# 9 Register Variation: A Corpus Approach

DOUGLAS BIBER AND SUSAN CONRAD

## 0 Introduction

Analyses of discourse context can be approached from two perspectives. First, they can focus on the textual environment, considering lexical, grammatical, and rhetorical features in the text. Alternatively, analyses can concentrate on the extratextual communicative situation. Furthermore, such extratextual analyses can differ in terms of their generality. For example, the communicative situation of a given interaction can be described in relation to the specific individuals involved, their precise relationship, their personal motivations for the interaction, etc. A different approach would be to focus on the general parameters defining the communicative situation of a text – for example, the mode, the level of interactiveness, the general purpose, etc.

Varieties defined in terms of general situational parameters are known as *registers*. We use the label *register* as a cover term for any variety associated with a particular configuration of situational characteristics and purposes. Thus, registers are defined in nonlinguistic terms. However, as illustrated in this chapter, there are usually important linguistic differences among registers as well.

There have been numerous studies that describe the situational parameters that are important for studies of discourse. As early as the 1930s, Firth identified crucial components of speech situations, applying principles from Malinowki's work. More recent and particularly well known is Hymes's (1974) framework for studying the ethnography of communication. In addition, a number of other anthropologists and sociolinguists have proposed frameworks or identified particularly important characteristics that can be applied to identifying registers (e.g. Basso 1974; Biber 1994; Brown and Fraser 1979; Crystal and Davy 1969; Duranti 1985). Throughout these discussions, the important characteristics that are identified include: the participants, their relationships, and their attitudes toward the communication; the setting, including factors such as the extent to which time and place are shared by the participants, and the level of formality; the channel of communication; the production and processing circumstances (e.g. amount of time available); the purpose of the communication; and the topic or subject matter. A register can be defined by its particular combination of values for each of these characteristics.

In many cases, registers are named varieties within a culture, such as novels, memos, book reviews, and lectures. However, registers can be defined at any level of generality, and more specialized registers may not have widely used names. For example, "academic prose" is a very general register, while "methodology sections in experimental psychology articles" is a much more highly specified one.

There are many studies that describe the situational and linguistic characteristics of a particular register. These studies cover diverse registers such as sports announcer talk (Ferguson 1983), note-taking (Janda 1985), personal ads (Bruthiaux 1994), classified advertising (Bruthiaux 1996), and coaching (Heath and Langman 1994). Analyses of register variation have also been conducted within a Hallidayan functional-systemic framework (see, e.g., the collection of papers in Ghadessy 1988, which include registers such as written sports commentary, press advertising, and business letters); several studies employing this approach are particularly concerned with describing school-based registers and their implications for education (e.g., Christie 1991; Martin 1993). Analysis of single registers has also been conducted for languages other than English, such as sports reporting in Tok Pisin (Romaine 1994). Atkinson and Biber (1994) provide an extensive survey of empirical register studies.

In addition to describing single registers, studies have also made comparisons across registers. These comparative studies have shown that there are systematic and important linguistic differences across registers, referred to as the patterns of *register variation*. This comparative register perspective is particularly important for two major arenas of research: (1) linguistic descriptions of lexical and grammatical features, and (2) descriptions of the registers themselves. With respect to traditional lexical and grammatical investigations, it turns out that functional descriptions based on texts without regard for register variation are inadequate and often misleading; we illustrate the importance of register for such analyses in section 1. For register descriptions, a comparative register perspective provides the baseline needed to understand the linguistic characteristics of any individual register. That is, by describing a target register relative to a full range of other registers, we are able to accurately identify the linguistic features that are in fact notably common in that register. We illustrate analyses of this type in section 2.1.

In recent years, studies of register variation have also been used to make cross-linguistic comparisons of registers. Such investigations are problematic because apparently similar linguistic features can have quite different functional roles across languages. However, from a comparative register perspective, researchers can first identify the configurations of linguistic features *within* each language that function to distinguish among registers; then, these parameters of variation can be used for cross-linguistic comparison. We briefly summarize an analysis of this type in section 2.2.

## 1 A Register Perspective on Traditional Linguistic Investigations

In general, any functional description of a linguistic feature will *not* be valid for the language as a whole. Rather, characteristics of the textual environment interact with register differences, so that strong patterns of use in one register often represent only

weak patterns in other registers. We illustrate such patterns of use with analyses taken from the *Longman Grammar of Spoken and Written English* (Biber et al. 1999).

For lexical analysis, we illustrate these associations by considering the most common "downtoners" in English (section 1.1). These words are roughly synonymous in meaning, but they have quite different distributions across registers. Further, many of these words have distinctive collocational associations with following adjectives, but those typical collocations also vary in systematic ways across registers.

Similarly distinctive register patterns are typical with grammatical features. We illustrate those associations here by considering the textual factors that influence the omission versus retention of the complementizer *that* in *that*-clauses (section 1.2). It turns out that textual factors are most influential when they run counter to the register norm. For example, the complementizer *that* is usually omitted in conversation, so textual factors favoring the retention of *that* are particularly influential in that register. In contrast, the complementizer *that* is usually retained in news reportage, and as a result, the textual factors favoring the omission of *that* are particularly influential in that register.

Analyses of this type show that there is no single register that can be identified as "general English" for the purposes of linguistic description. Further, dictionaries and grammars based on our intuitions about "general" or "core" English are not likely to provide adequate exposure to the actual linguistic patterns found in the target registers that speakers and writers use on a regular basis.

## *1.1 Register variation in lexical descriptions*

It is easy to demonstrate the importance of register variation for lexical analysis by contrasting the use of near-synonymous words. (See, for example, Biber et al. 1998: chs 2 and 4, on *big*, *large*, and *great*; *little* vs. *small*; and *begin* vs. *start*. See also Kennedy 1991 on *between* and *through*; and Biber et al. 1994 on *certain* and *sure*.)

We illustrate this association here by considering the use of *downtoners* (based on the analyses reported in Biber et al. 1999: ch. 7). Downtoners are adverbs that scale down the effect of a modified item, most often a following adjective. For example:

(1) It did look *pretty* bad. (Conversation)

(2) The mother came away *somewhat* bewildered. (News reportage)

(3) Different laboratories have adopted *slightly* different formulations. (Academic prose)

Downtoners show that the modified item is not to be taken in its strongest sense. For example, in (1)–(3) above, *the way it looked*, *the mother*, and *formulations* do not have the full qualities of *bad*, *bewildered*, and *different*.

