The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings

Author(s): Douglas Biber

Source: *Computers and the Humanities*, Vol. 26, No. 5/6, Common Methodologies in Humanities Computing and Computational Linguistics (Dec., 1992), pp. 331-345
Published by: Springer
Stable URL: https://www.jstor.org/stable/30204629
Accessed: 01-06-2020 07:44 UTC

The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings

Douglas Biber

Department of English, Northern Arizona University, Flagstaff, AZ 86011-6032, USA
e-mail: biber@nauvax.ucc.nau.edu

Abstract: The present paper summarizes the major methods and results of the multi-dimensional approach to genre variation. The approach combines the resources of computational tools, large text corpora, and multivariate statistical tools (such as factor analysis and cluster analysis). It has been used to address issues such as the relations among spoken and written genres in English, and the historical development of genres and styles. The approach has also been applied to other languages; in this regard it has been used to address broader theoretical issues, such as the extent to which genre and style variation are comparable cross-linguistically, and the linguistic consequences of literacy.

Key Words: linguistic variation, genre, style, register, dimension, factor analysis, cluster analysis, historical change, English, Somali

1. Introduction

Computer-based text corpora and computer programs for automated language processing are complementary research tools, and their combined use enables linguistic investigations of a type not otherwise feasible. For example, this combination has been used for stylistic investigations, since it enables analysis of a large number of linguistic features across many texts and text types (see, e.g., Francis and Kucera, 1982; Johansson and Hofland, 1989; Oakman, 1975; Ross, 1973). These tools have also been combined for lexicographic research; for example, the COBUILD Dictionary (Sinclair, 1987), which is based on a large text corpus, has been so successful that a number of other publishers and research teams are pursuing lexicographic projects based on computer-based corpora and automated computational techniques (see Walker, 1989). Similarly with respect to research on natural language understanding systems, researchers such as Garside, Leech and Sampson (1987) have combined corpus-based and automated-computational approaches to develop natural language understanding systems that are more robust than previously achieved.

The "multi-dimensional" approach to linguistic variation, described in the present paper, similarly exploits the research potential of automated computational analyses and large computer-based text corpora; however, it adds the power of multivariate statistical techniques to investigate the quantitative distribution of linguistic features across texts and text varieties. This approach has been used to address several sociolinguistic and stylistic issues relating to the lexical and syntactic characteristics of texts and text varieties, and the nature of genre, register and dialect variation.

In the present paper I provide a methodological overview of this approach and outline some of the major linguistic findings that have resulted to date.

Douglas Biber is an associate professor in the Department of English, Applied Linguistics Program, Northern Arizona University, Flagstaff, AZ. His research deals with the linguistic variation among registers and the diachronic evolution of registers, addressing both theoretical concerns and methodological issues relating to the design and use of computer-based text corpora.

Computers and the Humanities **26**: 331–345, 1993.

Section 2 presents the basic model of variation that has been developed for English, based on analysis of a large number of spoken and written genres; it also summarizes the major computational and statistical techniques used in the approach. Section 3 summarizes several analyses of more specialized text genres with respect to this model. Section 4 then presents various statistical extensions to the model, again summarizing the major linguistic findings resulting in each case. Finally, Section 5 summarizes the application of these techniques to corpus-based analyses of linguistic variation in other languages.

2. The Multi-Dimensional Approach and Basic Model of Variation

2.1. *General characteristics of the multi-dimensional approach*

The Multi-Dimensional (MD) approach to genre variation (elsewhere referred to as the Multi-feature/Multi-dimension approach) was first outlined in Biber (1986) and then developed more fully in Biber (1988). The MD approach has several salient characteristics:

1. It is corpus based, depending on analysis of a large number of naturally-occurring texts.
2. It is computer-based in that it depends on automated analyses of linguistic features in texts. This characteristic enables distributional analysis of many linguistic features across many texts and text varieties.
3. The research goal of the approach is the linguistic analysis of texts, genres, text types, styles or registers, rather than analysis of individual linguistic constructions.
4. The importance of variationist and comparative perspectives is assumed by the approach. That is, the approach is based on the assumption that different kinds of text differ linguistically and functionally, so that analysis of any one or two text varieties is not adequate for conclusions concerning a discourse domain (e.g., speech and writing in English).
5. The approach is explicitly multi-dimensional. That is, it is assumed that multiple parameters of variation will be operative in any discourse domain.
6. It is quantitative. Analyses are based on frequency counts of linguistic features, and multivariate statistical techniques are used to analyze the relations among linguistic features and among texts. (The major statistical techniques used in this approach were first used in text studies by researchers such as Carroll, 1960; and Marckworth and Baker, 1974.)
7. It synthesizes quantitative and functional approaches. That is, the statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each distributional pattern. The approach is based on the assumption that statistical co-occurrence patterns reflect underlying shared communicative functions.
8. It synthesizes macroscopic and microscopic approaches. That is, macroscopic investigations of the overall parameters of linguistic variation, which are based on analysis of the distribution of many linguistic features across many texts and genres, are complemented by detailed analyses of particular features in particular texts.

The notion of linguistic co-occurrence is central to linguistic analyses of genres and text types. Brown and Fraser (1979, pp. 38—39) emphasize the importance of this notion, observing that it can be "misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers." Ervin-Tripp (1972) and Hymes (1974) define "speech styles" as varieties that are defined by a shared set of co-occurring linguistic features. In the series of studies using the MD approach, I have used the term "genre" (or "register") for text varieties that are readily recognized and "named" within a culture (e.g., letters, press editorials, sermons, conversation), while I have used the term "text type" for varieties that are defined linguistically (rather than perceptually). Both genres and text types can be characterized by reference to co-occurring linguistic features, but text types are further defined quantitatively such that the texts in a type all share frequent use of the same set of co-occurring linguistic features. Because co-occurrence reflects shared function, the resulting types are coherent in their linguistic form and communicative functions.

In the MD approach, linguistic co-occurrence is

analyzed in terms of underlying "dimensions" of variation. There are three distinctive characteristics of the notion of "dimension." First, no single dimension is adequate in itself to account for the range of linguistic variation in a language; rather, a multidimensional analysis is required. Second, dimensions are continuous scales of variation rather than dichotomous distinctions. Third, the co-occurrence patterns underlying dimensions are identified quantitatively (by a statistical factor analysis) rather than on an a priori functional basis.

Dimensions have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., passives, nominalizations, prepositional phrases) that co-occur with a markedly high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these patterns are interpreted in terms of the situational, social and cognitive functions most widely shared by the co-occurring linguistic features.

