

Jean Aitchison <i>London</i>	Roger Fowler <i>Norwich</i>	Pieter Muysken <i>Amsterdam</i>
Laurie Bauer <i>Wellington</i>	John Green <i>Bradford</i>	Joachim Neuhaus <i>Münster</i>
Geoffrey Beattie <i>Sheffield</i>	François Grosjean <i>Boston, Mass.</i>	Frans Plank <i>Konstanz</i>
Johannes Bechert <i>Bremen</i>	Hartmut Haberland <i>Roskilde</i>	Rebecca Posner <i>Oxford</i>
Claire Blanche-Benveniste <i>Aix-en-Provence</i>	Stephen Harlow <i>York</i>	Geoffrey K. Pullum <i>Santa Cruz, Ca.</i>
Geert E. Booij <i>Amsterdam</i>	Leslie Henderson <i>Hatfield</i>	S. G. Pulman <i>Cambridge</i>
Peter Bosch <i>Stuttgart</i>	John Hinds <i>University Park, Pa.</i>	Tanya Reinhart <i>Tel Aviv</i>
Georg Bossong <i>München</i>	Richard Hudson <i>London</i>	Suzanne Romaine <i>Oxford</i>
Melissa Bowerman <i>Nijmegen</i>	Harry van der Hulst <i>Leiden</i>	Pieter Seuren <i>Nijmegen</i>
Joan Bybee <i>Buffalo, N.Y.</i>	Tore Janson <i>Stockholm</i>	Petr Sgall <i>Prague</i>
Eve Clark <i>Stanford, Ca.</i>	Eric Kellerman <i>Nijmegen</i>	Neil Smith <i>London</i>
Richard Coates <i>Brighton</i>	Ewan Klein <i>Newcastle</i>	Arnim von Stechow <i>Konstanz</i>
René Collier <i>Antwerp</i>	Ekkehard König <i>Berlin</i>	Mark Steedman <i>Edinburgh</i>
Greville G. Corbett <i>Guildford</i>	Robert Ladd <i>Edinburgh</i>	Keith Stenning <i>Edinburgh</i>
Andrew Crompton <i>Nottingham</i>	Aditi Lahiri <i>Nijmegen</i>	Theo Vennemann <i>München</i>
Anne Cutler <i>Cambridge</i>	Stephen Levinson <i>Cambridge</i>	Brian Wenk <i>Nijmegen</i>
Simon Dik <i>Amsterdam</i>	Per Linell <i>Linköping</i>	Dieter Wunderlich <i>Düsseldorf</i>
Elisabet Engdahl <i>Edinburgh</i>	James McCawley <i>Chicago, Ill.</i>	Arnold Zwicky <i>Columbus, Ohio</i>

*Abstract*

*This paper develops a typology of texts in English with respect to a five-dimensional model of variation. Each dimension comprises a set of lexical and syntactic features that cooccur frequently in texts, reflecting underlying shared communicative functions. Eight text types are identified with respect to these dimensions; each type represents a grouping of texts that are markedly similar to one another with respect to their dimension characterizations. Although the linguistic variation among texts is continuous, there are only a few linguistic characterizations that occur frequently, and the text types identified here represent the prototypical groups of texts having these characterizations. The types are interpreted by considering their predominant linguistic features and the general communicative characteristics of the texts grouped in each type, and by performing microanalyses of particular texts. Based on these interpretations, functional labels, such as 'Informational interaction', 'Learned exposition', and 'Involved persuasion', are proposed for each type.*

**1. Introduction**

Over the last several years, numerous studies have attempted to document the nature and extent of linguistic similarities and differences among various kinds of texts. A major goal of such research is to develop an overall typology of texts, to provide a theoretical and empirical foundation for comparative discourse research. This need has been emphasized by Tannen (1982: 1):

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.

It is easy to illustrate the need for a typology of texts; many discourse studies analyze particular sets of texts without specifying their relations to other kinds of texts, often making the unwarranted assumption that findings can be generalized to 'discourse' as a whole. For example, there have been numerous contradictory conclusions concerning the linguistic characteristics of speech and writing due to this methodological shortcoming (many studies compare only face-to-face conversation and academic exposition but assume that their results characterize all speech and writing; see Tannen 1982; Biber 1986). Similarly, contradictory claims are common concerning the linguistic characteristics of 'complex' versus 'simple' discourse or 'formal' versus 'informal' discourse (Finegan and Biber 1986; Besnier 1986). In fact, a typology of texts is a research prerequisite to any comparative register analysis, whether of speech and writing, formal and informal texts, restricted and elaborated codes, literary and colloquial styles, 'good' and 'bad' student compositions, early and late historical periods, or whatever, to situate particular texts relative to the range of texts in English.

There have been a number of text typologies proposed within linguistics and related fields. Researchers have typically developed typologies on a functional basis: first identifying one or two particular functional dichotomies, and then describing the 'types' defined by the poles of those distinctions. For example, much of the research on spoken/written differences can be considered as implicitly typological in this way, where the mode differences between speech and writing are claimed to distinguish between oral and literate text types. More explicit typologies have been proposed by Longacre (1976) and Chafe (1982). Longacre proposes a four-way distinction of 'monologic' texts with respect to the parameters of projected time and temporal succession: narrative, expository, procedural, and hortatory. Chafe proposes a four-way classification of texts with respect to the parameters of 'involvement-detachment' and 'integration-fragmentation'.

Within rhetorical theory, four basic 'modes' of discourse are traditionally distinguished: narration, description, exposition, and argumentation. Although there is wide agreement on the importance of these four discourse types, there is less agreement on the particular parameters distinguishing among them; for example, different definitions of exposition have focused on one or another of the following parameters: content type, organization, objectivity, purpose, or informational density (Grabe 1984).

All of these typologies have been proposed on functional rather than linguistic grounds. That is, in each of these cases researchers have isolated an important functional difference among texts and have subsequently attempted to identify the linguistic features associated with that difference. For this reason, these typologies are not well defined from a strictly linguistic perspective, considering questions such as (1) do the proposed sets of defining linguistic features actually cooccur systematically in texts? (2) are the texts in each type actually similar to each other in their linguistic form? (3) are the types clearly distinct in their linguistic form? and (4) does the typology characterize the full range of texts in English? For example, neither Longacre's typology (see Smith 1985) nor Chafe's (see Redeker 1984) captures the sets of linguistic features that actually cooccur in English texts, and therefore these typologies cannot identify text types that are linguistically well defined. The rhetorical modes of discourse, on the other hand, are not intended as linguistic types; the characteristic linguistic features of each mode have not been consistently defined, and there is considerable linguistic variation among the texts within each mode. These typologies are important to the extent that they identify salient functional differences among texts. (The functional distinctions identified by Chafe have been particularly useful in this regard.) Functionally based typologies have not been successful, however, in identifying the salient LINGUISTIC differences among texts in English.

The present study uses a new approach to identify the salient linguistic text types of English. The typology developed here is based on sets of syntactic and lexical features that cooccur frequently in texts, rather than assuming sets of features defined on a priori functional grounds. These feature sets, the 'dimensions' of variation, are identified empirically by multivariate quantitative methods, and the linguistic characteristics of any given text can be specified precisely with respect to each dimension. The dimensions and associated text characterizations provide the basis for the present typology: the types are defined such that the texts in each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another. These types represent important functional differences in English, because linguistic cooccurrence reflects shared function. The order of analysis is reversed from previous studies, however. The types are first identified on the basis of their linguistic characteristics and only subsequently interpreted functionally. The resulting text distinctions represent the functional types that are linguistically well defined in English.

There is one further text typology that should be considered here: the folk-typology of 'genres'. Genres are the text categories readily distinguished by mature speakers of a language; for example, the genres of

English include novels, newspaper articles, editorials, academic articles, public speeches, radio broadcasts, and everyday conversations. These categories are defined primarily on the basis of external format. Thus, newspaper articles are found in the news sections of newspapers; academic articles are found in academic journals. These distinctions are related to other differences in purpose and situation, and there are marked linguistic differences among the genres of English (Biber 1988). Genre distinctions do not adequately represent the underlying text types of English, however. Texts within particular genres can differ greatly in their linguistic characteristics; for example, newspaper articles can range from extremely narrative and colloquial in linguistic form to extremely informational and elaborated in form. On the other hand, different genres can be quite similar linguistically; for example, newspaper articles and popular magazine articles can be nearly identical in form. Linguistically distinct texts within a genre represent different text types; linguistically similar texts from different genres represent a single text type. In the present typology of texts, genres and text types are clearly distinguished, and the relations among and between them are identified and explained.

## 2. Background: five dimensions of variation

The notion of linguistic cooccurrence is central to linguistic analyses of style or register. Brown and Fraser (1979: 38–39) emphasize the importance of this notion, observing that it can be ‘misleading to concentrate on specific, isolated [linguistic] markers without taking into account systematic variations which involve the cooccurrence of sets of markers’. Ervin-Tripp (1972) and Hymes (1972) define ‘speech styles’ as varieties that are defined by a shared set of cooccurring linguistic features; the text ‘types’ of the present study are the salient varieties of English defined in these terms. That is, text types are identified quantitatively such that the texts in a type all share frequent use of the same set of cooccurring linguistic features. Because cooccurrence reflects shared function, the resulting types are coherent in their linguistic form and communicative functions.

In the present study, I analyze linguistic cooccurrence in terms of underlying ‘dimensions’ of variation. There are three distinctive characteristics of the notion of ‘dimension’ as I use it. First, no single dimension is adequate in itself to account for the range of linguistic variation in a language; rather, a multidimensional analysis is required. Second, dimensions are continuous scales of variation rather than dichotomous distinctions. Third, the cooccurrence patterns underlying dimensions are

identified quantitatively (by a statistical procedure known as factor analysis) rather than on an a priori functional basis.

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

To date, five major dimensions of variation have been identified in English. Biber (1988) presents a unified description of genre variation in English in terms of this five-dimensional model. The model is developed by analyzing the cooccurrence distributions of 67 linguistic features in 481 spoken and written texts of contemporary British English. The texts, which were taken from the Lancaster-Oslo-Bergen and the London-Lund Corpora, represent 23 different genres (for example, academic prose, press reportage, conversation, radio broadcasts). The linguistic features fall into 16 major grammatical categories: (A) tense and aspect markers, (B) place and time adverbials, (C) pronouns and pro-verbs, (D) questions, (E) nominal forms, (F) passives, (G) stative forms, (H) subordination features, (I) prepositional phrases, adjectives, and adverbs, (J) lexical specificity, (K) lexical classes, (L) modals, (M) specialized verb classes, (N) reduced forms and dispreferred structures, (O) coordination, and (P) negation. The features are identified automatically in texts by computer programs written in PL/I. The cooccurrence patterns among features (that is, the dimensions) are identified quantitatively by a statistical procedure known as factor analysis. Biber (1988) includes both a theoretical analysis of genre variation in terms of the model and a full discussion of the methodological approach (including situational descriptions of the texts, functional descriptions of the linguistic features, and technical descriptions of the computational and statistical techniques). I will begin here with a brief explication of the overall model of variation and then go on to use this model as the basis for a typology of texts in English.

