allowing lexicographers to more accurately identify and characterize the range of meanings for the word (see, e.g., Sinclair 1987, 1991).

However, the usefulness of corpus-based lexical analysis is not limited to dictionary-making. For example, several studies identify and characterize the use of relatively fixed lexical expressions (e.g., Altenberg 1991b, 1993; Kjellmer 1991; Renouf and Sinclair 1991). In addition, statistical measurements of word associations have been developed to further clarify the senses of words and identify the most important patterns of use (Biber 1993a, Church and Hanks 1990, Nakamura and Sinclair 1995).

One type of corpus-based investigation that is particularly interesting is the investigation of seemingly synonymous or near-synonymous words (e.g., Biber, Conrad, and Reppen 1994 on *certain* and *sure*; Kennedy 1991 on *between* and *through*). Dictionaries and thesauruses often list such words as equivalent in meaning. However, corpus-based investigations of association patterns show that there are important, patterned differences in the ways that native speakers use seemingly synonymous words. To illustrate, I briefly compare the association patterns for a pair of near-synonymous adjectives: *happy* and *glad*.

First, Table 2 shows that these adjectives are used to differing extents across registers:

Table 2. Distribution of adjective pairs across registers

	Conversation	Fiction	Academic prose
happy	******	********	*
glad	***	*****	-

[Each * marks approximately 20 occurrences per million words]
[- represents less than 10 occurrences per million words]

To understand why these register differences exist, it is useful to study the associated words that typically co-occur with each adjective, that is, the "collocations". Table 3 displays the most common nouns that occur as right collocates of each adjective. This table reports the strongest lexical associations as identified by t-scores computed for each collocational pair (see Church, Hanks, and Hindle 1991; Stubbs 1995).

Table 3. Preferred right collocates of ***happy*** *versus* ***glad***

happy	+	Nouns:	*man, family/families, couple, one, life, face, days*
		Punctuation:	. , ? ! ;
		Prepositions:	*with, as, about, at, in*
glad	+	Pronouns:	*I, you, he, she, we, they*
		Complementizers:	*to, that*
		Other:	*the, there, of*

Table 3 shows that each of these adjectives has a distinct pattern of lexical association. However, these lexical associations can also be analyzed as reflecting different grammatical associations. Thus, *happy* has a strong lexical association with several following nouns, indicating that it is relatively common as an attributive adjective; for example,

He was a *happy* man.
...like one big *happy* family.
She led a full and *happy* life there.

In contrast, none of the strong right collocates of *glad* are nouns, suggesting that this adjective is not common in attributive position. In fact, analysis of the grammatical distribution of *happy* and *glad* shows that both adjectives are much more common in predicative position rather than attributive position: the adjective *happy* occurs about 80% of the time in predicative position, while *glad* is almost always used in predicative position (over 98%).[3] For example:

1. I was comfortable and very *happy*.
2. You look *happy*!
3. She was *glad* to go.
4. He was *glad* that he could rest.

However, more detailed examination of the grammatical associations for these two adjectives uncovers a fundamental difference: although both adjectives usually occur in predicative position, *happy* is often used as the entire predication, and thus it is commonly followed by clause-final punctuation (as in examples #1 and #2 above). In contrast, *glad* is commonly followed by a *to-*

clause or a *that*-clause, which specifies what the person is glad about (as in examples #3 and #4 above).

In fact, nearly all of the preferred right collocates of *glad* represent a grammatical association of a following complement clause. For example, this pattern accounts for the strong lexical association that *glad* has to following pronouns (e.g., *I, you, he/she*). These cannot be analysed as attributive constructions, since pronouns cannot normally be modified by an attributive adjective. Instead, these are almost always the beginning of a following *that*-clause with the complementizer omitted, as in:

I'm *glad* I'll never finish it.
I'm so *glad* you could make it.
I'm *glad* she's here.

The collocate pairs *glad + the* and *glad + there* represent the same grammatical association (of *glad* followed by a *that*-clause):

I was just *glad* the abortion was over with.
I'm *glad* there isn't a radio.

The preferred grammatical associations of *happy* are quite different. As noted above, *happy* is frequently used as a predication standing on its own, as in SOMEBODY BE HAPPY. However, when *happy* does take a complement, it most commonly occurs with a prepositional phrase instead of a complement clause (thus accounting for the lexical association with following prepositions, such as *with*, *about*, and *at*). For example,

She used to be *happy* with her.
I'm not *happy* about this signature.
She did not appear *happy* at finding herself where she was.

This example illustrates the case where seemingly synonymous words actually have quite different patterns of use, reflected in their differing lexical and grammatical associations. A more complicated type of lexical-grammatical association involves words that are grammatical in the same range of structures; despite identical potentials, such words often have quite different typical associatons.