2.2. *Overview of computational methods*

Computational tools have been used in the MD approach to: "tag" the words in texts for various lexical, grammatical and syntactic categories; compile frequency counts of linguistic features within texts; and analyze the distribution of linguistic features within and across texts.

The grammatical tagging used here is richer than that used in the Brown and LOB Corpora, in that it marks both word class and syntactic information. Sixteen grammatical and functional classes of linguistic features are automatically identified on the basis of these tags:

1. tense and aspect markers
2. place and time adverbials
3. pronouns and pro-verbs
4. questions
5. nominal forms
6. passives
7. stative forms
8. subordination features
9. prepositional phrases, adjectives and adverbs
10. lexical specificity
11. lexical classes
12. modals
13. specialized verb classes
14. reduced forms and discontinuous structures
15. coordination
16. negation

Early versions of the tagging program used a large-scale dictionary together with a number of context-dependent disambiguating algorithms. The dictionary was compiled from a sorted version of the Brown Corpus; it contained 50,624 lexical entries from the four major grammatical categories of noun, verb, adjective and adverb. Grammatical ambiguities (e.g., *move* as noun and verb) were identified by the multiple entries in the dictionary. Biber (1988, Appendix II) provides a fuller description of this version of the tagging program and the algorithms used to identify each linguistic feature.

More recently, I have developed a tagging program that runs on desktop computers as part of a project to analyze diachronic genre variation in English.[1] This program also uses an on-line dictionary derived from a sorted version of a previously tagged corpus (in this case, the LOB Corpus). However, while the earlier program depended exclusively on a series of context-sensitive algorithms to resolve grammatical ambiguities, the more recent program exploits probabilistic information, following the approach developed for the CLAWS tagging system by Garside, Leech and Sampson (1987). This probabilistic approach has been extended in several ways, based on distributional analyses of the LOB Corpus. First, overall probabilities for the grammatical categories of each lexical item are used in the present project (CLAWS simply "scaled-down" the likelihood of rare tags). For example, *book* and *runs* are both noun-verb ambiguities, but *book* has a very high likelihood of being a noun (99% in the LOB expository genres), while *runs* has a high likelihood of being a verb (74%). Second, probabilities of tag pairs are computed differently depending on which members of the pair are ambiguous. In CLAWS, all tag sequence probabilities are computed as:

$$\text{freq}(A \text{ followed by } B) / \text{freq}(A) * \text{freq}(B)$$

In the present project, this formula is being used for the case where both A and B are ambiguous forms (e.g., *abstract forms*). In the case where A is not ambiguous, however, the probability of A

followed by *B* can be estimated with much higher certainty, as:

freq(*A* followed by *B*) / freq(*A*)

and when *B* is not ambiguous, the probability of *A* followed by *B* (or the probability of *B* preceded by *A*) can be estimated as:

freq(*A* followed by *B*) / freq(*B*)

Finally, in the present project I have compiled separate dictionaries for exposition and fiction, to reflect the differing lexical and grammatical preferences of the two, and to test the extent to which consideration of genre variation aids in automated grammatical tagging. Table 1 presents probabilities for several common lexical items that have markedly different distributions in fiction and exposition. These include both content words and function words. For example, many past participial forms (e.g., *admitted*, *expected*) are much more likely to mark past tense or perfect aspect in fiction and much more likely to mark passive constructions in exposition. Many noun/verb ambiguities (e.g., *trust*, *rule*, *-ing* forms) are much more likely to occur as verbs in fiction and as nouns in exposition. Among the function words, some preposition/subordinator ambiguities are more likely to occur as subordinators in fiction (e.g., *before*, *as*) than in exposition.

TABLE 1
Examples of marked differences between fiction and exposition in the grammatical functions of common words (based on their distribution in the LOB Corpus).

word	grammatical category	Fiction %	Exposition %
admitted	past tense	77	24
	passives	17	66
	perfects	6	0
	adjectives	0	7
remembered	past tense	89	20
	passives	2	72
expected	past tense	54	8
	passives	7	77
	perfects	34	6
known	passives	28	65
	perfects	77	13
	adjectives	6	15
trust	noun	18	85
	verb	82	15
rule	noun	31	91
	verb	69	9
thinking	noun	7	56
	verb	92	41
major	titular noun	69	11
	adjective	31	85
before	preposition	30	54
	subordinator	48	32
as	preposition	21	41
	subordinator	61	40
that	demonstrative	37	17
	complementizer	45	69
	relative pronoun	14	11

The tagging programs are being used at present to analyze the texts in a large historical corpus of English. When running the programs, the operator must specify whether the input texts are more similar to fiction or exposition, to choose the appropriate dictionary files. Although the programs have generally small error rates, we are post-editing texts to insure accuracy.

Based on the tagged texts, other programs tally frequency counts for each text, normalizing the counts to their frequency per 1,000 words. These frequencies form the basis for subsequent statistical and linguistic analyses, described in the following sections.

2.3. *Five basic dimensions of variation in English; analysis of spoken and written genres*

In Biber (1988) five major dimensions of variation are identified for English.[2] The dimensions represent the co-occurrence distributions of 67 linguistic features across 481 spoken and written texts of contemporary British English. The texts, which are taken from the Lancaster-Oslo-Bergen Corpus (Johansson, *et al.*, 1978) and the London-Lund Corpus (Svartvik, 1990), represent 23 major genre categories (e.g., academic prose, press reportage, fiction, letters, conversations, interviews, radio broadcasts, public speeches).

Factor analysis is used to identify the groups of linguistic features associated with each dimension; these are the sets of linguistic features that co-occur in texts with markedly high frequencies. The interpretation of the factors as functional "dimensions" is based on the assumption that co-occurrence reflects shared function; that is, features

co-occur frequently in texts because they serve some shared, underlying communicative functions associated with the situational contexts of the texts.

Table 2 summarizes the co-occurring features associated with each of the five dimensions. The decimal numbers on this table represent the factor "loadings" for each linguistic feature. Loadings can run from −1.0 to +1.0; the further from 0.0 a loading is, the more one can generalize from the factor in question to the particular linguistic feature. Features with higher loadings are thus better representatives of the dimension underlying a factor. In Table 2, only features with loadings larger than 0.35 (plus or minus) are included.