Many downtoners are roughly synonymous in meaning. For example, *pretty*, *somewhat*, and *slightly* could be interchanged in sentences (1)–(3) above with little change in meaning. However, it turns out that the most common downtoners have quite different distributions across registers. For the illustration here, we restrict our comparison

**Table 9.1** Distribution of most common downtoners (immediately preceding adjectives) across two registers

| | *Conversation (AmE)* | *Academic prose* |
|---|---|---|
| Pretty | ******** | . |
| Relatively | . | **** |
| Rather | . | ** |
| Fairly | . | ** |
| Slightly | . | ** |
| Almost | . | * |
| Somewhat | . | * |
| Nearly | . | . |

*Notes*: Frequencies are based on analysis of texts from the Longman Spoken and Written English Corpus: *c*.2.5 million words from American English conversation and *c*.5 million words from academic prose. See Biber et al. (1999: ch. 1) for a complete description of the corpus.
Each * represents 50 occurrences per million words; . represents less than 20 occurrences per million words.
*Source*: Adapted from Biber et al. 1999: table 7.13

to two registers defined in relatively general terms: conversation (American English only) and academic prose. As displayed in table 9.1, in conversation, the downtoner *pretty* is very common, while all other downtoners are quite rare.[1] In contrast, academic prose uses a wider range of common downtoners, although none of them is extremely frequent.

Further analysis shows that downtoners are also used for different purposes in conversation and academic prose. For example, the downtoner *pretty* in conversation often occurs as a modifier of evaluative adjectives, as in *pretty good*, *pretty bad*, *pretty cool*, *pretty easy*, *pretty sure*. Typical examples include:

I'm *pretty good* at driving in the snow in my car.

That looks *pretty bad*.

That's a *pretty cool* last name, huh?

Is it a system that would be *pretty easy* to learn?

In contrast, downtoners in academic prose occur with a much wider range of descriptive adjectives. For example, the downtoner *fairly* occurs repeatedly with adjectives such as *resistant*, *consistent*, *constant*, *simple*, *obvious*, *common*, *recent*, and *direct*. Many of the downtoner + adjective collocations in academic prose have to do with marking the extent of comparison between two items (e.g. *slightly smaller*, *somewhat lower*). The downtoner *relatively* always has an implied comparison, as in *relatively simple*, *relatively stable*, *relatively unimportant*. In addition, several downtoners in academic prose commonly occur modifying the adjective *different*, specifying a

comparison that gives the amount of difference (as in *rather different*, *slightly different*, *somewhat different*, etc.). Typical examples include:

> It does seem *fairly common* for children to produce project work consisting entirely of reiterations of knowledge they already have . . .
>
> . . . this regular periodicity of outbreaks suggests that the factors causing fluctuations in these populations are *relatively simple* and tractable . . .
>
> . . . the European study asked a *slightly different* question . . .

A complete description of downtoners obviously requires further analysis and interpretation, based on a fuller consideration of the individual items and a detailed analysis of particular downtoners in their discourse contexts. While it is not possible to undertake such an analysis here, the above discussion has illustrated the central importance of register differences in describing the meaning and use of related words.

## *1.2 Register variation in grammatical descriptions*

Similar to lexical analysis, investigations of grammatical features require a register perspective to fully describe the actual patterns of use. Most grammatical features are distributed in very different ways across registers. For example, among the various types of dependent clause in English, relative clauses are many times more common in academic writing than in conversation, while *that*-complement clauses have the opposite distribution (i.e. much more common in conversation).

There are numerous book-length treatments of grammatical structures from a corpus-based register perspective; for example, Tottie (1991) on negation; Collins (1991) on clefts; Granger (1983) on passives; Mair (1990) on infinitival complement clauses; Meyer (1992) on apposition; and several books on nominal structures (e.g. de Haan 1989; Geisler 1995; Johansson 1995; Varantola 1984). The importance of a register perspective can be further highlighted by considering the distribution and use of roughly equivalent structures (such as *that*-clauses versus *to*-clauses; see Biber et al. 1998: chs 3 and 4).

In the present section, we consider differences in the use of *that*-clauses with the complementizer *that* retained versus omitted (based on analyses reported in Biber et al. 1999: ch. 9). In most *that*-clauses, the complementizer can be freely omitted with no substantial change in meaning. For example, compare:

> I hope *I'm not embarrassing you*. (Conversation)

with

> I hope ***that*** *Paul tells him off*. (Conversation)

However, there are several characteristics of the textual environment that influence the retention versus omission of *that*, and these textual factors interact in important

**Table 9.2** Proportional retention versus omission of the complementizer *that*, by register

| | *% of* that-*clauses with* that *retained* | *% of* that-*clauses with* that *omitted* |
|---|---|---|
| Conversation | *** | ********** |
| Fiction | ******** | ********** |
| News reportage | ********** | ****** |
| Academic prose | ********** | * |

*Notes*: Frequencies are based on analysis of texts from the Longman Spoken and Written English Corpus: *c*.4 million words from British English Conversation, and *c*.5 million words each from Fiction, British News Reportage, and Academic Prose. See Biber et al. (1999: ch. 1) for a complete description of the corpus.
Each * represents 5 percent of the occurrences of *that*-clauses in that register.

ways with register differences. In the following discussion we first review the register patterns for *that* retention versus omission; we then explain textual factors influencing the use of *that*; and we then proceed to describe the association between the register patterns and textual factors.

As table 9.2 shows, different registers have different overall norms for *that* retention versus omission: in conversation, *that*-omission is the typical case, with the complementizer being omitted in *c*.85 percent of all occurrences. At the other extreme, academic prose almost always retains the complementizer *that*.

These overall distributional patterns correspond to the differing production circumstances, purposes, and levels of formality found across registers. Conversations are spoken and produced on-line; they typically have involved, interpersonal purposes; and they are casual and informal in tone. These characteristics are associated with omission rather than retention of *that* as the norm. Academic prose has the opposite characteristics: careful production circumstances; an expository, informational purpose; and a formal tone. Correspondingly, *that* retention is the norm in academic prose.

Textual factors influencing the choice between omission and retention can be divided into two groups:

1 *Textual factors favoring the omission of* that*:*
The omission of *that* is favored when the grammatical characteristics of the surrounding discourse conform to the most common uses of *that*-clauses. To the extent that a construction conforms to the characteristics typically used with *that*-clauses, listeners and readers can anticipate the presence of a *that*-clause without the explicit marking provided by the *that* complementizer.
Two of the most important typical characteristics are:
(a) the use of *think* or *say* as the main clause verb (these are by far the two most common verbs taking a *that*-clause);
(b) the occurrence of coreferential subjects in the main clause and the *that*-clause (which is more common than noncoreferential subjects).