For example, the verbs *tell* and *promise* are grammatical in the same valency patterns: intransitive, monotransitive, and ditransitive. In actual use, though, these two verbs typically occur in quite different grammatical patterns:

Table 4. Percentage of verb tokens for ***tell*** *and* ***promise*** *occurring with intransitive, monotransitive, and ditransitive valency patterns.*[4]

	SV	SV+O (np)	SV+O (Comp-cls)	SV+IO+O (np)	SV+O+IO	SV+IO+O (Comp-cls)	SV+IO
tell							
Conv	-	-	-	***	-	****	***
Acad	-	**	-	*	-	*****	*
promise							
Conv	*	*	****	*	-	***	-
Acad	-	**	*****	*	-	*	-

[each * represents c. 10% of the tokens for that verb
- marks patterns that occur less than 10% of the time]

The most striking difference between the grammatical associations of these two verbs concerns their use in a monotransitive pattern followed by a complement clause: this is the most common pattern for *promise* but the rarest pattern for *tell*:[5]

Academic prose:
In return the student *promises* to campaign for the politician.

Conversation:
I *promised* that I wouldn't play it.
We still *promised* to go to aunty's.

The intransitive pattern is also more common with *promise* than with *tell*, especially in conversation:

No I'm not gonna use it — I've *promised*.
I won't laugh — I *promise*.

In contrast, ditransitive valency patterns with an indirect object are most common with the verb *tell*:

Academic prose:
The central mark *tells* us which region we are in.
The first law of thermodynamics *tells* us that energy may be converted...

Conversation:
I'll *tell* you what it is.
I *told* him it might need a new switch.

These association patterns seem to reflect a fundamental difference between the typical discourse functions of *tell* and *promise*: With the verb *promise*, the content of the promise (given as the direct object) is the most important consideration, while the person to whom the promise was made is often irrelevant. In contrast, the person being addressed is much more important with the verb *tell*, while the content of the speech act is in some cases irrelevant. As a result of these typical discourse functions, the grammatical associations of these two verbs are strikingly different, even though they have identical grammatical potentials.

Obviously, findings such as these require further interpretation, based on a fuller consideration of the individual patterns and a detailed analysis of individual instances in their discourse contexts. While it is not possible to undertake such an analysis here, these examples have illustrated the importance of lexical and grammatical association patterns in describing the meaning and use of individual words.

3. Association patterns for grammatical features

Corpus-based analyses can also be used to investigate grammatical issues, addressing research questions such as: What discourse functions does a grammatical construction serve, and how are related constructions used differently? How rare or common are related constructions? Are particular constructions used more or less frequently in different registers? Are there particular words that a grammatical construction commonly co-occurs with? What factors in the discourse context are associated with the use of grammatical variants?

There are a number of book-length treatments reporting corpus-based grammatical investigations: for example, Tottie (1991) on negation, Collins (1991) on clefts, Granger (1983) on passives, Mair (1990) on infinitival complement clauses, Meyer (1992) on apposition, and several books on nominal structures (e.g., de Haan 1989, Geisler 1995, Johansson 1995, Varantola 1984). In addition, there have been numerous research papers using corpus-based techniques to study English grammar (see many of the papers in collected volumes such as Aarts and Meyer 1995, Aijmer and Altenberg 1991, Johansson and Stenström 1991).

Many of these studies use corpora to analyze the influence of contextual factors on the distribution of structural variants. Both lexical and grammatical

association patterns have been shown to be important. For example, Mair (1990) identifies a number of individual verbs that are particularly common with various infinitival constructions; de Haan (1989) identifies the association of relative clauses with head noun phrases having different grammatical roles.

To illustrate association patterns of this type, I briefly describe certain aspects of the grammar of complement clauses in English. The two most common types of complement clause are *that*-clauses and *to*-clauses. In some contexts, these two are similar in meaning. For example, compare:

I hope that I can go.
I hope to go.

However, corpus-based study shows that the actual use of these two structures is quite different. First, in terms of their overall distribution, *that*-clauses are very common in conversation but not so common in academic prose. In contrast, *to*-clauses are moderately common in both conversation and academic prose:

Table 5. Overall distribution of ***that****-clauses and* ***to****-clauses in conversation and academic prose*

	Conversation	Academic Prose
that-clauses	**********	****
to-clauses	********	*********

[each * represents 500 occurrences per million words]

This difference in overall distribution can be related in part to the differing lexical associations of the two types of complement clause. That is, while a few verbs can control both *that*-clauses and *to*-clauses (e.g., *hope*, *decide*, and *wish*), most verbs control only one or the other type of complement clause. For example, the verbs *imagine*, *mention*, *suggest*, *conclude*, *guess*, and *argue* can control a *that*-clause but not a *to*-clause; the verbs *begin*, *start*, *like*, *love*, *try*, and *want* can control a *to*-clause but not a *that*-clause.

These differential patterns of lexical association are even stronger when we consider relative frequency. Thus, Tables 6 and 7 show that the most common verbs controlling a *that*-clause constitute a completely separate set from the most common verbs controlling a *to*-clause, even though some of these verbs are grammatical with both types of complement clause.