Most of the dimensions consist of two groupings of features, having positive and negative loadings. Positive or negative sign does not indicate a more-or-less relationship; rather, these two groups represent sets of features that occur in a complementary pattern. That is, when the features in one group occur together frequently in a text, the features in the other group are markedly less frequent in that text, and vice versa. To interpret the dimensions, it is important to consider likely reasons for the complementary distribution of these two groups of features as well as the reasons for the co-occurrence pattern within each group.[3]

For example, consider Dimension 2. The features in the top group (the positive loadings above the dashed line on Table 2) are past tense verbs, perfect aspect verbs, third person pronouns and public verbs (primarily speech act verbs), while the features in the bottom group (the negative loadings) are present tense verbs and adjectives. Considering all of the features on Dimension 2, this dimension is interpreted as distinguishing narrative discourse from other types of discourse,

TABLE 2
Summary of the co-occurrence patterns underlying five major dimensions of English.

DIMENSION 1 (Informational vs. Involved)		DIMENSION 2 (Narrative versus Non-Narrative)		DIMENSION 3 (Elaborated vs. Situated Reference)		DIMENSION 4 (Overt Expression of Persuasion)		DIMENSION 5 (Abstract versus Non-Abstract Style)	
nouns	0.80	past tense verbs	0.90	WH relative clauses on object positions	0.63	infinitives	0.76	conjuncts	0.48
word length	0.58	third person pronouns	0.73	pied piping constructions	0.61	prediction modals	0.54	agentless passives	0.43
prepositional phrases	0.54	perfect aspect verbs	0.48	WH relative clauses on subject position	0.45	suasive verbs	0.49	past participial clauses	0.42
type / token ratio	0.54	public verbs	0.43	phrasal coordination	0.36	conditional subordination	0.47	BY-passives	0.41
attributive adjs.	0.47	synthetic negation	0.40	nominalizations	0.36	necessity modals	0.46	past participial WHIZ deletions	0.40
private verbs	−0.96	present participial clauses	0.39	time adverbials	−0.60	split auxiliaries	0.44	other adverbial subordinators	0.39
that deletions	−0.91	present tense verbs	−0.47	place adverbials	−0.49	possibility modals	0.37	[No complementary features]	
contractions	−0.90	attributive adjs.	−0.41	other adverbs	−0.46	[No complementary features]			
present tense verbs	−0.86								
2nd person pronouns	−0.86								
do as pro-verb	−0.82								
analytic negation	−0.78								
demonstrative pronouns	−0.76								
general emphatics	−0.74								
first person pronouns	−0.74								
pronoun *it*	−0.71								
be as main verb	−0.71								
causative subordination	−0.66								
discourse particles	−0.66								
indefinite pronouns	−0.62								
general hedges	−0.58								
amplifiers	−0.56								
sentence relatives	−0.55								
WH questions	−0.52								
possibility modals	−0.50								
non-phrasal coordination	−0.48								
WH clauses	−0.47								
final prepositions	−0.43								

suggesting the interpretive label "Narrative versus Non-narrative Concerns." Narrative concerns are marked by considerable reference to past time, third person animate referents, and reported speech (public verbs), together with the relative absence of present tense verbs and adjectives. Non-narrative concerns, whether expository, descriptive or other, are marked by immediate time (present tense) and attributive nominal elaboration, together with the relative absence of past tense verbs, third person pronouns, etc.

The complementary groupings on the other factors (shown in Table 2) reflect similar functional relations. To represent these communicative function(s), the dimensions are interpretively labeled as follows:

I: Involved versus Informational Production
II: Narrative versus Non-narrative Concerns
III: Elaborated versus Situation-Dependent Reference
IV: Overt Expression of Persuasion
V: Abstract versus Non-abstract Style

Biber (1988, Chapters 6—7) provides detailed justification for these interpretations based on the shared communicative functions of the co-occurring linguistic features on each dimension plus the distribution of registers along each dimension.

Although identification of the dimensions is theoretically important in itself, in that it isolates several of the basic parameters of text variation in English, the primary use of the dimensions is to analyze the linguistic characteristics of texts and text varieties. This can be achieved by computing "dimension scores" (or "factor scores") for each text: a summation of the frequencies for those features having salient loadings on a dimension.[4] For example, the Dimension 3 score for each text can be computed by adding together the frequencies of WH-relative clauses on object positions, pied-piping constructions, WH-relative clauses on subject positions, phrasal coordinators and nominalizations — the top group of features (on Table 2) — and then subtracting the frequencies of time adverbials, place adverbials and other adverbs — the bottom group of features.

A dimension score for each dimension is computed for each text; then, the mean of each dimension score for each genre is computed. Consideration of these dimension scores enables linguistic characterization of any given text or genre, comparison of the relations between any two genres, and a fuller functional interpretation of the underlying dimension.

Consider Table 3, which presents the mean dimension scores of fifteen spoken and written English genres for Dimension 1, "Informational versus Involved Production." The genres with large negative values (such as conversation) have high frequencies of present tense verbs, private verbs, first and second person pronouns, contractions, etc. (the bottom group of features on Dimension 1), together with markedly low frequencies of nouns, prepositional phrases, long words, etc. (the top group of features on Dimension 1). Genres with large positive values (such as official documents and academic prose) have very high frequencies of nouns, prepositional phrases, etc., plus very low frequencies of private verbs, contractions, etc. The characterizations of genres shown in Table 3 confirm the interpretation of Dimension 1 as distinguishing among texts according to the demands and possibilities of informational production. Conversational texts are largely interactive and involved, since participants typically do not have time for highly informational production, nor are they inclined to

TABLE 3
Dimension 1 scores for 15 spoken and written genres ("Involved versus Informational Production").

Genre	Mean	Std. Dev.
Official Documents	18.1	4.8
Press Reportage	15.1	4.5
Academic Prose	14.9	6.0
Biographies	12.4	7.5
Press Editorials	10.0	3.8
Broadcasts	4.3	10.7
Professional Letters	3.9	13.7
General Fiction	0.8	9.2
Prepared Speeches	−2.2	6.7
Romance Fiction	−4.3	5.6
Interviews	−17.1	10.7
Spontaneous Speeches	−18.2	12.3
Personal Letters	−19.5	5.4
Face-to-face Conv.	−35.3	9.1
Telephone Conversation	−37.2	9.9

$F = 111.9$; $df = 22,458$; $r^2 = 84.3\%$

highly informational purposes. Genres such as spontaneous speeches are intermediate because they are relatively informational in purpose, but participants are constrained by on-line production. Finally genres such as academic prose are extremely informational in purpose and produced under highly controlled and edited circumstances.

Considering all five dimensions together enables multi-dimensional analyses of the linguistic characteristics of particular genres and the linguistic differences among genres. Figure 1 plots the mean dimension scores for six spoken and written genres (conversations, spontaneous speeches, personal letters, professional letters, general fiction and official documents). For example, the leftmost side of the figure plots the mean Dimension 1 scores given in Table 3. (These scores are divided by 5.0 to put them on a scale comparable to the other four dimensions.)

