2 The comprehensive analysis of register variation

2.1 Requirements of a comprehensive analytical framework for studies of register variation

A comprehensive analysis of register variation in a language must be based on an adequate sampling of registers, texts, and linguistic features:

1 *registers*: the full range of the registers in the language should be included, representing the range of situational variation;

2 *texts*: a representative sampling of texts from each register should be included;

3 *linguistic features*: a wide range of linguistic features should be analyzed in each text, representing multiple underlying parameters of variation.

Although there have been many important register studies, few previous analyses of register variation are comprehensive in this sense.

The restrictions of previous research, and the need for more comprehensive analyses, have been noted by several scholars. Characterizing the state of research in 1974, Hymes notes that 'the fact that present taxonomic dimensions consist so largely of dichotomies – restricted vs. elaborated codes, . . . standard vs. non-standard speech, formal vs. informal scenes, literacy vs. illiteracy – shows how preliminary is the stage at which we work' (1974: 41). This state of affairs was still largely true in the 1980s. Thus, Schafer (1981: 12) finds it 'frustrating' that although previous studies 'are based on texts produced in particular circumstances by only a few subjects . . . speaking and writing in only one situation, this doesn't prevent researchers from offering their results as accurate generalizations of universal difference between speaking and writing'. And Tannen (1982a: 1) writes that:

> Linguistic research too often focuses on one or another kind of data, without specifying its relationship to other kinds. In order to determine which texts are appropriate for proposed research, and to determine the significance of past and projected research, a perspective is needed on the kinds of language and their interrelationships . . . discourse analysis needs a taxonomy of discourse types, and ways of distinguishing among them.

Specifically, an analytical framework for comprehensive register studies should provide tools for analysis of the linguistic characteristics of registers, analysis of the situational characteristics of registers, and analysis of the functional and conventional associations between linguistic and situational characteristics.

abrangência

1 A comprehensive framework should enable analysis of all salient linguistic characteristics of registers, including specification of the co-occurrence relations among the linguistic features themselves. As Crystal and Davy (1969: 13) put it: 'A definitive book on English stylistics would provide a specification of the entire range of linguistic features entering into the definition of what we have been calling a variety of language, as well as a theoretical framework capable of accounting for them'.

coocorrência

Two major types of linguistic characterization can be distinguished. First, there are *register markers*, which are distinctive linguistic features found only in particular registers: for example, the 'count' (balls and strikes) is a linguistic routine found only in broadcasts of baseball games (Ferguson 1983: 165–67). Second, registers are distinguished by *register features*, that is, differing quantitative distributions of core linguistic features (e.g., nouns, pronouns, subordinate clauses). A comprehensive framework should include a specification of the full range of such features, as well as mechanisms for analyzing the relations among features in terms of their patterns of co-occurrence and alternation.

propósito comunicativo

2 A comprehensive framework should enable a complete situational characterization of individual registers, as well as a precise specification of the similarities and differences among registers. Frameworks for the situational characterization of registers have been proposed by Hymes (1974), Duranti (1985), and Biber (1988, 1994).

3 A comprehensive framework should provide formal apparatus to specify the relationship between situational characteristics and linguistic characteristics, as mediated by communicative functions and conventions. Such mechanisms should be able to cope with the continuous nature of register variation. In the multidimensional approach, this requirement is met through the analysis of linguistic co-occurrence patterns, as described in section 2.3 below.

2.2 Linguistic features used for register analyses

Register markers – distinctive indicators of a register – are relatively rare. Many registers are restricted topically, making individual lexical items possible candidates as register markers: for example, the terms *home run* and *inning* are likely to occur in texts about baseball games. In practice, though, lexical choice itself does not typically distinguish a register. Thus the term *home run* could easily occur in a baseball game broadcast, a newspaper article, a personal letter, or a romance novel, among other registers. Grammatical routines, on the other hand, can sometimes serve as distinctive register markers; for example, the phrase *the count is two and one* would provide a fairly distinctive marker of a baseball game broadcast (see Ferguson 1983).[1] Most registers, though, are not reliably distinguished by the presence of register markers.

In contrast, register features are core lexical and grammatical characteristics found to some extent in almost all texts and registers. Register features are pervasive indicators of register distinctions because there are often large differences in their relative distributions across registers. In fact, many registers are distinguished only by a particularly frequent or infrequent occurrence of a set of register features.