The summary that follows on pages 8 and 9 lists the cooccurring features associated with each of the five dimensions:

Summary of the cooccurrence patterns underlying the dimensions:

Dimension 1. 'Involved versus informational production'

private verbs  
 THAT deletion  
 contractions  
 present-tense verbs  
 2nd person pronouns  
 DO as pro-verb  
 analytic negation  
 demonstrative pronouns  
 general emphatics  
 1st person pronouns  
 pronoun IT  
 BE as main verb  
 causative subordination  
 discourse particles  
 indefinite pronouns  
 general hedges  
 amplifiers  
 sentence relatives  
 WH questions  
 possibility modals  
 nonphrasal coordination  
 WH clauses  
 final prepositions  
 adverbs

---

nouns  
 word length  
 prepositions  
 type/token ratio  
 attributive adjectives  
 place adverbials

Dimension 2. 'Narrative versus nonnarrative concerns'

past-tense verbs  
 3rd person pronouns  
 perfect-aspect verbs  
 public verbs  
 synthetic negation  
 present-participial clauses

---

present-tense verbs  
 attributive adjectives

Dimension 3. 'Explicit versus situation-dependent reference'

WH relative clauses on object positions  
 pied-piping relative clauses  
 WH relative clauses on subject positions  
 phrasal coordination  
 nominalizations

---

time adverbials  
 place adverbials  
 adverbs

Dimension 4. 'Overt expression of persuasion'

infinitives  
 prediction modals  
 suasive verbs  
 conditional subordination  
 necessity modals  
 split auxiliaries  
 possibility modals

---

no complementary features

Dimension 5. 'Abstract versus nonabstract style'

conjuncts  
 agentless passives  
 past-participial clauses  
 BY passives  
 past-participial WHIZ deletions  
 other adverbial subordinators

---

no complementary features

Most of the dimensions consist of two groupings of features, which 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 cooccurrence pattern within each group.

For example, consider dimension 2. The features in the top group (above the line) are past-tense verbs, perfect-aspect verbs, third-person pronouns, and public verbs (primarily speech-act verbs), while the features in the bottom group are present-tense verbs and adjectives. Considering all of the features on dimension 2, this dimension is interpreted as distinguishing narrative discourse from other types of discourse, suggesting the interpretive label 'Narrative versus nonnarrative concerns'. Narrative concerns are marked by considerable reference to

past time, third-person animate referents, and reported speech (public verbs); nonnarrative concerns, whether expository, descriptive, or other, are marked by immediate time and attributive nominal elaboration. The complementary groupings on the other factors reflect similar functional relations.

To represent the communicative function(s) underlying each cooccurrence pattern, the dimensions are labeled as follows:

1. Involved versus informational production
2. Narrative versus nonnarrative concerns
3. Elaborated versus situation-dependent reference
4. Overt expression of persuasion
5. Abstract versus nonabstract style

Dimension 1 (see summary, page 8) represents a dimension marking high informational density and exact informational content (the bottom grouping of features), versus affective, interactional, and generalized content (the top group of features). Two communicative parameters seem to be involved here: (1) the primary purpose of the writer/speaker: informational versus involved; and (2) the production circumstances: those circumstances providing careful editing possibilities, enabling precision in lexical choice and an integrated textual structure, versus circumstances that are characterized by real-time constraints, resulting in generalized lexical choice and a generally fragmented presentation of information. To reflect these parameters, the interpretive label 'Involved versus informational production' is used for this dimension.

Considering both groups of defining features, dimension 3 is interpreted as characterizing highly explicit, context-independent reference versus nonspecific, situation-dependent reference. WH relative clauses (including pied-piping constructions) are used to specify the identity of referents within a text in an explicit and elaborated manner. Time and place adverbials, on the other hand, are usually used for text-external references, where the addressee must identify the intended place and time referents in the actual physical context of the discourse. Overall, the label 'Elaborated versus situation-dependent reference' captures the character of this dimension.

The interpretive label 'Overt expression of persuasion' is used for dimension 4. This dimension marks the degree to which persuasion is marked overtly, whether marking the speaker's point of view, or the speaker's attempt to persuade the addressee.

Finally, the cooccurrence of conjuncts, passive constructions, and past-participial clauses on dimension 5 marks informational discourse that is abstract, technical, and formal in style versus other types of discourse.

The label 'Abstract versus nonabstract style' can thus be proposed for dimension 5.

### 3. Multidimensional characterizations of texts

In the same way that the frequency of nouns in a text might be called the 'noun score' of that text, 'dimension scores' can be computed to characterize each text with respect to each dimension. First, the frequencies of all linguistic features are normalized to a text length of 1,000 words and standardized to a mean of 0.0 and a standard deviation of 1.0. On such a scale, a score of 1.0 marks a value that is one standard deviation higher than the overall mean score; a score of  $-1.0$  marks a value that is one standard deviation below the mean.<sup>1</sup> Standardized scores are used because they set frequency counts to a single scale, making the frequencies directly comparable across features.

After standardization, dimension scores are computed by summing, for each text, the frequencies of the salient defining features of the dimension. To illustrate, consider dimension 3 as given in the summary on page 8. The dimension score representing dimension 3 is computed by adding together the frequencies of WH relative clauses on object and subject positions, pied-piping relative clauses, phrasal coordination, and nominalizations (the features with positive loadings), and subtracting the frequencies of time adverbials, place adverbials, and general adverbs (the features with negative loadings) — for each text

The linguistic relations among texts can be considered by comparing their dimension scores, and the relations among text varieties can be considered by comparing the mean dimension score of each variety. For example, Figure 1 plots the mean dimension scores for nine English genres with respect to dimension 1, 'Involved versus informational production'. Face-to-face conversation has the highest value, marking it as extremely involved and interactive; this high score reflects high frequencies of present-tense verbs, private verbs, first- and second-person pronouns, contractions, etc., together with markedly low frequencies of nouns, prepositional phrases, long words, etc. Personal letters and interviews have moderately high scores on dimension 1, while general fiction and prepared speeches have intermediate values. Genres like press reportage, academic prose, and official documents have the lowest values on dimension 1, marking them as quite informational and noninvolved; these low scores reflect very high frequencies of nouns, prepositional phrases, etc., plus very low frequencies of private verbs, contractions, etc.

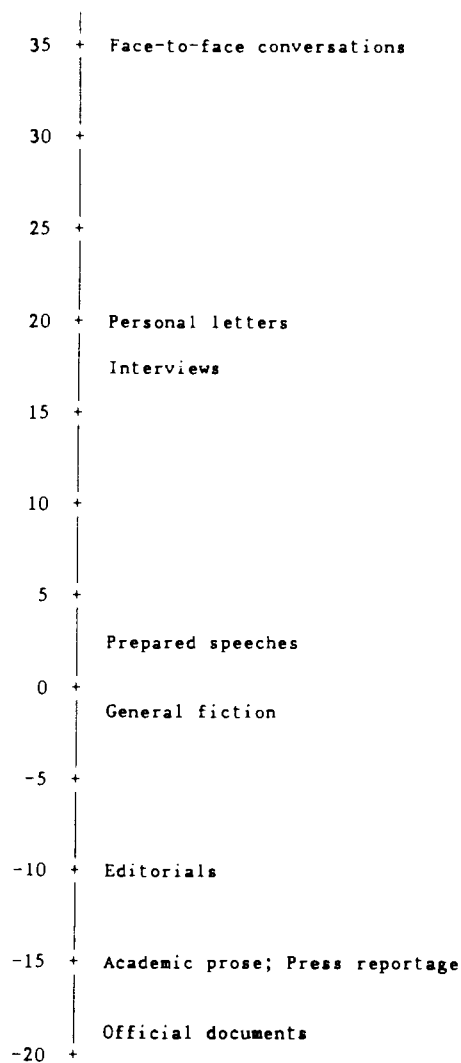


Figure 1. Mean scores of dimension 1 ('Involved versus informational production') for nine genres

The overall relations between any two texts or varieties can be analyzed by consideration of their relative scores on all five dimensions. In previous studies, I have used the dimensions to examine the relations among various genre classes (see for example 1987, 1988; Biber and Finegan 1988); the present paper develops a typology of English texts with respect to the five dimensions.

#### 4. A typology of texts in English

##### 4.1. Identification of formally distinct text types

Groupings of texts that are similar in their linguistic form can be identified empirically by using a statistical procedure known as cluster analysis. In the present case, the groupings are identified on the basis of similarities with respect to the five dimensions outlined above. Cluster analysis groups texts such that the texts within each cluster are maximally similar to each other in their exploitation of the textual dimensions, while each cluster is maximally distinct from the others. That is, those texts with the most similar dimension scores are grouped in each cluster.

To identify the major text types in English, it was necessary to analyze the similarities and differences among a large number of texts representing many different spoken and written genres. In all, 481 texts taken from 23 major genre categories were analyzed, as summarized in Table 1.<sup>2</sup> Some of these genre categories represent several distinct subgenres. For example, press reportage includes cultural, sports, and financial news reports; academic prose includes humanities, social sciences, and engineering expositions; broadcasts include sports reporting and reportage of less-exciting events, such as a funeral and a scientific demonstration. Taken together, these texts represent a large range of the communicative situations and purposes found in English.

A cluster analysis produces different solutions for different numbers of clusters (that is, a one-cluster solution, a two-cluster solution, etc.). Therefore, the first task for the researcher is to determine which solution provides the best 'fit' to the data, that is, in which solution the texts within each cluster are maximally similar while the clusters themselves are maximally distinct. The 'fit' of a solution can be assessed quantitatively, and in the present analysis, the solution for eight clusters provides the best fit to the data.<sup>3</sup>

The cluster analysis assigns every text in the study to some cluster. If each text is labeled with the number of its cluster, all 481 texts can be plotted in a way that illustrates the differences among clusters. Figure 2, for example, shows the distribution of texts with respect to dimension 1 (Involved versus informational production) and dimension 3 (Explicit versus situated reference). This plot represents the distribution of texts according to their exploitation of the linguistic features on these two dimensions; the horizontal axis plots the dimension score of each text for dimension 1; the vertical axis plots the scores for dimension 3. The numbers in the plot represent the cluster number of the texts having the given scores on these two dimensions. For example, the position held by

Table 1. Distribution of texts across 23 genres

Genre	Number of texts
<b>Written --- genres 1-15 from the LOB corpus</b>	
1. Press reportage	44
2. Editorials	27
3. Press reviews	17
4. Religion	17
5. Skills and hobbies	14
6. Popular lore	14
7. Biographies	14
8. Official documents	14
9. Academic prose	80
10. General fiction	29
11. Mystery fiction	13
12. Science fiction	6
13. Adventure fiction	13
14. Romantic fiction	13
15. Humor	9
16. Personal letters	6
17. Professional letters	10
<b>Spoken --- from the London-Lund corpus</b>	
18. Face-to-face conversation	44
19. Telephone conversation	27
20. Public conversations, debates, and interviews	22
21. Broadcast	18
22. Spontaneous speeches	16
23. Planned speeches	14
<b>Total</b>	<b>481</b>
<b>Approximate number of words</b>	<b>960,000</b>

the circled number 5 on Figure 2 locates the text that has a dimension score of 13 on dimension 1 (the horizontal axis) and a score of -5 on dimension 3 (the vertical axis), and that belongs to cluster 5; these dimension scores mark this text as moderately involved in focus and moderately situated in reference.

