14

Register and discourse analysis

Douglas Biber

Introduction

One major approach to discourse analysis focuses on the study of language use, describing the ways in which lexical and grammatical features are used in texts (see Schiffrin *et al.*, 2001: 1; Biber *et al.*, 2007: 1–4). Different kinds of texts have different linguistic characteristics, representing systematic patterns of variation that can be investigated under the rubric of *register*: text varieties of a language associated with particular situations of use.

The description of a register includes three major components: the situational context, the typical linguistic features, and the functional relationships between the first two components (Biber and Conrad, 2009: 6–11). The situational context involves description of the circumstances of text production and reception, as well as the relationships among participants. For example: Is the text produced in speech or writing? Is the addressee present, and is communication interactive? What are the primary communicative purposes?

The linguistic analysis includes all lexical and grammatical characteristics that are typical of the text variety. These are usually core linguistic features like nouns, past tense verbs, relative clauses, and so on. The linguistic description of a register requires quantitative analysis to identify the features that are "typical." That is, these linguistic features can occur in any text from any variety. What makes them register features is that they are especially frequent and pervasive in some text varieties in contrast to other varieties.

To give a simple example, nouns and pronouns can be found in any text. However, nouns are extremely frequent in written academic texts but comparatively rare in spoken conversations, while pronouns have the opposite distribution (extremely frequent in conversation; rare in academic writing). Thus compare:

Text sample 1: academic research article

Nouns are underlined; **pronouns are marked in bold italics**

This paper reports an analysis of Tucker's central-prediction-system model and an empirical comparison of ***it*** with two competing models. ***One*** of these competing models is a modification of Tucker's model developed by Bashaw. The other is the standard linear-regression model. The term "central-prediction system" refers to any centralized statistical system for the prediction of academic success at a given educational level from achievement at a previous level. The most common application has been the prediction of college-freshman grade averages from high-school performance for a particular school system. The application of interest to the writer is the

prediction of (college) junior-year achievement from lower-division achievement—especially in the case of the junior-college transfer student.

Text sample 2: conversation [two women with an infant]

Nouns are underlined; **pronouns are marked in bold italics**

A: ***She*** cut herself?
B: ***I***'m not sure
A: Yeah, ***she*** cut her lip.
B: Okay. Oh my gosh—a big fat lip.
A: <sighing> Oh, oh.
B: Oh, ***that*** hurts. <sighing> oww
A: ***You*** want a little ice? a little paper towel?
B: Yeah, ***that*** would be great. This orange juice is not gonna feel good. ***I***'m just gonna put some water in here. ***It*** won't feel good, ***it*** won't feel good, 'cause ***it***'s orange juice.
A: Here, ***it***'ll just help in a little.
B: Let***'s*** put some water in, 'cause maybe ***that*** won't hurt your mouth. 'Cause if ***I*** give ***her*** that bit of orange juice ***that*** really hurts if ***she*** drinks ***that***.
A: Um.

This sample from an academic research article uses only two pronouns (*it, one*), but it has numerous nouns, which often occur in complex noun phrases (e.g., *the prediction of college-freshman grade averages from high-school performance for a particular school system*). In contrast, nearly every utterance in the conversational sample includes one or more pronouns (e.g., *I, you, she, it, that*) but comparatively few nouns.

Linguistic differences of this type are the data that must be explained by the third component of a register analysis: the functional interpretation. That is, one of the central assumptions of register analysis is that linguistic features are always functional: linguistic features tend to occur in a register because they are particularly well-suited to the purposes and situational context of the register.

The functional interpretation attempts to explain linguistic preferences in terms of the situational characteristics. In the above example, there are several important situational differences between the registers, including:

Academic article	*Conversation*
written	spoken
separate physical setting	shared time/place
no interaction	interactive
professional background knowledge	personal background knowledge
time for planning/editing	real-time production
purposes: convey information; document past events	purposes: on-going actions and events; express feelings

With this many situational differences, it is easy to identify potential functional motivations for the linguistic differences described above. That is, pronouns are very common in conversation (as opposed to academic writing) because interlocutors make frequent reference to each other during the interaction (*I, you*) as well as to objects and people in their shared time and place (e.g. *it, he, she,*

that). Pronouns are also used in expressions of personal stance (e.g. *that's great*). From a production point of view, it takes more effort to produce a noun phrase with specific reference than a pronoun with situated reference. For example, the situated pronoun in the utterance *Oh that hurts* would need to be replaced by a fuller noun phrase like *Oh that bad sore on your lip hurts* if the speaker wanted to achieve a more explicit situation-independent reference. Academic writing has the opposite characteristics (e.g. no shared time/place; no interaction or individual addressees; but extensive planning time and a much more "informational" purpose). As a result, we see the dense use of nouns rather than pronouns in academic writing.

The linguistic component of register analysis requires identification of the *pervasive* linguistic features in the variety: linguistic characteristics that might occur in any text but are especially common in the target register. It is these pervasive linguistic features that are clearly functional. As a result, registers can be identified and described based on analysis of either complete texts or a collection of text samples.

Text varieties can also be described by analyzing language features that characterize complete texts, referred to as the *genre* perspective (see Biber and Conrad, 2009: 15–19). Genre analysis corresponds to a second major approach to discourse analysis: consideration of linguistic structure "beyond the sentence" and of the ways in which texts are constructed (see Schiffrin *et al.*, 2001: 1; Biber *et al.*, 2007: 4–6).

Genre features are not pervasive; rather, they might occur only one time in a complete text, often at the beginning or ending of a text. An oft-cited example of genre features is the rhetorical sections that are conventionally used with construct an academic research article: abstract, introduction, methods section, results/discussion, and bibliography (see e.g. Swales, 1990). By convention, these sections are found in most research articles (at least in experimental studies), occurring in this fixed order. Unlike the distribution of nouns and pronouns, genre features often occur only once in a text, and thus they can only be identified through analysis of complete texts.