Any linguistic feature having a functional or conventional association can be distributed in a way that distinguishes among registers. Such features come from many linguistic classes, including: phonological features (pauses, intonation patterns), tense and aspect markers, pronouns and pro-verbs, questions, nominal forms (nouns, nominalizations, gerunds), passive constructions, dependent clauses (complement clauses, relative clauses, adverbial subordination), prepositional phrases, adjectives, adverbs, measures of lexical specificity (once-occurring words, type–token ratio), lexical classes (hedges, emphatics, discourse particles, stance markers), modals, specialized verb classes (speech act verbs, mental process verbs), reduced forms (contractions, *that*-deletions), co-ordination, negation, and grammatical devices for structuring information (clefts, extraposition).

A comprehensive linguistic analysis of a register requires consideration of a representative selection of linguistic features. Analyses of these register features are necessarily quantitative, because the associated register distinctions are based on differences in the relative distribution of linguistic features. Register markers can be analyzed using qualitative methods, because the mere presence of the marker serves to identify a register. In contrast, register features must be analyzed using quantitative methods, because it is the relative frequency of the feature that serves to identify a register.

2.3 Co-occurrence in register analyses

On first consideration, it seems unlikely that the relative distribution of common linguistic features could reliably distinguish among registers. In fact, individual linguistic features do not provide the basis for such distinctions. However, when analyses are based on the co-occurrence and alternation patterns within a group of linguistic features, important differences across registers are revealed.

The importance of linguistic co-occurrence has been emphasized by linguists such as Firth, Halliday, Ervin-Tripp, and Hymes. Brown and Fraser (1979: 38–39) observe that it can be 'misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the co-occurrence of sets of markers'. Ervin-Tripp (1972) and Hymes (1974) identify 'speech styles' as varieties that are defined by a shared set of co-occurring linguistic features. Halliday (1988:162) defines a register as 'a cluster of associated features having a greater-than-random . . . tendency to co-occur'.

The notion of linguistic co-occurrence has been given formal status in the Multi-Dimensional approach to register variation, where different co-occurrence patterns are analyzed as underlying *dimensions* of variation. In this approach, the co-occurrence patterns comprising each dimension are identified quantitatively, rather than on an *a priori* functional basis. That is, based on the actual distributions of linguistic features in a large corpus of texts, statistical techniques (specifically factor analysis) are used to identify the sets of linguistic features that frequently co-occur in texts. The methods used to identify these co-occurrence patterns are fully described in chapter 5.

It is not the case, though, that quantitative techniques are sufficient in themselves for analyses of register variation. Rather, qualitative techniques are required to interpret the functional bases underlying each set of co-occurring linguistic features. The dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions. A simple example is the way in which first- and second-person pronouns, direct questions, and imperatives are all related to interactiveness. Contractions, false starts, and generalized content words (e.g., *thing*) are all related to the constraints imposed by real-time production. The functional bases of other co-occurrence patterns are less transparent, so that careful qualitative analyses of co-occurring features in particular texts are required to interpret the underlying functions. The functional interpretations of the four MD analyses considered in the present study are given in chapter 6.

2.4 Register as a continuous construct

One of the main distinguishing characteristics of the framework developed here is that it treats register as a continuous rather than discrete construct. From a linguistic perspective, this means that the focus of analysis is on the relative distribution of common linguistic features, in terms of the patterns of co-occurrence and alternation. Registers are not equally well defined in their linguistic characteristics. Some registers (e.g., legal documents) have well-defined norms so that there is relatively little variation among the texts within the register; other registers (e.g., academic prose) are less specified linguistically, so that there are considerable differences among texts within the register (see Biber 1988: chapter 8; Biber 1990). Similarly from a situational perspective, registers are distributed across a continuous range of variation, and they can be defined at different levels of generality (see Biber 1994). It is important to recognize these differences in register comparisons since they determine in part the extent to which any two registers are comparable.

