Conversation text types: A multi-dimensional analysis

Douglas Biber

Northern Arizona University, USA
douglas.biber@nau.edu

Abstract

Multi-dimensional (MD) analysis is a methodological approach that applies multivariate statistical techniques (especially factor analysis and cluster analysis) to the investigation of register variation in a language. The approach was originally developed to analyze the full range of spoken and written registers in a language. Early studies focused on English register variation (Biber 1985, 1986 and 1988), while later studies have applied the same approach to Somali, Korean, Tuvaluan, Taiwanese, and Spanish.

Surprisingly, these studies have found some striking similarities in the underlying 'dimensions' that distinguish among spoken and written registers in these diverse languages. It is even more surprising that MD studies of restricted discourse domains have also uncovered dimensions that are similar in linguistic form and function to the more general studies of register variation.

The present study presents an MD analysis of a single register: conversation. Three primary dimensions of variation are identified, and then cluster analysis is used to distinguish among six conversation text types. The dimensions and text types are interpreted in linguistic and functional terms.

The author's expectations were that a unique set of dimensions would emerge to characterize the variation among conversational texts. Instead, the three dimensions identified here turn out to be closely related to dimensions identified in previous analyses of general register variation. Taken together with previous studies, the present study of conversation raises the possibility of universal dimensions of variation.

1. Introduction

Multi-dimensional (MD) analysis is a methodological approach that applies multivariate statistical techniques (especially factor analysis and cluster analysis) to the investigation of register variation in a language. The approach was originally developed to analyze the range of spoken and written registers in English (Biber 1985, 1986 and 1988). There are two major quantitative steps in an MD analysis: (1) identifying the salient linguistic co-occurrence patterns in a language; and (2) comparing spoken and written registers in the linguistic space defined by those co-occurrence patterns. In a third step, it is possible to identify groupings of texts — 'text types' — that are maximally similar in their multi-dimensional profiles.

Almost any linguistic feature will vary in its distribution across registers, reflecting the discourse functions of the feature in relation to the situational characteristics of each register (see, e.g., the grammatical descriptions in the *Longman Grammar of Spoken and Written English*; Biber *et al.*, 1999). However, individual features cannot reliably distinguish among registers: There are simply too many different linguistic characteristics to consider, and individual features often have idiosyncratic distributions. Instead, analyses based on linguistic *co-occurrence* and *alternation* patterns are required to uncover general register differences.

The theoretical importance of linguistic co-occurrence has been emphasized by linguists such as Firth, Halliday, Ervin-Tripp, and Hymes. Brown and Fraser (1979: 38-39) observe that it

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

The MD approach gives formal status to the notion of linguistic co-occurrence, by providing empirical methods to identify and interpret co-occurrence patterns as underlying *dimensions* of variation. The co-occurrence patterns comprising each dimension are identified quantitatively through factor analysis. It is not the case, though, that quantitative techniques are sufficient in themselves for MD analyses of register variation. Rather, qualitative techniques are required to interpret the functional bases underlying each set of co-occurring linguistic features. The dimensions of variation have both linguistic and functional content. The linguistic content of a dimension comprises a group of linguistic features (e.g., nominalizations, prepositional phrases, attributive adjectives) that co-occur with a high frequency in texts. Based on the assumption that co-occurrence reflects shared function, these co-occurrence patterns are interpreted in terms of the situational, social, and cognitive functions most widely shared by the linguistic features. That is, linguistic features co-occur in texts because they reflect shared functions.

Several experiments have been carried out to evaluate the reliability (and to a lesser extent validity) of the original MD analysis of register variation in English. For example, Biber (1990) shows that factor analyses carried out on split corpora result in nearly the same dimensions of variation, as long as the texts in those corpora are sampled to include equivalent ranges of register variation. Biber (1993) shows how these dimensions can be used to predict the register category of unclassified texts with a high degree of accuracy (using discriminant analysis). And Biber (1992) uses confirmatory factor analysis to test the goodness of fit of several factorial models determined on theoretical grounds, confirming the basic structure identified using exploratory factor analysis in the 1988 analysis.