Genre features are often conventional rather than functional. That is, genre features conform to the social expectations of how a text of a particular type should be constructed, rather than having clear functional associations with the situational context. To give a simple example, by convention we expect the author/speaker to self-identify at the beginning of a text in many genres, including novels, textbooks, research articles, and even telephone conversations. However, in contrast, there is a strong conventional expectation that the author will self-identify at the *end* of a text in a personal letter, an e-mail message, or even a short note left for a friend. In cases like these it is not clear that the placement of the genre feature is directly functional. However, these are important aspects of textual structure.

The following sections will focus mostly on register analysis rather than genre analysis. Section "corpus-based analyses of registers" introduces corpus-based analysis as a research methodology that is particularly well suited for register studies. Section "e-mail messages as a register," then, presents a more detailed case study of a register analysis, focusing on email messages (adapted from Biber and Conrad, 2009, Chapter 7). This case study shows how registers can be investigated at different levels of generality. Thus emails as a general register are first compared to conversation and academic writing, but the case study also shows that it is possible to consider variation among sub-registers of email messages, depending on the relationship between the sender and recipient. This case study illustrates how even small situational differences among registers are associated with systematic linguistic differences.

Finally, section "multi-dimensional studies of register variation" describes the second major type of research question that arises in register studies: investigation of the overall patterns of register variation (rather than detailed descriptions of individual registers). Multi-dimensional analysis is introduced as a research approach designed for research questions of this type.

Corpus-based analyses of registers

Register analyses are often conducted using the methodologies of "corpus linguistics." There are several introductory textbooks that introduce this subfield of linguistics (e.g. McEnery *et al.*, 2006). According to Biber *et al.* (1998: 4), the essential characteristics of corpus-based analysis are:

- it is empirical, analyzing the actual patterns of use in natural texts;
- it utilizes a large and principled collection of natural texts, known as a "corpus," as the basis for analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques.

Several of the advantages of the corpus-based approach come from the use of computers. Computers make it possible to identify and analyze complex patterns of language use on the basis of the consideration of a much larger collection of texts than could be dealt with by hand. Furthermore, computers provide consistent, reliable analyses—they don't change their mind or become tired during a register analysis. Taken together, these characteristics result in a scope and reliability of analysis otherwise not possible. However, the quantitative and computational aspects of corpus analysis do not lessen the need for functional interpretations in register studies. Rather, corpus-based analyses must go beyond simple counts of linguistic features to include qualitative, functional interpretations of the quantitative patterns. In this regard, all register studies follow the same major methodological steps, whether they are corpus-based or not.

In sum, the main contributions of corpus-based research are that it is based on the empirical analysis of a large sample of texts representing a register and, as a result, descriptions are more reliable and valid than analyses based on only a few texts. For these reasons, the case studies illustrated in the following sections all employ corpus-analysis techniques.

E-mail messages as a register

From a register perspective, e-mail messages are interesting because they share some situational characteristics with both conversational registers and written informational registers. For the case study I compiled a mini-corpus of 76 messages that I had received, with a total of 15,840 words. (All proper names except my own have been changed in the examples below.) Like face-to-face conversation, e-mail messages can involve single or multiple recipients, and they can be motivated by many communicative purposes. The corpus used here includes both professional/academic as well as social e-mail messages. However, the corpus was restricted to include only personal/individual e-mail messages: messages written to a single specific person by another person (excluding mass advertising, fraudulent attempts by an anonymous person to obtain money, etc.).

Like conversation, personal e-mail messages are interactive. Addressors normally expect the addressee of a message to respond (at least acknowledging receipt of the message). In addition, addressors in both personal e-mail and conversation convey personal feelings and attitudes. In the mini-corpus studied here, even the authors of workplace e-mails often expressed personal stance, as in:

> It **would be great** to have a lesson on these structures.
> **Hope** you have a **great** trip!
> Well, I find our grammar discussions **very interesting** and **would love** to talk about Tom's writing sample ...

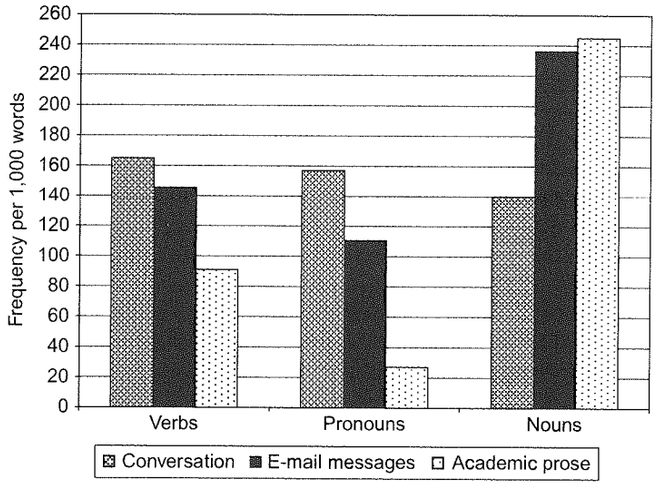

Figure 14.1 The use of major world classes in e-mail messages, compared with conversation and academic prose

At the same time, individual e-mail messages present some important differences from conversations. Conversation is spoken, while e-mail is written and then sent electronically. E-mail is therefore slower than conversation, but it has the potential to be more carefully planned, revised, and edited. In addition, time and space are shared to a lesser extent in e-mail messages than in face-to-face conversations. Physical space is rarely shared in e-mail messages, and an extended email interaction can occur over a period of many weeks, or even months.

In sum, e-mail messages are interpersonal and interactive (similar to conversation), but they are produced in writing, and the sender does not usually share time/place with the addressee (which makes e-mail more like other written registers). The linguistic characteristics of e-mail messages reflect this hybrid combination of situational characteristics.

Figure 14.1 compares the frequency of three basic grammatical features—lexical verbs (e.g. *run*, *want*), pronouns, and nouns—in e-mail messages, conversation, and academic prose. These three features were selected because they illustrate the range of distributions:

Linguistic feature	*Characterization of e-mail messages*
lexical verbs	similar to conversation
nouns	similar to academic prose
pronouns	intermediate

The frequency of lexical verbs in Figure 14.1 shows that e-mail messages incorporate frequent clauses, similar to conversation. For example, notice the relatively short clauses and numerous lexical verbs in the e-mail in Text sample 3:

Text sample 3: e-mail

Lexical verbs in bold

Dr. Biber --

I would love to meet with you in the afternoon on March 10. Anytime is fine. Just name the time and describe directions to your office. I appreciate all of your help in this. I have emailed Sandy Jackson to possibly meet about teaching placements and have been in contact with Andrea. See you in a few weeks!