2.5 Corpus linguistics and the computational analysis of register variation

Although automated analyses using computers are not absolutely necessary for comprehensive register analyses, the use of computers does greatly facilitate such analyses. In fact, given that comprehensive analyses require large text samples from many registers, analyzed for a wide range of linguistic features, it is not really feasible for an individual to undertake such a study without the aid of computers.

principaled

Currently there are very large computer-based text *corpora* available – systematic text collections that represent a domain of use within a language. Numerous software tools have been developed to process these text collections, enabling (semi-)automatic linguistic analyses. These tools include:

concordancing programs, which provide lists of the occurrences of some key word and its surrounding context;

large on-line dictionaries, which can provide many kinds of information about individual words: their grammatical category, the relative probabilities of grammatical categories for ambiguous items, and word senses;

part-of-speech taggers, which assign grammatical categories to words in texts, using rules that depend on the surrounding context, or using probabilistic information;

morphological analyzers, which determine the grammatical category of unknown words based on identifiable affixes;

- syntactic analyzers and parsers, which identify various syntactic constructions and determine their boundaries;
- simple counting programs, which compile frequency counts for various linguistic features in texts.

The ~~advantages~~ needs of corpus-based analysis include: → language in use

1 The adequate representation of naturally occurring discourse, including representative text samples from each register. Thus, corpus-based analyses can be based on long passages from each text, and multiple texts from each register.

2 The adequate representation of the range of register variation in a language; that is, analyses can be based on a sampling of texts from a large number of spoken and written registers.

3 The (semi-)automatic linguistic processing of texts, enabling analyses of much wider scope than otherwise feasible. With computational processing, it is feasible to entertain the possibility of a comprehensive linguistic characterization of a text, analyzing a wide range of linguistic features. Further, once the software tools are developed for this type of analysis, it is possible to process all available on-line texts.

4 Much greater reliability and accuracy for quantitative analyses of linguistic features; that is, computers do not become bored or tired – they will count a linguistic feature in the same way every time it is encountered. → reutilização de dimensões

5 The possibility of cumulative results and accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be analyzed using the same computational techniques. Such studies can verify the results of previous research, and findings will be comparable across studies, building a cumulative linguistic description of the language.

There are currently numerous computer-based corpora generally available, and the amount of corpus-based research is steadily increasing. Taylor, Leech, and Fligelstone (1991) survey thirty-six English machine-readable corpora (including information on the availability of each), while Altenberg (1991) has compiled a bibliography of approximately 650 studies based on the major text corpora.

Among the English corpora, the Brown Corpus, Lancaster–Oslo–Bergen Corpus, and London–Lund Corpus are the best known. The Standard sample of Present-Day American English (the Brown Corpus for short), which was completed in 1964, was the first large computer-based text corpus for any language. Work on this corpus began in 1962 at Brown University, supervised by Nelson Francis and Henry Kučera. A tagged version of this corpus, in which each word is marked for its grammatical category, was completed in 1979 (see Francis and Kučera 1979, 1982). This corpus was designed to provide a representative selection of published written texts in American English. Defining the universe of texts as the collection of books and periodicals published in 1961 in the Brown University Library and the Providence Athenaeum, texts were randomly selected from fifteen major registers (e.g., press reportage, press editorial, popular lore, learned and scientific writings, general fiction, science fiction, humor). Texts are about 2,000 words in length, and the entire corpus comprises 500 texts, or a total of approximately 1 million words of running text. Details on the specific texts included in the corpus are given in Francis and Kučera (1979).

In 1978, a parallel corpus of British English was completed, providing a broad sample of written texts published in Britain in 1961. Work on this corpus was carried out at three sites: the University of Lancaster, the University of Oslo, and the Norwegian Computing Centre for the Humanities at Bergen, and thus the corpus is known as the LOB (Lancaster–Oslo–Bergen) Corpus. This corpus has the same basic design as the Brown Corpus: the same fifteen registers, and 500 texts of about 2,000 words each. The LOB corpus manual (Johansson, Leech, and Goodluck 1978) describes the corpus as a whole and provides specifics on the particular texts included. In the case of the LOB Corpus, books were randomly selected from the 1961 publications listed in *The British National Bibliography Cumulated Subject Index, 1960–1964* (which is based on the subject divisions of the Dewey Decimal Classification system), and periodicals and newspapers were randomly selected from the publications listed in *Willing's Press Guide*, 1961. A tagged version of the LOB Corpus became available in the late 1980s.

These two corpora of written texts are complemented by the London–Lund Corpus of Spoken English (Svartvik and Quirk 1980; Svartvik 1990). Based on the spoken texts in the Survey of English Usage compiled at University College London (supervised by Randolph Quirk), this corpus was subsequently computerized at Lund University (supervised by Jan Svartvik). The London–Lund Corpus includes 100 spoken British English texts of about 5,000 words each. The total corpus contains approximately 500,000 words, representing six major spoken registers: private conversations, public conversations (including interviews and panel discussions), telephone conversations, radio broadcasts, spontaneous speeches, and prepared speeches.