Figure 2 shows relatively distinct groupings for clusters 1, 2, 5, 7, and 8, while the remaining three clusters (3, 4, and 6) are less well distinguished in terms of dimensions 1 and 3. The texts in cluster 1 (marked in the plot by the numeral 1) are characterized by quite high scores on dimension 1 and relatively low scores on dimension 3; cluster 2 is similar except it has lower scores on dimension 1. Clusters 5, 7, and 8, all have unmarked scores on dimension 1; they differ from one another along dimension 3:

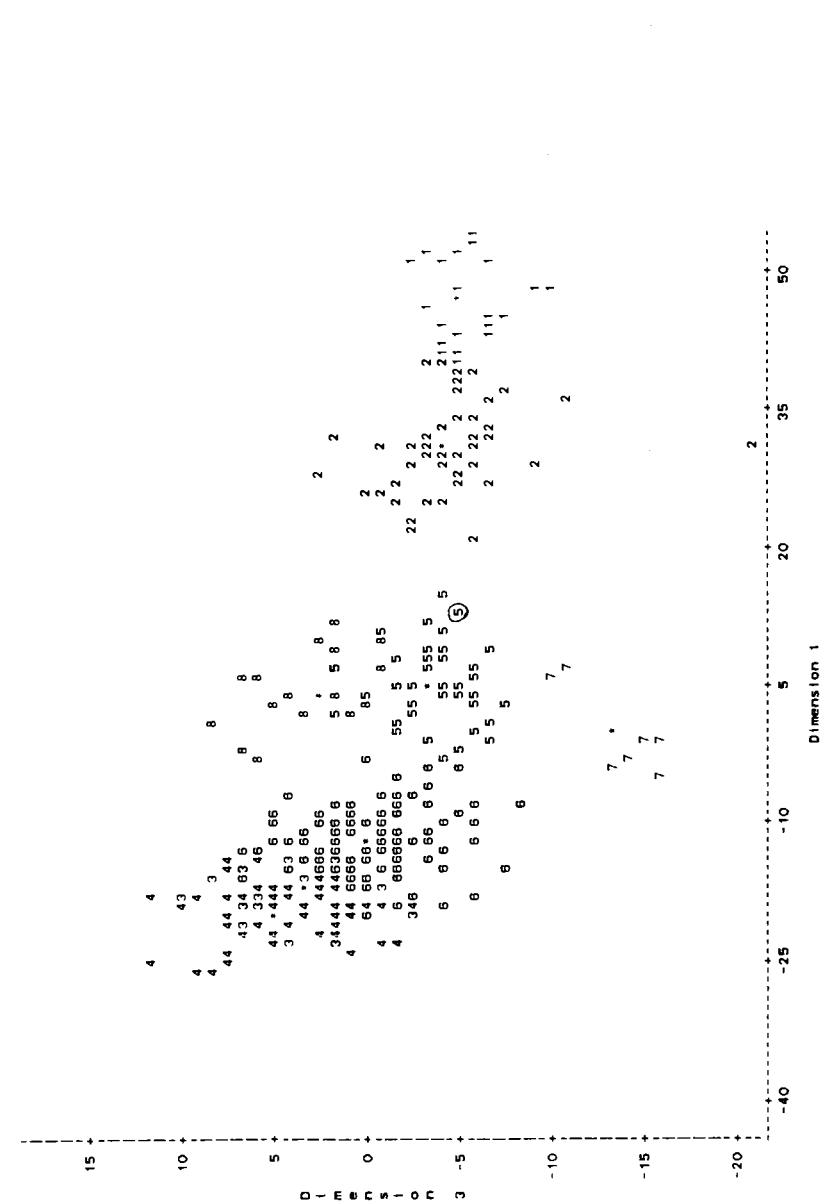


Figure 2. Plot of cluster distributions with respect to dimension 1 (Involved versus informational production) versus dimension 3 (Explicit versus situated reference)

cluster 8 texts have markedly high scores, cluster 7 texts have extremely low scores, cluster 5 texts have unmarked scores. Clusters 3, 4, and 6 all have low scores on dimension 1; on dimension 3, cluster 4 texts tend to have high scores, cluster 6 texts tend to have relatively low scores, and cluster 3 texts have unmarked scores.

The asterisks on Figure 2 plot the 'centroids' (the central characterizations) for each cluster with respect to the two dimensions. An overall summary of each cluster is given in Table 2, including the number of texts in the cluster, the nearest cluster, and the 'distance' to the nearest cluster. The 'distance' measures the cumulative difference between the cluster centroids with respect to the five dimensions. Table 2 confirms the impression given by Figure 2 that the types are not equally distinct in their linguistic characterizations. In particular, this table shows that clusters 3, 4, and 6 are relatively nondistinct: both cluster 3 and cluster 6 have cluster 4 as the nearest cluster, with a distance of only 8.3 between cluster 3 and cluster 4.

The cluster analysis identifies the 'core' text types in English: the groupings that contain very high concentrations of texts. There is a group of core texts and a group of peripheral texts associated with each cluster. Core texts are very similar to the central linguistic characterization of a cluster; peripheral texts are relatively dissimilar to the central cluster characterization, but even more dissimilar to other clusters.<sup>4</sup> Out of the 481 texts in this study, 345 are grouped into one of the core text types by the cluster analysis; Figure 2 plots only these core texts. Peripheral texts, however, are not aberrant; their existence rather reflects the fact that textual variation is continuous. Texts do not divide into sharply distinct 'types' — instead there is a continuous range of variation in linguistic form and use. The notion of 'text type' developed here is based on the

frequent and therefore typical clusterings of texts, which account for the majority of texts in English. In a sense, these can be considered the text 'prototypes' of English. There are, however, other texts that fall in between clusters, grading from one type to the next. I return to the continuous nature of variation among texts in section 5.

The grouping of texts into clusters is determined on the basis of their characterization with respect to all five dimensions. That is, texts that are similar with respect to one dimension but very different with respect to other dimensions are likely to be grouped into different clusters. Figure 2 shows the distribution of texts with respect to only two dimensions, but it can be used as an illustration of the way texts are grouped into clusters. For example, texts in clusters 1, 2, and 5 are very similar with respect to their dimension 3 scores (the vertical axis); texts in all three clusters generally have scores between 0 and -8. With respect to their dimension 1 scores (the horizontal axis), however, the texts in these three clusters are distinct: texts in cluster 1 have scores ranging from 40 to 54; texts in cluster 2 range from 22 to 40; texts in cluster 5 range from -3 to 15. Similar comparisons can be made for clusters 8, 5, and 7: texts in these clusters are quite similar with respect to their dimension 1 scores (ranging generally between -3 and 12), but quite distinct with respect to their dimension 3 scores (cluster 8 ranging from 8 to -1; cluster 5 ranging from 2 to -8; cluster 7 ranging from -10 to -16). The picture given by Figure 2 is incomplete because only two dimensions are considered. When all five dimensions are considered, it is possible to identify the salient distinguishing characteristics of all eight clusters.

Figures 3 and 4 summarize the distinguishing characteristics of the eight clusters, plotting the centroid score of each cluster with respect to each dimension. These two figures present the same information: Figure 3 highlights clusters 1-4, while Figure 4 highlights clusters 5-8. The information presented in these figures overlaps the information presented in Figure 2; the centroid values for dimensions 1 and 3, which are given by asterisks on Figure 2, are repeated on the respective scales of Figures 3 and 4.

On the basis of Figures 3 and 4, it is possible to describe the distinguishing linguistic characteristics of each of the eight text types. Using the interpretive dimension labels, cluster 1 is situated, nonabstract, and extremely involved, but not marked for narrative concerns or persuasion; cluster 2 is similar to cluster 1, except it is less involved. Clusters 3 and 4 are also similar to each other: both are extremely informational, highly elaborated, nonnarrative, and nonpersuasive. These two clusters differ primarily with respect to dimension 5, where cluster 3 is extremely abstract in style while cluster 4 is only moderately abstract.

Table 2. *General summary of the clusters*

Cluster number	Frequency of core texts	Frequency of peripheral texts	Nearest cluster	Centroid distance to nearest cluster
1	22	1	2	15.3
2	49	24	1	15.3
3	28	15	4	8.3
4	53	18	3	8.3
5	47	13	8	11.0
6	117	33	4	10.2
7	7	5	5	15.3
8	22	27	5	11.0

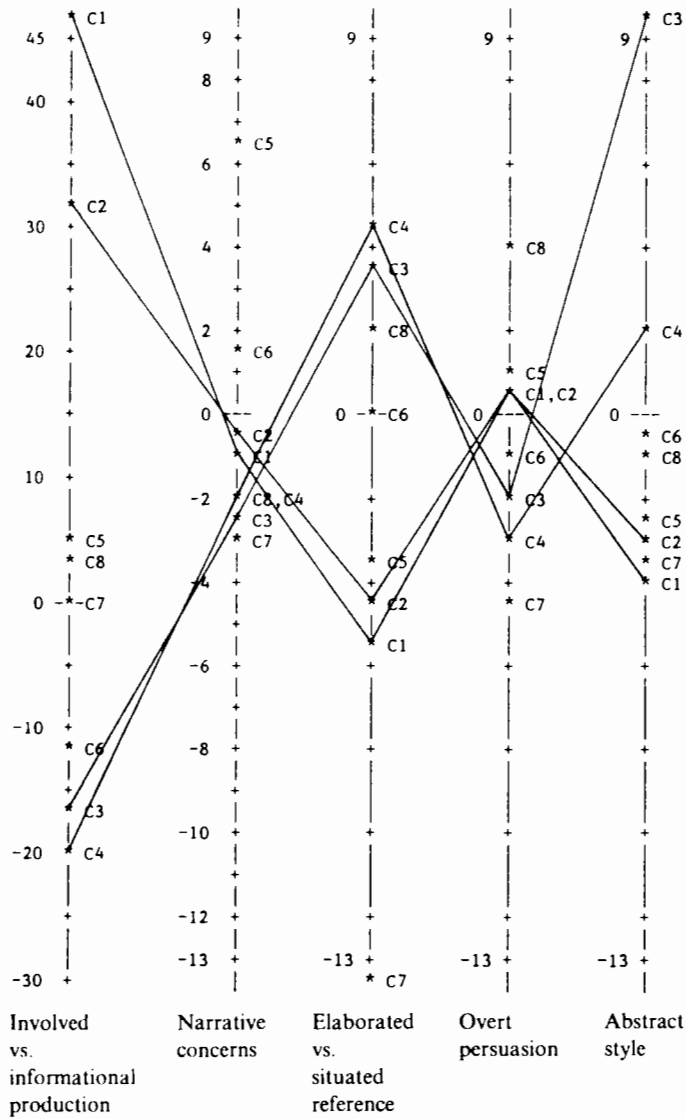


Figure 3. Dimensional characterization of the eight text types, highlighting clusters 1, 2, 3, and 4: \* marks the centroid score of each cluster on each of the dimensions; dimensions 2-5 use the same scale (from -13 to +9), while dimension 1 uses a compressed scale (from -30 to +48)