-- Dora

This linguistic pattern is similar to the conversation sample (Text sample 2, repeated below), but dramatically different from the academic writing sample (Text sample 1, repeated below), which employs only three lexical verbs in a quite long passage:

Text sample 2 [repeated]: conversation

Lexical verbs in bold

A: She **cut** herself?
B: I'm not sure
A: Yeah, she **cut** her lip.
B: Okay. Oh my gosh – a big fat lip.
A: <sighing> Oh, oh.
B: Oh, that **hurts**. <sighing> oww
A: You **want** a little ice? a little paper towel?
B: Yeah, that would be great. This orange juice is not gonna **feel** good. I'm just gonna **put** some water in here. It won't **feel** good, it won't **feel** good, 'cause it's orange juice.
A: Here, it'll just **help** in a little.
B: Let's **put** some water in, 'cause maybe that won't **hurt** your mouth. 'Cause if I **give** her that bit of orange juice that really **hurts** if she **drinks** that.
A: Um.

Text sample 1 [repeated]: academic research article

This paper **reports** an analysis of Tucker's central-prediction-system model and an empirical comparison of it with two competing models. One of these competing models is a modification of Tucker's model **developed** by Bashaw. The other is the standard linear-regression model. The term "central-prediction system" **refers** to any centralized statistical system for the prediction of academic success at a given educational level from achievement at a previous level. The most common application has been the prediction of college-freshman grade averages from high-school performance for a particular school system. The application of interest to the writer is the prediction of (college) junior-year achievement from lower-division achievement—especially in the case of the junior-college transfer student.

Fast production and a focus on specific tasks, activities, and personal stance (rather than concepts) all contribute to the high frequency of lexical verbs in e-mail messages. However, given those characteristics, the higher frequencies of nouns and pronouns in e-mails is surprising. Because e-mail messages are interactive, we might predict that pronouns would be used to the same extent as in conversation. Instead, we find more pronouns in conversation but more nouns in e-mail messages.

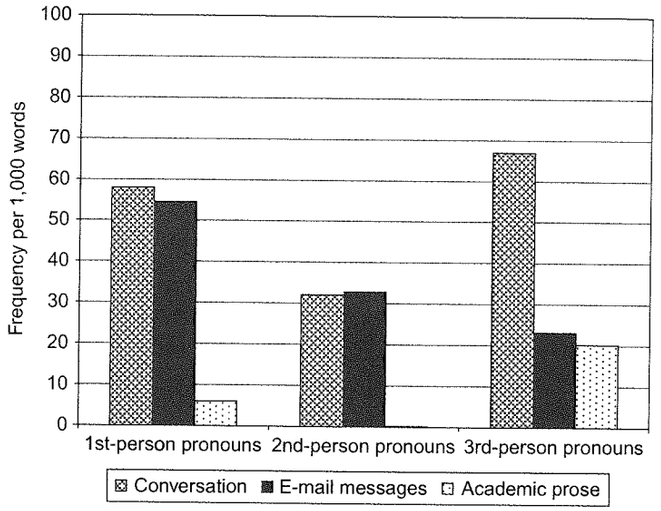

Figure 14.2 The use of pronoun classes, comparing conversation to e-mail messages

More detailed linguistic analyses help to explain these patterns. For example, Figure 14.2 considers the use of pronouns for each person separately: first, second, and third person.

Figure 14.2 shows that e-mail messages are actually very similar to conversations in the use of first-person pronouns (*I, we*) and second-person pronouns (*you*), indicating that these two registers are very similar in their overall interactivity. Text sample 3 above illustrates this dense use of *I* and *you*. In contrast, first-person pronouns are much less common in academic prose, while second-person pronouns are extremely rare in that register.

However, the pattern of use for third-person pronouns is completely different: common in conversation, but relatively rare in both e-mail messages and academic prose. Thus the conversation sample (Text 2) contains numerous occurrences of third-person pronouns (*she, it, that*), while there are few third-person pronouns in either the email or the academic writing passage (samples 1 and 3).

Instead of third-person pronouns, e-mail messages and academic prose both tend to rely on full nouns for third-person references. Sample 1 (above) illustrates this pattern for academic writing, while sample 3 is repeated below highlighting the dense use of nouns in everyday email messages:

Text sample 3 [repeated]: e-mail

nouns in bold

Dr. Biber --

I would love to meet with you in the **afternoon** on March 10. Anytime is fine. Just name the **time** and describe **directions** to your **office**. I appreciate all of your **help** in this. I have emailed **Sandy Jackson** to possibly meet about teaching **placements** and have been in contact with **Andrea**. See you in a few **weeks**!

--Dora

First- and second-person pronouns are common in conversation and individual e-mail messages because both registers have a specific addressor and a specific addressee, and the two interact directly

with one another. However, the frequent use of third-person pronouns in conversation reflects a different situational characteristic: shared time and place. Participants usually do not share the same physical space in e-mail interactions, and often they do not share a temporal context either. As a result, these situated uses of third-person pronouns are much less common in e-mail messages, and full nouns are used instead. Text samples 3 (above) and 4 (below) both illustrate this pattern of use:

Text sample 4: professional e-mail

[third-person pronouns marked in ***bold italics***; nouns underlined]

> Dear Professor Biber,
> Things are moving on for IALCC2004. The Program Committee met yesterday: we received 140 submissions and we have accepted around 90 papers for oral presentation. There will be also some poster presentations, but I do not know the number yet, because the "call for posters" is still open.
>
> I believe we have not talked about the proceedings yet. We plan to publish as usual two volumes of proceedings before the conference (Proceedings are usually distributed at the conference). ***This*** means that the delay is quite short for the editing work and we will have several people working on ***it***. Of course, we would like to include the text of your talk in this book. Would it be possible for you to send us your text by the end of January? I am sorry I did not mention ***that*** to you earlier. I hope the delay will be ok for you.
> <...>

Notice first of all that this message incorporates numerous first- and second-person pronouns, referring directly to the writer (*I*) and the addressee (*you*). However, the message uses comparatively few third-person pronouns, and the ones that do occur are directly anaphoric, referring to the preceding proposition or a noun phrase in the preceding discourse. There are no third-person pronouns in this message that have a vague reference to the general situation or that refer directly to some entity in the writer's physical context. In contrast, there are numerous full nouns, referring to many entities and concepts in an explicit manner. The use of pronouns and nouns thus corresponds to the situational characteristics of high interactivity coupled with the lack of shared physical context.