While early MD studies focused on register variation in English, subsequent studies have applied the same approach to Somali, Korean, Tuvaluan, Taiwanese, and Spanish (see, e.g., Biber, 1995; Jang, 1998). Although these studies all apply the same methodological approach, they are carried out independently. In each case, a corpus was designed to represent the range of spoken and written registers found in the target culture, and a computational tagger was written to capture the grammatical structure of the target language. The set of linguistic variables used in each analysis includes the full range of lexical/grammatical distinctions that are relevant in the target language. Despite this fact, the resulting MD analyses have turned out to be strikingly similar in some respects. In particular, the analyses of all languages have uncovered dimensions relating to interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse (see Biber, 1995, especially Chapter 7).

The MD methodological framework has also been applied to more restricted discourse domains.[1] These include analyses of elementary school registers (Reppen, 1994 and 2001),

[1] There have also been several studies of specific registers that apply the dimensions that were identified and interpreted in the 1988 MD analysis of spoken and written variation in English (see, for example, the collection of studies in Conrad and Biber, 2001). It is important to note that these studies do *not* entail separate MD analyses. That is, these studies apply the dimensions identified in the 1988 MD analysis of English to some new discourse domain, but they do not undertake new MD analyses (i.e., involving a new factor analysis).

job interview language (White, 1994), television talk shows (Connor-Linton, 1989), 18th century written and speech-based registers (Biber, 2001), university spoken and written registers (Biber, 2003), and academic subregisters (e.g., Grabe, 1987; Kanoksilapatham, 2003). Many of these studies have identified dimensions of variation similar to those found in the cross-linguistic studies, especially relating to the same functional concerns of interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse.

This result is surprising for two reasons. First, the statistical technique of factor analysis — like all correlational techniques — requires variability. Two linguistic variables cannot be shown to correlate unless the texts included in an analysis represent a wide range of variation for those variables. Similarly, factor analysis cannot reliably identify sets of co-varying linguistic features unless the texts included in the analysis represent a wide range of variation for the full set of features. Thus, factor analysis is most appropriate for general analyses of spoken and written texts, which represent an extensive range of variation for almost any linguistic feature (see the detailed analyses in Biber *et al.*, 1999). In contrast, it might be assumed that factor analysis is less appropriate for analyses of texts from a single, restricted discourse domain, because that domain will represent a much smaller range of variation.

Second, to the extent that there is linguistic variability among the texts in a restricted discourse domain, there is no reason to assume that it would be similar to the patterns of variation found in a general-purpose corpus. We would rather expect to find different linguistic features varying in a restricted domain, reflecting the specific functional differences found in that domain. In the MD analyses, these specific patterns of linguistic variation should result in dimensions of variation that are unique to each discourse domain.

Previous MD analyses have shown that restricted discourse domains represent sufficient linguistic variability for the successful application of this methodological approach. More surprisingly, these analyses show that some of the same basic dimensions of variation seem to be fundamentally important across restricted and general discourse domains. (In addition, there are other dimensions that are unique to a particular domain.) This repeated finding — that some dimensions occur across languages and across general and restricted discourse domains — raises the possibility of universal dimensions of register variation.

The present study further explores this possibility by undertaking an MD analysis of linguistic variation within a single spoken register: conversation. Factor analysis is used to identify the linguistic dimensions of variation operating in this discourse domain, and then cluster analysis is used to identify conversation 'text types' that are well-defined in that multi-dimensional space. The following sections describe these analyses, followed by discussion of the more general theoretical implications for the study of register variation.

2. Overview of methodology in the Multi-Dimensional approach

A Multi-Dimensional analysis follows eight methodological steps:

1. An appropriate corpus is designed based on previous research and analysis. Texts are collected, transcribed (in the case of spoken texts), and input into the computer. The situational characteristics of each spoken and written register are noted (e.g., communicative purpose, production circumstances, etc.).

2. Research is conducted to identify the linguistic features to be included in the analysis, together with functional associations of the features.

3. Computer programs are developed for automated grammatical analysis, to identify — or 'tag' — all relevant linguistic features in texts.

4. The entire corpus of texts is tagged automatically by computer, and all texts are edited interactively to insure that the linguistic features are accurately identified.

5. Additional computer programs compute normed counts of each linguistic feature in each text of the corpus.

6. The co-occurrence patterns among linguistic features are analyzed, using factor analysis.

7. The factors are interpreted functionally as underlying dimensions of variation.

8. Dimension scores for each text are computed; the mean dimension scores for each register are then compared to analyze the salient linguistic similarities and differences among the registers being studied. The functional interpretation of each dimension is refined based on the distribution among registers.