Cluster 5 is extremely narrative, moderately involved, situated, and nonabstract, and not marked for persuasion. Cluster 6 combines the distinctive characteristics of clusters 4 and 5: it is informational as well as narrative, while it is not marked on dimensions 3, 4, and 5. Cluster 7 is

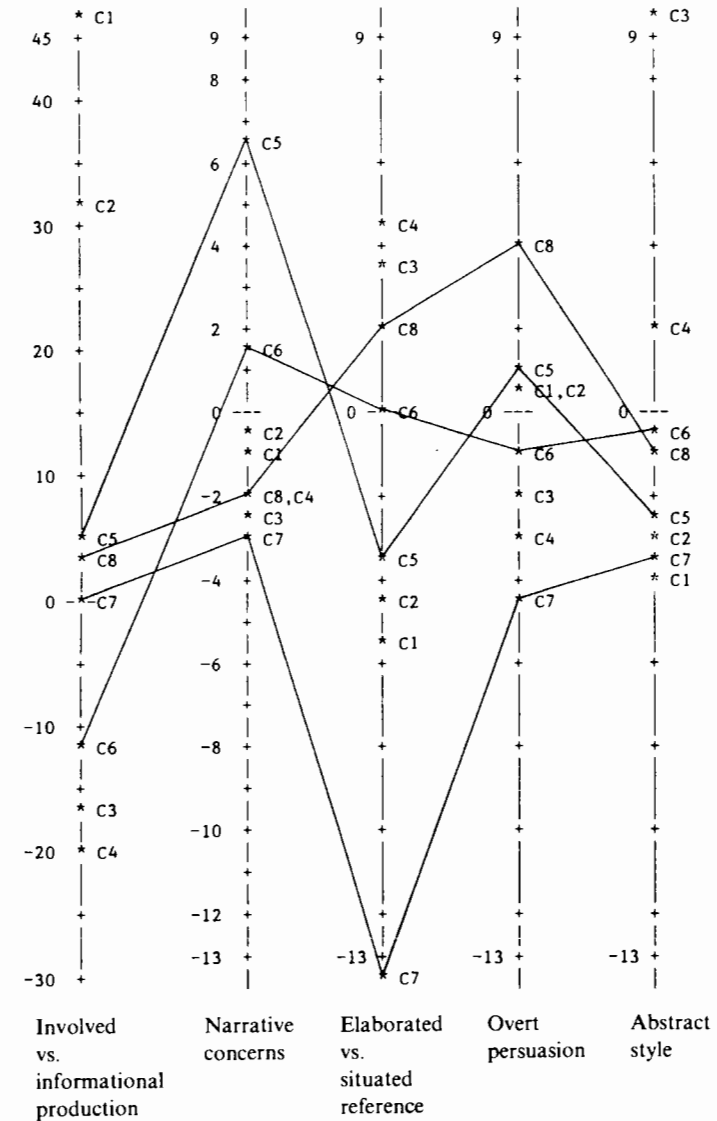


Figure 4. Dimensional characterization of the eight text types, highlighting clusters 5, 6, 7, and 8: \* marks the centroid score of each cluster on each of the dimensions; dimensions 2-5 use the same scale (from -13 to +9), while dimension 1 uses a compressed scale (from -30 to +48)

extremely situated in its reference, in addition to being markedly nonnarrative, nonpersuasive, and nonabstract. Finally, cluster 8 is distinctive in being extremely persuasive, in addition to being moderately involved, nonnarrative, and elaborated.<sup>5</sup>

As discussed above, the dimensions represent frequency scales for sets of linguistic features, and the clusters of texts are determined on the basis of their similarities in the use of the features grouped on each dimension (summarized on pages 8 and 9). Clusters are marked with respect to particular dimensions to the extent that they have large positive or negative scores for those dimensions. For example, cluster 1 is marked for dimensions 1, 3, and 5. Ranking extremely high on dimension 1, the texts in this cluster tend to have very frequent occurrences of private verbs, contractions, first- and second-person pronouns, etc. (the upper features on dimension 1), together with shorter words and lesser lexical variety (the bottom features on dimension 1). Ranking low on dimension 3, the texts in this cluster are characterized by frequent time and place adverbials (the bottom features on dimension 3) and markedly few relative clauses, phrasal coordinators, and nominalizations (the upper features on dimension 3). Finally, ranking very low on dimension 5, these texts tend to have very few conjuncts, agentless passives, past-participial clauses, etc.

The following is a breakdown of the texts in each cluster by genre. For each cluster, the total number of texts in the cluster (both core texts and peripheral texts), an interpretive label, the number of texts from each genre occurring in the cluster, and (in parentheses) the percentage of texts from each genre that occurs in the cluster are given. The asterisks identify cases where a majority of the texts from a single genre occur in a particular cluster (for example, cluster 1 contains 62% of the telephone conversations between personal friends).

Composition of the text types by genre, giving the number of texts in the core cluster, the number of texts peripheral to the cluster, and the percentage of texts from each genre occurring in that cluster ('\*' marks the case where more than 50% of a genre occurs in a single cluster):

Text type 1. Intimate interpersonal interaction (22 core + 1 peripheral texts)

- 12 + 1 Face-to-face conversations (29%)
- 8 + 0 Telephone conversations — personal friends (62%)\*
- 1 + 0 Telephone conversations — disparates (17%)
- 1 + 0 Telephone conversations — business associates (13%)

Text type 2. Informational interaction (49 core + 24 peripheral texts)

- 26 + 5 Face-to-face conversations (70%)\*
- 7 + 4 Interviews (50%)\*
- 5 + 2 Telephone conversations — business associates (88%)\*
- 4 + 1 Telephone conversations — personal friends (39%)
- 3 + 1 Telephone conversations — disparates (67%)\*
- 2 + 7 Spontaneous speeches (56%)\*

- 2 + 1 Personal letters (50%)\*
- 0 + 1 Nonsports broadcasts (13%)
- 0 + 1 Professional letters (10%)
- 0 + 1 General fiction (3%)

Text type 3. 'Scientific' exposition (28 core + 15 peripheral texts)

- 23 + 12 Academic prose (44%)
- 3 + 1 Official documents (29%)
- 1 + 0 Biographies (7%)
- 1 + 0 Press reviews (6%)
- 0 + 1 Hobbies (7%)
- 0 + 1 Press reportage (2%)

Text type 4. Learned exposition (53 core + 18 peripheral texts)

- 17 + 8 Academic prose (31%)
- 10 + 1 Press reportage (25%)
- 6 + 2 Official documents (57%)\*
- 6 + 2 Press reviews (47%)
- 5 + 0 Popular lore (36%)
- 4 + 0 Biographies (29%)
- 2 + 1 Hobbies (21%)
- 2 + 1 Religion (18%)
- 1 + 0 Press editorials (4%)
- 0 + 3 Professional letters (30%)

Text type 5. Imaginative narrative (47 core + 13 peripheral texts)

- 12 + 0 Romance fiction (92%)\*
- 12 + 3 General fiction (51%)\*
- 7 + 2 Mystery fiction (70%)\*
- 7 + 2 Adventure fiction (70%)\*
- 4 + 3 Prepared speeches (50%)\*
- 2 + 0 Interviews (9%)
- 1 + 1 Science fiction (33%)
- 1 + 0 Popular lore (7%)
- 1 + 0 Biography (7%)
- 0 + 1 Personal letters (17%)
- 0 + 1 Religion (6%)

Text type 6. General narrative exposition (117 core + 33 peripheral texts)

- 30 + 2 Press reportage (73%)\*
- 15 + 8 Press editorials (86%)\*
- 12 + 0 General fiction (41%)
- 8 + 0 Biographies (57%)\*
- 7 + 1 Humor (89%)\*
- 7 + 1 Press reviews (47%)
- 7 + 6 Academic prose (17%)

- 6+4 Religion (59%)\*
- 4+2 Hobbies (43%)
- 3+2 Nonsports broadcasts (63%)\*
- 3+0 Science fiction (50%)\*
- 3+1 Adventure fiction (31%)
- 3+0 Mystery fiction (23%)
- 3+0 Popular lore (21%)
- 2+3 Prepared speeches (35%)
- 2+0 Official documents (14%)
- 1+1 Professional letters (20%)
- 1+0 Romance fiction (8%)
- 0+2 Sports broadcasts (20%)
- Text type 7. Situated reportage (7 core + 5 peripheral texts)
  - 7+1 Sports broadcasts (80%)\*
  - 0+1 Nonsports broadcasts (13%)
  - 0+1 Science fiction (17%)
  - 0+1 Mystery fiction (8%)
  - 0+1 Hobbies (7%)
- Text type 8. Involved persuasion (22 core + 27 peripheral texts)
  - 5+4 Interviews (41%)
  - 4+3 Spontaneous speeches (44%)
  - 4+1 Popular lore (36%)
  - 2+2 Professional letters (40%)
  - 2+1 Religion (18%)
  - 2+0 Prepared speeches (14%)
  - 1+0 Telephone conversations — disparates (17%)
  - 1+0 Humor (11%)
  - 1+2 Editorial letters (11%)
  - 0+7 Academic prose (9%)
  - 0+3 Hobbies (21%)
  - 0+2 Personal letters (33%)
  - 0+1 Nonsports broadcasts (13%)
  - 0+1 General fiction (3%)

The labels in this breakdown summarize the interpretations of the clusters as text types, based on consideration of the predominant linguistic features in each cluster (summarized by the centroid dimension score of each cluster given in Figures 3 and 4), the communicative characteristics of the texts grouped in each cluster, and microanalyses of individual texts. I turn now to a detailed consideration of each type.

## 4.2. Interpretation of the clusters as text types

4.2.1. *Clusters 1 and 2: 'Intimate interpersonal interaction' and 'Informational interaction'*. I noted above that clusters 1 and 2 are quite similar to each other (see Figure 3). Both are characterized by 'situated reference' (dimension 3), 'a nonabstract style' (dimension 5), and relatively unmarked scores on dimensions 2 and 4. Even on dimension 1, they are both characterized as 'involved production'; the major difference between these two clusters is that cluster 1 has an extreme characterization on this dimension, while cluster 2 has a more moderate characterization. The linguistic features associated with these dimension scores are summarized on pages 8–9: for example, very frequent private verbs, contractions etc., plus markedly infrequent nouns, prepositions, etc., associated with the high score on dimension 1; frequent time and place adverbials and infrequent WH relative clauses associated with the low score on dimension 3; and markedly infrequent conjuncts and past-participial forms associated with the low score on dimension 5.