Variation among sub-registers of e-mail messages

The linguistic characteristics described above apply generally to individual e-mail messages regardless of particular communicative purpose, because those messages are all interactive (with a specific addressor and addressee) but not produced in a shared physical context. In other respects, though, there are important situational differences among sub-registers of e-mail messages, and those differences correspond to systematic linguistic differences. Two parameters that are especially important in this case are the primary purpose/topic of communication, and the social relationship between the addressor and addressee.

To investigate these sub-registers, all e-mail messages in the mini-corpus were classified into three sub-categories: e-mails from friends and family on non-professional topics; e-mails from colleagues/friends on professional topics; and e-mails from "strangers" on professional topics. Table 14.1 shows the breakdown of messages across these categories:

One difference in these e-mail types is immediately clear from Table 14.1: text length. E-mail messages to friends and family on personal topics tend to be much shorter than e-mails on

Table 14.1. Composition of the mini-corpus of individual e-mail messages, classified according to addressee and purpose

Category	*# of messages*	*Total words*	*Average length of message*
friends and family; personal topics	23	2,852	124 words
colleagues/friends; professional topics	32	7,360	230 words
strangers; professional topics	21	5,628	268 words
Total	76	15,840	

professional topics; professional e-mails to strangers tend to be the longest. This difference exists in part because e-mails to friends can assume much more background knowledge, and therefore require much less explanatory prose. At one extreme, there are e-mail exchanges like the following—where people, places and contexts require no explanation:

Text sample 5: two e-mails between friends planning a social get-together

> Doug, climbing gym tomorrow night, 6-ish, Scott
> ok—see you then—Doug

In contrast, professional e-mails to strangers tend to be much longer, because the writers need to introduce themselves (or remind the recipient of who they are), state the reason for writing, provide any necessary background, and frame the whole discussion in a polite manner. Even a quick reminder about a meeting generally has more context than the exchange between friends, for example:

Text sample 6: e-mail from stranger confirming a meeting

> Dr. Biber,
>
> Just wanted to email and confirm that we were still on for meeting at 2:00 tomorrow. Hope to see you then. I don't know if I had CCd you, but I will be meeting with Dr. Bock at 1:30 and Dr. Edwards at 2:30, so it will be a whirlwind tour of the hallway!
>
> If there are any problems, please call me at (111) 241–1925, as I will not have access to email until then. Thanks and I look forward to meeting with you.
>
> Sincerely,
> Donna Johansson

Not surprisingly, workplace e-mails between colleagues/friends tend to fall between these two extremes. Colleagues who interact regularly often write short messages that get directly to the point and assume a great deal of shared background, yet they still require more explanation than close friends continuing a social interaction.

Overall, there is a continuum of linguistic variation among these e-mail sub-registers. For example, Figure 14.3 repeats the information in Figure 14.1, but it distinguishes among the three e-mail sub-registers. Although the linguistic differences among the sub-registers are small, they are entirely consistent: "friends and family" emails are closest to conversation; "professional stranger" emails are closest to academic prose. Figure 14.4 plots the register distributions for a selection of other linguistic features, showing the same consistent patterns, but with the differences among email sub-registers

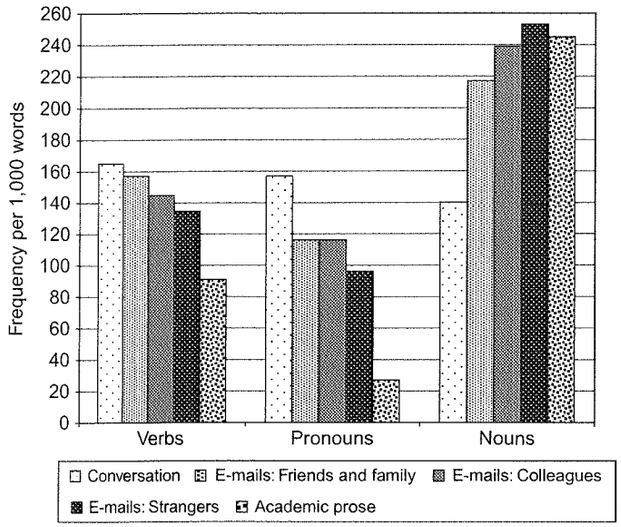

Figure 14.3 The use of major world classes, comparing conversation to e-mail sub-registers

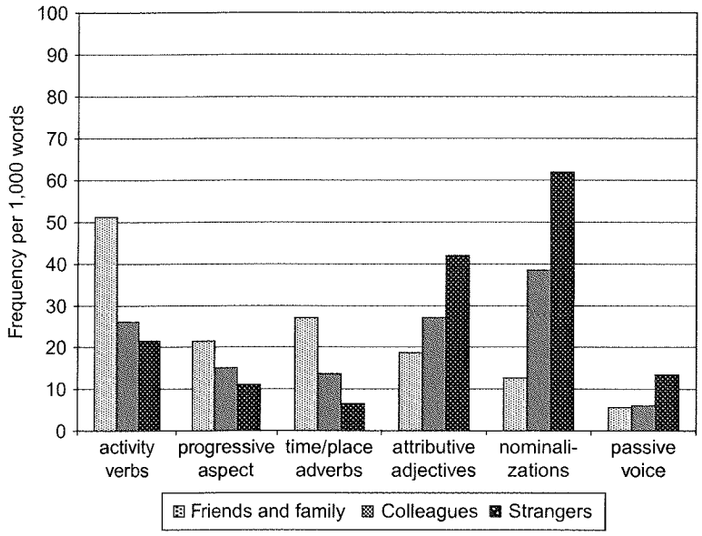

Figure 14.4 The use of selected grammatical characteristics across email sub-registers, depending on the relationship between addressor and addressee

being relatively large for some features. For example, activity verbs and time/place adverbs are much more common in the "friends and family" emails than in the other categories, reflecting the primary focus on everyday activities rather than conceptual discussions. In contrast, attributive adjectives and nominalizations are much more common in the professional emails, especially those written by "strangers," reflecting their informational focus (similar to academic prose).