2 *Textual factors favoring the retention of* that:
The retention of *that* is favored with grammatical characteristics that are not typical of *that*-clauses, making these structures difficult to process if the *that* were omitted. Three of the most important such factors are:
(a) the use of coordinated *that*-clauses;
(b) the use of a passive voice verb in the main clause;
(c) the presence of an intervening noun phrase between the main clause verb and the *that*-clause.

For the present discussion, the most interesting aspect of these discourse factors is that they are mediated by register considerations. That is, textual factors are most influential when they operate *counter* to the overall register norm. Table 9.3 describes these patterns for conversation and news reportage.

For instance, because conversation has a strong register norm favoring the omission of *that*, the factors favoring omission have little influence in that register. In contrast, the factors favoring *that* retention are very powerful in conversation. For example:

- with coordinated *that*-clauses:

  Cos every time they use it, she reminds them *that it's her television* **and** *that she could have sold it*.

  I'm sure *they think I'm crazy* **and** *that I'm in love with him or something*.

- with a passive voice verb in the matrix clause:

  I **was told** *that Pete was pissed*.

  About two weeks after that it **was diagnosed** *that she had cancer of the ovary*.

- with an intervening noun phrase between the matrix clause verb and the *that*-clause:

  Then I told **him** *that I'm not doing it anymore*.

  I was busy trying to convince **him** *that he had to go to the doctor*.

  I promised **her** *that I wouldn't play it*.

News reportage shows the opposite tendencies: the overall register norm favors *that* retention and thus the contextual factors favoring retention have comparatively little influence. In contrast, the factors favoring *that* omission are relatively influential in news. The following sentences from news reportage illustrate the most common main verbs, together with coreferential subjects, co-occurring with *that*-omission:

After a month she said (0) *she couldn't cope with it*.

He thought (0) *he was being attacked*.

**Table 9.3** Departure from the register norms for retention versus omission of the complementizer *that*, depending on textual factors

| | ← *Greater proportion of* that *retained than the register norm* | *Greater proportion of* that *omitted than the register norm* → |
|---|---|---|
| **Conversation**: | | |
| A Factors favoring omission: | | |
| Main verb: | | |
| *think* or *say* as matrix verb | | >> |
| Other matrix verb | <<< | |
| Reference of subject: | | |
| Coreferential | | |
| Not coreferential | < | |
| B Factors favoring retention: | | |
| Complex complement: | | |
| Coordinated *that*-clauses | <<<<<<<<<<<<<<< | |
| Simple *that*-clause | | |
| Active/passive main verb: | | |
| Passive | <<<<<<<<<<< | |
| Active | | |
| Presence of indirect object: | | |
| V + NP + *that*-clause | <<<<<<<<<<<<<<< | |
| V + *that*-clause | | |
| **News reportage**: | | |
| A Factors favoring omission: | | |
| Matrix verb: | | |
| *think* or *say* as matrix verb | | >>>> |
| Other matrix verb | << | |
| Reference of subject: | | |
| Coreferential | | >>>>>>>> |
| Not coreferential | <<< | |
| B Factors favoring retention: | | |
| Complex complement: | | |
| Coordinated *that*-clauses | <<<<< | |
| Simple *that*-clause | | |
| Active/passive main verb: | | |
| Passive | <<<<< | |
| Active | < | |
| Presence of indirect object: | | |
| V + NP + *that*-clause | <<<< | |
| V + *that*-clause | < | |

*Notes*: Each < or > represents 5 percent departure from the register norm, for all occurrences of *that*-clauses in that register with the stated textual factor.
< marks proportionally greater use of *that* retention than the register norm.
> marks proportionally greater use of *that* omission than the register norm.

The present section has illustrated several ways in which a register perspective is important for grammatical analysis. First, grammatical features are used to differing extents in different registers, depending on the extent to which the typical functions of the feature fit the typical communicative characteristics of the register. However, there are also much more complex patterns of association, with textual factors interacting with register patterns in intricate ways. Although patterns such as those described here must be interpreted much more fully, the present section has illustrated the systematicity and importance of register patterns in describing the use of related grammatical features.

## 2 Register Comparisons

A major issue for discourse studies since the early 1970s concerns the relationship between spoken and written language. Early research on this question tended to make global generalizations about the linguistic differences between speech and writing. For example, researchers such as O'Donnell (1974) and Olson (1977) argued that written language generally differs from speech in being more structurally complex, elaborated, and/or explicit. In reaction to such studies, several researchers (including Tannen 1982, Beaman 1984, and Chafe and Danielewicz 1986) argued that it is misleading to generalize about overall differences between speech and writing, because communicative task is also an important predictor of linguistic variation; therefore equivalent communicative tasks should be compared to isolate the existence of mode differences.

Multidimensional (MD) analyses of register variation (e.g. Biber 1986, 1988) took this concern one step further by analyzing linguistic variation among the range of registers within each mode, in addition to comparing registers across the spoken and written modes. Further, these analyses included consideration of a wide range of linguistic characteristics, identifying the way that these features configured themselves into underlying "dimensions" of variation. These studies show that particular spoken and written registers are distinguished to differing extents along each dimension.

One potential biasing factor in most early studies of register variation is that they tended to focus on western cultures and languages (especially English). More recently, the MD approach has been used to investigate the patterns of register variation in nonwestern languages. Three such languages have been studied to date: Besnier's (1988) analysis of Nukulaelae Tuvaluan; Kim's (1990; Kim and Biber 1994) analysis of Korean; and Biber and Hared's (1992a, 1992b, 1994) analysis of Somali. Taken together, these studies provide the first comprehensive investigations of register variation in nonwestern languages. Biber (1995) synthesizes these studies, together with the earlier MD analyses of English, to explore cross-linguistic patterns of register variation, and to raise the possibility of cross-linguistic universals governing the patterns of discourse variation across registers.

In the following sections, we briefly describe and compare the patterns of register variation for three of these languages: English, Korean, and Somali.[2] These three languages represent quite different language types and social situations. Thus, they provide a good basis for exploring systematic cross-linguistic patterns of register

variation. In section 2.1 we introduce the multidimensional approach to register variation with specific reference to the MD analysis of English. In 2.2 we then briefly summarize the major patterns of register variation across English, Korean, and Somali.