3. Preliminary steps for the MD analysis of conversation: Corpus and linguistic features

3.1. *The conversation corpus*

The corpus used for the present analysis is taken from the Longman Spoken and Written English Corpus (LSWE Corpus; see Biber *et al.*, 1999: chapter 1). Only the British English sub-corpus of conversation was analyzed here; this sub-corpus includes 164 texts containing c. 4 million words. (A large part of this corpus was also included in the BNC sample of conversation.) Texts were collected by asking participants to carry tape recorders for several days, recording their daily interactions. The language collected in this way is conversational, but most text files are very large and actually include many different conversations. Participants would generally turn the tape recorder off in between conversations, but each text file in the corpus includes all the conversations that were recorded on a single tape.

The LSWE/BNC corpus of conversation is large enough to provide the basis for a multi-dimensional analysis. Texts were collected over many days by many different participants, representing a wide range of social backgrounds. As a result, the corpus should represent the range of linguistic variability found within conversation. However, much of that variability is lost when the corpus is analyzed in its current form, with each text file combining multiple conversations. Individual conversations can vary with respect to situation and purpose, and to the extent that there is linguistic variability among conversational texts, it will be associated with those situational/communicative differences. As a result, the first step in the present analysis was to segment text files into individual conversations (based on the internal headers included in each text file). Table 1 shows that the 164 text files in the LSWE conversational corpus were segmented into 2,926 individual conversations. 760 of these conversations were shorter than 200 words. Because of the difficulties in obtaining reliable rates of occurrence for linguistic features in shorter texts, these shorter conversations were excluded from subsequent analysis.

Table 1 shows that the conversations included in the analysis are on average quite long (1,775 words), with the longest conversations being almost 14,000 words.

# of text files: 164 (conversation transcripts from the LSWE Corpus) # of words: 3,930,000
Individual conversations longer than 200 words: 2,166 Individual conversations shorter than 200 words: 760 (dropped from subsequent analyses) Total individual conversations: 2,926
Length of individual conversations included in the analysis: mean = 1,775 words min = 200 words max = 13,776 words

Table 1. Initial segmentation of the conversation corpus into individual conversations

3.2. *Linguistic features used for the analysis*

After the conversation corpus was segmented, each conversation was automatically 'tagged' using the Biber grammatical tagger. The current version of this tagger incorporates the corpus-based research carried out for the *Longman Grammar of Spoken and Written English* (Biber *et al.*, 1999). The tagger identifies a wide range of grammatical features, including word classes (e.g., nouns, modal verbs, prepositions), syntactic constructions (e.g., WH relative clauses, conditional adverbial clauses, that-complement clauses controlled by nouns), semantic classes (e.g., activity verbs, likelihood adverbs), and lexico-grammatical classes (e.g., that-complement clauses controlled by mental verbs, to-complement clauses controlled by possibility adjectives). Appendix A lists the full set of features analyzed here.

4. Identifying and interpreting the dimensions of variation in English conversation

As noted above, the Multi-Dimensional approach to register variation uses factor analysis to reduce a large number of linguistic variables to a few basic parameters of linguistic variation. In MD analyses, the distribution of individual linguistic features is analyzed in a corpus of texts. Factor analysis is then used to identify the systematic co-occurrence patterns among those linguistic features — the 'dimensions' — and then texts and registers are compared along each dimension.

Table 2 gives the full factorial structure for the analysis in this case, while Table 3 summarizes the important linguistic features defining each dimension (i.e., features with factor loadings over + or –.3). Only 27 of the original 120+ linguistic features were retained in the final factor analysis. Several features were dropped because they were redundant or overlapped to a large extent with other features. For example, the counts for common verbs, nouns, and adjectives overlapped extensively with the semantic categories for those word classes, even though the counts were derived independently. In other cases, features were dropped because they were extremely rare in conversation. Several of these features were combined into a more general class. For example, the seven phrasal verb types were combined into a single feature. Similarly, the five specific types of postnominal modifying clause were combined into a single 'relative clause' feature. Finally, many features were dropped either because they did not vary across conversational texts, or because they shared little variance with the overall factorial structure of this analysis (as shown by the communality estimates). The solution for three factors was selected as optimal. These three factors account for only 36% of the shared variance, but they are readily interpretable, and subsequent factors accounted for relatively little additional variance. Given that only 27 linguistic variables were retained in the final factor analysis, the solution with 3 factors was considered optimal.