Table 6. Most common verbs controlling a ***that****-clause*

	Conversation	Academic Prose
think	**********	*
say	**********	**
know	*********	*
see	**	**
show	*	***
find	*	**
believe	*	*
feel	*	-
suggest	-	**

[each * represents 100 occurrences per million words]
[- represents less than 50 occurrences per million words]

Table 7. Most common verbs controlling a ***to****-clause*

	Conversation	Academic Prose
want	**********	-
try	***	*
like	**	**
seem	*	*
tend	-	**
appear	-	**
begin	-	*
attempt	-	*
continue	-	*
fail	-	*

[each * represents 100 occurrences per million words]
[- represents less than 50 occurrences per million words]

Some of these verbs (such as *want* and *try*) are grammatical controlling only one type of complement clause, and they have strong lexical associations with that structural type. Other verbs — such as *think*, *say*, and *know* — can control both types of complement clause; however, these verbs have strong association patterns with only one type of complement clause.[6] Thus, although there is some overlap between the two types of complement clause in the controlling verbs that are grammatical, corpus-based analysis shows that there is in fact very little overlap in the commonly occurring lexical associations.

Further, *that*-clauses and *to*-clauses are productive in different ways. *That*-clauses combine with relatively few verbs, from only a few semantic domains — mostly mental verbs (e.g., *think*, *know*, *feel*, *hope*) or communication verbs (e.g., *say*, *suggest*). However, a few of those verbs are extremely common controlling *that*-clauses, especially the verbs *think*, *say*, and *know* in conversation. The verb *say* controlling a *that*-clause is also extremely common in written registers such as fiction and news reportage.

In contrast, apart from the verb *want* in conversation, no individual verb is extremely common controlling *to*-clauses. However, there are a large number of different verbs that can control a *to*-clause, and those verbs come from many different semantic domains: mental verbs (e.g., *expect*, *learn*), communication verbs (e.g., *ask*, *promise*), verbs of desire (e.g., *want*, *like*), verbs of decision (e.g., *decide*, *intend*), verbs of effort or facilitation (e.g., *try*, *attempt*, *allow*, *enable*), aspectual verbs (e.g., *begin*, *continue*), and likelihood verbs (e.g., *seem*, *appear*).

These differing patterns of lexical association help to account for the overall differences in register distribution between *that*-clauses and *to*-clauses. Conversational partners tend to use a relatively restricted range of vocabulary, but it is almost always appropriate to report one's own thoughts (*I think that...*, *I know that...*) or the speech of others (*he/she said that...*). Because of the extremely heavy reliance on a few combinations of this type, *that*-clauses are generally very common in conversation.

In contrast, the more frequent use of *to*-clauses in academic prose can be attributed in part to the wide range of different verbs controlling *to*-clauses. That is, academic prose is characterized by a much higher degree of lexical diversity than conversation. Thus, although no single verb is extremely common with *to*-clauses in academic prose, there are a large number of different verbs from different semantic domains used in combination with *to*-clauses. As a result, the overall frequency of *to*-clauses is higher than *that*-clauses in academic prose.

That-clauses and *to*-clauses also differ in their grammatical associations, and these differences also contribute to the overall register association pattern. One reflection of this difference is their use in extraposed versus non-extraposed constructions. Both types of complement clause have the grammatical potential to occur in either extraposed or non-extraposed constructions:

That-clauses

Non-extraposed:

I think that you might be wrong.

Extraposed:

It's possible that it'll happen again.

To-clauses

Non-extraposed:

I want to sleep here.

Extraposed:

It's possible to adjust the limit upwards.

However, *to*-clauses are in fact used much more commonly in extraposed constructions than *that*-clauses, especially in academic prose:

Table 8. Use of ***that****-clauses and* ***to****-clauses in extraposed constructions*

	Conversation	Academic Prose
Extraposed *that*-clauses	**	******
Extraposed *to*-clauses	**	**********

[each * represents 100 occurrences per million words]

It further turns out that well over 80% of the extraposed *to*-clauses in academic prose are controlled by adjectival predicates rather than verbs, as in the following examples:

It is also possible to make more subtle combinations...

It is important to note that it is formed from tissue...

It is therefore essential to insist that true communities must be bare communities as well.

It is hard to resist the temptation...

These grammatical association patterns further explain the overall differences in register distribution between *that*-clauses and *to*-clauses. Extraposed *to*-clauses are by definition impersonal, since they do not have a referential subject. Further, extraposed *to*-clauses controlled by an adjectival predicate are typically used to present a stance that is not directly attributed to any human agent, as in the above examples. These characteristics fit well with the static, impersonal presentation of information typical of academic prose. In contrast, non-extraposed *that*-clauses controlled by verbs more commonly include dynamic predicates attributed to a personal agent or experiencer, and these functions fit well with the typical communicative purposes of conversation.

In sum, both lexical associations and grammatical associations influence the extent to which a grammatical feature is used in different registers, depending on the extent to which those associations fit the typical communicative characteristics of the register. Although patterns such as these must be interpreted much more fully, the present section has illustated the systematicity and importance of these association patterns in describing the use of related grammatical features.

4. Using Textual Co-occurrence Patterns to Analyze Register Variation

Research on discourse and the linguistic characteristics of particular varieties tends to be empirical, based on analysis of some collection of texts. There is a long tradition of such research on "registers", "genres", and "styles", dating from the work of Ferguson, Halliday, Leech, Crystal, and others in the early 1960s. In recent years, most analysts studying registers have begun to use corpus-based techniques; recent edited collections with studies of this kind include Ghadessy (1988) and Biber and Finegan (1994).[7]

In addition to descriptions of a single register, a corpus- based approach enables a variationist perspective. Using computational (semi)automatic techniques to analyze large text corpora, it is possible to investigate the patterns of variation across a large number of registers, with respect to a wide range of relevant linguistic characteristics.

