Using Register-Diversified Corpora for General Language Studies

Douglas Biber*
Northern Arizona University

The present study summarizes corpus-based research on linguistic characteristics from several different structural levels, in English as well as other languages, showing that register variation is inherent in natural language. It further argues that, due to the importance and systematicity of the linguistic differences among registers, diversified corpora representing a broad range of register variation are required as the basis for general language studies.

First, the extent of cross-register differences are illustrated from consideration of individual grammatical and lexical features; these register differences are also important for probabilistic part-of-speech taggers and syntactic parsers, because the probabilities associated with grammatically ambiguous forms are often markedly different across registers. Then, corpus-based multidimensional analyses of English are summarized, showing that linguistic features from several structural levels function together as underlying dimensions of variation, with each dimension defining a different set of linguistic relations among registers. Finally, the paper discusses how such analyses, based on register-diversified corpora, can be used to address two current issues in computational linguistics: the automatic classification of texts into register categories and cross-linguistic comparisons of register variation.

1. Introduction

As the use of computer-based text corpora has become increasingly important for research in natural language processing, lexicography, and descriptive linguistics, issues relating to corpus design have also assumed central importance. Two main considerations are important here: 1) the size of the corpus (including the length and number of text samples), and 2) the range of text categories (or *registers*) that samples are selected from.[1] Within social science, these considerations are associated with the two main kinds of error that can threaten 'external validity' (the extent to which it is possible to generalize from a sample to a larger target population): 'random error' and 'bias error.' Random error occurs when a sample is not large enough to accurately estimate the true population; bias error occurs when the selection of a sample is systematically

* Department of English, Northern Arizona University, P. O. Box 6032, Flagstaff, AZ 86011-6032; biber@nauvax.bitnet.

1 Corpus designs can differ along several other parameters, including:

1. bounded/static versus unbounded/dynamic;
2. richly encoded versus minimally encoded (e.g., grammatical tagging, phonological/prosodic encoding, tagging of social characteristics (of participants) and situational characteristics);
3. complete texts versus samples from texts;
4. selection of texts: convenience versus purposeful versus random within strata versus proportional random.

different from the target population it is intended to represent. Both kinds of error must be minimized to achieve a representative corpus.

Recent debates concerning the design of general purpose corpora have often divided into two opposing camps emphasizing one or the other source of error: those advocating large corpora versus those advocating "balanced" corpora (i.e., including a wide range of registers). For example, the ACL Data Collection Initiative (DCI) and the Linguistic Data Consortium are focusing on the rapid collection and dissemination of very large corpora, with relatively little attention to the range of registers; older corpora, such as the Brown Corpus and LOB Corpus, are small by present-day standards but are explicitly structured to 'represent a wide range of styles and varieties' (Francis and Kučera 1964 [cf. Johansson, Leech, and Goodluck 1978]). Projects such as the COBUILD Corpus, Longman/Lancaster Corpus, and British National Corpus (BNC) combine both emphases to varying extents.

Although all of these corpora could be considered "representative" of at least some varieties of English, it is important to address the question of whether the varieties represented match the intended uses of a corpus. For example, studies of a single sublanguage are legitimately based on corpora representing only that variety, such as journal articles on lipoprotein kinetics (Sager 1986), Navy telegraphic messages (Fitzpatrick et al. 1986), weather reports (Lehrberger 1982), and aviation maintenance manuals (Kittredge 1982).

One of the main issues addressed here, though, is whether general language studies must be based on a corpus that is register-diversified as well as large. Some proponents of very large corpora have suggested that size can compensate for a lack of diversity—that if a corpus is large enough, it will represent the range of linguistic patterns in a language, even though it represents only a restricted range of registers. Given this assumption, linguistic analyses of any very large corpus could be generalized to the entire language.

In contrast, I argue here that analyses must be based on a diversified corpus representing a wide range of registers in order to be appropriately generalized to the language as a whole, as in a dictionary or grammar of English, or a general purpose tagging program for English.[2] In fact, global generalizations are often not accurate at all, because there is no adequate overall linguistic characterization of the entire language; rather, there are marked linguistic differences across registers (or sublanguages; cf. Kittredge 1982). Thus a complete description of the language often entails a composite analysis of features as they function in various registers. Such analyses must be based on corpora representing the range of registers.

In the following discussion, I first briefly illustrate the extent of cross-register differences from consideration of individual grammatical and lexical features. Section 2.1 focuses on the marked differences in the distribution of dependent clauses across registers. This section then shows that register differences are also important for probabilistic part-of-speech taggers and syntactic parsers, because the probabilities associated with grammatically ambiguous forms are often markedly different across registers. Section 2.2 focuses on adjectives marking 'certainty' to illustrate how lexical patterns are also distributed differently across registers.

Section 3 makes this point more strongly by describing a multidimensional analysis of numerous linguistic features in a register-diversified corpus of English. The analysis shows that linguistic features from all levels function together as underly-

2 In other papers (Biber 1990, in press, a), I have explored issues of representativeness relating to corpus size.

Table 1
Mean frequencies of three dependent clause types (per 1,000 words) in four registers (from Biber 1988, Appendix III).

Register	Relative Clauses	Causative Adverbial Subordinate Clauses	*that* Complement Clauses
Press reports	4.6	.5	3.4
Official documents	8.6	.1	1.6
Conversations	2.9	3.5	4.1
Prepared speeches	7.9	1.6	7.6

ing dimensions of variation, and that there are systematic and important linguistic differences among registers with respect to these dimensions. The extent of those differences clearly shows that linguistic analyses based on a restricted corpus cannot be generalized to the language as a whole.

Section 4, then, shows how multidimensional analyses of register variation can be used to address additional computational issues. First, Section 4.1 discusses the application of the multidimensional model of English to predict the register category of texts automatically with a high degree of accuracy. The predictive power of the model at three levels of abstraction is tested. In Section 4.2, then, I turn to cross-linguistic patterns of register variation, illustrating how multidimensional analyses of register-diversified corpora in English, Nukulaelae Tuvaluan, Korean, and Somali enable register comparisons of a kind not otherwise possible, providing the background for more detailed investigations of particular subregisters or sublanguages.

2. Particular Grammatical and Lexical Features

The analyses in this section illustrate the fact that there are systematic and important grammatical and lexical differences among the registers of English; Section 2.1 treats grammatical features and Section 2.2 discusses lexical features.

2.1 Grammatical Issues

2.1.1 Descriptive Analyses. One of the main uses of general text corpora has been to provide grammatical descriptions of particular linguistic features, such as nominal premodification structures, relative clauses, verb and particle combinations, and clefts and pseudoclefts (see the numerous entries in the bibliography of corpus-based studies compiled by Altenberg [1991]). Two findings repeatedly come out of this literature: first, individual linguistic features are distributed differently across registers, and second, the same (or similar) linguistic features can have different functions in different registers.