Other important English corpora include the Helsinki Diachronic Corpus of English, the Birmingham (COBUILD) Corpus, the Longman/Lancaster English Language Corpus, the British National Corpus, and the Data Collection Initiative. There are also several text collections for other European languages. Engwall (1992) provides a brief survey of French language corpora, including the *Trésor de la langue française*, a collection of literary works totaling well over 70 million words. There are also numerous Swedish language computer-based text collections. Gellerstam (1992) describes eighteen different Swedish corpora, which together comprise more than 20 million words of text from a wide range of spoken and written registers. The SUC corpus is a comprehensive Swedish corpus with a similar design to the Brown Corpus. The Danish–English–French Corpus in Contract Law (Faber and Lauridsen 1991) is a parallel corpus, containing a carefully stratified selection of texts relating to contract law for these three languages.

There are far fewer computer-based corpora of non-western languages, and the three corpora of spoken and written registers used for the present study are among the most comprehensive of these: the corpus of Nukulaelae Tuvaluan compiled by Niko Besnier, the corpus of Korean compiled by Yong-Jin Kim, and the corpus of Somali compiled by the present author and Mohamed Hared. These text collections are described in chapters 3 and 5.

2.6 Theoretical overview of the Multi-Dimensional approach to register variation

The Multi-Dimensional approach to register variation (elsewhere referred to as the Multi-feature/Multi-dimensional approach) was used in the analysis of all four languages considered here. The approach was first used in Biber (1984c, 1985, 1986) and then developed more fully in Biber (1988). Since this approach provides the basis for the present book, I provide a theoretical overview here. Methodological aspects are discussed in detail in chapter 5, and the MD analyses of all four languages are presented in chapter 6.

Some of the general characteristics of the MD approach are:

1 It is corpus-based, depending on analysis of a large number of naturally occurring texts.

2 It is computer-based in that it depends on automated and interactive analyses of linguistic features in texts. This characteristic enables distributional analysis of many linguistic features across many texts and text varieties.

3 The research goal of the approach is the linguistic analysis of texts, registers, and text types, rather than analysis of individual linguistic constructions.

base de
análise = texto

4 The importance of variationist and comparative perspectives are assumed by the approach. That is, the approach is based on the assumption that different kinds of text differ linguistically and functionally, so that analysis of any one or two text varieties is not adequate for conclusions concerning a discourse domain (e.g., speech and writing in English).

5 The approach is explicitly multidimensional. That is, it is assumed

that multiple parameters of variation will be operative in any discourse domain.

6 The approach is quantitative. Analyses are based on frequency counts of linguistic features, describing the relative distributions of features across texts. Multivariate statistical techniques are used to analyze the relations among linguistic features and among texts.[2]

7 The approach synthesizes quantitative and qualitative/functional methodological techniques. That is, the statistical analyses are interpreted in functional terms, to determine the underlying communicative functions associated with each distributional pattern. The approach is based on the assumption that statistical co-occurrence patterns reflect underlying shared communicative functions.

8 The approach synthesizes macroscopic and microscopic analyses. That is, macroscopic investigations of the overall parameters of linguistic variation, which are based on analysis of the distribution of many linguistic features across many texts and registers, are complemented by detailed analyses of particular linguistic features in particular texts.

dimensions

texts where certain dimensions are strongly represented

As noted above, several sociolinguists have emphasized the centrality of linguistic co-occurrence for analyses of registers, genres, or text types (e.g., Ervin-Tripp 1972; Hymes 1974; Brown and Fraser 1979; Halliday 1988). Surprisingly, despite these theoretical discussions, few empirical investigations are based on the analysis of co-occurring linguistic features. Rather the norm has been to compare varieties with respect to a few apparently unrelated linguistic features, with no analysis of the relations among the linguistic characteristics.[3] In part, this shortcoming is due to the fact that the empirical identification of co-occurrence patterns has proven to be quite difficult.