Research into the patterns of register variation is based on a different kind of association pattern: sets of linguistic features that tend to co-occur in texts. In previous studies, I refer to each grouping of linguistic features as a "dimension".

Studies of this kind (e.g., Biber 1988, 1995) have shown that there are systematic patterns of variation among registers; that these patterns can be analyzed in terms of underlying "dimensions" of variation; and that it is necessary to recognize the existence of a multidimensional space in order to capture the overall relations among registers.

In Biber (1988), six major dimensions of variation are identified from a quantitative analysis of the distribution of linguistic features in the LOB and

London-Lund Corpora. Each dimension comprises a distinct set of co-occurring linguistic features; each defines a different set of similarities and differences among spoken and written registers; and each has distinct functional underpinnings. One of the major findings coming out of this study relates to the marking of discourse complexity: contradicting the view that complexity is a homogeneous construct, early multi-dimensional studies showed that complexity features were distributed across several dimensions of variation.

To further investigate the association patterns comprising the dimensions of discourse complexity in English, Biber (1992) used confirmatory factor analysis to study the distribution of 33 linguistic markers of complexity across 23 spoken and written registers. Confirmatory factor analysis is a theory-based statistical approach: different models are hypothesized on theoretical grounds and then compared statistically to determine which best fits the observed patterns of variation. This study showed that discourse complexity is a multi-dimensional construct, that different types of structural elaboration reflect different discourse functions, and that different kinds of texts are complex in different ways (in addition to being more or less complex).

In particular, a five-dimensional model was identified as the most adequate representation of the associations among these complexity features. Each of the dimensions is labeled to reflect its functional and grammatical underpinnings: Reduced Structure and Specificity, Structural Elaboration of Reference, "Framing" Structural Elaboration, Integrated Structure, and Passive Constructions.

To illustrate, Table 9 presents the defining linguistic features for two of the complexity dimensions: "Integrated Structure" (Dimension C) and "Framing Elaboration" (Dimension D). Each of these dimensions represents a text-based association pattern. That is, the groupings of features listed for each dimension represent linguistic characteristics that commonly co-occur in texts.

The linguistic features grouped on Dimension C represent integrated structure. Features such as nouns, prepositional phrases, attributive adjectives, and nominalizations reflect a high informational focus and a relatively dense integration of information in a text; long words and diversified vocabulary (i.e., high type/token ratio) reflect a careful, precise word choice. Together, these features represent a dimension marking integrated structure.

The linguistic features grouped on Dimension D are all dependent clauses that represent "framing elaboration". (Framing dependent clauses should be distinguished from postnominal modifying clauses, which comprise a separate

Table 9. Summary of the linguistic features grouped on Dimension C and Dimension D (based on Biber 1992; Table 5)

Dimension C: Integrated Structure:
nouns
prepositional phrases
attributive adjectives
nominalizations
phrasal coordination
word length
type/token ratio
Dimension D: Framing Elaboration:[8]
WH complement clauses
THAT complement clauses controlled by verbs
conditional adverbial subordination
causative adverbial subordination
sentence relatives
(THAT clauses controlled by adjectives)
(infinitives)
(concessive adverbial subordination)

dimension.) These clause types can be considered "framing" in that they commonly serve one of two major functions: they either provide a discourse frame for a portion of text (as in the case of many types of adverbial subordination; see, e.g., Thompson 1983, 1985; Ford and Thompson 1986); or they provide an overt assessment of the speaker/writer's stance (in the case of sentence relatives, *that* complement clauses, and WH clauses; see Beaman 1984, Quirk et al. 1985, Winter 1982).

Dimensions C and D represent two quite different parameters of discourse complexity, with respect to both their defining linguistic characteristics and their underlying functions. These differences can be studied further by considering the kinds of texts that make extensive use of each complexity dimension.

To compare registers with respect to each of these text-based association patterns, it is necessary to compute "dimension scores" (explained in Biber 1988.93-97). Dimension scores for each text are computed by summing the occurrences of the linguistic features grouped on each dimension; then mean dimension scores for each register can be compared to analyze the salient linguistic similarities and differences among spoken and written registers.

To illustrate register comparisons of this type, Figure 1 presents the differences among twelve spoken and written registers within the two-dimensional space defined by Dimension C and Dimension D. The distribution of scores along the vertical axis represents the Integrated Structure Dimension (C). Registers with large scores along this dimension have frequent occurrences of nouns, prepositional phrases, attributive adjectives, long words, etc. This dimension distinguishes the expository (informational) written registers — which show a very frequent use of integrative features — from all other registers.