The linguistic description of dependent clauses in English illustrates these patterns. Although these constructions are often treated as a single coherent system, the various types of structural dependency actually have quite different distributions and functions in English (cf. Biber 1988, 1992). For example, Table 1 shows that relative clauses are quite frequent in official documents and prepared speeches but quite rare in conversation. In contrast, causative adverbial subordination occurs most frequently in conversation and is quite rare in official documents and press reports. Finally, *that* complement clauses occur most frequently in prepared speeches, and with moderate frequencies in conversations and press reports, but they are rare in official documents. There is further variation within these structural categories. For example, most rel-

ative clauses in official documents have WH rather than *that* relative pronouns (7.7 out of 8.6 total), while the two types are evenly split in conversation (1.6 WH relative clauses versus 1.3 *that* relatives). Although causative adverbial clauses are generally more frequent in spoken registers, clauses headed by *as/since* occur almost exclusively in writing (although they are relatively rare in both modes); clauses headed by *because* are much more frequent in speech (Tottie 1986).

Biber (1992) uses confirmatory factor analysis, building on corpus-based frequency counts of this type, to show that discourse complexity is itself a multi-dimensional construct; different types of structural elaboration reflect different discourse functions, and different registers are complex in different ways (in addition to being more or less complex). The analysis further identifies a fundamental distinction between the discourse complexities of written and spoken registers: written registers exhibit many complexity profiles, differing widely in both the extent and the kinds of complexity, while spoken registers manifest a single major pattern differing only in extent.

2.1.2 Probabilistic Tagging and Parsing. The differing distributions illustrated in Section 2.1.1 have important implications for probabilistic tagging and parsing techniques, which depend on accurate estimates of the relative likelihood of grammatical categories in particular contexts. Two kinds of probabilistic information are commonly used in part-of-speech taggers: 1) for ambiguous lexical items, the relative probability of each grammatical category (e.g., *abstract* as a noun, adjective, and verb); and 2) for groups of ambiguous words, the relative probability of various tag sequences (e.g., the likelihood of a noun being followed by a verb, adjective, or another noun).

To investigate whether grammatically ambiguous words have different distributions across registers, I compiled two separate on-line dictionaries from the LOB Corpus: one based on the expository registers, and one from the fiction registers. Table 2 presents descriptive statistics from a comparison of these two dictionaries. The first observation is that many words occurred in only one of the dictionaries. This is not surprising in the case of exposition, since there were over twice as many lexical entries in the exposition dictionary. However, it is more surprising that there were over 6000 words that occurred only in fiction. These included many common words, such as *cheek, kissed, grandpa, sofa, wallet, briefcase, intently,* and *impatiently.*

A comparison of the probabilities of words occurring in both dictionaries is even more revealing. One thousand ten words had probability differences greater than 50%, while another nine hundred eighty words had probability differences greater than 30%. These words represented many lexical types. For example, the first group of words listed on Table 2 are past participle forms. There is a strong likelihood that the *-ed* forms (i.e., *admitted, observed, remembered, expected*) will function as past tense verbs in fiction (probabilities of 77%, 91%, 89%, and 54%) but as passive verbs in exposition (probabilities of 67%, 45%, 72%, and 77%). The word *observed* shows a slightly different pattern in that it has a relatively high probability of occurring as an adjective in exposition (33%). In fiction, the *-ed* forms never occur as adjectives, apart from the 4% likelihood for *remembered.* Finally, the two true participles (*known* and *given*) are most likely to occur as perfect aspect verbs in fiction (65% and 77% likelihood), but they are similar to the *-ed* forms in typically occurring as passive verbs in exposition (65% and 71%).

The words in the second group on Table 2 represent noun/verb/adjective ambiguities. These include noun/verb ambiguities (*trust, rule*), *-ing* participles (*thinking, breathing*), verb/adjective ambiguities (*secure*), and noun/adjective ambiguities (*major, representative*). Apart from the last two types, these forms all show a strong likelihood to function as verbs in fiction (62% to 92% probability). In contrast, these forms are

Table 2
Comparison of the probabilities of grammatically ambiguous words in a dictionary based on Exposition versus a dictionary based on Fiction. (Both dictionaries are derived from the one million words of the LOB Corpus.)

Overall Statistics:

Total lexical entries in the Fiction dictionary = 22,043
Total lexical entries in the Expository dictionary = 50,549

Total words occurring in the Fiction dictionary only: 6,204
Total words occurring in the Expository dictionary only: 31,476

Words having probability differences of > 50%: 1,010
Words having probability differences of > 30%: 980

Examples of marked differences in probabilities for common words:
(Note: The probabilities for some words do not add up to 100% because minor categories are not listed.)

Word	Grammatical Category	Fiction %	Exposition %
Past Participle	Forms		
admitted	past tense	77	24
	passives	17	67
	perfects	6	0
	adjectives	0	9
observed	past tense	91	22
	passives	9	45
	adjectives	0	33
remembered	past tense	89	20
	passives	2	72
	perfects	4	0
	adjectives	4	4
expected	past tense	54	8
	passives	7	77
	perfects	34	6
	adjectives	0	8
known	passives	26	65
	perfects	65	13
	adjectives	6	15
given	passives	19	71
	perfects	77	13
	adjectives	0	9

more likely to occur as nouns in exposition (all greater than 80% likelihood except for *thinking*). *secure* shows a similar likelihood of having an adjectival function in exposition (80%), and the two noun/titular noun/adjective ambiguities are much more likely to occur as adjectives in exposition than fiction.

The third group of ambiguous forms are function words. The probability differences here are less large than in the other two groups, but they are still important given the central grammatical role that these items serve. The first three of these words (*until, before, as*) are considerably more likely to occur as a subordinator in fiction than in exposition, while they are more likely to occur as a preposition in exposition. The

Table 2
Continued.

Word	Grammatical Category	Fiction %	Exposition %
Noun/Verb/Adjective Ambiguities			
trust	noun	18	85
	verb	82	15
rule	noun	31	91
	verb	69	9
thinking	noun	7	56
	verb	92	41
breathing	noun	33	85
	verb	62	15
	adjective	5	0
secure	verb	82	20
	adjective	18	80
major	titular noun	69	11
	adjective	31	85
representative	titular noun	0	7
	noun	100	45
	adjective	0	48
Function Word Ambiguities			
until	preposition	19	38
	subordinator	81	62
before	preposition	30	54
	subordinator	48	32
	adverb	22	14
as	preposition	21	41
	subordinator	61	40
that	demonstrative	37	17
	complementizer	45	69
	relative pronoun	14	11

word *that* is quite complex. It has roughly the same likelihood of occurring as a relative pronoun in fiction and exposition, but it is more likely to occur as a demonstrative in fiction, and more likely as a complementizer in exposition.