The few researchers who have recognized the importance of co-occurrence relations, such as Chafe (1982; Chafe and Danielewicz 1986) and Longacre (1976), have been forced to resort to their intuitions to posit basic groupings of co-occurring features. Chafe thus identifies two parameters – integration/fragmentation and detachment/involvement – and posits a number of linguistic features associated with each parameter. Longacre also identifies two underlying parameters – projected time and temporal succession – and posits a group of features associated with each. These studies are important in that they recognize the need for analyses based on the co-occurrence relations in texts, and they attempt to identify basic sets of co-occurring linguistic features.

In fact, a large number of comparative sociolinguistic investigations identify a basic dichotomy among registers and propose a set of linguistic features associated with the dichotomy, thus giving at least implicit recognition to the importance of co-occurrence relations. Studies of this type include the following: Ferguson (1959) on 'high' and 'low' diglossic varieties; Bernstein (e.g., 1970) on restricted and elaborated codes; Irvine (1984) on formal and informal registers; Ochs (1979) on planned and unplanned discourse; and numerous studies on speech versus writing.

There are three major theoretical differences between the MD approach and these earlier investigations of register variation. First, apart from the Chafe and Longacre frameworks, most studies have analyzed linguistic variation in terms of a single parameter, suggesting that there is a single basic situational distinction in language (e.g., formality or attention paid to speech) and that all other distinctions are derivative. In contrast, MD studies have demonstrated that no single parameter or dimension is adequate in itself to capture the full range of variation among registers in a language. Rather, different dimensions are realized by different sets of co-occurring linguistic features, reflecting different functional underpinnings (e.g., interactiveness, planning, informational focus and explicitness).

A related difference is that most previous studies have assumed that register variation can be analyzed in terms of simple, dichotomous distinctions, so that varieties are either formal or informal, planned or unplanned, etc. Empirical investigations do not support the existence of such dichotomous distinctions, however. Rather, registers differ from one another by being more or less formal, more or less planned, more or less interactive, etc.; and MD studies have shown that there is a continuous range of linguistic variation associated with each of these distinctions. The dimensions used in MD studies are thus quantitative, continuous parameters of variation, and each dimension is able to distinguish among a continuous range of texts or registers. For this reason, dimensions can be used to analyze the *extent* to which registers are similar (or different).

Finally, in the MD approach dimensions are identified empirically using quantitative statistical techniques, providing a solution to the methodological problem of identifying the salient co-occurrence patterns in a language. There is no guarantee that groupings of features proposed on intuitive grounds actually co-occur in texts: for example, neither Longacre's parameters (see Smith 1985) nor Chafe's parameters (see Redeker 1984) accurately describe sets of linguistic features that actually co-occur regularly in English texts. In contrast, the statistical techniques used in MD studies provide a precise quantitative specification of the co-occurrence patterns among linguistic features in a corpus of texts.

The use of quantitative techniques, however, does not replace the need for qualitative/functional analysis. Rather, the co-occurrence patterns defining each dimension are identified using quantitative/statistical techniques, but the shared functions underlying these co-occurrence patterns must be determined through qualitative analyses of particular texts.

The studies of the four languages considered in the present book show that the MD approach enables comprehensive comparative analyses of the registers within a language. In contrast, most earlier register variation studies had restricted research designs: they typically analyzed only a few registers (and a few texts), with respect to a few, selected linguistic features, and with no empirical investigation of the co-occurrence relations among features. Even though such studies do not provide an adequate research basis for global generalizations concerning the patterns of register variation in a language, global conclusions are commonly presented in a confident manner. As a result, the conclusions of earlier studies have often been contradictory (see Biber 1986; 1988: chapter 3). These contradictions can be reconciled when the full range of registers, linguistic features, and dimensions are considered together in a comprehensive analysis of register variation.

Similarly, the present book shows that conclusions concerning cross-linguistic register differences cannot be based on analysis of individual linguistic features or pairwise comparisons of selected registers. To an even greater extent than for register comparisons within a language, adequate cross-linguistic comparisons must be based on prior analyses of the co-occurrence patterns among linguistic features in each language. As chapter 4 shows, cross-linguistic register comparisons based on individual features are doomed to failure due to indeterminacy concerning the appropriate level of structure to be used in the analysis. The main point here is more basic: even if it were methodologically possible to compare registers across languages with respect to individual features, such comparisons would not provide the basis for general conclusions concerning the cross-linguistic patterns of register variation – for this purpose, a multidimensional analysis incorporating the full range of linguistic features is required.