In sum, the descriptions in this case study illustrate how register can be studied at any level of specificity. At the highest level, register differences can be studied between very general text categories, such as conversation versus academic prose. However, sub-registers can also be defined much more specifically, by focusing on particular situational parameters. The present case has shown how there are systematic patterns of linguistic variation among sub-registers within the general category of email message, depending on the role relation between sender and receiver, and depending on the primary communicative purpose of the message.

Multi-dimensional studies of register variation

The sections above have focused on the description of a particular register (and related sub-registers) with respect to both situational and linguistic characteristics. The second major type of research question that arises in register studies relates to the general patterns of register variation. That is, the distribution of individual linguistic features cannot reliably distinguish among a large set of registers: there are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. Instead, sociolinguistic research has argued that register descriptions must be based on linguistic co-occurrence patterns (see e.g. Ervin-Tripp, 1972; Hymes, 1974; Brown and Fraser, 1979: 38–39; Halliday, 1988: 162).

Multi-dimensional (MD) analysis is a corpus-driven methodological approach that identifies the frequent linguistic co-occurrence patterns in a language, relying on inductive empirical/quantitative analysis (see e.g. Biber, 1988, 1995). The set of co-occurring linguistic features that comprise each dimension is identified quantitatively. That is, on the basis of 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 original MD analyses investigated the relations among general spoken and written registers in English, based on analysis of the LOB Corpus (15 written registers) and the London-Lund Corpus (6 spoken registers). Six/seven different linguistic features were analyzed computationally in each text of the corpus. Then the co-occurrence patterns among those linguistic features were analyzed using factor analysis, identifying the underlying parameters of variation: the factors or "dimensions." In the 1988 MD analysis, the 67 linguistic features were reduced to 7 underlying dimensions. (The technical details of the factor analysis are given in Biber, 1988, Chapters 4–5; see also Biber, 1995, Chapter 5.)

The dimensions are interpreted functionally, on the basis of the assumption that linguistic co-occurrence reflects underlying communicative functions. That is, linguistic features occur together in texts because they serve related communicative functions. For example, Table 14.2 lists the most important features on dimensions 1 and 2 in the 1988 MD analysis.

Each dimension can have "positive" and "negative" features. Rather than reflecting importance, positive and negative signs identify two groupings of features that occur in a complementary pattern as part of the same dimension. That is, when the positive features occur together frequently in a text, the negative features are markedly less frequent in that text, and vice versa.

On dimension 1, the interpretation of the negative features is relatively straightforward. Nouns, word length, prepositional phrases, high type/token ratio, and attributive adjectives all reflect an informational focus, a careful integration of information in a text, and precise lexical choice. Text sample 1 (above) illustrates these co-occurring linguistic characteristics in an academic article.

The set of positive features on dimension 1 is more complex, although all of these features have been associated with interpersonal interaction, a focus on personal stance, and real-time production circumstances. For example first- and second-person pronouns, WH-questions, emphatics, amplifiers, and sentence relatives can all be interpreted as reflecting interpersonal interaction and

Table 14.2. Summary of the major linguistic features co-occurring in dimensions 1 and 2 from the 1988 MD analysis of register variation

Dimension 1: involved vs. informational production
Positive features:
mental (private) verbs, *that* complementizer deletion, contractions, present tense verbs, WH-questions, 1st and 2nd person pronouns, pronoun *it*, indefinite pronouns, *do* as pro-verb, demonstrative pronouns, emphatics, hedges, amplifiers, discourse particles, causative subordination, sentence relatives, WH-clauses
Negative features:
nouns, long words, prepositions, high type/token ratio, attributive adjectives
Dimension 2: narrative vs. non-narrative discourse
Positive features:
past tense verbs, third-person pronouns, perfect aspect verbs, communication verbs
Negative features:
present tense verbs, attributive adjectives

the involved expression of personal stance (feelings and attitudes). Other positive features are associated with the constraints of real-time production, resulting in a reduced surface form, a generalized or uncertain presentation of information, and a generally "fragmented" production of text; these include *that*-deletions, contractions, pro-verb DO, the pronominal forms, and final (stranded) prepositions. Text sample 2 above illustrates the use of many positive dimension 1 features in conversation.

Overall, factor 1 represents a dimension marking interactional, stance-focused, and generalized content (the positive features in Table 14.1) versus high informational density and precise word choice (the negative features). Two separate communicative parameters seem to be represented here: the primary purpose of the writer/speaker (involved versus informational), and the production circumstances (those restricted by real-time constraints versus those enabling careful editing possibilities). Reflecting both of these parameters, the interpretive label "Involved versus Informational Production" was proposed for the dimension underlying this factor.

The second major step in interpreting a dimension is to consider the similarities and differences among registers with respect to the set of co-occurring linguistic features. To achieve this, *dimension scores* are computed for each text, by summing the individual scores of the features that co-occur on a dimension (see Biber, 1988: 93–97). For example, the dimension 1 score for each text was computed by adding together the frequencies of private verbs, *that* deletions, contractions, present tense verbs, etc. – the features with positive loadings (from Table 14.1)—and then subtracting the frequencies of nouns, word length, prepositions, and so on—the features with negative loadings.

Once a dimension score is computed for each text, the mean dimension score for each register can be computed. Plots of these mean dimension scores allow linguistic characterization of any given register, comparison of the relations between any two registers, and a fuller functional interpretation of the underlying dimension. For example, Figure 14.5 plots the mean dimension scores of registers along dimension 1 from the 1988 MD analysis.