Table 3 illustrates the same kinds of comparison for tag sequences. Although the differences are not as striking, several of them are large enough to be relevant for automatic tagging. For example, prepositions and nouns are considerably more likely to follow singular nouns in exposition than in fiction. Similarly, nouns are more likely to follow adjectives in exposition than in fiction. Passive verbs are considerably more likely to follow the copula *be* in exposition than in fiction, while progressive verb forms are more likely to follow *be* in fiction. Other differences are not great, but they are consistent across tag sequences.

Finally, comparisons of this type are also important for syntactic ambiguities, which are the bane of probabilistic parsers. To illustrate the importance of register differences in this arena, Table 4 presents frequency counts for prepositional phrase

Table 3
Comparison of probabilities for selected tag-sequences in Exposition versus Fiction (derived from analysis of tags in the LOB Corpus). (Note: The probabilities do not add up to 100% because minor categories are not listed.)

Grammatical Category of First Word	Grammatical Category of Second Word	Fiction %	Exposition %
singular noun	preposition	23	31
	singular noun	4	8
	plural noun	1	4
	. (full stop)	18	12
	, (comma)	15	11
adjective	singular noun	42	47
	plural noun	12	20
	. (full stop)	7	3
	, (comma)	8	4
copula *be*	passive verb	13	31
	progressive verb	11	4
past tense verb	indefinite article	11	17
	adverb	18	11
	preposition	15	18
	. (full stop)	7	3
	, (comma)	6	3
present tense verb	indefinite article	12	18
	adverb	13	9
	preposition	15	19

Table 4
Average frequency counts (per 1,000 words of text) and percentages for prepositional phrases attached as noun modifiers and verb modifiers in Editorials versus Fiction (based on analysis of ten-text subsamples from the LOB Corpus).

	Editorials		Fiction	
	Frequency	%	Frequency	%
Prepositions as Noun Modifiers	38.2	46.4%	15.2	21.3%
Prepositions as Verb Modifiers	44.1	53.6%	56.2	78.7%
Total Prepositions	82.3	100.0%	71.4	100.0%

Table 5
Comparison of average normalized frequencies of certainty adjectives in the Longman/Lancaster English Language Corpus (written texts), London/Lund Corpus (spoken texts), and two specific text categories from the Longman/Lancaster Corpus (Social Science and Fiction). (per 1 million words of text).

	Frequency of XXX + *certain*	Frequency of XXX + *sure*	Frequency of XXX + *definite*
Longman/Lancaster Corpus (Written Texts)	259.0	234.0	34.9
London/Lund Corpus (Spoken Texts)	292.5	426.9	19.4
Selected Registers from the Longman/Lancaster Corpus			
Social Science	358.7	73.8	114.2
Fiction	178.5	353.1	10.8

attachment, as a nominal versus verbal modifier, in ten-text subsamples taken from editorials and fiction. (The texts are from the LOB Corpus; these counts were done by hand.) For editorials, this table shows that there is nearly a 50/50 split for prepositional phrases attached as nominal versus verbal modifiers, with a slight preference for nominal modifiers. In fiction, on the other hand, there is a much greater likelihood that a prepositional phrase will be attached as a verbal modifier (78.7%) rather than a nominal modifier (21.3%).

For any automated language processing that depends on probabilistic techniques, whether part-of-speech tagging or syntactic parsing, the input probabilities are crucial. The analyses in this section suggest that it might be advantageous to store separate probabilities for different major registers, rather than using a single set of probabilities for a general-purpose tool. Minimally, these analyses show that input probabilities must be based on the distribution of forms in a diversified corpus representing the major register distinctions; probabilities derived from a single register are likely to produce skewed results when applied to texts from markedly different registers.

2.2 Lexicographic Issues

Text corpora have proven to be invaluable resources for research on word use and meaning, as in Sinclair's pioneering work on the COBUILD Dictionary (Sinclair 1987). In fact, corpus-based research shows that our intuitions about lexical patterns are often incorrect (Sinclair 1991; 112 ff). However, similar to the patterns for grammatical structures, for many words there is no general pattern of use that holds across the whole language; rather, different word senses and collocational patterns are strongly preferred in different registers.

This point can be illustrated from an analysis of certainty adjectives in English (exploring in more detail some of the findings of Biber and Finegan 1989). Table 5 presents overall frequencies for three certainty adjectives—*certain, sure,* and *definite*—in two text corpora: the Longman/Lancaster Corpus, including written texts from ten

Table 6
Comparison of frequencies of collocations for certainty adjectives in two text categories of the Longman/Lancaster English Language Corpus: Social Science and Fiction. (Counts are normalized per 1,000,000 words of text.)

	Frequency of *certain* in Social Science	Frequency of *certain* in Fiction	Frequency of *sure* in Social Science	Frequency of *sure* in Fiction
Preceding Words				
a +	101.2*	66.3	5.8	1.7
of/in/for +	56.4*	9.4	5.8	3.7****
there BE +	13.0*	1.1	0.0	0.0
PRO BE + (excluding *am/I'm*)	5.8	21.3**	18.8	66.0**
I am/I'm +	0.0	4.5**	0.0	79.9**
MAKE +	2.9	4.0	30.4	38.3
quite/so/ pretty +	4.3	5.4	4.3	27.2**
Following Words				
+ *kind(s)*	30.4*	0.6	0.0	0.0
+ *amount*	10.1	7.1	0.0	0.0
+ *of*	11.6	6.8	11.6	28.9
+ *that*	14.5	12.5	27.5	38.8
+ SUBJ PRO***	0.0	4.8**	10.1	129.2**
+ *the****	1.4	2.3	1.4	10.2**
+ *enough*	0.0	0.0	1.4	10.2**

* Collocations much more common in social science than in fiction.
** Collocations much more common in fiction than in social science.
*** These collocations represent *that* complement clauses where the complementizer has been deleted.
**** These collocations are all tokens of the idiom *for sure.*

major text categories, and the London/Lund Corpus, made up of spoken texts from six major text categories. The overall pattern shows *certain* and *sure* occurring with approximately the same frequency in the written (Longman/Lancaster) corpus. In the spoken (London/Lund) corpus, *sure* occurs more frequently than *certain*, and both words are more common than in the written corpus. The word *definite* is relatively rare in both corpora, although it is slightly more common in the written corpus.

Further, there are striking differences across written registers in the use of these words. In social science, *certain* is quite common, *sure* is relatively rare, and *definite* is common relative to its frequency in the whole written corpus. Fiction shows the opposite pattern: *certain* is relatively rare, *sure* is relatively common, and *definite* is quite rare. These patterns alone show that the semantic domain of certainty in English could not be adequately described without considering the patterns in complementary registers.