The texts grouped into clusters 1 and 2 in the breakdown at the end of the last section reflect the shared linguistic characteristics of the two clusters as well as the major difference between them on dimension 1. Cluster 1 comprises strictly 'involved', interpersonal conversations, for the most part face-to-face conversations and telephone conversations between personal friends. Cluster 2, on the other hand, comprises person-to-person interactions that have an informational concern, such as interviews, business telephone conversations, and face-to-face conversations in professional contexts. In both of these text types, the interaction is primary — speakers address individual listeners who are immediately present and personal. The main difference between these types relates to the primary purpose of interaction: to convey information in cluster 2 and to maintain the interpersonal relationship in cluster 1. This difference can be illustrated by comparison of text sample 1, representing cluster 1, with text samples 2, 3, and 4, representing cluster 2.<sup>6</sup>

Cluster 1 texts:

Text sample 1 (LL:1.8.; informal face-to-face conversation between friends)

- A: there are cups # [pause] Nescafe # [pause]
- B: shall we have a cup of coffee # [pause]
- A: yes certainly # yes certainly # [pause] yes #
- B: I see # they're all [??] [long pause]
- A: some of them are rather large # [pause]
- B: mm #

- A: some of them are rather large # [pause]  
 ....  
 B: want any sugar # [pause]  
 A: yes please Brenda # [pause]  
 B: one # [pause]  
 A: that's about right # yes that's enough thank you # [long pause] not yours is it #  
 B: oh no # those are my scripts # I just saw the note # and I know that's all right # [long pause]  
 A: just put my glasses on # I can't see a thing without them # [long pause] well after all they're too dark to be inspiring # aren't they # [pause]  
 B: I don't want one # I'm afraid #  
 A: I think I'd rather substitute #  
 B: yes # I haven't space # I don't want [pause] portraits #  
 A: no # [pause]

## Cluster 2 texts:

## Text sample 2 (LL:5.5; panel discussion)

Question: Do you think that there is any chance that the Labour Party will provide an effective opposition in the foreseeable future?

- ....  
 A: Christopher Chataway #  
 C: I've seldom heard a string of sentences # that I really do believe # to [pause] contain quite so many # [pause] faulty analyses # of the present situation # [long pause] I don't believe # that this country is swinging to unilateralism #  
 ....  
 A: Lord Boothby #  
 B: well # [pause] I don't think you know # that Tony Wedgwood Benn can seriously say that personalities [pause] don't matter # [long pause] because I think they do matter tremendously # in [pause] politics today # and especially in the politics of the Left # [long pause] what has happened is ...

## Text sample 3 (LL:1.1; face-to-face conversation between academic colleagues, concerning student comprehensive exams)

- A: well # [pause] may I ask # what goes into that paper now # because I have to advise # [pause] a couple of people who are doing the [mm]  
 B: well what you do # is to [long pause] this is sort of between the two of us # what you do # is to make sure that your own [pause] candidate

[mm] # is [pause] that your [pause] there's something that your own candidate can handle # [long pause]

- ....  
 A: you mean that the the the papers are more or less set ad hominem # are they # [pause]  
 B: [mm] [long pause] they shouldn't be # [long pause] but [mm] [pause] I mean one # sets [long pause] one question # now I mean this fellow's doing the language of advertising # [pause] so very well #  
 A: yeah #  
 B: give him one on  
 A: is this a spare paper <change of topic>  
 B: yeah ....

## Text sample 4 (LL:11.1; spontaneous speech — specifically a court examination of a witness)

- A: Mr Potter # did you # [long pause] arrive # about two o'clock # on the [pause] Sunday # [pause] the date the will was [pause] signed # [pause]  
 B: yes # [long pause]  
 A: and [pause] did you [pause] go # and see your mother straight away #  
 B: yes I did #  
 A: what was she then doing # [pause]  
 B: she was having her lunch # [long pause]  
 A: what about the brandy bottle # where was that # [long pause]  
 B: I don't know # I didn't [pause] I didn't see #  
 A: you didn't see it #  
 B: well # [pause] no I didn't #  
 ....  
 A: I would think I ought to tell your lordship # [pause] so that nothing # is [pause] is to be hidden # from our side # [pause] that the plaintiffs supplied # certain statements # to us # [long pause] several months ago [pause] one of which # was a short statement # from the doctor # [pause] ....

Text samples 1–4 illustrate the shared characteristics of clusters 1 and 2, as well as the major difference between these two types. All four samples are interactional, but text sample 1, which represents cluster 1, is nearly exclusive in its interpersonal focus, while samples 2–4, which represent cluster 2, all have a specific informational purpose in addition to their interpersonal purpose. In sample 1, there is no particular topic that is the focus of discussion; rather, the participants change topic freely and place primary emphasis on the interaction itself instead of on the exchange of

information. In contrast, samples 2–4 are all informational to some extent. Sample 2 is from a panel discussion, where a group of discussants interact with one another debating a series of specific issues; this speech event is thus explicitly informational and interactional at the same time. Sample 3 is from a face-to-face conversation in a professional context for professional purposes, and it thus has a markedly informational as well as interactional focus. The text that contains sample 4 is from the genre labeled ‘spontaneous speeches’ in the London-Lund Corpus, and it is therefore surprising that this text is grouped into the ‘interactional’ cluster 2. As sample 4 shows, however, this text is in fact an interaction — a courtroom examination of a witness which is both informational and interactional. The grouping of this text into cluster 2 highlights the fact that the types are defined on the basis of linguistic characteristics rather than any external criteria.

As noted above, clusters 1 and 2 are very similar with respect to dimensions 2–5, and text samples 1–4 illustrate these shared characteristics. All four texts are marked by nonelaborated reference in that they have very few WH relative clauses, phrasal coordinators, and nominalizations (the upper features on dimension 3). In fact, all of these texts contain implicit referents that can be understood only by reference to the immediate situation or shared background of the participants; for example, *there*, *they*, *some of them*, *that*, *those* in sample 1; *the present situation*, *this country*, *today* in sample 2; *that paper*, *this fellow* in sample 3; *the brandy bottle*, *several months ago* in sample 4. In addition, all four texts are nonabstract in that they contain very few conjuncts, passives, or other past-participial constructions (the features on dimension 5). Like their respective clusters, text samples 1–4 are relatively unmarked with respect to dimension 2, ‘Narrative concerns’ and dimension 4, ‘Expression of persuasion’. Thus, these texts show relatively few past-tense forms, perfect-aspect verbs, third-person pronouns, etc. (dimension 2) and few infinitives, prediction modals, suasive verbs, etc. (dimension 4).

The major linguistic difference between clusters 1 and 2 relates to their characterization on dimension 1: cluster 1 is extremely involved, while cluster 2 is more moderately involved. The situational characteristics of the texts grouped into clusters 1 and 2, together with this linguistic characterization, lead to the interpretation of cluster 1 as highly interpersonal, noninformational interaction, and cluster 2 as informational interaction. The dimension 1 characteristics of text samples 1–4 support this interpretation. Sample 1 has by far the highest dimension 1 score of these four, reflected by frequent private verbs (such as *know*, *think*), *that* deletions (*I know <that> that’s all right*; *I think <that> I’d rather substitute*), contractions (such as *they’re*, *that’s*, *can’t*), present-tense verbs

(*are*, *see*), first- and second-person pronouns, demonstrative pronouns (such as *that’s about right*, *those are my scripts*), occurrences of *it* as pronoun, *be* as main verb, etc. At the same time, text sample 1 shows a marked absence of nouns, prepositions, attributive adjectives, and long words, and it has markedly low lexical variety. This sample is thus strikingly ‘involved’, shown by the very high frequencies of the features with positive weights on dimension 1, and strikingly noninformational, shown by the very low frequencies of the features with negative weights.

Text samples 2–4 are also ‘involved’ and interactional, but much less so than sample 1. Some dimension 1 features occur relatively frequently in these texts; for example, first and second person pronouns, contractions, and private verbs. At the same time, though, these texts show relatively many nouns, prepositions, attributive adjectives, and long words in comparison to sample 1, reflecting the less involved and more informational focus of these texts.

In summary, text type 1 (cluster 1) is a type of interaction that is situated in reference, nonabstract in style, and extremely interpersonal and involved in focus. Text type 2 (cluster 2) is also a type of interaction that is situated and nonabstract, but the texts of this type have specific informational as well as interpersonal purposes. The interpretation of these two clusters illustrates the way in which text-type categories can cut directly across genre classifications. For example, some texts classified as ‘spontaneous speeches’ are highly interactive and belong to text type 2; texts classified as ‘face-to-face conversation’ can be either highly interpersonal interaction (text type 1) or relatively informational interaction (text type 2). Text types 3–8 provide many other examples of linguistic characterizations that cut across external genre classifications.

4.2.2. *Clusters 3 and 4: “Scientific” exposition’ and ‘Learned exposition’.* Clusters 3 and 4 form a second pair of related text types. Figure 3 shows that these clusters have very similar characterizations on dimensions 1–4, differing primarily on dimension 5. Both clusters are markedly nonnarrative (dimension 2) and nonpersuasive (dimension 4), and both are extremely informational in production (dimension 1) and explicit in reference (dimension 3). On dimension 5, both clusters are characterized by an abstract style; they differ in that cluster 3 is extremely abstract in style, while cluster 4 is only moderately so.

The texts grouped into clusters 3 and 4 at the end of section 4.1 show that both of these clusters represent types of informational exposition. Cluster 3 comprises primarily academic prose texts and a few official documents; the academic prose texts are primarily from natural science, engineering/technology, and medicine. Cluster 4, on the other hand,

comprises a relatively broad range of texts, including academic prose, press reportage, official documents, press reviews, popular lore, biographies, hobbies, and religion. Academic prose texts make up approximately a third of the texts in this cluster; but unlike those in cluster 3, the academic prose texts in cluster 4 are primarily from the humanities, social sciences, education, and law. Both of these clusters are expository with an extreme informational focus. The difference between them relates to the extreme technical content and style found in cluster 3 versus the more 'learned' presentation of information in cluster 4. Text samples 5 and 6 illustrate the shared characteristics of clusters 3 and 4 as well as the primary difference between them.

#### Cluster 3:

##### Text sample 5 (LOB:J.8; physics journal article)

Thus the first few atomic layers deposited during the gettering period are highly oxidized, and when the chamber has been 'cleaned up' the deposit is more metallic. After the evaporation ceases, the deposited film remains open to oxidation. Thus the deposited film is inhomogeneous and approximates to a sandwich layer of oxide/metal/oxide, in which the outer layers are more highly oxidized than the inner layer.

The exact state of oxidation of the deposited film is unknown and a further effect of oxidation can be observed upon baking in air. ...

#### Cluster 4:

##### Text sample 6 (LOB:J.27; sociology text)

Government in Spain continues to rest on the three institutions of an hereditary monarchy (rejected by two short-lived republics), the parliament of the old Castilian Cortes, and an extensive Civil Service, with a permanent staff except for its highest officials. Spain is at the moment a kingdom without a king. The Franco regime has committed itself to the maintenance of the monarchy as an institution by the 1947 Law of Succession and the Referendum of the following year. Meanwhile the regime, in its own words, is a representative, organic democracy in which the individual participates in government through the natural representative organs of the family, the city council and the syndicate.