The relations among registers shown in Figure 14.5 confirm the interpretation of dimension 1 as distinguishing among texts along a continuum of involved versus informational production. There is a large range of variation among spoken registers along this dimension, and an even larger range of variation among written registers. For example, expository informational registers, like

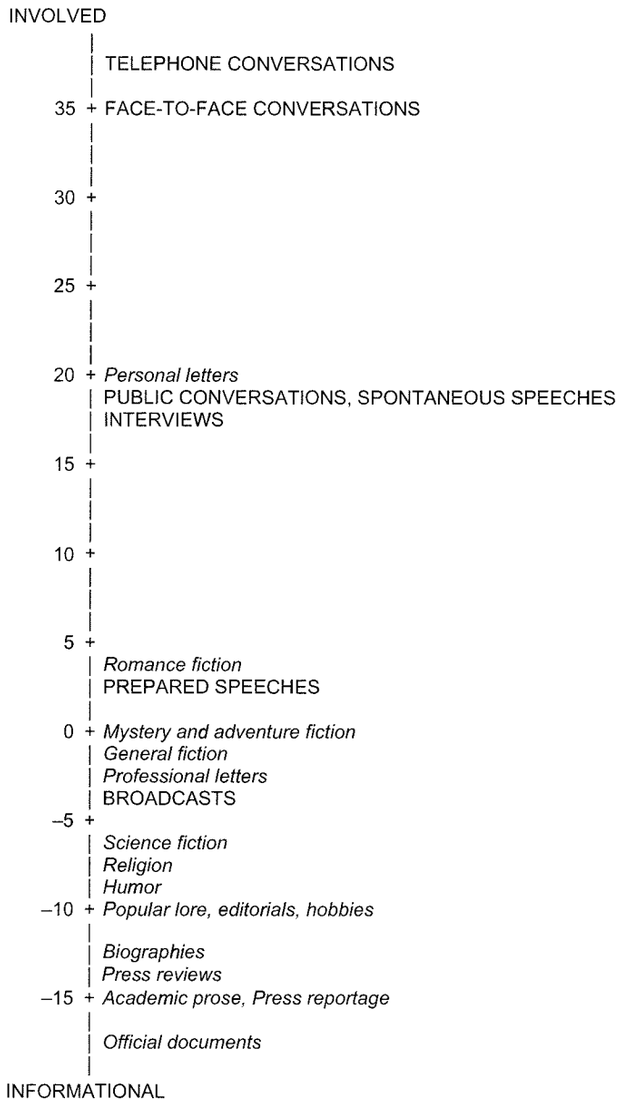

Figure 14.5. Mean scores of registers along dimension 1: involved vs informational production written registers are in *italics*; spoken registers are in CAPS. ($F = 111.9$, $p < .0001$, $r^2 = 84.3\%$) (adapted from Figure 7.1 in Biber, 1988)

official documents and academic prose, have very large negative scores; the fiction registers have scores around 0.0; while personal letters have a relatively large positive score.

This distribution shows that no single register can be taken as representative of the spoken or written mode. At the extremes, written informational prose is dramatically different from spoken conversation with respect to dimension 1 scores. But written personal letters are relatively similar

to spoken conversation, while spoken prepared speeches share some dimension 1 characteristics with written fictional registers. Taken together, these dimension 1 patterns indicate that there is extensive overlap between the spoken and written modes in these linguistic characteristics, while the extremes of each mode (i.e. conversation versus informational prose) are sharply distinguished from each other.

The overall comparison of speech and writing resulting from the 1988 MD analysis is actually much more complex, because six separate dimensions of variation were identified, and each of these defines a different set of relations among spoken and written registers. For example, dimension 2 is interpreted as "narrative vs. non-narrative concerns." The positive features—past tense verbs, third-person pronouns, perfect aspect verbs, communication verbs, and present participial clauses—are associated with past time narration. In contrast, the positive features—present tense verbs and attributive adjectives—have non-narrative communicative functions.

Each of the dimensions in the analysis can be interpreted in a similar way. Overall, the 1988 MD analysis showed that English registers vary along several underlying dimensions associated with different functional considerations, including: interactiveness, involvement and personal stance, production circumstances, informational density, informational elaboration, narrative purposes, situated reference, persuasiveness or argumentation, and impersonal presentation of information.

Many studies have applied the 1988 dimensions of variation to study the linguistic characteristics of more specialized registers and discourse domains. For example:

Present-day registers:	*Studies:*
spoken and written university registers	Biber *et al.* (2002)
AmE versus BrE written registers	Biber (1987)
AmE versus BrE conversational registers	Helt (2001)
student vs. academic writing (biology, history)	Conrad (1996)
direct mail letters	Connor and Upton (2003)
oral proficiency interviews	Connor-Linton and Shohamy (2001)
academic lectures	Csomay (2005)
conversation versus TV dialogue	Quaglio (2009)
female/male conversational style	Rey (2001); Biber and Burges (2000)
Historical registers:	*Studies:*
written and speech-based registers; 1650-present	Biber and Finegan (1989; 2001)
medical research articles and scientific research articles; 1650-present	Atkinson (1992, 1999)

Numerous other studies have undertaken new MD analyses, using factor analysis to identify the dimensions of variation operating in a particular discourse domain in English rather than applying the dimensions from the 1988 MD analysis (e.g. Biber, 2001, 2006, 2008; Reppen, 2001; Biber and Jones, 2005; Biber *et al.*, 2007; Friginal, 2009).

Given that each of these studies is based on a different corpus of texts, representing different registers, it is reasonable to expect that they would each identify a unique set of dimensions. This expectation is reinforced by the fact that the more recent studies have included additional linguistic features not used in earlier MD studies (e.g. semantic classes of nouns and verbs). However, despite these differences in design and research focus, there are certain striking similarities in the set of dimensions identified by these studies.

Most importantly, in nearly all of these studies the first dimension identified by the factor analysis is associated with an informational focus versus a personal focus (personal involvement/stance, interactivity, and/or real time production features). This parameter of variation has emerged in the study of many different discourse domains, including general spoken and written registers (Biber, 1988), university spoken and written registers (Biber, 2006), and eighteenth-century speech-based and written registers (Biber, 2001). Surprisingly, a similar dimension has emerged in studies restricted to only spoken registers, such as White's (1994) study of job interviews and Biber's (2008) study of conversational sub-registers.