Table 6 shows, however, that the actual patterns of use are even more complex. This table presents normalized frequencies (per 1 million words of text) of the major collocational patterns for *certain* and *sure*, comparing the distributions in social science and fiction. A single * is used to mark collocations that are much more common in social science, while ** is used to mark collocations that are much more common

in fiction. The collocational patterns (confirmed by concordance listings) identify two surprising facts about these words: 1) *certain* is commonly used to mark uncertainty rather than certainty; and 2) certainty is rarely expressed in social science at all. Thus, the most common collocations for *certain* in social science reflect a kind of vagueness, marking a referent as possibly being known (by someone) but not specified in the text (e.g., *a certain kind of...*, *in certain cases...*, *there are certain indications that...*). These collocations are relatively rare in fiction. In contrast, those collocational patterns for *certain* that directly state that someone or something is certain—*he/she/they/you/it + BE + certain* and *I/we + BE + certain*—are extremely rare in social science but relatively common in fiction.

Unlike the word *certain*, the term *sure* is most typically used to express certainty. This is apparently the reason why the overall frequency of *sure* is so low in social science. The difference between *certain* and *sure*, and between social science and fiction, is perhaps most striking for the collocations of *pronoun + BE + certain/sure*. These collocations are very common for *sure* in fiction; moderately common for *certain* in fiction; but quite rare for either *sure* or *certain* in social science.

These examples illustrate the point that a corpus restricted to only one register would enable at best a partial analysis of lexical use; and if the results were generalized to the entire language, they would be incorrect. Thus, the major registers of English must be treated on their own terms in order to provide a comprehensive analysis of either grammatical structures or lexical patterns of use.

3. Multidimensional Differences among Registers in English

The inherent nature of register variation in English is illustrated even more clearly in a series of studies using the multidimensional framework (e.g., Biber 1988, 1989, 1992). These studies 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.

Each dimension comprises a set of linguistic features that co-occur frequently in texts. The dimensions are identified from a quantitative analysis of the distribution of 67 linguistic features in the texts of the LOB and London-Lund Corpora. There is space here for only a brief methodological overview of this approach; interested readers are referred to Biber (1988; especially Chapters 4 and 5) for a more detailed presentation.

First, texts were automatically tagged for linguistic features representing several major grammatical and functional characteristics: tense and aspect markers, place and time adverbials, pronouns and pro-verbs, nominal forms, prepositional phrases, adjectives, adverbs, lexical specificity, lexical classes (e.g., hedges, emphatics), modals, specialized verb classes, reduced forms and discontinuous structures, passives, stative forms, dependent clauses, coordination, and questions. All texts were post-edited by hand to correct mis-tags.

The frequency of each linguistic feature in each text was counted, and all counts were normalized to their occurrence per 1000 words of text. Then a factor analysis was run to identify the major co-occurrence patterns among the features. (Factor analysis is a statistical procedure that identifies groupings of linguistic features that co-occur frequently in texts.) So that texts and registers could be compared with respect to the dimensions, dimension scores were computed for each text by summing the major linguistic features grouped on each dimension. Finally, the dimensions were interpreted functionally based on the assumption that linguistic features co-occur in texts because

they share underlying communicative functions. Similarly, the patterns of variation among registers were interpreted from both linguistic and functional perspectives.

Five major dimensions are identified and interpreted in Biber (1988; especially Chapters 6 and 7).[3] Each 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. The five dimensions are interpretively labeled:

1. Informational versus Involved Production[4]
2. Narrative versus Nonnarrative Concerns
3. Elaborated versus Situation-Dependent Reference
4. Overt Expression of Persuasion
5. Abstract versus Nonabstract Style

The primary communicative functions, major co-occurring features, and characteristic registers associated with each dimension are summarized in Table 7. As this table shows, registers differ systematically along each of these dimensions, relating to functional considerations such as interactiveness, involvement, purpose, and production circumstances, all of which have marked correlates in linguistic structure.[5] To illustrate these differences more concretely, Figure 1 presents the differences among nine spoken and written registers within the two-dimensional space defined by Dimension 1: 'Involved versus Informational Production' and Dimension 3: 'Elaborated versus Situation-Dependent Reference.'

The register characterizations on Figure 1 reflect different relative frequencies of the linguistic features summarized in Table 7. For example, academic prose and newspaper reportage have the largest positive scores on Dimension 1, reflecting very frequent occurrences of nouns, adjectives, prepositional phrases, long words, etc. (the 'informational' features grouped on Dimension 1), together with markedly infrequent occurrences of 1st and 2nd person pronouns, questions, reductions, etc. (the 'involved' features on Dimension 1). On Dimension 3, academic prose and professional letters have the largest positive scores, reflecting very frequent occurrences of WH relative clause constructions (the features associated with 'elaborated reference'), together with markedly infrequent occurrences of time and place adverbials (the 'situation-dependent' features). At the other extreme, conversations have the largest negative score on Dimension 1, reflecting very frequent occurrence of the 'involved' features grouped on that dimension (1st and 2nd person pronouns, questions, etc.) together with markedly few occurrences of the 'informational' features (nouns, adjectives, etc.). Conversations also have a quite large negative score on Dimension 3, although broadcasts have the largest negative score, reflecting very frequent occurrences of time and place adverbials together with markedly few WH relative clauses, etc.

3 The factorial structure was derived from a common factor analysis with a Promax rotation. A seven-factor solution was extracted as the most adequate, although only the first five factors are presented here. The first factor in the analysis accounts for 26.8% of the shared variance, while all seven factors together account for 51.9% of the shared variance. Further details are given in Biber (1988).

4 The polarity of Dimension 1 has been reversed to aid in the comparison to Dimensions 3 and 5.

5 See Biber (in press, b) for a comprehensive framework comparing registers with respect to their situational characteristics, such as the relations among participants, purposes, production circumstances, and typical topics.