Both of these samples show the characteristics of extreme informational production and explicit reference. Both have a very high concentration of nouns, prepositions, attributive adjectives, long words, and a quite varied vocabulary — the bottom group of features on dimension 1; both have essentially none of the top group of features on dimension 1, such as

private verbs, first- and second-person pronouns, and contractions. Although the verbs in these texts are consistently in the present tense, there is a striking absence of verbs in general, coupled with a preference for noun and prepositional phrases. Taken together, these features result in an extreme 'informational' characterization on dimension 1. On dimension 3, neither text makes direct reference to items in the external situation, and in both cases the full texts make frequent use of WH relative clauses to elaborate intended references (although these samples illustrate only one relative construction each). Both samples are written entirely in the present tense, resulting in their marked nonnarrative focus on dimension 2. Neither sample makes use of the persuasive features associated with dimension 4, resulting in their low score on that dimension.

The striking difference between these two samples relates to their dimension 5 characterization. Sample 5 is extremely passive in form, with the agent of predicates being deleted in every case. These forms include both main verbs (for example, *are highly oxidized*, *has been 'cleaned up'*, *can be observed*) and postnominal modifying clauses (such as *layers deposited during the gettering*). This sample also illustrates the frequent use of conjuncts to mark the logical relations in a text (in this case a repeated use of *thus*). Sample 6, on the other hand, is consistently in the active voice; the only passive form is in the postnominal clause *monarchy rejected by two short-lived republics*, and even this case is unlike sample 5 in that the agent is specified. This sample also illustrates a lesser use of conjuncts, counting on the reader to infer the logical relations among propositions.

In summary, both text type 3 and text type 4 are expository, extremely informational, and explicit in reference. The difference between them relates to their use of an abstract, technical style. This seems to be both a content and a stylistic distinction. On the one hand, the texts in type 3 focus on highly abstract and technical information; they are therefore much more concerned with the entities being acted on (the patients) than with any active agents. They further depend on a frequent use of conjuncts to specify the logical relations among propositions. Type 4 texts tend to be less technical in content. In addition, the differences between these types seem to reflect attitudinal preferences. Thus, the texts in type 3, coming for the most part from engineering and natural sciences, show a stylistic preference for a presentation of information apart from active agents, possibly to give the appearance of scientific rigor. In contrast, the texts in type 4, coming from a broad range of 'literate' prose, show a preference for a more active style, perhaps reflecting the influence of prescriptive notions of 'good' style. As was the case with text types 1 and

2, the classification of texts into types 3 and 4 cuts across genre categories. For example, several social science and humanities academic texts are grouped into type 3 because they are relatively technical in content and adopt the abstract and technical style of that type; conversely, a few natural science and engineering academic texts are grouped into type 4, adopting an active, nonabstract style in contrast to the norms for their subgenres.

4.2.3. *Cluster 5: 'Imaginative narrative'*. Figure 4 shows that cluster 5 is situated in reference (dimension 3), nonabstract in style (dimension 5), and slightly involved (dimension 1), but the primary distinguishing characteristic of this cluster is its extreme narrative emphasis, shown by its high score on dimension 2. Not surprisingly, the texts grouped into this cluster are mostly fiction, or 'imaginative narrative'. Text sample 7 illustrates the involved type of narrative common in fiction, while sample 8 illustrates a nonfictional type of involved narrative from a judge's final statement in a court case.

Cluster 5:

Text sample 7 (LOB:L.12; Mystery fiction)

I'd finished making the bed by then. As I pushed it back against the wall I heard something drop on the floor.

That was when the percolator in the living-room started making bubbling noises. There was nothing on the floor that I could see. I told myself it must've fallen down between the bed and the wall.

... Wasn't urgent anyway. Maybe my cigarette-case ... or Sonia's powder compact ... I'd look for it later.

So I got up from my hands and knees, went into the living room and fixed myself a cup of coffee.

Text sample 8 (LL:12.4b; prepared speech — court case)

A:

I have to decide in this case # [pause] what # [pause] if any maintenance # [pause] should be paid # [pause] by the husband as I shall call him # [pause] to the wife # [long pause] he's in fact # no longer the husband # [long pause] he was originally petitioner # [pause] because there's been a decree # [pause] absolute # [long pause] and he has remarried # [pause] the decree # [long pause] was pronounced in favour # of the respondent wife # [pause] on the grounds of the husband's admitted adultery # [pause] his charge # of adultery # [pause] against her # with the main correspondent # [long pause] failed # after a [pause] somewhat lengthy [pause] hearing # [pause] her charges of cruelty # against him # [pause] likewise failed # [long pause]

Sample 7 is typical of the majority of texts grouped into cluster 5. Most of this text is simple narration in the past, which uses frequent past-tense forms, third-person personal pronouns, and perfect-aspect verbs, resulting in a high score on dimension 2. The sample is also relatively 'involved' even though there is no direct interaction of participants; for example, this sample shows a very frequent use of first-person pronouns and contractions. A sample that included direct dialogue between participants would have even more dimension 1 'involved' features than sample 7. Most cluster 5 texts are similar to sample 7, being fictional with an extreme narrative focus and a moderately involved characterization.

Text sample 8 illustrates how nonfictional texts can have a similar mixing of narrative focus and involved presentation. This sample is from a final summation and judgment in a court case. As background to the final judgment, the judge summarizes the events that are relevant to the case. As such, this speech event is largely narrative, and sample 8 contains many of the features characteristic of a high dimension 2 score, such as frequent past-tense verbs, perfect-aspect verbs, and third-person pronouns. In addition, this text is moderately 'involved', with the speaker using first-person pronouns, contractions, private verbs, etc. There are only a few nonfiction texts grouped in cluster 5, but sample 8 illustrates how these other texts can also have a primary narrative focus combined with an involved presentation.

4.2.4. *Cluster 6: 'General narrative exposition'*. Cluster 6 is the largest cluster, with 117 core texts and another 33 peripheral texts. Figure 4 shows that this cluster combines expository/informational and narrative features, making it similar to clusters 3 and 4 in some respects and to cluster 5 in other respects. With respect to dimension 1, cluster 6 is similar to clusters 3 and 4 in being markedly informational and noninvolved; with respect to dimension 2, cluster 6 is similar to cluster 5 in that it has a moderately high narrative concern. On dimensions 3 and 5, however, cluster 6 is not similar to any of these other three clusters; it is unmarked on these dimensions, rather than 'elaborated' and 'abstract' like clusters 3 and 4, or 'situated' and 'nonabstract' like cluster 5. The distinctive characteristics of cluster 6 are thus a marked informational focus (dimension 1) and a moderate narrative concern (dimension 2).

The texts grouped into this cluster likewise combine these two concerns. They are primarily informational and expository but often use narration to convey information. Unlike cluster 5, the narrative portions in these texts are not imaginary or for entertainment; they are rather an integral part of the expository information being conveyed.

Cluster 6 is the most general of the text types identified by the present

study. As noted above, there is a total of 150 texts grouped into this cluster; these texts represent 19 different genres, including press reportage, press editorials, general fiction, biographies, humor, press reviews, academic prose, and religion. This is thus a very general type of exposition; it is not markedly learned or technical, not markedly elaborated in reference or abstract in style, and it often uses narration as part of its exposition.

Text samples 9, 10, and 11 illustrate the distinctive characteristics of this cluster. Sample 9 is from an editorial and illustrates the use of narrative forms to convey expository information. Sample 10 is from press reportage, in which the information being conveyed comprises a narration of past events. Finally, sample 11 is from a humor text and is representative of the fictional and biographical types of writing that use the features of this cluster for entertainment purposes.

#### Cluster 6:

##### Text sample 9 (LOB:B.20; editorial letter)

Communism had little or nothing to do with the riots in South Africa or the more recent disorders in Rhodesia. In fact, former leaders of the Communist Party in the Union have left the country. Some are now in the Rhodesian copper belt and at least one of them is in London.

In contrast, Moscow has embarked upon a special operation in Ruanda-Urundi, which borders on the Belgian Congo. This state of some 21,000 square miles and a population of 4,630,000 has been a United Nations trust territory under the administration of Belgium, but a few days ago she announced that she was giving up the trusteeship.

##### Text sample 10 (LOB:A.24; press reportage)

Four hundred angry Soccer fans chanted 'Sack the manager' outside Newcastle United Football Club's ground yesterday.

United had just been thrashed 4-0 by Everton, and now look certain to be relegated to the Football League's Division Two. Newcastle's manager is ex-winger Charlie Mitten.

At half-time, with United two goals down, one disgusted fan climbed the club's flagpole and hauled the Union Jack to half mast.

It was a riotous day for soccer....

##### Text sample 11 (LOB:R.2; humor)

He had long sensed injustice in the distinctions drawn between ordinary wage-earners and those self-employed. By the time his monthly salary arrived, the Inland Revenue had already taken their share, and there were precious few reductions in tax except for wives, children, life-insurances or any of the other normal encumbrances which Cecil had

so far avoided. He read the film star's sorry story and frowned at the provisions of Schedule D taxation which not only allowed her to claim relief on the most unlikely purchases, but also postponed demanding the tax until her financial year was ended, audited and agreed by the Inspector.

All three of these text samples illustrate the informational features associated with dimension 1: frequent occurrences of the bottom features (such as nouns, prepositional phrases, attributive adjectives) plus markedly infrequent use of the upper features (such as private verbs, contractions). This is true of the editorial (sample 9), which is primarily informative and expository in purpose, as well as the humor text (sample 11), which is primarily entertaining and narrative in purpose. Further, despite the different purposes of these texts, they all use narrative forms associated with dimension 2 (such as past-tense forms, perfect-aspect verbs, third-person pronouns). This tendency is most pronounced in the humor text sample, but it is found in all three samples. On the other three dimensions, these samples illustrate the unmarked characterization of cluster 6: not markedly 'elaborated' or 'situated' in reference, and not marked with respect to persuasion or abstract style.

The text type represented by this cluster has a special place in the present typology: it is the most general and nondistinct of the eight types. Although the texts in this type share a general linguistic characterization, having a carefully crafted, informational presentation and making relatively frequent use of narrative forms, the linguistic characterization of this type tends to be relatively unmarked on all five dimensions. As text samples 9-11 show, the texts in this type can have different purposes, although the underlying logical development used to achieve those purposes seems relatively similar. That is, all of these samples use a narrative line and careful informational elaboration to achieve their end. In the case of editorials, that end is analysis of some political or social situation; in the case of press reportage, that end is informing through a factual report of events; in the case of humor (as well as fiction and biography), that end is entertainment through a report of events. These texts belong to the same type in their surface characterizations and, to a lesser extent, in their underlying organizations; they show considerable variation, however, with respect to their specific purposes. I will return to the discussion of text type 6 relative to the other types in section 5.