	Factor1	Factor2	Factor3
Major Factor 1 Features:			
wrdlngth	0.75638	0.06122	-0.08615
n_nom	0.52756	0.02817	-0.08707
prep	0.46987	0.04783	0.05038
abstrctn	0.46893	0.13866	-0.11843
rels	0.44690	0.16282	-0.09168
adj_attr	0.35287	-0.17778	-0.08123
allpasv	0.29913	0.08863	0.09965
contrac	-0.43589	0.24817	-0.18706
pro1	-0.39020	0.23735	0.08420
pro2	-0.36418	-0.07955	-0.25500
actv	-0.31900	-0.15923	0.03750
Major Factor 2 Features:			
that_del	-0.02783	0.67341	0.38349
mentalv	-0.01213	0.66432	-0.07223
fact_vth	0.14676	0.54415	0.08570
lkly_vth	0.01636	0.42516	0.01091
lklyadvl	0.38093	0.40397	-0.07723
sub_all	0.07570	0.35545	-0.00143
gen_hdg	0.33564	0.34281	-0.08784
factadvl	0.28542	0.33556	0.01889
n	0.16827	-0.52004	-0.04842
wh_ques	-0.23266	-0.34870	-0.10075
Major Factor 3 Features:			
pasttnse	-0.04272	-0.04189	0.79494
nonf_vth	-0.12568	0.14077	0.60910
commv	-0.13631	0.06964	0.58285
pro3	0.00932	0.10243	0.52077
pres	-0.45977	0.26021	-0.51128
allmodal	-0.23413	0.20474	-0.26521
Inter-Factor Correlations			
	Factor1	Factor2	Factor3
Factor1	1.00000	-0.24213	0.09272
Factor2	-0.24213	1.00000	0.08185
Factor3	0.09272	0.08185	1.00000

Table 2. Results of the factor analysis: 3 factor solution; Promax rotation.

Each factor comprises a set of linguistic features that tend to co-occur in the conversations from the conversation corpus. Factors are interpreted as underlying 'dimensions' of variation based on the assumption that linguistic co-occurrence patterns reflect underlying communicative functions. That is, particular sets of linguistic features co-occur frequently in texts because they serve related communicative functions. Features with positive and negative loadings represent two distinct co-occurrence sets. These comprise a single factor because the two sets tend to occur in complementary distribution: when a conversation has high frequency

of the positive set of features, that same conversation will tend to have low frequencies of the negative set of features, and vice versa. In the interpretation of a factor, it is important to consider the likely reasons for the complementary distribution between positive and negative feature sets as well as the reasons for the co-occurrence patterns within those sets.

Dimension 1: Information-focused vs. interactive discourse

Features with positive loadings: word length, nominalizations, prepositional phrases, abstract nouns, relative clauses, attributive adjectives, passive verb phrases, (likelihood adverbs, general hedges)

Features with negative loadings: present tense verbs, contractions, 1st person pronouns, 2nd person pronouns, activity verbs

Dimension 2: Stance vs. context-focused discourse

Features with positive loadings: *that*-deletions, mental verbs, factual/mental verb + *that*-clause, likelihood/mental verb + *that*-clause, likelihood adverbs, adverbial clauses, general hedges,

factual adverbs

Features with negative loadings: nouns, *WH*-questions

Dimension 3: Narrative-focused discourse

Features with positive loadings: past tense verbs, 3rd person pronouns, non-factual/communication verb + *that*-clause, communication verbs, *that*-deletions

Features with negative loadings: present tense verbs

Table 3. Summary of the factorial structure

For example, the positive features on Factor 1 (e.g., long words, nominalizations, prepositional phrases, abstract nouns, relative clauses, etc.) all relate to informational purposes. These features are mostly associated with elaborated noun phrases and a dense integration of information in a text; previous MD studies have shown these features to be typical of written non-fictional registers intended for specialist audiences (see, e.g., Biber, 1995; Biber and Finegan, 2001).

In contrast, the negative features on Dimension 1 reflect a focus on the immediate interaction and activities: present tense verbs, contractions, 1st and 2nd person pronouns, and activity verbs. The overall interpretation of Dimension 1 is thus relatively straightforward, showing that conversations tend to be either 'informational' or 'interactive', but not both. The functional label 'Information-focused versus interactive discourse' is proposed for this dimension.