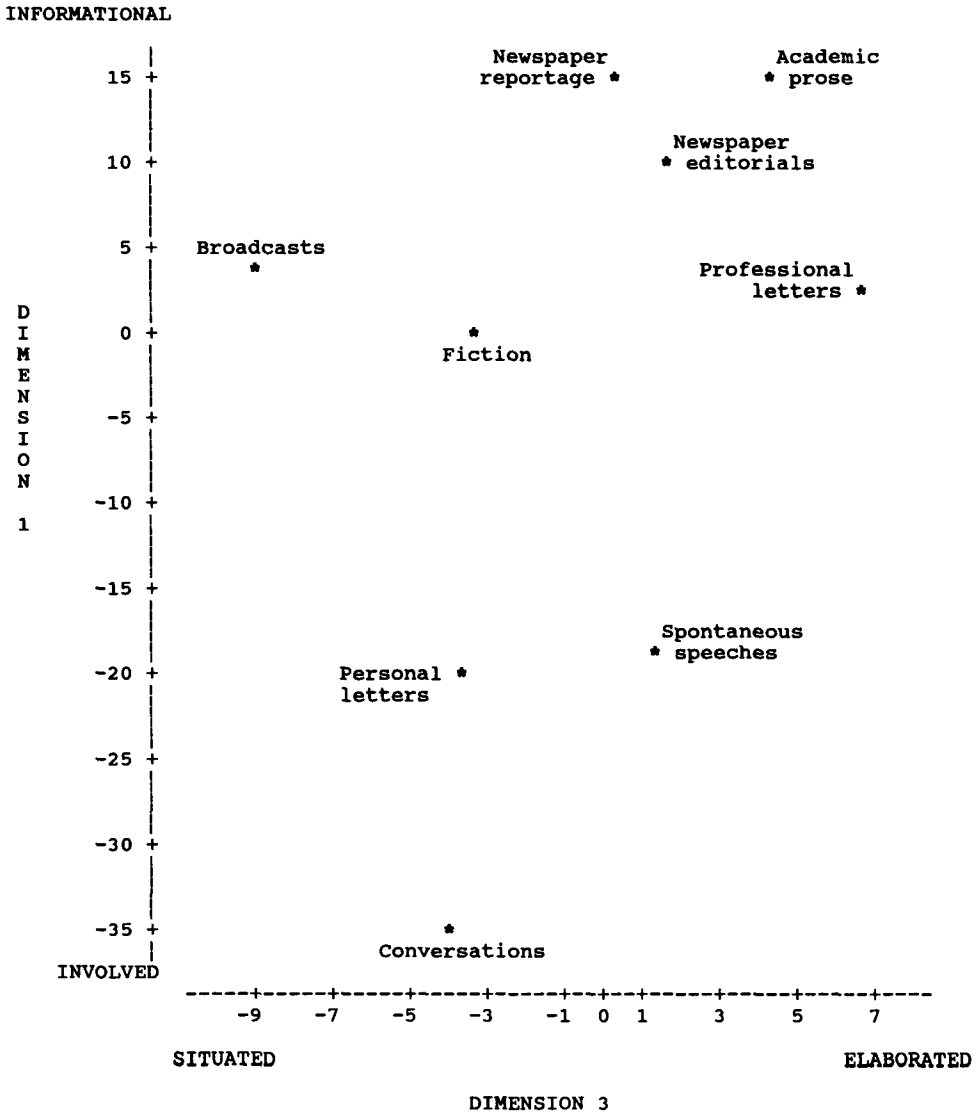

Figure 1
Linguistic characterization of nine spoken and written registers with respect to Dimension 1 ('Informational versus Involved Production') and Dimension 3 ('Elaborated versus Situation-Dependent Reference').

Table 7
Summary of functions, linguistic features, and characteristic registers for the five major English dimensions identified in Biber (1988). *Continued next page.*

Functions	Linguistic features	Characteristic registers
Dimension 1 'Informational versus Involved Production'		
Monologue Careful Production Informational Faceless	nouns, adjectives, prepositional phrases long words	informational exposition e.g., official documents academic prose
Interactive (Inter)personal Focus Involved Personal Stance On-Line Production	1st and 2nd person pronouns questions, reductions stance verbs, hedges, emphatics adverbial subordination	conversations (personal letters) (public conversations)
Dimension 2 'Narrative versus Nonnarrative Concerns'		
Narrative	past tense perfect aspect 3rd person pronouns speech act (public) verbs	fiction
Nonnarrative	present tense attributive adjectives	exposition, broadcasts professional letters telephone conversations
Dimension 3 'Elaborated versus Situation-Dependent Reference'		
Situation-Independent Reference (Elaborated)	WH relative clauses pied-piping constructions phrasal coordination	official documents professional letters (exposition)
Situation-Dependent Reference On-Line Production	time and place adverbials	broadcasts (conversations) (fiction) (personal letters)

As can be seen from Figure 1, these nine registers are strikingly different in their linguistic characteristics, even within this two-dimensional space. When all six dimensions are considered, these differences are even more notable. Table 8 further shows that a significant and important amount of variation among texts can be accounted for based on the register distinctions. It is important to emphasize here that the register categories were not considered when the dimensions were originally identified; rather, the dimensions represent the linguistic co-occurrence patterns across texts, regardless of their register category. However, Table 8 shows that there are important differences across registers with respect to each dimension. The F values and probabilities report

Table 7
Continued.

Functions	Linguistic features	Characteristic registers
Dimension 4 'Overt Expression of Persuasion'		
Overt Argumentation and Persuasion	modals (prediction, necessity, possibility) suasive verbs conditional subordination	professional letters editorials
Not Overtly Argumentative	—	broadcasts (press reviews)
Dimension 5 'Abstract versus Non-abstract Style'		
Abstract Style	agentless passives *by* passives passive dependent clauses	technical prose (other academic prose) (official documents)
Non-abstract	—	conversations, fiction personal letters public speeches public conversations broadcasts

Table 8
F scores and correlations for dimension score differences across 23 spoken and written registers (df = 22,459).

Dimension	F value	Probability	$r2$
1	111.9	$p < .001$	84.3%
2	32.3	$p < .001$	60.8%
3	31.9	$p < .001$	60.5%
4	4.2	$p < .001$	16.9%
5	28.8	$p < .001$	58.0%

the results of an Analysis of Variance showing that the registers are significant discriminators for each dimension, and the r^2 values show their strength (r^2 is a direct measure of the percentage of variation in the dimension score that can be predicted on the basis of the register distinctions). Four of the five dimensions have r^2 values over 50%, with only Dimension 4 having a relatively small r^2 value of 16.9%. These

values show that the registers are strong predictors of linguistic variability along all five dimensions.

Registers are defined in terms of their situational characteristics, and they can be analyzed at many different levels of specificity.[6] However, there are important linguistic differences even among many closely related subregisters (Biber 1988; Chapter 8). A complementary perspective is to analyze the total space of variation within a language in terms of *linguistically* well-defined text categories, or *text types* (Biber 1989).[7] Given a text type perspective, linguistically distinct texts within a register represent different types, while linguistically similar texts from different registers represent a single text type.

In sum, all of these analyses show that there are extensive differences among English registers with respect to a wide array of linguistic features. A corpus restricted to any one or two of these registers would clearly be excluding much of the English language, linguistically as well as situationally.

4. Further Applications of Corpus-Based Analyses of Register Variation

The multidimensional model of register variation summarized in Section 3 can be used to address additional computational issues. In this section, I focus on two of these: the automated prediction of register category and cross-linguistic comparisons.

4.1 Automated Prediction of Registers

One issue of current relevance within computational linguistics is the automated prediction of register category, as a preliminary step to work in information retrieval, machine translation, and other kinds of text processing. Because the model of register variation summarized in the last section is multidimensional, with each dimension comprising a different set of linguistic features and representing a different set of relations among registers, it is well suited to this research question.