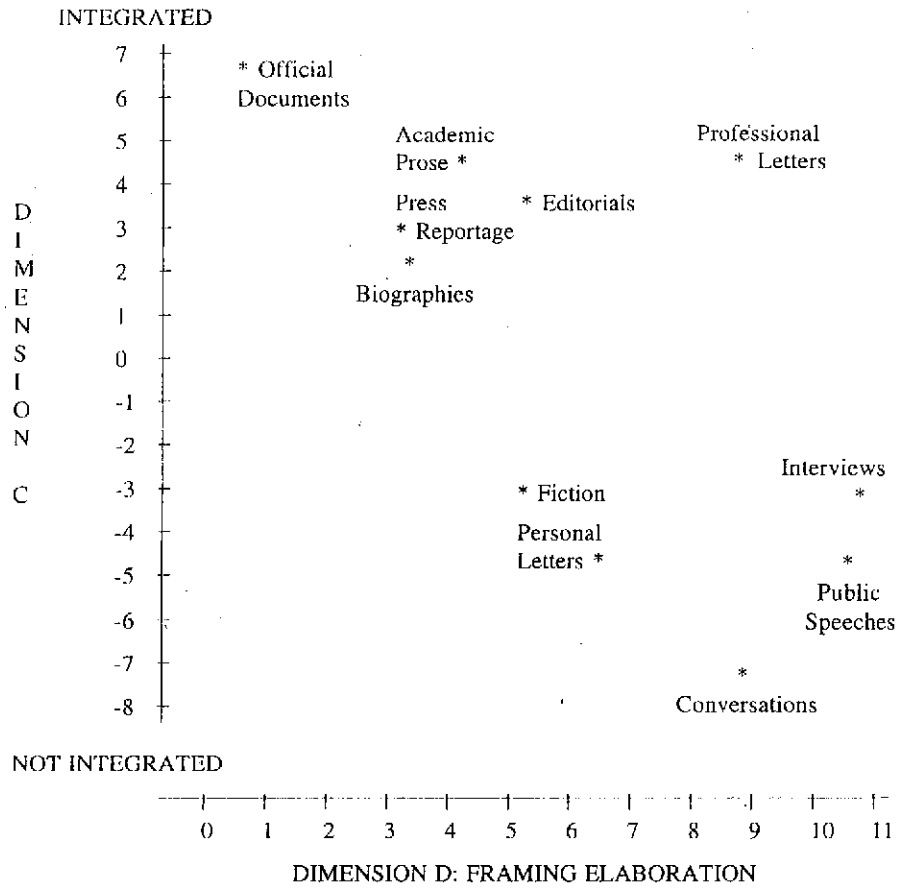

Figure 1. Linguistic characterization of eleven spoken and written registers with respect to Dimension C: "Integrated Structure", and Dimension D: "Framing Elaboration"

The extremely dense use of Integrated Structure features in written informational prose is illustrated by Text Sample 1 (from an official document). These features include frequent nouns, often in noun-noun sequences (e.g., *family income, tax reliefs, family allowances*), attributive adjectives (e.g., *average, incomplete, younger, lower*), and phrasal coordination (e.g., *indices of domestic food expenditure per head* ***and*** *quantities purchased by older* ***and*** *younger couples* ***and*** *families*). The extensive embedding of prepositional phrases is indicated by italics in the text excerpt.

Text Sample 1: Excerpt from official document.
[prepositions are italicized]

Nevertheless, average net family income was appreciably higher *in* families *with* several children than *in* those *with* only one, many *of* which were incomplete families *of* younger parents *with* lower earnings, and of course *with* lower tax reliefs and no family allowances. ... Table 24 gives indices *of* domestic food expenditure *per* head and quantities purchased *by* older and younger couples and families *with* different numbers *of* children, *with* 1954 *as* the base year.

All non-expository registers show an infrequent use of Integrated Structure complexity; these include all spoken registers, as well as fiction and personal letters. With regard to the spoken informational registers (e.g., speeches and interviews), production constraints apparently limit the extent to which information can be carefully integrated. With respect to the non-informational written registers (e.g., fiction and personal letters), the primary communicative purposes do not require a dense integration of information.

The distribution of scores along the horizontal axis in Figure 1 represents the "Framing" Structural Elaboration Dimension. Registers with large scores along this dimension have frequent occurrences of *wh* complement clauses, *that* complement clauses, and various types of adverbial subordination. This type of complexity is most common in those registers that promote the expression of personal attitudes, justifications, and feelings: interviews, public speeches, conversations, and professional letters. Interestingly, many of these registers are spoken and interactive. Framing complexity features are much less common in those registers having informational, impersonal, "factual" purposes, such as press reportage, biographies, and academic prose. These features are especially rare in official documents, which are typically direct statements of "fact" with no acknowledged author (see Text Sample 1 above).

Text Sample 2 illustrates the use of framing elaboration features in a public speech; Sample 3 illustrates these features in a type of interview (court testimony), and Sample 4 illustrates these features in a professional letter. These samples differ in spoken versus written mode, but they share a focus on the expression of personal attitudes and opinions with justification for those positions.

Text samples illustrating Framing Elaboration complexity (Framing Elaboration features are italicized.)

Sample 2: Excerpt from a public speech.

If you look at the steel industry, you see *that* the steel industry, *when* nationalized, was not nationalized on a basis that...

Sample 3: Excerpt from court testimony.

My wife phoned the doctor *when* we arrived, *because* my mother said to my wife will you phone Richard, *because* she wanted *to* go into a nursing home, and the doctor said *[0]* it wasn't necessary.

Sample 4: Professional letter.