## 2.1 *Overview of the multidimensional (MD) approach to register variation*

The MD approach to register variation was developed to provide comprehensive descriptions of the patterns of register variation in a language. An MD analysis includes two major components: (1) identification of the underlying linguistic parameters, or *dimensions*, of variation; and (2) specification of the linguistic similarities and differences among registers with respect to those dimensions.

Methodologically, the MD approach has three major distinguishing characteristics: (1) the use of computer-based text corpora to provide a broad representation of the registers in a language; (2) the use of computational tools to identify linguistic features in texts; and (3) the use of multivariate statistical techniques to analyze the co-occurrence relations among linguistic features, thereby identifying underlying *dimensions* of variation in a language. MD studies have consistently shown that there are systematic patterns of variation among registers; that these patterns can be analyzed in terms of the underlying dimensions of variation; and that it is necessary to recognize the existence of a multidimensional space (rather than a single parameter) to adequately describe the relations among registers.

The first step in an MD analysis is to obtain a corpus of texts representing a wide range of spoken and written registers. If there are no pre-existing corpora, as in the case of the Korean and Somali analyses, then texts must be collected and entered into computer. The texts in these corpora are then automatically analysed (or "tagged") for linguistic features representing several major grammatical and functional characteristics, such as: tense and aspect markers, place and time adverbials, pronouns and nominal forms, prepositional phrases, adjectives, adverbs, lexical classes (e.g. hedges, emphatics, speech act verbs), modals, passives, dependent clauses, coordination, and questions. All texts are postedited interactively to correct mis-tags.

Next, the frequency of each linguistic feature in each text is counted. (All counts are normalized to their occurrence per 1000 words of text.) A statistical factor analysis is then computed to identify the co-occurrence patterns among linguistic features, that is, the dimensions. These dimensions are subsequently interpreted in terms of the communicative functions shared by the co-occurring features. Interpretive labels are posited for each dimension, such as "Involved versus Informational Production" and "Narrative versus Non-narrative Concerns." In addition, dimension scores for each text are computed by summing the major linguistic features grouped on each dimension; this score provides a cumulative characterization of a text with respect to the co-occurrence pattern underlying a dimension. Then, the mean dimension scores for each register are compared to analyze the salient linguistic similarities and differences among spoken and written registers.

To illustrate, consider English Dimension 1 in figure 9.1. This dimension is defined by two groups of co-occurring linguistic features, listed to the right of the figure. The top group (above the dashed line) consists of a large number of features, including

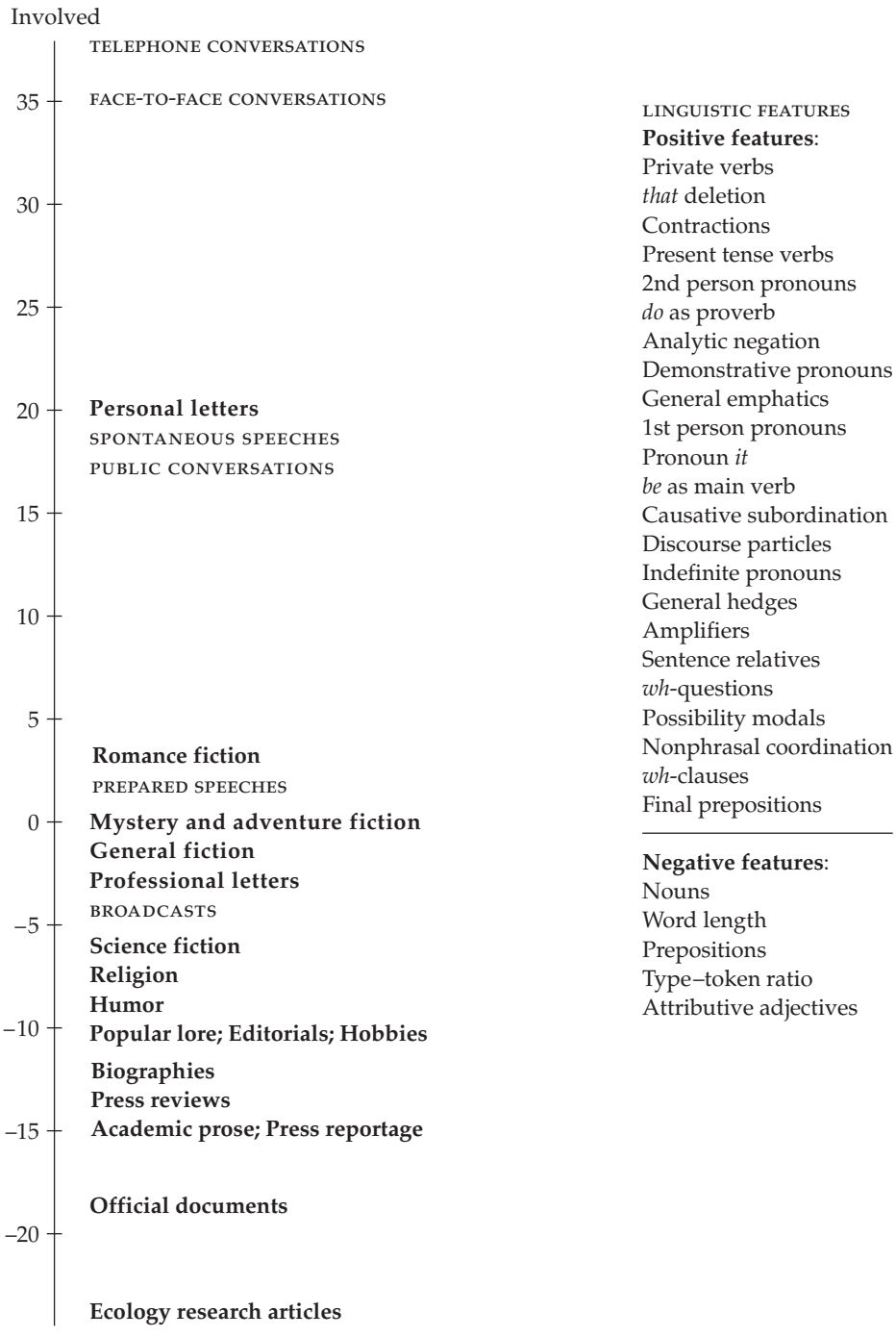


**Figure 9.1** Mean scores of English Dimension 1 for twenty-three registers: "Involved versus Informational Production"

first and second person pronouns, questions, "private" verbs (such as *think* or *know*), and contractions. The bottom group has fewer features, including nouns, attributive adjectives, and prepositional phrases. The statistical analysis shows that these two groups have a complementary relationship and thus constitute a single dimension: when a text has frequent occurrences of the top group of features, it will tend to have few occurrences of the bottom group, and vice versa.