One statistical procedure commonly used for classificatory purposes is *discriminant analysis*. This procedure computes the *generalized squared distance* between a text and each text category (or register), and the text is then automatically classified as belonging to the closest category.[8]

To illustrate, Figure 2 plots the five-dimensional profiles of three target registers (academic prose, fiction, and newspaper reportage) together with an unclassified text

6 Because registers can be specified at many different levels of generality, there is no "correct" set of register distinctions for a language; rather, I have argued elsewhere that registers should be seen as semi-continuous (rather than discrete) constructs varying along multiple situational parameters (Biber, in press, b). Further, since registers are defined situationally rather than on a linguistic basis, they are not equally coherent in their linguistic characteristics. Some registers have quite focused norms and therefore show little internal linguistic variation (e.g., science fiction). Registers such as popular magazine articles, on the other hand, include a wide range of purposes, and thus show extensive linguistic differences among the texts within the register (cf. the investigations in Biber 1988, Chapter 8; Biber 1990).

7 The statistical procedure used to identify linguistically well-defined text types is called cluster analysis. This procedure identifies concentrations of texts such that the texts within each cluster, or text type, are maximally similar to one another in their linguistic characteristics, while the types are maximally distinct from one another. Biber (1989) identifies eight major text types in English, which are interpretively labeled: Intimate Interpersonal Interaction, Informational Interaction, "Scientific" Exposition, Learned Exposition, Imaginative Narrative, General Narrative Exposition, Situated Reportage, and Involved Persuasion. Biber (in press, c) compares the text type distinctions in English and Somali.

8 The SAS procedure PROC DISCRIM was used for the discriminant analyses in this section.

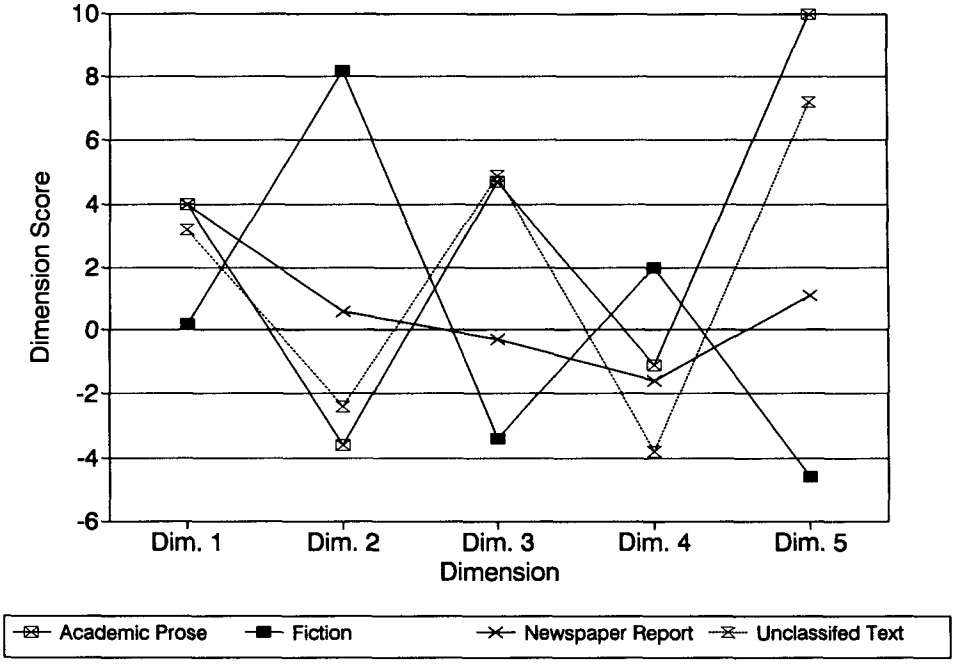

Figure 2
Five-dimensional distance profile of unclassified text to three target registers.

(which is in fact a social science academic article).[9] With respect to Dimensions 1 and 4, the unclassified text is about equidistant to the mean scores for academic prose and newspaper reportage, but along Dimensions 2, 3, and 5, the unclassified text is much closer to the target means for academic prose. Generalizing over all five dimensions, this text has the smallest distance to academic prose and would be classified into that category.

The predictive power of this five-dimensional model of variation was tested at three levels of abstraction. The first test was based on texts from three high-level register categories: newspaper articles (including press reportage, editorials, and reviews), academic prose (including humanities, social science, medicine, natural science, and engineering), and fiction. Table 9 presents the discriminant analysis results for these categories. The top part of the table presents the calibration results: the model was trained on 118 texts from these three categories and then used to classify the same

9 To aid in comparison across dimensions, all dimension scores in Figure 2 have been converted to a common scale of plus-or-minus 10. The scaling coefficients are:

Dimension 1: .27

Dimension 2: 1.39

Dimension 3: 1.11

Dimension 4: 2.27

Dimension 5: 1.82

Table 9
Automatic classification of texts into three general register categories (fiction, academic prose, newspaper articles), based on a discriminant analysis using five underlying dimensions.

Calibration results: Classification of the 118 texts used to derive the discriminant function

	Number of Observations and Percent Classified into:			
	Newspapers	Academic Prose	Fiction	Total
From Newspapers	37 86.05%	6 13.95%	0 0.00%	43 100.00%
Academic Prose	9 22.50%	31 77.50%	0 0.00%	40 100.00%
Fiction	3 8.57%	0 0.00%	32 91.43%	35 100.00%

Test results: Classification of 124 'unknown' texts

	Number of Observations and Percent Classified into:			
	Newspapers	Academic Prose	Fiction	Total
From Newspapers	31 68.89%	13 28.89%	1 2.22%	45 100.00%
Academic Prose	11 27.50%	29 72.50%	0 0.00%	40 100.00%
Fiction	6 15.38%	0 0.00%	33 84.62%	39 100.00%

texts. The rates for successful prediction are high in all three cases. This model was then tested on a new set of 124 'unknown' texts; the rates for successful classification are still high (ranging from 68.89% to 84.62%), although not as high as in the calibration data. Academic prose and fiction are clearly distinguished, with no misclassifications between these two groups. Newspaper texts are less sharply distinguished from the other two categories: 28.89% of the newspaper texts are incorrectly classified as academic prose; 27.5% of the academic prose texts are incorrectly classified as newspaper texts; and 15.38% of the fiction texts are incorrectly classified as newspaper texts. Overall, though, the large majority of texts in these three categories are correctly classified by the five-dimensional model.

Table 10 shows that roughly the same success rate can be achieved at a more specific level of prediction: distinguishing among press reportage, editorials, and reviews within newspapers. The calibration model shows success rates ranging from 69.23% to 87.5% for these three categories, and the test data are correctly predicted at comparable rates (ranging from 68.18% to 92.86%).