A second parameter found in most MD analyses corresponds to narrative discourse, reflected by the co-occurrence of features like past tense, third-person pronouns, perfect aspect, and communication verbs (see e.g. the Biber, 2006 study of university registers; Biber, 2001 on eighteenth century registers; and the Biber, 2008 study of conversation text types). In some studies a similar narrative dimension emerged, with additional special characteristics. For example, in Reppen's (2001) study of elementary school registers, "narrative" features like past tense, perfect aspect, and communication verbs co-occurred with once-occurring words and a high type/token ratio; in this corpus history textbooks rely on a specialized and diverse vocabulary to narrate past events. In Biber and Kurjian's (2007) study of web text types, narrative features co-occurred with features of stance and personal involvement on the first dimension, distinguishing personal narrative web pages (e.g. personal blogs) from the various kinds of more informational web pages.

At the same time, most of these studies have identified some dimensions that are unique to the particular discourse domain. For example, Biber's (2006) study of university spoken and written registers identified two specialized dimensions: "procedural vs. content-focused discourse" (distinguishing between classroom management talk and course syllabi versus textbooks), and "academic stance" (especially prevalent in classroom teaching and classroom management talk). A second example comes from Biber's (2008) MD analysis of conversational text types, which identified a dimension of "stance-focused versus context-focused discourse."

In sum, MD studies of English registers have uncovered both surprising similarities and notable differences in the underlying dimensions of variation. Two parameters seem to be fundamentally important, regardless of the discourse domain: a dimension associated with informational focus versus (inter)personal focus, and a dimension associated with narrative discourse. At the same time, these MD studies have uncovered dimensions particular to the communicative functions and priorities of each different domain of use.

These same general patterns have emerged from MD studies of languages other than English, including Nukulaelae Tuvaluan (Besnier, 1988), Korean (Kim and Biber, 1994), Somali (Biber and Hared, 1992), and Spanish (Biber *et al.*, 2006; Parodi, 2007). Taken together, these studies provide the first comprehensive investigations of register variation in non-Western languages.

Biber (1995) synthesizes several of these studies to investigate the extent to which the underlying dimensions of variation and the relations among registers are configured in similar ways across languages. These languages show striking similarities in their basic patterns of register variation, as reflected by:

- the co-occurring linguistic features that define the dimensions of variation in each language;
- the functional considerations represented by those dimensions; and
- the linguistic/functional relations among analogous registers.

For example, similarly to the full MD analyses of English, these MD studies have all identified dimensions associated with informational versus (inter)personal purposes and with narrative discourse.

At the same time, each of these MD analyses has identified dimensions that are unique to a language, reflecting the particular communicative priorities of that language and culture. For example, the MD analysis of Somali identified a dimension interpreted as "distanced, directive interaction," represented by optative clauses, first- and second-person pronouns, directional pre-verbal particles, and other case particles. Only one register is especially marked for the frequent use of these co-occurring features in Somali: personal letters. This dimension reflects the particular communicative priorities of personal letters in Somali, which are typically interactive as well as explicitly directive.

Conclusion

This chapter has surveyed the ways in which situational and linguistic differences distinguish among registers. Registers differ with respect to a wide array of situational characteristics relating to purpose, topic, physical setting, production circumstances, and the relations among participants. These situational differences are associated with important linguistic differences at the lexical, grammatical, and lexico-grammatical levels. Further, corpus-based analytical techniques can be employed to identify the linguistic co-occurrence patterns that regularly occur in texts from different registers, providing the basis for comprehensive analyses of register variation.

All language users adapt their language to different situations of use. It would be nearly impossible to spend an entire day using only one register – only participating in conversations, only listening to radio broadcasts, only reading a newspaper, or only writing an academic paper. Rather, switching among registers is as natural as human language itself. As a result, understanding register variation is not a supplement to the description of grammar, discourse, and language use; it is central.

Further reading

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

This is the first major study of register variation to apply multi-dimensional analysis. The book identifies and interprets the major dimensions of variation among spoken and written registers in English.

Biber, Douglas and Susan Conrad (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

This book describes the most important kinds of texts in English and introduces the methodological techniques used to analyse them. Three analytical approaches are introduced and compared throughout the book, describing texts from the perspective of register, genre and style.

Friginal, Eric (2009). *The Language of Outsourced Call Centers*. Amsterdam: John Benjamins.

This is one of the first books to undertake a comprehensive linguistic description of an emerging register. The book describes the register of call-center discourse at multiple linguistic levels, including a survey of lexico-grammatical features, detailed descriptions of stance features, and a multi-dimensional analysis that captures the underlying parameters of variation.

Quaglio, Paulo (2009) *Television Dialogue: The Sitcom Friends versus Natural Conversation*. Amsterdam: John Benjamins.

This book presents a corpus-based description of the popular TV sitcom Friends compared to normal face-to-face conversations. The book offers a thorough linguistic description of the television sitcom register, including in-depth chapters that focus on vague language, the expression of personal emotion, informal language (including slang and expletives), and a comparison of narrative features in *Friends* versus natural conversation.

References

Atkinson, D. (1992) 'The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal', *Applied Linguistics*, 13: 337–374.

Atkinson, D. (1999) *Scientific Discourse in Sociohistorical Context: The Philosophical Transactions of the Royal Society of London, 1675–1975*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Biber, D. (1987) 'A textual comparison of British and American writing', *American Speech*, 62: 99–119.

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

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

Biber, D. (2001) 'Dimensions of variation among eighteenth-century speech-based and written registers', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 200–214.

Biber, D. (2006) *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D. (2008) 'Corpus-based analyses of discourse: dimensions of variation in conversation', in V. Bhatia, J. Flowerdew, and R. Jones (eds.) *Advances in Discourse Studies*. London: Routledge, pp. 100–114.

Biber, D., and Burges, J. (2000). 'Historical change in the language use of women and men: gender differences in dramatic dialogue', *Journal of English Linguistics*, 28: 21–37.