The positive features on Dimension 2 are mostly linguistic features that express 'stance': personal attitudes or indications of likelihood. In the 1988 MD study of spoken and written register variation, several of these features were shown to co-occur typically with interactive and reduced structure features (on Dimension 1). In contrast, the analysis here shows that stance-focused discourse is not necessarily highly interactive discourse, and vice versa. (This dimension also includes several specific features that were not distinguished in the feature set used for the 1988 analysis, such as likelihood/mental verb + that-clause and factual adverbs).

The negative pole of Dimension 2 shows a surprising co-occurrence of only two features: nouns and WH-questions. In past analyses, nouns have co-occurred with other stereotypically 'literate' features (like adjectives, prepositional phrases, etc.), while WH-questions have co-occurred with stereotypically 'oral' and interactive features. The interpretation here must consider why these two features would tend to co-occur in conversations, and why they would

tend to occur in a complementary distribution to stance features. Consideration of texts with a high frequency of these two features indicates that they are used together to reflect a focus on the larger context. WH-questions — the 'what', who', 'where', 'when' and 'how' — directly ask about that context, and nouns are the primary device used to refer to it. Thus, considering both positive and negative poles, we propose the interpretive label 'Stance-focused versus context focused discourse' for Dimension 2.

Finally, Dimension 3 is composed of stereotypically narrative features — past tense verbs, 3rd person pronouns, and communication verbs controlling that-clauses; the only negative feature on this dimension is present tense verbs. Given this grouping of features, the interpretation as 'Narrative-focused discourse' is uncontroversial.

5. Identifying and interpreting conversation text types

Most MD studies have been undertaken to investigate the patterns of variation among 'registers': varieties of language that are defined by their situational (i.e. non-linguistic) characteristics (see Biber, 1994). Conversation is an example of a register according to this definition, as is newspaper reportage, classroom lectures, personal letters, and academic research articles. Registers can be defined at any level of specificity, depending on the extent to which the situational characteristics are specified. For example, academic prose is a very general register, while academic research articles, psychology research articles, and methodology sections in experimental psychology research articles are registers defined at increasing levels of specificity. The original MD studies (Biber, 1986 and 1988) analyzed a wide range of general spoken and written registers in English, while many subsequent analyses have applied those dimensions to the analysis of other more specialized registers (see, e.g., the studies in Conrad and Biber 2001).

These analyses have shown that there are important, systematic linguistic differences among registers. Those linguistic differences exist because of the functional basis of MD analysis: linguistic co-occurrence patterns reflect underlying communicative functions. Registers differ in their situational/communicative characteristics, and as a result, the dimensions identify important linguistic differences among registers. However, it is important to note that the register categories are defined in situational rather than linguistic terms.

A complementary perspective on textual variation is to identify and interpret the text categories that are **linguistically** well defined, referred to as **text types**. Text type distinctions have no necessary relation to register distinctions. Rather, text types are defined such that the texts within each type are maximally similar in their linguistic characteristics, regardless of their situational/register characteristics. However, because linguistic features have strong functional associations, text types can be interpreted in functional terms.

Text types and registers thus represent complementary ways to dissect the textual space of a language. Text types and registers are similar in that both can be described in linguistic and in situational/functional terms. However, the two constructs differ in their primary bases: registers are defined in terms of their situational characteristics, while text types are defined linguistically.

In the MD approach, text types are identified quantitatively using Cluster Analysis, with the dimensions of variation as predictors. Cluster analysis groups texts into 'clusters' on the basis of shared multi-dimensional/linguistic characteristics: the conversations grouped in a cluster are maximally similar linguistically, while the different clusters are maximally distinguished. This approach has been used to identify the general text types in English and Somali (see

Biber, 1989 and 1995). The present section describes the text types that can be distinguished linguistically within the single register of conversation.

The dimensions of variation (see Section 4 above) are used as linguistic predictors for the clustering of conversations. The individual feature counts are first standardized so that each feature has a comparable scale with a mean of 0.0 and a standard deviation of 1. (The standardization was based on the overall means and standard deviations for each feature in the conversation corpus.) Then, 'dimension scores' were computed by summing the standardized frequencies for the features comprising each of the three dimensions. The cluster analysis is based on the three dimension scores for each conversation.

The methodology in this analytical step can be illustrated conceptually by the 2-dimensional plot in Figure 1. Each point on Figure 1 represents a conversation, plotting the scores for that conversation on Dimensions 1 and 2. The numbers in the figure show the cluster number for each conversation, based on the results of the cluster analysis. Conversations that are similar in their dimension scores are grouped together as a cluster, or 'text type'. For example, the conversations labelled with a '1' on Figure 1 all have large positive scores on Dimension 1 (the vertical axis) and large negative scores on Dimension 2 (the horizontal axis). In contrast, Cluster 2 has positive scores on both Dimensions 1 and 2.