Finally, Table 11 applies this technique at a much more specific level, attempting to discriminate among four kinds of press reportage that differ primarily in their content domains: political, sports, spot news, and financial reportage. In this case, only the

Table 10
Automatic classification of texts into three specific press registers (press reportage, press reviews, press editorials), based on a discriminant analysis using five underlying dimensions.

Calibration results: Classification of the 43 newspaper texts used to derive the discriminant function				
	Number of Observations and Percent Classified into:			
	Reportage	Editorials	Reviews	Total
From Reportage	16 72.73%	4 18.18%	2 9.09%	22 100.00%
Editorials	1 7.69%	9 69.23%	3 23.08%	13 100.00%
Reviews	1 12.50%	0 0.00%	7 87.50%	8 100.00%
Test results: Classification of 45 'unknown' texts				
	Number of Observations and Percent Classified into:			
	Reportage	Editorials	Reviews	Total
From Reportage	15 68.18%	2 9.09%	5 22.73%	22 100.00%
Editorials	0 0.00%	13 92.86%	1 7.14%	14 100.00%
Reviews	1 11.11%	1 11.11%	7 77.78%	9 100.00%

calibration results are reported because of the small sample size. The results indicate, however, a high success rate (ranging from 61.54% to 85.71%), suggesting that this approach can be profitably used for prediction among closely related subregisters or sublanguages.

The predictive power of this technique is only as robust as the underlying model of variation. In the present case, that model represents multiple dimensions of linguistic variation derived from analysis of a register-diversified corpus, and the results achieved are generally robust for the successful prediction of different kinds of text at quite different levels of abstraction.

4.2 Cross-Linguistic Comparisons

The analysis of parallel text corpora in different languages has received considerable attention in recent years, usually in relation to research on information retrieval and machine translation. Many researchers dealing with these issues from a register perspective have focused on the computational analysis of *sublanguages*, a subsystem of a language that operates within a particular domain of use with restricted subject matter (see Kittredge and Lehrberger 1982; Grishman and Kittredge 1986). Processing

Table 11
Automatic classification of 34 newspaper reportage texts into four content areas (political reportage, sports reportage, spot news reportage, financial reportage), based on a discriminant analysis using five underlying dimensions.

	Number of Observations and Percent Classified into:				
	Political	Sports	Spot News	Financial	Total
From Political	8 61.54%	2 15.38%	2 15.38%	1 7.69%	13 100.00%
Sports	0 0.00%	6 85.71%	0 0.00%	1 14.29%	7 100.00%
Spot News	1 10.00%	0 0.00%	8 80.00%	1 10.00%	10 100.00%
Financial	0 0.00%	1 25.00%	0 0.00%	3 75.00%	4 100.00%

research in this area has achieved high levels of success by focusing on very restricted textual domains.

Kittredge (1982) adopts a variation perspective, comparing the extent of sublanguage differences within and across languages. Some of the provocative conclusions of that study are:

- "the written style of English and French tended to be more similar in specialized technical texts than in general language texts" (1982, p. 108).
- "parallel sublanguages of English and French are much more similar structurally than are dissimilar sublanguages of the same language. Parallel sublanguages seem to correspond more closely when the domain of reference is a technical one" (1982, p. 108).

The multidimensional framework provides a complementary approach to these issues. From a linguistic perspective, dimensions are more readily compared cross-linguistically than individual features, since structurally similar features often serve quite different functional roles across languages. Similarly, cross-linguistic comparisons of individual registers are more readily interpretable when they are situated relative to the range of other registers in each language, since the 'same' registers can serve quite different functions across languages when considered relative to their respective register systems.

To date, there have been multidimensional analyses of register variation in four languages: English (summarized in Section 3), Nukulaelae Tuvaluan (Besnier 1988), Korean (Kim and Biber in press), and Somali (Biber and Hared 1992, in press). In each case, the description is based on analysis of a diversified corpus representing a wide range of spoken and written registers. The cross-linguistic patterns of variation represented by these four languages, both synchronic and diachronic, are discussed in Biber (in press, c).

Table 12
Summary of functions, linguistic features, and characteristic registers for the five major Somali dimensions identified in Biber and Hared (1992).

Functions	Linguistic Features	Characteristic Registers
Dimension 1		
Interactive (Inter)personal focus Involved Personal Stance (On-Line Production)	main clause features questions, imperatives contractions stance adjectives downtoners 1st and 2nd person pronouns	conversations family meetings conversational narratives
Monologue Informational Faceless (Careful Production)	dependent clauses relative clauses clefts, verb complements nouns, adjectives	written expository registers
Dimension 2		
On-Line Production (Situation Dependent)	—	sports broadcast (other spoken registers)
Careful Production Informational	once-occurring words high type/token ratio nominalizations compound verbs	editorials written political speeches and pamphlets analytical press
Dimension 3		
Overt Argumentation Persuasion	present tense, adjectives possibility modals concession conjuncts conditional clauses	family and formal meetings general interest and analytical press (invited editorials)
Reported Presentation	past tense proper and agentive nouns future modals	press reportage (folk stories)
Dimension 4		
Narrative Discourse	3rd person pronouns past tense verbs temporal clauses clefts, habitual modals	folk stories (serial stories) (general fiction)
Non-narrative Discourse	compound nouns gerunds, agentive nouns	petitions announcements memos

To illustrate, the multidimensional analysis of English (discussed above; see Table 7) can be compared with the multidimensional patterns of variation in Somali, summarized in Table 12. Both languages represent many of the same functional considerations in their dimensional structure, including interactiveness, involvement, produc-

Table 12
Continued.

Functions	Linguistic Features	Characteristic Registers
Dimension 5		
Interactive Distanced and Directive Communication	optative clauses 1st and 2nd person pronouns directional particles imperatives	personal letters (family meetings) (Quranic exposition)
Non-interactive Non-directive	—	press reportage and editorials written expository registers

tion circumstances, informational focus, personal stance, and narrative purposes. There are also many similarities in the co-occurrence patterns among linguistic features. For example, first and second person pronouns, downtoners, stance features, contractions, and questions group together in both languages as markers of involvement; nouns and adjectives group together in both languages as markers of an informational focus; third person pronouns and past tense verbs group together in both languages as markers of narration. In other respects, though, the multi-dimensional structure of the two languages differ. For example, Somali has two dimensions marking different kinds of interaction plus a third dimension relating to production circumstances; all of these functions are combined into a single dimension in English (Dimension 1). Conversely, English Dimension 5 marks a passive, abstract style, which has no counterpart in Somali.