...Any such drastic change would ultimately require the action of the board of directors *because* it would involve a change in the constitution...and *because* any such change would in turn require a vote... *If* it is not possible *to* add your concern this year, it would certainly be possible *to* add it next year... Please understand *that while* I am sympathetic to what you are trying to achieve, and *that while* I understand *that* certain XYZ populations are ...impacted..., I am not at present entirely in sympathy...

Several of the framing functions of these elaboration features are illustrated in these text samples. For example, conditional clauses and WH clauses are used to contrast various possible actions or points-of-view (e.g., *If you look at the steel industry, If it is not possible, when nationalized, while I am sympathetic, while I understand*). Causative adverbial clauses are used to justify attitudes or actions (e.g., *because my mother said, because she wanted to go*). Similarly, in the professional letter sample, causative adverbial clauses are used to explain the opinion that: *any such drastic change would ultimately require the action of the board of directors because it would involve a change in the constitution...and because any such change would in turn require a vote*. *That* complement clauses are often used to overtly frame an attitude or position relative to a "stance" verb or adjective (e.g., *you see that...* from the public speech; and *I understand that...* from the professional letter). Thus, although these elaboration features vary in their particular functions, they share general "framing" uses common in more personal, attitudinal registers.

Similar analytical techniques have been used to study the dimensions of register variation in other languages (e.g., Besnier (1988) on Nukulaelae Tuvaluan; Biber and Hared (1992) on Somali; and Kim and Biber (1994) on Korean). All of these studies focus on text-based association patterns — i.e., the linguistic co-occurrence patterns defining the multi-dimensional space of variation among texts and registers in a language. Biber (1995) synthesizes these earlier studies to investigate the possibility of cross-linguistic universals governing the patterns of register variation.[9]

A comparison of text-based association patterns cross-linguistically shows that structurally complex features can serve a number of different functions, associated with both oral and literate registers. There are, however, systematic generalizations concerning the marking of discourse complexity that hold across the four languages compared in Biber (1995 — see especially Chapter 7). For example, relative clauses, and nominal modifiers generally, are characteristic of literate registers, being used for informational elaboration. In contrast, adverbial subordination is used most commonly in oral registers, often to mark some aspect of personal stance; adverbial clauses often co-occur with involved, reduced, or fragmented features. Complement clauses and infinitives occur frequently in both oral and literate registers, but they frequently co-occur with other features marking personal stance or persuasion.

Overall, such comparisons clearly show that it is not adequate to treat structural complexity as an undifferentiated whole. Rather, there are different kinds of complexities having quite different distributions and functional associations. Corpus-based analyses of linguistic co-occurrence patterns — i.e., text-based association patterns — enable register comparisons of this type, resulting in a more adequate understanding of the interacting discourse systems that define the range of register variation in a language.

5. Conclusion

One fruitful area for future research is to integrate the methodologies developed for corpus-based research with those developed for quantitative

studies of sociolinguistic variation. That is, the inter-relations among linguistic association patterns, together with an assessment of their relative importance, could be studied in more detail using variable rule methodologies (see, e.g., Sankoff 1987). Although most sociolinguistic variation studies have been restricted to phonological variants that are semantically equivalent, a number of studies have extended these methods to consider lexical, grammatical, and discourse variables that represent "equivalent-in-discourse" relations (e.g., Sankoff, Thibault, and Bérubé; 1978, Tottie 1991; Dines 1980; Horvath 1985, Helt 1996). These techniques provide probabilistic estimates of the extent to which each contextual factor favors or disfavors a linguistic variant, when considered relative to the influence of other factors.

As the present paper shows, the study of linguistic variability can also be extended to include systematic text-based association patterns. In this case, texts and registers are characterized and compared, rather than the variants for a linguistic feature. For both types of research question (linguistic and text-based), quantitative corpus-based analyses regularly uncover important patterns of use that are highly systematic but often inaccessible to intuitions. Such language use patterns must be interpreted functionally, with respect to a number of inter-related influences, including:

production and processing factors;
communicative purpose and topic;
situational context and interactiveness;
social identity;
textual connectivity.

Recent linguistic theory has generally favored discrete/categorical descriptions over those that allow for continuous/quantitative relations. In large part, this is due to the preconception that individual linguistic competence cannot accommodate systematic tendencies in addition to discrete categories and structures. Further, language competence has been regarded as an independent mental faculty that is not influenced by situational, social, or textual factors.

However, these exclusionary views are not well-grounded: First, there is no empirical evidence suggesting that mental processes cannot involve systematic tendencies. Second, there is no a priori reason to suppose that mental competencies concerning situational/social appropriateness or textual connectivity should not interact with linguistic production. Finally, neither the categorical formalisms found in generative linguistics nor the specific probabilities identified in variation studies are likely to have any direct representation in actual mental processes. However, it is reasonable to suppose that both types of description correspond to aspects of linguistic competence (cf. Sankoff 1988).

Obviously, future research is required to investigate the relative importance of use factors and the ways in which particular functional considerations relate to particular kinds of association patterns. The goals of the present paper have been more modest: to set out a framework for describing the various kinds of association patterns and to illustrate the highly systematic nature of each type.