When dimension scores are computed for English Dimension 1, conversation texts are identified as the register that makes the most frequent use of the top group of features. Figure 9.1 plots the Dimension 1 score for several English registers, providing a graphic representation of the relations among registers with respect to this group of linguistic features. Conversation texts, with the largest positive Dimension 1 score, tend to have frequent occurrences of first and second person pronouns, questions, stance verbs, hedges, and the other features above the dashed line; at the same time, relative to the other registers, conversation texts have notably few occurrences of nouns, adjectives, prepositional phrases, and long words. At the other extreme, registers such as official documents and academic prose have the largest negative score, showing that they are marked for the opposite linguistic characteristics: very frequent occurrences of nouns, adjectives, prepositional phrases, and long words, combined with notably few occurrences of first and second person pronouns, questions, stance verbs, etc.

Considering both the defining linguistic features together with the distribution of registers, each dimension can be interpreted in functional terms. Thus, the top group of linguistic features on English Dimension 1, associated most notably with conversation, is interpreted as reflecting interactiveness, high involvement, and on-line production. For example, interactiveness and involvement are reflected in the frequent use of *you* and *I*, and the private verbs that convey the thoughts and feelings of the participants, as well as many other features. The reduced and vague forms – such as contractions, *that* deletions, and general emphatics and hedges – are typical of language produced under real-time constraints. The bottom group of linguistic features, associated most notably with informational exposition, is interpreted as reflecting careful production and an informational focus. That is, as exemplified below, nouns, prepositional phrases, and attributive adjectives all function to convey densely packed information, and the higher type–token ratio and longer words reflect a precise and often specialized choice of words. Such densely informational and precise text is nearly impossible to produce without time for planning and revision.

As noted earlier, one of the advantages of a comparative register perspective is to understand the linguistic characteristics of a particular register relative to a representative range of registers in the language. This advantage can be illustrated with respect to the specific register of research articles in biology (in the subdiscipline of ecology). Figure 9.1 shows that this register is extremely marked on Dimension 1, with a considerably larger negative score than academic prose generally.

Even a short extract from an article shows the high density of informational features from Dimension 1 (nouns are underlined, prepositions italicized, and attributive adjectives capitalized):

> There were MARKED differences *in* root growth *into* regrowth cores *among* the three communities, both *in* the distribution of roots *through* the cores and *in* the

response to ELEVATED $CO^2$. *In* the Scirpus community, root growth was evenly distributed *throughout* the 15-cm profile, *with* no SIGNIFICANT differences *in* root biomass *among* the 5-cm sampling intervals *within* a treatment.

All three of these features serve the purpose of densely packing the text with information about specific referents. Nouns refer to entities or concepts, and are then further specified by prepositional phrases, attributive adjectives, or other nouns which function as premodifiers (e.g. *root growth*). Clearly, the emphasis in this text is on transmitting information precisely and concisely, not on interactive or affective concerns.

Furthermore, by considering the scores of other registers on Dimension 1, we can see that such densely packed informational features are not typical in more colloquial registers of English. For this reason, it is not surprising that many novices experience difficulty when asked to read biology research articles or write up research reports like a professional (cf. Walvoord and McCarthy 1990; Wilkinson 1985). Even with this very brief examination of just one dimension in the MD model of English, we can see why, linguistically, these texts are challenging and why students are unlikely to have had practice with such densely informational prose.

## 2.2 *Comparison of the major oral/literate dimensions in English, Korean, and Somali*

The MD methodological approach outlined in the last section has been applied to the analysis of register variation in English, Korean, and Somali. Biber (1995) provides a full description of the corpora, computational and statistical techniques, linguistic features analyzed, and multidimensional patterns of register variation for each of these languages. That book synthesizes these studies to focus on typological comparisons across languages. Here we present only a summary of some of the more striking cross-linguistic comparisons.

Table 9.4 presents a summary of the major "oral/literate" dimensions in English, Korean, and Somali. Oral/literate dimensions distinguish between stereotypical speech – i.e. conversation – at one pole, versus stereotypical writing – i.e. informational exposition – at the other pole. However, as discussed below, each of these dimensions is composed of a different set of linguistic features, each has different functional associations, and each defines a different set of relations among the full range of spoken and written registers.

The first column in table 9.4 lists the co-occurring linguistic features that define each dimension. Most dimensions comprise two groups of features, separated by a dashed line on table 9.4. As discussed above for Dimension 1 in English, 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.

It should be emphasized that the co-occurrence patterns underlying dimensions are determined empirically (by a statistical factor analysis) and not on any a priori

**Table 9.4** Overview of the major oral/literate dimensions in English, Korean, and Somali

| *Linguistic features* | *Characteristic registers* | *Functional associations* |
|---|---|---|
| **English:** | | |
| Dimension 1: | | |
| 1st and 2nd person pronouns; questions; reductions; stance verbs; hedges; emphatics; adverbial subordination | Conversations<br>(Personal letters)<br>(Public conversations) | Interactive<br>(Inter) personal focus<br>Involved<br>Personal stance<br>On-line production |
| Nouns; adjectives; prepositional phrases; long words | Informational exposition, e.g. official documents, academic prose | Monologue<br>Careful production<br>Informational focus<br>Faceless |
| Dimension 3: | | |
| Time and place adverbials | Broadcasts<br>(Conversations)<br>(Fiction)<br>(Personal letters) | Situation-dependent reference<br>On-line production |
| *wh*-relative clauses; pied-piping constructions; phrasal coordination | Official documents<br>Professional letters<br>Exposition | Situation-independent reference<br>Careful production |
| Dimension 5: | | |
| [No features] | Conversations, fiction<br>Personal letters<br>Public speeches<br>Public conversations<br>Broadcasts | Nonabstract |
| Agentless passives; *by* passives; passive dependent clauses | Technical prose<br>(Other academic prose)<br>(Official documents) | Abstract style<br>Technical, informational focus |
| **Korean:** | | |
| Dimension 1: | | |
| Questions; contractions; fragmentary sentences; discourse conjuncts; clause connectors; hedges | Private conversations<br>TV drama<br>(Public conversations)<br>(Folktales) | Interactive<br>On-line production<br>Interpersonal focus |
| Postposition–noun ratio; relative clauses; attributive adjectives; long sentences; nonfinite and noun complement clauses | Literary criticism<br>College textbooks<br>Scripted speeches<br>Written exposition<br>(Broadcast news and TV documentary) | Monologue<br>Informational focus<br>Careful production |
| Dimension 2: | | |
| Explanative conjuncts; explanative, conditional, coordinate, and discourse clause connectors; | Folktales<br>(Conversations)<br>(Speeches)<br>(Public conversations) | Overt logical cohesion |