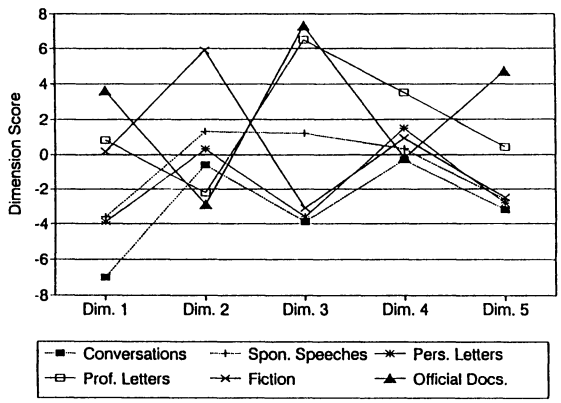

Figure 1. Five-dimensional plot of six spoken and written genres.

The figure shows that conversations tend to have scores near the bottom of each scale: markedly "involved" (Dimension 1), "situated" (Dimension 3), "non-persuasive" (Dimension 4), and "non-abstract" (Dimension 5). On Dimension 2, though, conversations have an intermediate score, reflecting the use of both narrative and non-narrative strategies. Official documents have the opposite characteristics with respect to being "informational" (Dimension 1), markedly "non-narrative" (Dimension 2), "elaborated" (Dimension 3), and "abstract" (Dimension 5), but they have the same characterization as conversations on Dimension 4 (being "non-persuasive"). Professional letters are relatively similar to official documents; they are slightly less "informational" on Dimension 1, but they differ primarily in being markedly "persuasive" (Dimension 4) and relatively "non-abstract" (Dimension 5). Fiction texts are similar to professional letters on Dimension 1, but these two genres are quite different on all other dimensions. In fact, fiction is quite similar to conversation in being "situated" (Dimension 3), "non-persuasive" (Dimension 4), and "non-abstract" (Dimension 5). It can be seen that the characterization of each genre is relatively complex, reflecting the particular purposes and situational characteristics of the variety. Similarly, comparisons among genres must be multi-dimensional. Figure 1 illustrates the fact that there are no simple overall differences between speech and writing, and that it is not adequate to claim simple overall similarities or differences between any two genres. Rather, depending on their functional correlates, genres can be markedly similar on some dimensions while markedly different on others.

3. Analyses of More Specialized Genres with Respect to the Five-Dimensional Framework

The five-dimensional framework described in the last section can be used to characterize and compare texts in a number of more specialized discourse domains. For example, Biber (1987) compares nine British and American written genres along underlying dimensions (actually using an earlier three-dimensional framework developed in Biber, 1986). The study found that American written genres consistently permit a more colloquial, interactive style than corresponding British genres, but at the same time, American written genres are consistently more abstract and nominal than corresponding British genres.

Grabe (1987) uses a multi-dimensional analysis to investigate the parameters of variation among expository prose genres. This study analyzed the distribution of 33 linguistic features across expository genres from six topical areas (e.g., natural science, social science, humanities) and three target audiences (academic, introductory university level and popular). Four major dimensions are identified and used to analyze the relations among

expository genres. Some of these dimensions are closely associated with audience-level differences, while others distinguish among the topical classes, suggesting that both of these situational parameters are important linguistically.

Connor-Linton (1988) uses the five-dimensional framework of Biber (1988) to highlight differences among four authors writing about nuclear weapons. The observed linguistic differences are interpreted in terms of different perspectives on the (in)appropriate use of nuclear weapons as well as differing overall world views.

Biber (1991) also uses the 1988 framework to analyze the linguistic characteristics of selected primary school reading materials. The study shows that there are marked differences among primary school texts, and that overall these texts are notably different from adult genres. For example, the second grade science text analyzed in that study was markedly involved and interactive (on Dimension 1) and situated (on Dimension 3), in contrast to science texts from more advanced levels. In fact, the second grade science text was considerably more involved and interactive than the second grade basal reader, raising serious questions about the adequacy of that text as a model of scientific literacy.

In a larger study, Biber and Finegan (1989a) use the 1988 framework to analyze the historical development of English genres from the late 17th century to the present, focusing on three genres: essays, fiction and personal letters. The study shows that there has been a steady "drift" to more "oral" styles in all three genres, as measured by changes towards more "involved" styles (on Dimension 1), more "situated" styles (on Dimension 3), and less "abstract" styles (on Dimension 5). The study further found a reaction to this general pattern of drift in the 18th century, when there was an extreme range of variation among the texts within genres (i.e. some texts were markedly "oral" while others were extremely "literate"), and the average genre characterizations actually became somewhat more "literate." These patterns are interpreted by reference to conscious and unconscious attitudes towards style in the various periods, together with larger developments relating to standardization, popular literacy and mass education.

Biber and Finegan (1992) extend this diachronic analysis to include two speech-based genres: dialogue from novels and drama. Figure 2 illustrates the findings of these two studies, plotting the changes in these five genres with respect to the "elaborated versus situated reference" dimension (Dimension 3).[5] The figure shows the general drift to more "situated" styles in all five genres as well as the reaction to include more "elaborated" styles in the 18th century.

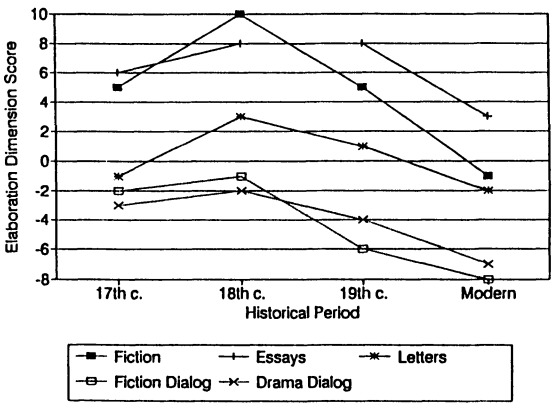

Figure 2. Plot of the elaboration dim. for five genres across four periods.