Notes

1. A carefully designed, representative corpus is crucial for studies of this type. Some projects have used extremely large corpora, with relatively little consideration for the kinds of texts included; other projects have used a very careful corpus design (regarding the kinds of text) but relatively small sample size. Both types of skewing are likely to influence research findings. That is, a representative corpus must pay equal attention to both composition and size. (See Biber 1990, 1993; Leech 1991; Fries, Tottie, and Schneider 1994 for more detailed discussions of corpus design issues.)

2. Similar research goals have been investigated in sociolinguistic variation studies, and variable rules can be regarded as a formal statement of association patterns for "equivalent" variants (see, e.g., the discussions in Sankoff and Labov 1979; Sankoff 1987, 1988; and several papers in Sankoff 1978).

3. The grammatical associations for these two adjectives are determined from automatic analysis using a grammatical tagger; these counts were confirmed and adjusted slightly based on interactive analysis of 200 randomly selected tokens for each adjective.

4. These percentages are based on interactive analysis of 200 randomly selected tokens for each verb.

5. This pattern is attested for the verb *tell*, as in:
   You can *tell* she's from London. (Conv)

6. Although rare, these verbs can take a *to*-clause as well as a *that*-clause. This pattern is most often found when the matrix verb is in the passive voice; for example,
   The follow-up action can be taken if the initial response *is thought* to be unsatisfactory.
   Volvo *is known* to be keen to strengthen its manufacturing base.
   The deal *was said* to enable LTCB to gain information and knowledge in international asset management.

7. Within computational linguistics, research on "sublanguages" uses corpus-based analyses to address many of these same issues, with the ultimate goal of automatically processing texts from particular varieties with a high degree of accuracy (see Grishman and Kittredge 1986, Kittredge and Lehrberger 1982).

8. The linguistic features listed in parentheses on Dimension D do not have strong positive loadings on this dimension. Two other features — present participle adverbial clauses and other adverbial subordination — had negative loadings on this dimension, indicating that they do not function as hypothesized.

9. Multi-dimensional register comparisons have also been used to study diachronic register variation (e.g., Biber and Finegan 1989, Atkinson 1992).

References

Aarts, B. and C. Meyer (eds). 1995. *The Verb in Contemporary English: Theory and Description*. Cambridge: Cambridge University Press.

Aijmer, K. and B. Altenberg (eds). 1991. *English Corpus Linguistics*. London: Longman.

Altenberg, B. 1991a. "A Bibliography of Publications Relating to English Computer Corpora". In S. Johansson and A-B. Stenström (eds). *English Computer Corpora: Selected Papers and Research Guide*, 355-95. Berlin: Mouton.

Altenberg, B. 1991b. "Amplifier Collocations in Spoken English". In S. Johansson and A-B. Stenström (eds). *English Computer Corpora: Selected Papers and Research Guide*, 127-147. Berlin: Mouton.

Altenberg, B. 1993. "Recurrent Verb-complement Constructions in the London-Lund Corpus". In J. Aarts, N. Oostdijk, and P. de Haan (eds). *English Language Corpora: Design, Analysis, and Exploitation*, 227-46. Amsterdam: Rodopi.

Armstrong, S. (ed.). 1994. *Using Large Corpora*. Cambridge, MA: MIT Press.

Atkinson, D. 1992. "The Evolution of Medical Research Writing from 1735 to 1985: The Case of the Edinburgh Medical Journal". *Applied Linguistics* 13: 337-374.

Beaman, K. 1984. "Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse". In D. Tannen (ed.). *Coherence in Spoken and Written Discourse*. Norwood, N.J.: Ablex.

Besnier, N. 1988. "The Linguistic Relationships of Spoken and Written Nukulaelae Registers". *Language* 64:707-736.

Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. 1990. "Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation". *Literary and Linguistic Computing* 5:257-269.

Biber, D. 1992. "On the Complexity of Discourse Complexity: A Multidimensional Analysis". *Discourse Processes* 15:133-163.

Biber, D. 1993a. "Co-occurrence Patterns Among Collocations: A Tool for Corpus-based Lexical Knowledge Acquisition". *Computational Linguistics* 19:549-556.

Biber, D. 1993b. "Representativeness in Corpus Design". *Literary and Linguistic Computing* 8:1-15.

Biber, D. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D., S. Conrad, and R. Reppen. 1994. "Corpus-based Approaches to Issues in Applied Linguistics". *Applied Linguistics* 15:169-189.

Biber, D. and E. Finegan. 1989. "Drift and the Evolution of English Style: A History of Three Genres". *Language* 65:487-517.

Biber, D. and E. Finegan (eds). 1994. *Sociolinguistic Perspectives on Register*. New York: Oxford University Press.

Biber, D. and M. Hared. 1992. "Dimensions of Register Variation in Somali". *Language Variation and Change* 4:41-75.

Church, K. and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics* 16:22-29.

Church, K., W. Gale, P. Hanks and D. Hindle. 1991. "Using Statistics in Lexical Analysis". In U. Zernik (ed.). *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, 115-164. Hillsdale, NJ: Lawrence Erlbaum.