**Table 9.4** (*cont'd*)

| *Linguistic features* | *Characteristic registers* | *Functional associations* |
|---|---|---|
| adverbial subordination<br>Nouns; possessive markers; passive constructions | Written expository registers<br>Broadcast news | Implicit logical cohesion<br>Informational focus |
| Dimension 3: | | |
| Verb and NP complements; emphatics; hedges; attitudinal expressions; private verbs; 1st person pronouns | TV drama<br>(Private and public conversations)<br>(Personal letters)<br>(Personal essays) | Personal stance |
| Nouns | Newspaper reportage<br>Official documents<br>(Broadcast news) | Nonpersonal focus |
| **Somali:** | | |
| Dimension 1: | | |
| Main clause features; questions; imperatives; contractions; stance adjectives; downtoners; 1st and 2nd person pronouns | Conversations<br>Family meetings<br>Conversational narratives | Interactive<br>(Inter) personal focus<br>Involved<br>Personal stance<br>On-line production |
| Dependent clauses; relative clauses; clefts; verb complements; nouns; adjectives | Written expository registers | Monologue<br>Informational focus<br>Faceless<br>Careful production |
| Dimension 2:<br>[No features] | Sports broadcast<br>(Other spoken registers) | On-line production<br>Situation-dependent |
| Once-occurring words; high type–token ratio; nominalizations; compound verbs | Editorials<br>Written political speeches and pamphlets<br>Analytical press | Careful production<br>Informational focus |
| Dimension 5: | | |
| Optative clauses; 1st and 2nd person pronouns; directional particles; imperatives | Personal letters<br>(Family meetings)<br>(Quranic exposition) | Interactive<br>Distanced and directive communication |
| [No features] | Press reportage and editorials<br>Written expository registers | Noninteractive<br>Nondirective |

basis. Thus, the dimensions represent those groupings of linguistic characteristics that most commonly co-occur in the spoken and written texts of each corpus. Subsequent to the statistical identification of these co-occurrence patterns, each grouping is interpreted in functional terms, to assess the underlying communicative forces associated with each cluster of linguistic features. The functional associations for each dimension are summarized in the third column of table 9.4.

The dimensions can be used to compare spoken and written registers by computing a "dimension score" for each text (described in 2.1 above). The second column on table 9.4 lists those registers that have the most extreme dimension scores; that is, the registers that use the co-occurring linguistic features on a dimension to the greatest extent.

Table 9.4 summarizes only those dimensions that are closely associated with speech and writing. (Several additional dimensions in each of these languages have little or no association with physical mode.) Each of the dimensions listed in table 9.4 is defined by a different set of co-occurring linguistic features, and each identifies a different overall pattern of relations among registers. However, these dimensions are similar in that they all isolate written expository registers at one extreme (referred to below as the "literate" pole). These registers are formal, edited kinds of text written for informational, expository purposes: for example, official documents and academic prose in English; literary criticism and college textbooks in Korean; and editorials and analytical press articles in Somali.

The opposite extreme along these dimensions (referred to below as the "oral" pole) characterizes spoken registers, especially conversational registers. In addition, colloquial written registers, such as personal letters, are shown to have characteristics similar to spoken registers along several of these dimensions.

Table 9.4 shows that the two extremes of these dimensions are not equally associated with speech and writing: while the "literate" pole of each dimension is associated exclusively with written expository registers, the "oral" pole of many dimensions characterizes written registers, such as letters and fiction, as well as a range of spoken registers. Thus, written registers are characterized by both the "oral" and "literate" poles of English Dimensions 1, 3, and 5, Korean Dimension 3, and Somali Dimension 5.

These patterns indicate that the spoken and written modes provide strikingly different potentials. In particular, writers can produce dense expository texts as well as texts that are extremely colloquial, but speakers do not normally produce texts that are similar to written expository registers. This basic difference holds across all three languages considered here.

It should be emphasized that cross-linguistic similarities are found despite the fact that the statistical techniques used in MD analysis result in independent dimensions: each dimension is defined by a different set of co-occurring linguistic features, and each dimension defines a different set of overall relations among registers. Further, the MD analysis of each language is carried out independently, so there are no methodological factors favoring the identification of analogous dimensions across registers.

Despite this methodological independence, strong similarities emerge across these three languages. For example, three major patterns occur cross-linguistically with respect to the kinds of linguistic expression found exclusively in written expository registers:

1 frequent nouns, adjectives, and prepositional/postpositional phrases, reflecting an extremely dense integration of referential information;
2 high type–token ratio, frequent once-occurring words, and frequent long words, reflecting extreme lexical specificity and complex vocabulary;
3 greater use of nominal structural elaboration, including relative clauses and other nominal modifiers, reflecting elaboration of referential information.[3]

The existence of these linguistic characteristics particular to written exposition can be attributed to the cumulative influence of three major communicative factors (cf. Chafe 1982; Tannen 1982; Biber 1988): (1) communicative purpose, (2) physical relation between addressor and addressee, and (3) production circumstances:

1 *Communicative purpose:* Written expository registers have communicative purposes different from those found in most other registers: to convey information about non-immediate (often abstract) referents with little overt acknowledgement of the thoughts or feelings of the addressor or addressee. Spoken lectures are similar in purpose, but most other spoken registers (and many written registers) are more personal and immediately situated in purpose.
2 *Physical relation between addressor and addressee:* Spoken language is commonly produced in face-to-face situations that permit extensive interaction, opportunity for clarification, and reliance on paralinguistic channels to communicate meaning. Written language is typically produced by writers who are separated in space (and time) from their readers, resulting in a greater reliance on the linguistic channel by itself to communicate meaning.
3 *Production circumstances:* The written mode provides extensive opportunity for careful, deliberate production; written texts can be revised and edited repeatedly before they are considered complete. Spoken language is typically produced on-line, with speakers formulating words and expressions as they think of the ideas.

With respect to the last two of these factors, writing has a greater range of variability than speech. That is, while writing can be produced in circumstances similar to speech, it can also be produced in circumstances quite different from those possible in speech.