In addition, this framework can be used to analyze the linguistic characteristics of authors' "styles" (Biber and Finegan, 1991). Following Leech and Short (1981) this analysis is based on the assumption that styles are best analyzed by relative comparison to other texts. Given this position, one of the fundamental problems for stylistic research involves identification of an adequate frame of comparison. A multi-dimensional analysis provides such a frame, since it permits analysis of a particular text relative to a broad range of other texts — within the same period and genre, across periods and genres, and across multiple linguistic dimensions. For example, Figure 3 plots the range of Dimension 5 scores, reflecting "abstract versus non-abstract style," for the 51 essays and 33 fiction texts analyzed in Biber and Finegan (1989a). The figure shows the spread of scores within each period, the overall movement towards less abstract scores in more recent periods (similar to the development for the "elaboration" dimension, shown in Figure

Key to texts:
Defoe:
  Essays
    "Some Thoughts on . . . Commerce", 1713, # 1
    "Essays on Projects: Of Academies", 1698, # 2
  Fiction
    "Life of Robinson Crusoe" (Chap. 1), 1719, # 1
    "Moll Flanders" (Chap. 1), 1722, # 2

Johnson:
  Essays
    "On Fiction" (Rambler # 4), 1750, # 1
    "Preface to Dictionary of the English Language", 1755, # 2
  Fiction
    "The History of Rasselas: Prince of Abyssinia", 1759
      narrative sample, # 1
      descriptive sample, # 2

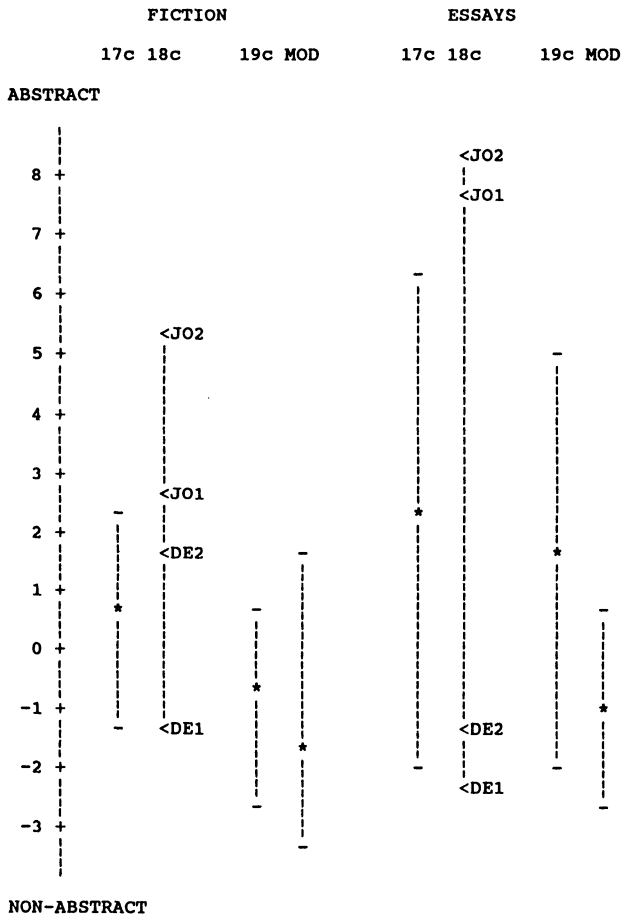

Figure 3. Characterization of texts from Defoe and Johnson along Dimension 5 ("Abstract versus Non-abstract Style").

2), and the extreme range of variation characterizing the 18th century. Against this backdrop, the dimension scores of eight texts from two 18th century authors are plotted: two fiction texts and two essays from Defoe and Johnson. This dimension shows a striking contrast between these two authors as represented by these particular text samples. The difference is especially marked with respect to the essays, but it is also noteworthy between the fiction samples — the Johnson texts, in both fiction and essays, tend to be markedly "abstract" in style, while the Defoe samples tend to be markedly "non-abstract." Similar characterizations can be made with respect to the other dimensions, and with respect to other text samples and authors. Such analyses can provide a solid linguistic foundation for the stylistic characterization of texts and authors, which can be complemented by detailed analysis of the salient linguistic features in particular texts.

4. Statistical Extensions to Investigate Other Linguistic Issues

As noted above, I have distinguished between "genres" and "text types" in this research program. Genres have a perceptual basis in a given culture, but they are not necessarily linguistically coherent. For example, adult speakers of English would presumably have no difficulty recognizing "academic articles" as a text category of English and therefore as a "genre."[6] However, Biber (1988, pp. 170—98) shows that this particular category has a wide range of internal variation linguistically. That is, academic articles can range from philosophical and literary essays to reports of scientific experiments, and these various sub-genres encompass a wide range of linguistic variation along each dimension.

Text types, on the other hand, are identified on linguistic grounds, so that the texts included in a type are linguistically coherent by definition. In Biber (1989) eight text types of English are identified on the basis of similarities along the five dimensions, by a statistical procedure known as cluster analysis. This procedure groups texts such that the texts within each cluster are maximally similar with respect to their dimension scores, while each cluster is maximally distinct from the others.

The "text types" identified in Biber (1989) do not correspond neatly to a priori expectations. For example, there is no single interactive or dialogue text type. Rather, the analysis identifies two major interactive types: Intimate Interpersonal Interaction (Type 1), concerned primarily with immediate interactions, and Informational Interaction (Type 2), which has an informational emphasis. Similarly, there is no single expository text type. Instead, the analysis identifies three expository types: Scientific Exposition (Type 3), which is extremely informational, elaborated in reference, and technical and abstract in style; Learned Exposition (Type 4), which is similar to Scientific Exposition except that it is markedly less abstract and less technical in style; and General Narrative Exposition (Type 6), which is a very general text type that combines narrative forms with expository, informational elaboration. In the same way, there is no single narrative text type. Instead, the analysis identifies General Narrative Exposition (Type 6) and Imaginative Narrative (Type 5), which is a relatively involved text type having a primary narrative focus.

Cluster analyses have also been used to identify the "stance styles" of English by Biber and Finegan (1988, 1989b). "Stance" in these studies refers to the lexical and grammatical expression of attitudes, feelings, judgments or commitment concerning the propositional content of a message. The 1988 study considered only adverbs marking stance, while the 1989 study extended the analysis to include 12 linguistic classes of stance, such as: adverbs, verbs and adjectives marking certainty; adverbs, verbs and adjectives marking doubt; hedges and emphatics; and positive and negative affect markers. This latter study identifes six stance styles, which are given interpretive labels such as "Emphatic Expression of Affect," "Expository Expression of Doubt," and "Faceless."