Biber, D., Connor, U., and Upton, T. A. (2007) *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Biber, D. and Conrad, S. (2009) *Register, Genre, and Style*. Cambridge: Cambridge University Press.

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

Biber, D., Conrad, S., Reppen, R., Byrd, P., and Helt, M. (2002) 'Speaking and writing in the university: a multi-dimensional comparison', *TESOL Quarterly*, 36: 9–48.

Biber, D., Davies, M., Jones, J. K., and Tracy-Ventura, N. (2006) 'Spoken and written register variation in Spanish: a multi-dimensional analysis', *Corpora*, 1: 7–38.

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. (2001) 'Diachronic relations among speech-based and written registers in English', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 66–83.

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

Biber, D. and Jones, J. K. (2005) 'Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles', *Corpus Linguistics and Linguistic Theory*, 1: 151–182.

Biber, D. and Kurjian, J. (2007) 'Towards a taxonomy of web registers and text types: a multi-dimensional analysis', in M. Hundt, N. Nesselhauf, and C. Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi, pp. 109–132.

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

Connor, U. and Upton, T. A. (2003) 'Linguistic dimensions of direct mail letters', in C. Meyer and P. Leistyna (eds.) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, pp. 71–86.

Connor-Linton, J. and Shohamy, E. (2001) 'Register variation, oral proficiency sampling, and the promise of multi-dimensional analysis', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 124–137.

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

Csomay, E. (2005) 'Linguistic variation within university classroom talk: a corpus-based perspective', *Linguistics and Education*, 15: 243–274.

Ervin-Tripp, S. (1972) 'On sociolinguistic rules: alternation and co-occurrence', in J. Gumperz and D. Hymes (eds.) *Directions in Sociolinguistics: The Ethnography of Communication*. New York: Holt, pp. 213–250.

Friginal, E. (2009) *The Language of Outsourced Call Centers*. Amsterdam: John Benjamins.

Halliday, M. A. K. (1988) 'On the language of physical science', in M. Ghadessy (ed.) *Registers of Written English: Situational Factors and Linguistic Features*. London: Pinter, pp. 162–178.

Helt, M. E. (2001) 'A multi-dimensional comparison of British and American spoken English', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 157–170.

Hymes, D. (1974) *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia, PA: University of Pennsylvania Press.

Kim, Y. -J and Biber, D. (1994) 'A corpus-based analysis of register variation in Korean', in D. Biber and E. Finegan (eds.) *Sociolinguistic Perspectives on Register*. New York: Oxford University Press, pp. 157–181.
McEnery, A., Xiao, R., and Tono, Y. (2006) *Corpus-Based Language Studies*. London: Routledge.
Parodi, G. (2007) 'Variation across registers in Spanish', in G. Parodi (ed.) *Working with Spanish Corpora*. London: Continuum, pp. 11–53.
Quaglio, P. (2009) *Television Dialogue: The Sitcom Friends Versus Natural Conversation*. Amsterdam: John Benjamins.
Reppen, R. (2001) 'Register variation in student and adult speech and writing', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 187–199.
Rey, J. M. (2001) 'Historical shifts in the language of women and men: gender differences in dramatic dialogue', in S. Conrad and D. Biber (eds.) *Variation in English: Multi-Dimensional Studies*. London: Longman, pp. 138–156.
Schiffrin, D., D. Tannen, and Hamilton, H. E. (eds.) (2001) *The Handbook of Discourse Analysis*. Oxford: Blackwell.
Swales, J. (1990) *Genre Analysis: English for Academic and Research Settings*. Cambridge: Cambridge University Press.
White, M. (1994). 'Language in job interviews: differences relating to success and socioeconomic variables', Ph.D. Dissertation, Northern Arizona University.

15

Genre in the Sydney school

David Rose

Genre and register: a stratal model of language in social context

Genre is the coordinating principle and starting point for discourse analysis in what has become known as the Sydney School (Martin, 2000, 2006; Martin and Rose, 2005). The approach has been designed over the past three decades with three major influences (among others): Halliday's (1975, 1994/2004) theory of language as a social semiotic (discussed by Schleppegrel in this volume; Martin, 1992; Martin and Rose, 2007, 2008); the sociological theory of Basil Bernstein (1990, 2000; see Christie and Martin, 1997); and a series of large-scale action research projects in literacy education (Martin, 1999, 2000; Rose, 2008; Rose and Martin, in press). The functional linguistic perspective on genre analysis distinguishes the Sydney School approach along several lines. With respect to linguistic models, its perspective is social rather than cognitive, its analysis of social contexts is social semiotic rather than ethnographic commentary, and it is designed along multiple dimensions as a stratified, metafunctional, multimodal theory of text in social context rather than eclectic. In relation to other fields, it is integrated in a functional theory of language rather than interdisciplinary, and its social goals are interventionist and focused on redistributing semiotic resources through education, rather than merely critical of those in power. With respect to the breadth and detail of its linguistic focus and its uniquely designed teaching strategies, Hyland (2007: 153) describes the Sydney School as 'perhaps the most clearly articulated approach to genre both theoretically and pedagogically' (see also Hyon, 1996; Johns, 2002).

As a working definition, genres have been characterized in this research tradition as staged, goal oriented social processes: social since texts are always interactive events; goal oriented in that a text unfolds towards its interactants' purposes; staged, because it usually takes more than one step to reach the goal. In functional linguistics terms this means that genres are defined as a recurrent configuration of meanings, which enact the social practices of a culture. Such a social semiotic interpretation necessitates going beyond individual genres, to consider how they relate to one another. For example, genres can be related and distinguished by recurrent global patterns. Thus story genres can be distinguished according to the presence or absence of sequence in time (news reports vs other stories) and the presence or absence of a complicating event (recount vs narrative); factual genres, according to whether they explain processes or describe things (explanation vs report); argument genres according to whether they argue for a point of view or discuss two or more points of view (exposition vs discussion). Secondly, the organization of each genre can be distinguished by recurrent local patterns, such as the narrative stages Orientation^Complication^Resolution, or the exposition stages Thesis^Arguments^Reiteration.

The range of genres described in the Sydney School research is large and diverse, but it is still just a fraction of the repertoire of genres available to members of a culture. This chapter presents a