With regard to the relation between addressor and addressee, it is possible for readers and writers to be directly interactive (as in personal letters) and even to share the same place and time (e.g. passing notes in class). At the other extreme, though, writers of expository prose typically do not address their texts to individual readers; they rarely receive written responses to their messages; and they do not share physical and temporal space with their readers. In contrast, speaker and hearer must share the same place and time (apart from the use of telephones or tape recorders), and they typically interact with one another to some extent.

Written language is similarly adaptive with respect to production circumstances. At one extreme, written language can be produced in an on-line manner with little preplanning or revision (as in a hasty note or letter). At the other extreme, written texts can be carefully planned and allow for extreme levels of editing and revision. In contrast, while utterances in spoken language can be restated (as with false starts), it is not possible to edit and revise a spoken text.

The written mode thus provides the potential for kinds of language production not possible in typical speech.[4] Written language can be produced at any speed, with any amount of planning, and it can be revised and edited as much as desired. As a result, it is possible to package linguistic structures in writing in ways that cannot be sustained in spoken production.

The linguistic patterns of variation described in this section, taken from three widely different languages, show that the unique production potential of the written mode can be exploited to result in styles of linguistic expression not found in any spoken register. Specifically, expository registers seem to be the kind of writing that develops to maximally exploit the production potential of the written mode, apparently in response to the highly informational communicative purposes. In addition, these unique expository styles have similar linguistic correlates across languages: a dense packaging of nouns, adjectives, and prepositional/postpositional phrases; careful word choice and lexical elaboration; and extensive nominal modification. Further research is required to determine the extent to which these generalizations hold across a broader sample of languages.

## *2.3 Register variation in more specialized domains*

The above discussion of register variation has focused on comparisons between broadly defined spoken and written registers across languages. In addition, MD analysis has also been applied to more specialized domains.

Conrad (1996a, 1996b) applies the MD model of variation in English to a study of disciplinary texts, comparing professional research articles, university-level textbooks, and university student papers in biology and history. The multiple perspectives provided by this analysis highlight similarities between all of these academic texts versus other nonacademic registers, as well as identifying systematic differences across the disciplines and types of texts. The study also highlights discipline-specific literacy demands and trends in writing development as students become more experienced in a discipline.

Reppen (1994, 1995; cf. Biber et al. 1998: ch. 7) uses MD analysis for a study of the spoken and written registers used by elementary school students in English. The study identifies and interprets the dimensions that characterize student registers, finding some dimensions with no counterparts in other MD analyses (such as one interpreted as "Projected scenario"). In addition, comparison of this student MD model and the adult English model discussed in the previous section provides a register perspective on the development of literacy skills.

The MD approach has also been used to study diachronic patterns of register variation in English and Somali. Biber and Finegan (1989, 1997; cf. Biber et al. 1998: ch. 8) trace the development of English written registers (e.g. letters, fiction, newspapers, science prose) and speech-based registers (e.g. drama, dialog in fiction) from 1650 to the present, along three different dimensions of variation. These studies describe a major difference in the historical evolution of popular registers (e.g. fiction, letters, drama) and specialized expository registers (e.g. science prose and medical prose): while popular registers have followed a steady progression toward more "oral" styles (greater involvement; less nominal elaboration; lesser use of passive constructions),

the written expository registers have evolved in the opposite direction, developing styles of expression that were completely unattested in earlier historical periods (e.g. with extremely dense use of elaborated nominal structures and passive constructions). Biber and Finegan (1994b) use this same framework to compare the written styles of particular eighteenth-century authors (Swift, Defoe, Addison, and Johnson) across different registers.

In addition, two studies by Atkinson use the MD approach to trace the evolution of professional registers in English. Atkinson (1992) combines a multidimensional approach with a detailed analysis of rhetorical patterns to study the development of five subregisters of medical academic prose from 1735 to 1985, focusing on the *Edinburgh Medical Journal*. Atkinson (1996) employs a similar integration of multidimensional and rhetorical methodologies to analyze the evolution of scientific research writing, as represented in the *Philosophical Transactions of the Royal Society of London* from 1675 to 1975.

Biber and Hared (1992b, 1994) extend the MD analysis of Somali to study historical change following the introduction of native-language literacy in 1973. Finally, Biber (1995: ch. 8) integrates these diachronic analyses of English and Somali to discuss cross-linguistic similarities and differences in the patterns of historical register change.

## 3 Conclusion

In a chapter of this size, it is impossible to give complete accounts and interpretations of register analyses. Nevertheless, the chapter has illustrated the importance of register variation for diverse aspects of discourse study – whether more traditional descriptions of lexical and grammatical features, or more comprehensive characterizations of registers within a language or across multiple languages. The register perspective illustrated here has repeatedly shown that patterns of language use vary systematically with characteristics of the situational context. As a result, attempts to characterize a language as a whole are likely to misrepresent the actual language use patterns in any particular register.

Clearly, comparisons among registers will play an important role in any thorough description of a language. Furthermore, control of a range of registers is important for any competent speaker of a language. Thus, not only our understanding of discourse but also our understanding of language acquisition and issues within educational linguistics can also benefit from the analysis of register variation.

## NOTES

1 The downtoner *pretty* is much less common in British English (BrE) conversation than in American English (AmE) conversation. In contrast, the adverb *quite* functioning as a modifier is very common in BrE conversation, where it often has a meaning similar to the other downtoners.

2 Nukulaelae Tuvaluan is spoken in a relatively isolated island community

and has a quite restricted range of register variation (only two written registers – personal letters and sermon notes – and five spoken registers). For these reasons, we have not included this study in our discussion here.

3 It is not the case that structural elaboration is generally more prevalent in written registers. In fact, each of these languages shows features of structural dependency distributed in complex ways. Certain types of structural complexity (e.g. adverbial clauses and complement clauses) can be found in conversational registers to a greater extent than written exposition, while nominal modifiers are by far more common in written informational registers (cf. Biber 1992, 1995).

4 Oral literature, such as oral poetry in Somali, represents a spoken register that runs counter to many generalizations concerning speech. The original production of oral poetry depends on exceptional intellectual and verbal ability. While such texts can be extremely complex in their lexical and grammatical characteristics, they also conform to rigid restrictions on language form, including requirements for alliteration, rhythm, and number of syllables per line.

## REFERENCES


Atkinson, D. (1992) The evolution of medical research writing from 1735 to 1985: the case of the *Edinburgh Medical Journal*. *Applied Linguistics*, 13, 337–74.