A second statistical extension involves the use of confirmatory factor analysis (and the LISREL statistical package). The analyses in Biber (1986, 1988) are based on exploratory factor analyses, which are appropriate when little is known about the underlying constructs in a discourse domain. Subsequent research, though, can profitably use confirmatory factor analysis, which enables hypothesis testing concerning the relative strengths of competing discourse models. Biber (1992b) uses this technique to test various models of discourse complexity. The study identifies a particular five-dimensional model as the most adequate with respect to the 33 surface linguistic markers of complexity analyzed. Labels are proposed for each dimension to reflect their functional and grammatical underpinnings; for example: "Structural Elaboration of Reference," " 'Framing' Structural Elaboration," and "Integrated Structure." Analysis of the relative complexities of spoken and written genres with respect to these dimensions identifies a fundamental distinction between the discourse complexities of written and spoken genres: while written genres exhibit many complexity profiles, differing widely in both the extent and the kinds of complexity, spoken genres manifest a single major pattern differing only in extent.

5. Multi-Dimensional Analyses of Other Languages

The first study to use a multi-dimensional approach for the investigation of genre variation in a non-western language is Besnier's (1988) analysis of Nukulaelae Tuvaluan (an Austronesian language spoken on a small atoll in the Central Pacific). This is a restricted literacy situation, with only two written genres (personal letters and sermon notes). Besnier's study identifies three primary dimensions of variation in Tuvaluan, interpretively labeled as:

1. Attitudinal versus Authoritative Discourse
2. Focus on Information versus Interaction
3. Rhetorical Manipulation versus Structural Complexity

Kim and Biber (1993) also use a multi-dimensional analysis to investigate the variation among 22 spoken and written genres in Korean, identifying five major dimensions:

1. Informational Interaction versus Explicit Elaboration
2. Discourse Chaining versus Discourse Fragmentation
3. Stance
4. Narrative Concern
5. Honorification

More recently, Biber and Hared (1991, 1993) use the multi-dimensional approach to analyze the patterns of variation among Somali spoken and written genres, from both synchronic and diachronic perspectives.[7] The recent history of Somali (a Cushitic language spoken by 5—6 million people in East Africa) makes it an ideal arena for the investigation of these issues: although it is currently a national language, used for a large number of professional purposes, it has existed as a written language only since 1972. Over that short time span, written varieties in government, education, entertainment and the press have evolved in Somali. These multidimensional analyses of Somali are thus able to trace the historical genesis and evolution of written varieties, in addition to providing a synchronic description of the relations among spoken and written genres.

For the analyses of Somali, a text corpus of 556 texts (c. 600,000 words) was assembled and tagged for 65 lexical, grammatical and syntactic characteristics. The size and scope of this corpus make it one of the largest and most comprehensive computer-based collections of tagged texts for a non-western language. The corpus includes 436 written texts from 22 written genres, collected from three time periods: 1973—75, 1978—82, 1987—89. The corpus also includes 120 texts (c. 150,000 words) from 10 spoken genres.

Since there were no pre-existing materials available for Somali, texts were tagged using an interactive program and an on-line dictionary. As new words were encountered in texts, they were automatically entered in the dictionary for future reference.

Five major dimensions of variation have been identified by the multi-dimensional analysis of Somali:

1. "Structural elaboration: Involvement versus exposition"
2. "Lexical elaboration: On-line versus planned/ integrated production"
3. "Argumentative versus reported presentation of information"
4. "Narrative versus non-narrative discourse organization"
5. "Distanced, directive interaction"

As with the MD studies of other languages, each of these dimensions comprises a distinct set of co-occurring linguistic features, and each defines a different set of relations among spoken and written genres.

The multi-dimensional studies of Somali are important for two main reasons. First, they provide a detailed description of genre (or register) variation in a non-western language, laying the foundation for cross-linguistic generalizations. Second, they enable an empirical investigation of the linguistic conequences of literacy, addressing fundamental issues such as: 1) What linguistic changes accompany the initial introduction of written genres in a language? and 2) How do written genres evolve relative to spoken genres in the early history of literacy in a language?

Cross-linguistic comparisons enable investigation of the extent to which there are universal patterns of variation among styles and genres. There are some general conclusions that hold across all four languages analyzed to date (English, Nukulaelae Tuvaluan, Korean and Somali): the linguistic relations among spoken and written genres are quite complex; a multi-dimensional analysis is required, because no single dimension by itself adequately captures the similarities and differences among genres; there is no absolute dichotomy between speech and writing (rather, situational factors such as purpose, topic and interactiveness work together with the physical mode distinction to define the salient linguistic differences among genres).

There are also certain specific cross-linguistic generalizations that hold across these four languages. The most notable is that all of these languages have one or more "oral/literate" dimensions. These do not define absolute differences between speech and writing, but they are associated with stereotypical spoken and written genres, and they are defined linguistically by interactive/involved features versus features reflecting structural/lexical elaboration and complexity. In addition, Somali, English and Korean all have a "narrative" dimension, defined by past tense and temporal features, and distinguishing fiction and traditional stories from other genres.

Regarding the linguistic influence of literacy, Biber and Hared (1991) show that the introduction of Somali written genres in 1973 greatly

TABLE 4
One dimension from the analysis of spoken and written registers in Somali.

DIMENSION 2 "Lexical elaboration"	
Positive features with strong loadings:	Less important positive features:
hapax legomena	demonstrative relatives
type-token ratio	clitic topic-coordination
nominalizations	gerunds
compound verbs	purpose clauses
case particles	word length
No Negative Features	

extended the pre-existing range of variation (i.e. among spoken genres) along several dimensions. Table 4 presents the defining linguistic features for one of those dimensions, interpreted as reflecting "lexical elaboration"; and Figure 4 shows the striking increase in the range of variation along this dimension due to the introduction of written genres. All spoken genres (in capital letters) are relatively unelaborated lexically, while written genres (underlined) in 1973 range from relatively unelaborated (e.g., press announcements) to the extremely elaborated style found in press commentary articles.

Interestingly, this range of variation continued to be extended over the early history of literacy in Somali, as written genres evolved to become even more distinct from spoken genres. Biber and Hared (1993) analyze the early development of six press genres with respect to several dimensions of variation. These developmental patterns can be illustrated through plots of three individual linguistic characteristics: T-unit length, reflecting a type of structural elaboration, in Figure 5; and nominalizations versus first person pronouns, reflecting aspects of reference and lexical elaboration, in Figure 6.