4.2.5. *Cluster 7: 'Situated reportage'*. Cluster 7 is the smallest and most distinct text type identified in the study. Figure 4 shows that it is not marked with respect to dimension 1, but it is markedly nonnarrative,

nonpersuasive, and nonabstract (with respect to dimensions 2, 4, and 5). The most distinctive characteristic of this cluster is on dimension 3, where it is characterized as extremely situated in reference. The core texts grouped into this cluster are all sports broadcasts. This text type thus characterizes the on-line reportage of events which are in progress and occur in a fairly rapid succession. Text sample 12, taken from a broadcast of a soccer game, illustrates the distinctive characteristics of this text type.

#### Cluster 7:

Text sample 12 (LL:10.2; sports broadcast — soccer)

A:

Dunn # down the line # a bad one # it's Badger that gets it # he's got time to control it # [pause] he feeds in fact # Tom Curry # one of the midfield players ahead of him # [pause] Curry has got the ball # on that far side # chips the ball down the centre # [pause] again # a harmless one # [pause] no danger # out comes Stepney # [pause] and now left-footed # his clearance # [pause] is again # a long [pause] high # probing ball # down centrefield # onto the head of [long pause] Flynn # Flynn to Badger # Badger on the far side #

Sample 12 illustrates the distinctive characteristics of cluster 7: neither involved nor informational (that is, relatively few occurrences of either upper or bottom features from dimension 1), markedly nonnarrative (no past-tense verbs, perfect-aspect verbs, or other features from dimension 2), nonpersuasive (none of the features associated with dimension 4), and markedly nonabstract in style (none of the passive constructions associated with dimension 5). Although these characterizations all represent features that are markedly infrequent, they reflect the very specialized purpose and production situation of these texts: a speech event that describes events actually in progress to a large audience that is not present. For example, the distant relationship between broadcaster and audience results in the lack of involvement features; the rapid on-line production of text results in relatively few informational features; and the reportage of events in progress results in the nonnarrative characterization of these texts. The most distinctive positive characterization of type 7 texts is the extremely high use of expressions referring directly to the physical and temporal situation of communication. Thus, sample 12 contains numerous expressions such as *down the line*, *ahead of him*, *on that far side*, *down the centre*, which require direct reference to the playing field for understanding. The very frequent use of these expressions results in the extremely situated characterization of type 7 on dimension 3.

We might wonder why this text type is much more 'situated' than type

1, 'Intimate interpersonal interaction'. Both text types are noninformational, and in both types the participants share the same temporal context. In fact, in some respects we might expect type 1 to be more situated in reference than type 7 texts: in type 1 texts, participants actually share the same physical situation, and addressees can request clarification in cases of misunderstanding; in typical type 7 texts, the speaker (a radio broadcaster) does not actually share a physical situation with the listeners, and there is no possibility of clarification. There seem to be two reasons for the observed characterization of type 7 as extremely situated relative to type 1. First, the interactive texts in cluster 1 do not involve the same informational demands and are not produced under the same time constraints as those in cluster 7. That is, on-line reportage of sports events involves a rapid production of speech describing all relevant events as they occur. In such a situation, there is great demand for situated reference, because there are very many different referents to keep track of but very little opportunity for elaborated referring expressions. Interpersonal interaction (type 1), on the other hand, involves considerably fewer referents and provides considerably more opportunity for elaborated reference. Second, the expected style of sports broadcasts is one that gives the impression of an extremely rapid and exciting succession of events, even if this is not actually the case. Thus, even in the reportage of a baseball game, where events occur much more slowly than in a soccer match, many of the same features seen in sample 12 are frequently used (see Ferguson 1983). Due to the distinctive linguistic characteristics associated with these unusual production demands, sports broadcasts are isolated as a separate text type of 'situated reportage' by the present analysis.

4.2.6. *Cluster 8: 'Involved persuasion'*. Cluster 8 also represents a relatively specialized text type. In their linguistic characterization, the texts in this cluster are moderately involved, nonnarrative, elaborated in reference, and nonabstract in style (Figure 4). The most distinctive characteristic of cluster 8 occurs on dimension 4, where these texts are markedly persuasive in form. The texts grouped into this cluster (see the breakdown at the end of section 4.1) are also characteristically argumentative or persuasive in their primary purpose. The majority of these texts are spoken: some are interactional and informational, such as the interviews and the telephone conversation between disparates; others are informational monologues, such as the spontaneous and prepared speeches. The remaining texts are written, informational texts, such as popular lore, professional letters, religion, humor, and editorials. Overall, the linguistic characterization of these texts is primarily persuasive and

secondarily involved, while the texts themselves are primarily argumentative or persuasive in purpose, leading to the interpretive label 'Involved persuasion'.

Text samples 13-15 illustrate several of the different ways that texts can be persuasive as well as involved. Text sample 13 is from a session of parliament, in which a number of MPs interact with the Secretary of State. This text contains a number of short interactive monologues, with each speaker attempting to persuade the others. Text sample 14 is from a sermon. This is strictly a monologue, with the preacher attempting to 'exhort' and persuade the audience. Sample 15 is from a professional letter. This text is moderately interactive in that it is a response to a previous letter from a specific individual. It is also persuasive, in that it responds to specific questions and proposes a specific course of action.

#### Cluster 8:

Text sample 13 (LL:11.4; spontaneous speech — MPs in Parliament — interacting with each other and the Secretary of State)

Q: would he not agree that it is essential at the moment # that more [pause] should be free for exports and less absorbed within our public sector # [long pause]

A: well # I think I would accept on the latter point # that more of our resources must go # into [mm] into the balance of payments # ....

Q: would he agree that [mm] # [pause] an absence of such a statement # [pause] continues to generate uncertainty in the industry # and perhaps he might like to take this opportunity to [mm] # re-emphasize his support # for the second force airline # [long pause]

A: well I would certainly # [pause] regret it if # [pause] parts or # or indeed the whole of the [mm] review # [pause] was to dribble out # that's not my intention at all # [pause] we shall of course # [pause] indeed we are # [pause] studying it [pause] very carefully # [pause] ....

Text sample 14 (LL:12.1c; prepared speech — sermon)

A:  
we must # [long pause] have our corporate life together # as a church # [long pause] .... we can fight # [pause] and we must fight # [pause] against the world # the flesh # and the Devil # [pause] as individuals # [pause] but we must also fight # [pause] as the whole church of God # [long pause] .... we must have God's guidance # and grace # [pause] .... we must go out realizing # [pause] that without God's grace # [pause] we are utterly powerless # [long pause]

Text sample 15 (Z.1; professional letter)

Furthermore it really would be inappropriate for me to put words in your mouth. In short, you should really take the format of the resolution and put in your own thoughts .... the association is already sampling opinion on a number of other matters and it may be possible to add this one. If it is not possible to add your concern this year, it would certainly be possible to add it next year.

In all three of these samples, there is a high frequency of the 'persuasive' features associated with dimension 4. In sample 13, there are prediction modals (*would, shall*), necessity modals (*should, must*), possibility modals (*might*), suasive verbs (*agree*), and conditional subordination. Sample 14 shows an extreme use of necessity modals (*must*) as well as the use of possibility modals (*can*). Sample 15 also shows frequent use of modals (*should, may, would*) as well as conditional subordination. In addition, all three samples are relatively involved, as shown by features such as first- and second-person pronouns, emphatics and hedges, private verbs, etc. These samples thus illustrate a relatively specialized text type having a primary persuasive or argumentative purpose. It is interesting that the speakers and writers of these texts are fairly uniform in adopting an involved presentation rather than a strictly informational presentation. That is, these texts use both overt persuasive markers (dimension 4) and identification with the listener/reader and an informal, colloquial style (dimension 1) to make their point.

In contrast, some of the peripheral texts grouped in this cluster are strictly informational while being overtly argumentative and persuasive. Text sample 16, from a philosophy journal, illustrates this type of text.

Text sample 16 (LOB:J.54; academic prose — philosophy)

....the impression must be describable without reference to any event or object distinct from it. It must be possible to characterize that internal impression without invoking any reference to the so-called object of desire.... The supposition, then, that desiring or wanting is a Humean cause, some sort of internal tension or uneasiness, involves the following contradiction: As Humean cause or internal impression, it must be describable without reference to anything else ...; but as desire this is impossible. Any description of the desire involves a logically necessary connection with the thing desired. No internal impression could possibly have this logical property. Hence, a desire cannot possibly be an internal impression.

This text is directly persuasive. It overtly considers arguments and counterarguments and forcefully argues in favor of a point of view. In the

above sample, this is shown principally by an extremely frequent use of modals (*must, could, can*). At the same time, this text has a very informational characterization with respect to dimension 1 (frequent nouns, prepositional phrases, etc.). Texts like sample 16 are grouped into cluster 8 because they are overtly persuasive, but they are peripheral to this text type because they do not combine the typical use of persuasive and involved features.

In summary, then, the texts in text type 8 are primarily distinguished by their persuasive and argumentative emphases. This orientation is typically combined with an involved, often interactive, style, which aids the persuasive force of the text by developing a sense of solidarity with the listener or reader. In other cases, though, these texts can be overtly persuasive while having a marked informational focus, as in text sample 16.

## 5. Discussion and conclusion

The typology developed here is relatively complex, and sometimes the resulting 'text types' are surprising. For example, there is no single interactive or dialogue text type. Rather, the analysis identifies two major interactive types: Intimate interpersonal interaction (type 1), concerned primarily with the immediate interpersonal interaction, and Informational interaction (type 2), which has a primary informational emphasis. Similarly, there is no single expository text type. Instead the analysis identifies three expository types: Scientific exposition (type 3), which is extremely informational, elaborated in reference, and technical and abstract in style and content; Learned exposition (type 4), which is similar to Scientific exposition except that it is markedly less abstract and less technical in style; and General narrative exposition (type 6), which is a very general text type that combines narrative forms with expository, informational elaboration. In the same way, there is no single narrative text type. Instead, the analysis identifies General narrative exposition (type 6) and Imaginative narrative (type 5), which is a relatively involved text type having a primary narrative focus. The remaining two text types are relatively distinct. Type 7 is labeled Situated reportage, a text type reporting events actually in progress. Type 8 is labeled Involved persuasion; these are texts with a primary argumentative and persuasive purpose and style, which are also typically involved in presentation.

This typology replicates the major text distinctions identified in Biber and Finegan (1986). That study, which was based on an earlier three-dimensional model of textual variation (Biber 1986), is primarily methodological, exploring the feasibility of the approach adopted here for

identifying underlying text types. Although this earlier study considered a narrower range of texts and variation along only three dimensions, it shows a very high degree of convergence with the salient functional distinctions made in the present analysis.<sup>7</sup> In particular, both studies identify text types having the following functions: (1) intimate or immediate interaction, (2) informational interaction, (3) imaginative narrative, (4) and (5) two varieties of formal exposition ('scientific' and 'learned' in the present typology; with and without narrative in the 1986 study), (6) general, informal exposition, and (7) situated, immediate reportage. Type 8 of the present study ('Involved persuasion') was not identified in the earlier study, and a type labeled 'Interactional narrative' in the 1986 study was not replicated here. These latter two types require further study.