Collins, P. 1991. *Cleft and Pseudo-cleft Constructions in English*. London: Routledge.

de Haan, P. 1989. *Postmodifying Clauses in the English Noun Phrase: A Corpus-Based Study*. Amsterdam: Rodopi.

Dines, E.R. 1980. "Variation in Discourse and Stuff Like That". *Language in Society* 9:13-31.

Firth, J.R. 1952. "Linguistic Analysis as a Study of Meaning". In Palmer (ed.), 12-26.

Ford, C.E., and S. Thompson. 1986. "Conditionals in Discourse: A Text-based Study from English". In E. Traugott, C. Ferguson, J. Snitzer Reilly, and A. ter Meulen (eds). *On Conditionals*. Cambridge: Cambridge University Press.

Fries, U., G. Tottie and P. Schneider (eds). 1994. *Creating and Using English Corpora*. Amsterdam: Rodopi.

Geisler, C. 1995. *Relative Infinitives in English*. Uppsala: Uppsala University.

Ghadessy, M. (ed.). 1988. *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter.

Granger, S. 1983. *The Be+Past Participle Construction in Spoken English with Special Emphasis on the Passive*. Amsterdam: Elsevier Science Publishers.

Grishman, R. and R. Kittredge (eds). 1986. *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Hillsdale, NJ: Lawrence Erlbaum.

Helt, M. 1996. "A Corpus-based Approach to Defining Discourse Markers". Ms. Northern Arizona University.

Horvath, B. 1985. *Variation in Australian English*. Cambridge: Cambridge University Press.

Johansson, C. 1995. *The Relativizers Whose and of Which in Present-day English: Description and Theory*. Uppsala: University of Uppsala.

Johansson, S. and A-B. Stenström (eds). 1991. *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton.

Kennedy, G. 1991. "Between and Through: The Company They Keep and the Functions They Serve". In K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*, 95-127. London: Longman.

Kim, Y. and D. Biber. 1994. "A Corpus-based Analysis of Register Variation in Korean".

In D. Biber and E. Finegan (eds). *Sociolinguistic Perspectives on Register*, 157-181. New York: Oxford University Press.
Kittredge, R. and J. Lehrberger. 1982. *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: DeGruyter.
Kjellmer, G. 1991. "A mint of phrases". In K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*, 111-127. London: Longman.
Leech, G. 1991. "The State of the Art in Corpus Linguistics". In K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*, 8-29. London: Longman.
Mair, C. 1990. *Infinitival Complement Clauses in English*. New York: Cambridge University Press.
Meyer, C. 1992. *Apposition in Contemporary English*. Cambridge: Cambridge University Press.
Nakamua, J., and J. Sinclair. 1995. "The World of *Woman* in the Bank of English: Internal Criteria for the Classification of Corpora". *Literary and Linguistic Computing* 10:99-110.
Oostdijk, N. and P. de Haan (eds). 1994. *Corpus-based Research into Language*. Amsterdam: Rodopi.
Palmer, F.R. (ed.). 1968. *Selected Papers of J.R. Firth, 1952-59*. Bloomington: Indiana University Press.
Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Renouf, A. and J. Sinclair. 1991. "Collocational Frameworks in English". In K. Aijmer and B. Altenberg (eds). *English Corpus Linguistics*, 128-143. London: Longman.
Sankoff, D. (ed.). 1978. *Linguistic Variation: Models and Methods*. New York: Academic Press.
Sankoff, D. 1987. "Variable Rules". In U. Ammon, N. Dittmar, and K. Mattheier (eds). *Sociolinguistics: An International Handbook of the Science of Language and Society*, 984-997. Berlin: de Gruyter.
Sankoff, D. 1988. "Sociolinguistics and Syntactic Variation". In F. J. Newmeyer (ed.). *Linguistics: The Cambridge Survey*, (Vol. IV), 140-161. Cambridge: Cambridge University Press.
Sankoff, D. and W. Labov. 1979. "On the Uses of Variable Rules". *Language in Society* 8:189-222.
Sankoff, D., P. Thibault and H. Bérubé. 1978. "Semantic Field Variability". In D. Sankoff (ed.). *Linguistic Variation: Models and Methods*, 23-43. New York: Academic Press.
Sinclair, J. (ed.) 1987. *Looking Up*. London: Collins.
Sinclair, J. 1991. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.
Stubbs, M. 1995. "Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies". *Functions of Language* 2:23-55.
Svartvik, J. (ed.) 1990. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press.
Svartvik, J. (ed.) 1992. *Directions in Corpus Linguistics: Proceedings from the Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin: Mouton.
Thompson, S.A. 1983. "Grammar and Discourse: The English Detached Participial Clause". In F. Klein-Andreu (ed.). *Discourse Perspectives on Syntax*, 43-65. New York: Academic Press.
Thompson, S.A. 1985. "Grammar and Written Discourse: Initial Versus Final Purpose Clauses in English": *Text* 5:55-84.
Tottie, G. 1991. *Negation in English Speech and Writing: A Study in Variation*. San Diego: Academic Press.
Varantola, K. 1984. *On Noun Phrase Structures in Engineering English*. Turku: University of Turku.
Winter, E. 1982. *Towards a Contextual Grammar of English: The Clause and Its Place in the Definition of Sentence*. London: George Allen and Unwin.