T-unit length measures the average length, in number of words, of main clauses and their associated dependent clauses. Figure 5 shows marked differences among genres with respect to this linguistic feature. For example, press reportage, editorials, and book introductions have quite long t-units, while novels and personal letters have

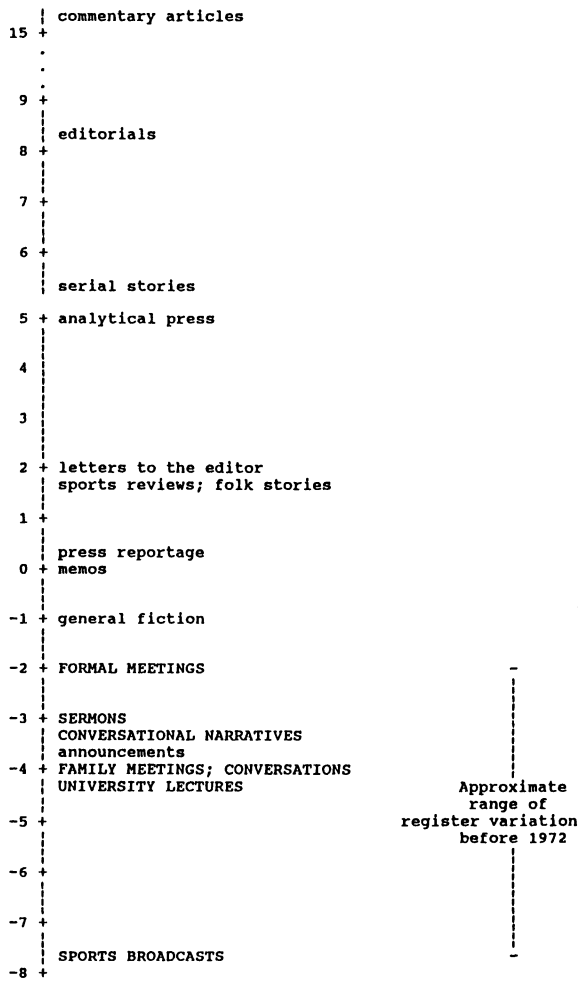

Figure 4. Distribution of Somali registers along the Lexical Elaboration Dimension, showing the extension in the range of variation due to the addition of written registers.

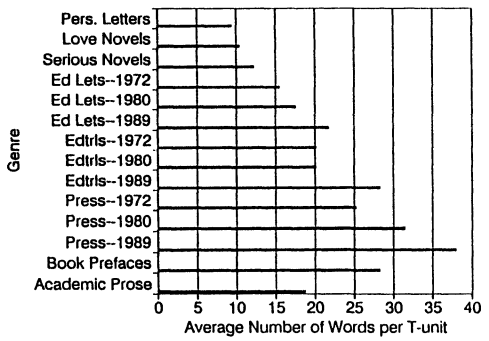

Figure 5. Plot of T-unit length for 8 Somali genres (+3 periods).

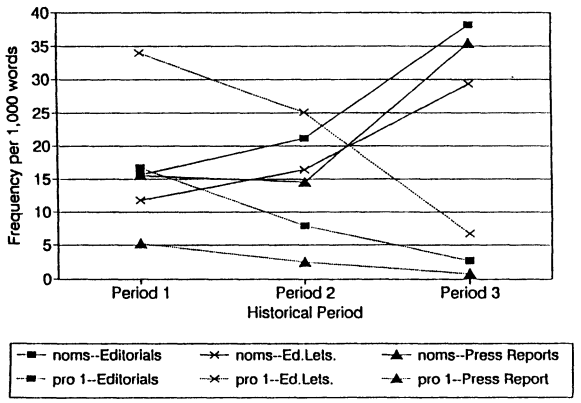

Figure 6. Plot of linguistic features for 3 Somali genres across 3 periods.

notably short t-units. The figure also plots the change in t-unit length over time for editorial letters, institutional editorials, and press reportage. All three of these press genres show the same pattern of a steady increase in t-unit length over time, reflecting a steady development towards more structural elaboration and greater differentiation from typical spoken genres.

Figure 6 shows similar diachronic patterns: a steady increase over time in the use of nominalizations and a steady decrease in the use of first person pronouns. (The solid lines show the frequency of nominalizations, while the dotted lines plot the frequency of first person pronouns.) As with t-unit length, the direction of these changes represents a continuous trend away from the linguistic characteristics of typical spoken genres.

The multi-dimensional analyses presented in Biber and Hared (1991, 1993) show similar patterns of change with respect to two "oral/literate" dimensions ("Structural elaboration: Involvement versus exposition" and "Lexical elaboration: On-line versus planned production"). These developments are interpreted as reflecting increased awareness of the differing purposes and potentials of written registers. In addition, the analyses show a quite different pattern of change with respect to a third dimension ("Argumentative versus reported presentation of information"). This latter trend is interpreted as reflecting changes in the intended purposes of certain press registers, perhaps tied to changes in the political scene of Somalia and the possibility for genuine "argumentative" expression.

In sum, multi-dimensional studies of non-western languages are being used to provide insights into some of the basic issues of humanities and social science research, such as the extent to which patterns of style and genre are shared cross-linguistically, and the linguistic consequences of literacy (relating to both the initial genesis of written varieties and their subsequent evolution).

6. Summary and Conclusion

The present paper provides an overview of the main methodological procedures and theoretical results of the multi-dimensional approach to genre variation. This approach is characterized by distributional analysis of a wide range of linguistic features across many texts and text varieties, the use of computer programs to automatically identify linguistic features in texts, and the use of multivariate statistical techniques to analyze the co-occurrence patterns among linguistic features, identifying underlying linguistic and textual constructs. The paper has illustrated the application of this approach to a number of textual domains and to genre variation in other languages. Further work within this framework is needed in several areas, including: analyses of additional kinds of linguistic features (e.g., cohesion markers, rhetorical patterns) to identify other dimensions of variation in English, analyses of the patterns of variation in additional discourse domains and languages, and a synthesis of cross-domain and cross-linguistic studies into a single coherent model of genre variation.

Notes

[1] Support for this project has been provided by the National Science Foundation, grant #BNS-9010893. As part of the project, we are compiling and tagging a corpus of c. 3 million words composed of texts from 15 written and speech-based genres across the last four centuries.

[2] In addition, two less important dimensions are identified and given prelimiary interpretations in that study. Biber (1986) presents an earlier analysis based on fewer linguistic features, in which three primary dimensions are identified. Although the five dimensions discussed in this section have been replicated in several studies, they are not intended to be exhaustive. For example, Biber (1992a) analyzes the distribution of linguistic features marking cohesion, given and new information, and informational packaging; that study tentatively identifies two additional dimensions based on these linguistic characteristics. It is likely that other dimensions will be identified as additional linguistic characteristics are analyzed.

[3] Although the co-occurrence patterns underlying the dimensions are identified empirically through factor analysis, the interpretations depend on a researcher's assessment of shared function, based on: 1) previous research studies, 2) further statistical analysis of the distribution of co-occurring features across texts and genres, and 3) further micro-analyses of particular features in particular texts. It is possible, however, to empirically test these interpretations through confirmatory factor analyses; Section 4 discusses an example of a confirmatory analysis that compares the relative strengths of several hypothesized models of discourse complexity.