One of the surprising findings from the comparison of all four languages (including Nukulaelae Tuvaluan and Korean) is the extent of the cross-linguistic similarities (see Biber, in press, c). Thus, all four languages have multiple dimensions reflecting oral/literate differences, interactiveness, production circumstances, and an informational focus; these dimensions are defined by similar kinds of linguistic features, and analogous registers have similar cross-linguistic characterizations along these dimensions. In addition, two functional domains that relate to purpose are marked in all four languages: personal stance (toward the content) and narration. These dimensions also have similar structural correlates across the languages.

In contrast, there are fewer major differences among these languages in their patterns of register variation. Dimensions relating to argumentation/persuasion are found in only some languages, and there are other dimensions particular to a single language (such as abstract style in English, and honorification in Korean). Analogous registers show some differences cross-linguistically with respect to these latter dimensions plus the purpose-related dimensions mentioned above (e.g., marking personal stance or narration).

Findings such as these are directly relevant to several of the issues raised in recent studies of sublanguages, since they can be used to specify the linguistic relations among sublanguages both within and across languages. In particular, these analyses support Kittredge's (1982) conclusion that parallel sublanguages across languages are more similar in their linguistic structure than are dissimilar sublanguages within the same language. Romaine (in press) discusses similar findings in a comparison of sports

reportage in Tok Pisin and English. The multidimensional comparisons summarized here show that even when registers are defined at a high level of generality (e.g., conversation, fiction, academic prose), and even when comparisons are across markedly different language families and cultures, parallel registers are indeed more similar cross-linguistically than are disparate registers within a single language.

5. Conclusion

In the present paper I have presented evidence from several different structural levels, as well as different languages, showing that there are important and systematic linguistic differences among registers. These data are used to argue for the general point that corpora representing a broad range of register variation are required as the basis for general language studies. In fact, the extent of register differences reported here suggests that overall linguistic characterizations of a language are often inadequate (or even incorrect). That is, since overall generalizations represent a kind of averaging of the linguistic patterns in a language, they often do not accurately represent the actual patterns of any register; in fact, such generalizations can conceal the systematic patterns found across registers. An alternative approach is to recognize the centrality of register differences and work toward a composite linguistic description of a language in those terms.

Corpus-based analyses of register variation obviously need to be extended in several ways. Future research should be based on larger corpora and include a wider representation of linguistic features and registers. In addition, register distinctions can be made at several levels of abstraction, and the intersection of register and text type analyses needs to be further explored (see notes 6 and 7). The analyses summarized here, though, clearly show the importance of a register perspective, supporting the continuing development and analysis of large register-diversified corpora as the basis for general language studies.

Acknowledgments

Parts of this paper were presented at the First Workshop of the Consortium for Lexical Research, New Mexico State University (January 1992) and at the Pisa Workshop on Textual Corpora, University of Pisa (January 1992). I would like to thank the workshop participants, as well as two anonymous reviewers for *CL*, for their helpful suggestions for revision.

References

Altenberg, Bengt (1991). "A bibliography of publications relating to English computer corpora." In *English Computer Corpora: Selected Papers and Research Guide*, edited by S. Johansson and A.-B. Stenström, 355–396. Mouton.

Besnier, Niko (1988). "The linguistic relationships of spoken and written Nukulaelae registers." *Language*, 64, 707–736.

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

Biber, Douglas (1989). "A typology of English texts." *Linguistics*, 27, 3–43.

Biber, Douglas (1990). "Methodological issues regarding corpus-based analyses of linguistic variation." *Literary and Linguistic Computing*, 5, 257–269.

Biber, Douglas (1992). "On the complexity of discourse complexity: A multidimensional analysis." *Discourse Processes*, 15, 133–163.

Biber, Douglas (in press, a). "Representativeness in corpus design." In *Proceedings of the Pisa Workshop on Textual Corpora*, edited by A. Zampolli. University of Pisa.

Biber, Douglas (in press, b). "Towards a comprehensive analytical framework for register studies." In *Perspectives on Register: Situating Register Variation within Sociolinguistics*, edited by D. Biber and E. Finegan. Oxford University Press.

Biber, Douglas (in press, c). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, Douglas, and Finegan, Edward (1989). "Styles of stance in English: Lexical and grammatical marking of evidentiality and affect." *Text*, 9, 93–124.

Biber, Douglas, and Hared, Mohamed (1992). "Dimensions of register variation in Somali." *Language Variation and Change*, 4, 41–75.

Biber, Douglas, and Hared, Mohamed (in press). "Linguistic correlates of the transition to literacy in Somali: Language adaptation in six press registers." In *Perspectives on Register: Situating Register Variation within Sociolinguistics*, edited by D. Biber and E. Finegan. Oxford University Press.

Fitzpatrick, Eileen; Bachenko, Joan; and Hindle, Don (1986). "The status of telegraphic sublanguages." In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, edited by Ralph Grishman and Richard Kittredge, 39–54. Lawrence Erlbaum.

Francis, W. N., and Kučera, H. (1964). *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Department of Linguistics, Brown University.

Grishman, Ralph, and Kittredge, Richard, editors (1986). *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*. Lawrence Erlbaum.

Johansson, Stig; Leech, Geoffrey N.; and Goodluck, Helen (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo.

Kim, Yong-Jin, and Biber, Douglas (in press). "A corpus-based analysis of register variation in Korean." In *Perspectives on Register: Situating Register Variation within Sociolinguistics*, edited by D. Biber and E. Finegan. Oxford University Press.

Kittredge, Richard (1982). "Variation and homogeneity of sublanguages." In *Sublanguage: Studies of Language in Restricted Semantic Domains*, edited by Richard Kittredge and John Lehrberger, 107–137. De Gruyter.

Kittredge, Richard, and Lehrberger, John, editors (1982). *Sublanguage: Studies of Language in Restricted Semantic Domains*. De Gruyter.

Lehrberger, John (1982). "Automatic translation and the concept of sublanguage." In *Sublanguage: Studies of Language in Restricted Semantic Domains*, edited by Richard Kittredge and John Lehrberger, 81–106. De Gruyter.

Romaine, Suzanne (in press). "On the creation and expansion of registers: Sports reporting in Tok Pisin." In *Perspectives on Register: Situating Register Variation within Sociolinguistics*, edited by D. Biber and E. Finegan. Oxford University Press.

Sager, Naomi (1986). "Sublanguage: Linguistic phenomenon, computational tool." In *Analyzing Language in Restricted Domains: Sublanguage Description and Processing*, edited by Ralph Grishman and Richard Kittredge, 1–18. Lawrence Erlbaum.

Sinclair, John, editor (1987). *Looking Up*. Collins.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Tottie, Gunnel (1986). "The importance of being adverbial: Adverbials of focusing and contingency in spoken and written English." In *English in Speech and Writing: A Symposium*, edited by G. Tottie and I. Bäcklund, 93–118. Almqvist and Wiksell.