Atkinson, D. (1996) The *Philosophical Transactions of the Royal Society of London*, 1675–1975: a sociohistorical discourse analysis. *Language in Society*, 25, 333–71.

Atkinson, D., and Biber, D. (1994) Register: a review of empirical research. In Biber and Finegan (1994a), 351–85.

Basso, K. (1974) In R. Bauman and J. Sherzer (eds), *Explorations in the Ethnography of Speaking*. Cambridge: Cambridge University Press, 425–32.

Beaman, K. (1984) Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In D. Tannen (ed.), *Coherence in Spoken and Written Discourse*. Norwood, NJ: Albex, 45–80.

Besnier, N. (1988) The linguistic relationships of spoken and written Nukulaelae registers. *Language*, 64, 707–36.

Biber, D. (1986) Spoken and written textual dimensions in English: resolving the contradictory findings. *Language*, 62, 384–414.

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

Biber, D. (1992) On the complexity of discourse complexity: a multidimensional analysis. *Discourse Processes*, 15, 133–63.

Biber, D. (1994) An analytical framework for register studies. In Biber and Finegan (1994a), 31–56.

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

Biber, D., and Finegan, E. (1989) Drift and the evolution of English style: a history of three genres. *Language*, 65, 487–517.

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

Biber, D., and Finegan, E. (1994b) Multi-dimensional analyses of authors'

styles: some case studies from the eighteenth century. In D. Ross and D. Brink (eds), *Research in Humanities Computing 3*. Oxford: Oxford University Press, 3–17.

Biber, D., and Finegan, E. (1997) Diachronic relations among speech-based and written registers in English. In T. Nevalainen and L. Kahlas-Tarkka (eds), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Helsinki: Societe Neophilologique, 253–75.

Biber, D., and Hared, M. (1992a) Dimensions of register variation in Somali. *Language Variation and Change*, 4, 41–75.

Biber, D., and Hared, M. (1992b) Literacy in Somali: linguistic consequences. *Annual Review of Applied Linguistics*, 12, 260–82.

Biber, D., and Hared, M. (1994) Linguistic correlates of the transition to literacy in Somali: language adaptation in six press registers. In Biber and Finegan (1994a), 182–216.

Biber, D., Conrad, S., and Reppen, R. (1994) Corpus-based approaches to issues in applied linguistics. *Applied Linguistics*, 15, 169–89.

Biber, D., Conrad, S., and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

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

Brown. P., and Fraser, C. (1979) Speech as a marker of situation. In K. Scherer and H. Giles (eds), *Social Markers in Speech*. Cambridge: Cambridge University Press, 33–62.

Bruthiaux, P. (1994) Me Tarzan, You Jane: linguistic simplification in "personal ads" register. In Biber and Finegan (1994a), 136–54.

Bruthiaux, P. (1996) *The Discourse of Classified Advertising*. New York: Oxford University Press.

Chafe, W. (1982) Integration and involvement in speaking, writing, and oral literature. In D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex, 35–54.

Chafe, W., and Danielewicz, J. (1986) Properties of spoken and written language. In R. Horowitz and S. Samuels (eds), *Comprehending Oral and Written Language*. New York: Academic Press, 82–113.

Christie, F. (1991) Pedagogical and content registers in a writing lesson. *Linguistics and Education*, 3, 203–24.

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

Conrad, S. (1996a) Academic discourse in two disciplines: professional writing and student development in biology and history. Unpublished PhD dissertation. Northern Arizona University.

Conrad, S. (1996b) Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education*, 8, 299–326.

Crystal, D. and Davy, D. (1969) *Investigating English Style*. London: Longman.

Duranti, A. (1985) Sociocultural dimensions of discourse. In T. van Dijk (ed.), *Handbook of Discourse Analysis, Vol. 1*. London: Academic Press, 193–230.

Ferguson, C. (1983) Sports announcer talk: syntactic aspects of register variation. *Language in Society*, 12, 153–72.

Firth, J. (1935) The technique of semantics. *Transactions of the Philological Society*, 36–72.

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

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

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

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

Heath, S., and Langman, J. (1994) Shared thinking and the register of coaching. In Biber and Finegan (1994a), 82–105.

Hymes, D. (1974) *Foundations in Sociolinguistics*. Philadelphia: University of Philadelphia Press.

Janda, R. (1985) Note-taking English as a simplified register. *Discourse Processes*, 8, 437–54.

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

Kennedy, G. (1991) *Between* and *through*: the company they keep and the functions they serve. In K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics*. London: Longman, 95–127.

Kim, Y. (1990) Register variation in Korean: a corpus-based study. Unpublished PhD dissertation. University of Southern California.

Kim. Y., and Biber, D. (1994) A corpus-based analysis of register variation in Korean. In Biber and Finegan (1994a), 157–81.

Mair, C. (1990) *Infinitival Complement Clauses in English*. Cambridge: Cambridge University Press.

Martin, J. (1993) Genre and literacy – modeling context in educational linguistics. In *Annual Review of Applied Linguistics, Vol. 13*. New York: Cambridge University Press, 141–72.

Meyer, C. (1992) *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

O'Donnell, R. (1974) Syntactic differences between speech and writing. *American Speech*, 49, 102–10.

Olson, D. (1977) From utterance to text: the bias of language in speech and writing. *Harvard Educational Review*, 47, 257–81.

Reppen, R. (1994) Variation in elementary student language: a multi-dimensional perspective. Unpublished PhD dissertation. Northern Arizona University.

Reppen, R. (1995) A multi-dimensional comparison of spoken and written registers produced by and for students. In B. Warvik, S. Tanskanen, and R. Hiltunen (eds), *Organization in Discourse* (Proceedings from the Turku Conference). Turku, Finland: University of Turku, 477–86.

Romaine, S. (1994) On the creation and expansion of registers: sports reporting in Tok Pisin. In Biber and Finegan (1994a), 59–81.

Tannen, D. (1982) Oral and literate strategies in spoken and written narratives. *Language*, 58, 1–21.

Tottie, G. (1991) *Negation in English Speech and Writing: A Study in Variation*. San Diego: Academic Press.

Varantola, K. (1984) *On Noun Phrase Structures in Engineering English*. Turku: University of Turku.

Walvoord, B., and McCarthy, L. (1990) *Thinking and Writing in College: A Naturalistic Study of Students in Four Disciplines*. Urbana, IL: National Council of Teachers of English.

Wilkinson, A. (1985) A freshman writing course in parallel with a science course. *College Composition and Communication*, 36, 160–5.