In the introduction, I noted that a typology of texts is needed as a theoretical basis for discourse and register studies. The typology developed here is immediately useful in this regard. For instance, numerous studies have described discourse characteristics of 'narrative' or 'exposition' — but the present typology shows that there is no single narrative or expository type. Rather, the typology identifies expository characteristics in three different text types and narrative characteristics in two text types. (In addition, Involved persuasion constitutes a fourth text type that has some expository features.) These text types have different linguistic and communicative characteristics, and each deserves study on its own terms. Texts chosen from one or another of these types will not adequately represent 'narrative' or 'exposition' as wholes; such a study would require analysis of texts chosen from all of the relevant types.

The typology also shows that the relationship between genres and text types is not straightforward, but it is not intended to invalidate the genre distinctions. I regard genre and text type categorizations as having different theoretical bases, so that they are both valid but distinct text constructs. As noted in the introduction, genres correspond directly to the text distinctions recognized by mature adult speakers, reflecting differences in external format and situations of use. The present study does not attempt to identify all of the basic genre distinctions in English; it only shows that the theoretical bases of genres are independent from those for text types. Genres are defined and distinguished on the basis of systematic nonlinguistic criteria, and they are valid in those terms. Text types, on the other hand, are defined on the basis of strictly linguistic criteria (similarities in the use of cooccurring linguistic features). As noted, text types often cut across genre categorizations. For example, large numbers of face-to-face conversations are grouped into both text type 1 and text type 2; academic prose texts are split among four text types: 3, 4, 6, and 8.

Although the texts in these genres are similar in their nonlinguistic characteristics, they belong to different 'types' in terms of their linguistic characterizations. The two perspectives are thus complementary.

By focusing on linguistic cooccurrence patterns, rather than on the distribution of individual features, the present study identifies some surprisingly subtle functional distinctions among texts. For instance, relative differences in the use of the cooccurring features that define dimension 1 correspond to the subtle functional differences between type 1 and type 2: both types are interactional, but 1 is strictly interpersonal while 2 is more informational. Similarly, differences in the use of the defining features of dimension 5 reflect the specialized functional distinction between type 3 and type 4: both are informational and expository, but 3 is abstract in style and represents technical specializations while 4 is nonabstract in style and represents specializations in humanities and the social sciences. Further, these surface cooccurrence patterns seem to reflect differences in rhetorical organization as well as communicative function. For example, the texts in types 3 and 4 all tend to have expository developments, while the texts in types 5 and 6 tend to have narrative developments, even though type 6 tends to have an informational or expository purpose. More research is required on the extent to which underlying rhetorical structure correlates with these functional types, but a quick survey indicates a surprisingly close relationship.

There are, however, sometimes diverse purposes among the texts in a type. This is especially the case for type 6, which is labeled General narrative exposition. I noted above that this is the most general and nondistinct of the text types. The texts in this type share the use of narrative forms and informational presentation for general expository purposes. There is some variation, however, among the specific purposes of these texts. Returning to text samples 9–11, we can see an analytical purpose in sample 9 (editorial), a simple factual reporting purpose in sample 10 (press reportage), and an entertainment purpose in sample 11 (humor). Type 5 (Imaginative narrative) is similar in this respect. The texts in this type have a primary narrative purpose and use narrative forms frequently, and most of these texts have fictional entertainment as their specific purpose. A few of the texts in type 5, though, are intended to inform rather than entertain (for example, the court case summary illustrated in sample 8 as well as two sermons). The other text types seem relatively consistent even in terms of their specific purposes, but types 5 and 6 show that the functional unity of the types is strongest at the level of general communicative functions (such as narration, exposition, interaction) rather than specialized intent (such as informing versus entertaining).

Finally, it must be emphasized that the text types identified here are in fact 'prototypes'. That is, these types represent the 'typical' text forms and functions of English rather than absolute distinctions among texts. The linguistic variation among texts was studied here in terms of a continuous five-dimensional space, where the types are dense concentrations of texts within that space. The types are based primarily on the areas of markedly high density, the 'core' texts, and secondarily on groupings of 'peripheral' texts. Because the peripheral texts do not occur in dense concentrations, they are assigned to the closest type; they are sometimes relatively dissimilar to that type, although they are even less similar to any other type.

Even if we limit the discussion to the core text types, the analysis here shows that the differences among types must be considered in relative terms. I noted in the discussion of Table 2 in section 4.1 that the types are not equally distinct. Some text types, like types 1, 2, and 7, are quite distinct from the other types; others, like types 3 and 4, are relatively similar to each other. In fact, there is a continuous range of variation among texts. It is theoretically possible for a text to have any score on each dimension, defining a continuous, multidimensional space of variation. It turns out, though, that there are regions that have very high concentrations of texts within that space, and these regions are identified as the text prototypes in English. In between these prototypes, there are particular texts that combine functional emphases and linguistic forms in complex and relatively idiosyncratic ways. These texts are not aberrations; they rather reflect the fact that speakers and writers exploit the linguistic resources of English in a continuous manner.

There are thus two complementary perspectives on linguistic variation among texts. One perspective focuses on the continuous nature of text variation; the other perspective, which forms the basis of the present study, identifies the relatively few distinct types that are frequently used in English. In theory, texts could be evenly distributed across possible linguistic and functional characterizations. This is not the case, however. Rather, the majority of texts are distributed across a few sets of linguistic form/function classes, and these marked concentrations of texts are interpreted as the major text 'types' of English. These types reflect marked tendencies of speakers and writers to construct texts around a limited set of functions and cooccurring linguistic forms. The typology thus gives structure to the multidimensional space of textual variation, even though it does not negate the continuous nature of that space.

Additional research on the dimensions of variation in English might help identify other, more specialized text types. The typology developed here, however, presents eight basic prototypes of texts in English. As such,

the typology provides an important step toward modeling the ways in which texts can differ from one another, providing the theoretical basis for discourse comparisons in English and a foundation for cross-linguistic research to identify universal dimensions of variation among texts.

Received 2 December 1987

University of Southern California

Revised version received

29 March 1988

## Notes

\* I would like to thank Pat Clancy, Ed Finegan, and an anonymous *Linguistics* reviewer for their many helpful comments on an earlier draft of this paper. Correspondence address: Department of Linguistics, University of Southern California, University Park, Los Angeles, CA 90089-1693, USA.

1. For example, past tense has a mean value of 40.1 and a standard deviation of 30.4 across all of the texts, and thus an absolute frequency of 113 translates into a standardized score of 2.4:

$$(113 - 40.1) / 30.4 = 2.4$$

That is, a frequency of 113 is 2.4 standard deviations from the mean of 40.1.

2. The same text corpus was used to determine the dimensions of variation and to develop the present typology. The written texts are taken from the Lancaster-Oldo-Bergen Corpus of British English (known as the LOB Corpus); the spoken texts are taken from the London-Lund Corpus of Spoken English. These two corpora are supplemented by private collections of personal and professional letters.
3. The FASTCLUS procedure from SAS was used for the clustering. Disjoint clusters were produced since there was no theoretical reason to expect a hierarchical structure. Peaks in the cubic clustering criterion and the pseudo F statistic, both produced by the FASTCLUS procedure, were used to determine the number of clusters to extract for analysis. These statistics provide a measure of the similarities among texts within each cluster in relation to the differences between the clusters. In the present case, both measures showed a peak for the eight-cluster solution, indicating that this solution provided the best fit to the data.
4. Core texts are those that have a distance of ten or less from their cluster centroid; peripheral texts have distances greater than ten. This distance was chosen because it excluded the major outliers in each cluster.
5. Clusters can have intermediate mean dimension scores for two reasons: the cluster is characterized by frequent occurrences of both positive and negative linguistic features on that dimension, or the cluster is characterized by the marked absence of both positive and negative features. Either distribution of features results in an unmarked characterization with respect to the dimension in question.
6. Text samples are labeled as follows:

CORPUS:GENRE.TEXT-NUMBER

For example, text sample 1 is labeled LL:1.8, because it is from the London-Lund Corpus, genre type 1 (face-to-face conversation), and text 8 number within that genre. In the spoken-text samples, # marks intonation unit boundaries.

7. There is considerable overlap in the texts used in these two studies (approximately 60–70%), which biases the results in favor of converging typologies.

## References

- Besnier, Niko (1986). Register as a sociolinguistic unit: defining formality. In *Social and Cognitive Perspectives on Language*, Jeff Connor-Linton, Christopher Hall, and Mary McGinnis (eds.), 25–63. Los Angeles: University of Southern California.
- Biber, Douglas (1986). Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62, 384–414.
- (1987). A textual comparison of British and American writing. *American Speech* 62, 99–119.
- (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- , and Finegan, Edward (1986). An initial typology of English text types. In *Corpus Linguistics II*, Jan Aarts and Willem Meijs (eds.), 19–46. Amsterdam: Rodopi.
- , and Finegan, Edward (1988). The drift of English genres from the 18th to the 20th centuries. Paper presented at the Georgetown University Round Table on Languages and Linguistics, Georgetown. (To appear in conference proceedings, edited by Thomas J. Walsh.)
- Brown, Penelope, and Fraser, Colin (1979). Speech as a marker of situation. In *Social Markers in Speech*, Klaus R. Scherer and Howard Giles (eds.), 33–62. Cambridge: Cambridge University Press.
- Chafe, Wallace L. (1982). Integration and involvement in speaking, writing, and oral literature. In *Spoken and Written Language: Exploring Orality and Literacy*, Deborah Tannen (ed.), 35–54. Norwood, N.J.: Ablex.
- Connor-Linton, Jeff, Hall, Christopher, and McGinnis, Mary (eds.), (1986). *Social and Cognitive Perspectives on Language*. Southern California Occasional Papers in Linguistics 11. Los Angeles: University of Southern California.
- Ervin-Tripp, S. M. (1972). On sociolinguistic rules: alternation and co-occurrence. In *Directions in Sociolinguistics*, John J. Gumperz and D. Hymes (eds.), 213–250. New York: Holt, Rinehart, and Winston.
- Ferguson, Charles A. (1983). Sports announcer talk: syntactic aspects of register variation. *Language in Society* 12, 153–172.
- Finegan, Edward, and Biber, Douglas (1986). Two dimensions of linguistic complexity in English. In *Social and Cognitive Perspectives on Language*, Jeff Connor-Linton, Christopher Hall, and Mary McGinnis (eds.), 1–24. Los Angeles: University of Southern California.
- Grabe, William (1984). Towards defining expository prose within a theory of text construction. Unpublished Ph.D. dissertation, University of Southern California.
- Hymes, Dell (1972). *Foundations of Sociolinguistics: An Ethnographic Approach*. Philadelphia: University of Pennsylvania Press.
- Longacre, Robert (1976). *An Anatomy of Speech Notions*. Lisse: de Ridder.
- Redeker, Gisela (1984). On differences between spoken and written language. *Discourse Processes* 7, 43–55.
- Smith, Edward L. (1985). Text type and discourse framework. *Text* 5, 229–247.
- Tannen, Deborah (1982). Oral and literate strategies in spoken and written narratives. *Language* 58, 1–21.