[4] All frequencies are standardized to a mean of 0.0 and a standard deviation of 1.0 before the dimension scores are computed (see Biber 1988.93—95). The polarity of Dimension 1 has been reversed here from that presented in Biber (1988), to highlight the textual parallels across dimensions.

[5] Overall, 163 texts were analyzed for this study.

[6] Faigley and Meyer (1983) is one of the few studies that have investigated the salient text categories of English from a perceptual perspective.

[7] Support for research on the investigation of genre variation in Somali has been provided by the National Science Foundation, grant #BNS-8811720.

References

Aitken, A.J., R.W. Bailey, and N. Hamilton-Smith, eds. *The Computer and Literary Studies.* Edinburgh: Edinburgh University Press, 1973.

Besnier, Niko. "The Linguistic Relationships of Spoken and Written Nukulaelae Registers." *Language*, 64 (1988), 707—36.

Biber, Douglas. "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings." *Language*, 62 (1986), 384—414.

Biber, Douglas. "A Textual Comparison of British and American Writing." *American Speech*, 62 (1987), 99—119.

Biber, Douglas. *Variation across Speech and Writing.* Cambridge: Cambridge University Press, 1988.

Biber, Douglas. "A Typology of English Texts." *Linguistics*, 27 (1989), 3—43.

Biber, Douglas. "Oral and Literate Characteristics of Selected Primary School Reading Materials." *Text*, 11 (1991), 73—96.

Biber, Douglas. "Using Computer-based Text Corpora to Analyze the Referential Strategies of Spoken and Written Texts." In *Directions in Corpus Linguistics*, (Proceedings of the Nobel Symposium on Corpus Linguistics, Stockholm, Sweden, August 1991). Ed. Jan Svartvik. Mouton, 1992a, 213—52.

Biber, Douglas. "On the Complexity of Discourse Complexity: A Multidimensional Analysis." *Discourse Processes*, 15 (1992b), 133—63.

Biber, Douglas, and Edward Finegan. "Adverbial Stance Types in English." *Discourse Processes*, 11 (1988), 1—34.

Biber, Douglas, and Edward Finegan. "Drift and the Evolution of English Style: A History of Three Genres." *Language*, 65 (1989a), 487—517.

Biber, Douglas, and Edward Finegan. "Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect." *Text*, 9 (1989b), 93—124.

Biber, Douglas, and Edward Finegan. "The Linguistic Evolution of Five Written and Speech-based English Genres from the 17th to the 20th Centuries." In *History of Englishes: New Methods and Interpretations in Historical Linguistics.* Ed. M. Rissanen, O. Ihalainen, T. Nevalainen, and I. Taavitsainen. Mouton, 1992, 688—704.

Biber, Douglas, and Edward Finegan. "Multi-dimensional Analyses of Author's style: Some Case Studies from the 18th century." To appear in *Research in Humanities Computing*, 1991.

Biber, Douglas, and Mohamed Hared. "Literacy in Somali: Linguistic Consequences." *Annual Review of Applied Linguistics*, 12 (1991), 260—82.

Biber, Douglas, and Mohamed Hared. Linguistic Correlates of the Transition to Literacy in Somali: Language Adaptation in Six Press Registers. In *Sociolinguistic Perspectives on Register.* Ed. D. Biber and E. Finegan. 1993. (In press, OUP.)

Brown, Penelope, and Colin Fraser. "Speech as a Marker of Situation." In *Social Markers in Speech.* Ed. Klaus R. Scherer and Howard Giles. Cambridge: Cambridge University Press, 1979, pp. 33—62.

Carroll, J.B. "Vectors of Prose Style." In *Style in Language.* Ed. T.A. Sebeok. Cambridge, MA: MIT Press, 1960, pp. 283—92.

Connor-Linton, Jeff. "Author's Style and World-view in Nuclear Discourse: A Quantitative Analysis." *Multilingua*, 7 (1988), 95—132.

Ervin-Tripp, Susan M. "On Sociolinguistic Rules: Alternation and Co-occurrence." In *Directions in Sociolinguistics.* Ed. John J. Gumperz and Dell Hymes. New York: Holt, Rinehart, and Winston, 1972, pp. 213—50.

Faigley, Lester, and Paul Meyer. "Rhetorical Theory and Readers' Classifications of Text Types." *Text*, 3 (1983), 305—25.

Francis, W.N., and H. Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin, 1982.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson. *The Computational Analysis of English: A Corpus-based Approach.* London: Longman, 1987.

Grabe, William. "Contrastive Rhetoric and Text-type Research." In *Writing across Languages: Analysis of L2 Text.* Ed. Ulla Connor and Robert B. Kaplan. Reading, MA.: Addison-Wesley, 1987, 115—37.

Hymes, Dell. *Foundations in Sociolinguistics: An Ethnographic Approach.* Philadelphia: University of Pennsylvania Press, 1974.

Johansson, Stig (with G.N. Leech and H. Goodluck). Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Department of English, University of Oslo, 1978.

Johansson, Stig, and Knut Hofland. *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus.* Oxford: Clarendon Press, 1989.

Kim, Yong-Jin, and Douglas Biber. "A Corpus-based Analysis of Register Variation in Korean." In *Sociolinguistic Perspectives on Register.* Ed. D. Biber and E. Finegan. 1993. (In press, OUP.)

Leech, Geoffrey N., and Michael H. Short. *Style in Fiction: A Linguistic Introduction to English Fictional Prose.* London: Longman, 1981.

Marckworth, M.L., and Baker, W.J. "A Discriminant Function Analysis of Co-variation of a Number of Syntactic Devices in Five Prose Genres." *American Journal of Computational Linguistics,* Microfiche 11 (1974).

Oakman, R.L. "Carlyle and the Machine: A Quantitative Analysis of Syntax in Prose Style." *ALLC Bulletin,* 3 (1975), 100—14.

Ross, Donald. "Beyond the Concordance: Algorithms for Description of English Clauses and Phrases." In Aitken *et al.,* 1973, pp. 85—99.

Sinclair, John M., ed. *Looking Up: An Account of the COBUILD Project in Lexical Computing.* Glasgow: Collins, 1987.

Svartvik, Jan, ed. *The London-Lund Corpus of Spoken English: Description and Research.* Lund: Lund University Press, 1990.

Walker, Donald E. "Developing Lexical Resources." *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary.* Waterloo, Ontario: University of Waterloo Centre for the New Oxford English Dictionary, 1989.