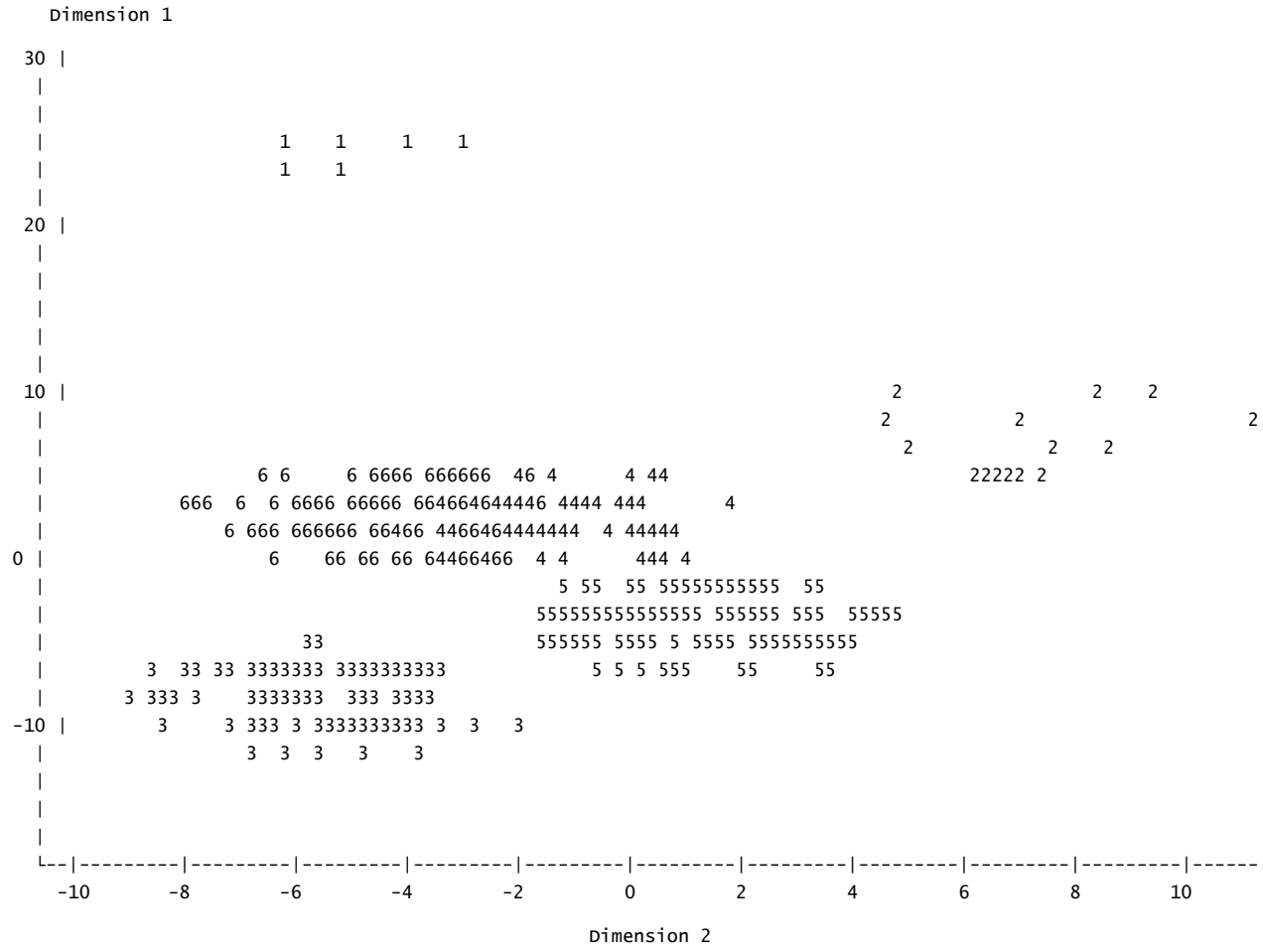


Figure 1. Plot of VBDUs along Dimension 1 vs. Dimension 2 (showing all DUs with a distance < 3 from the cluster centroid. Symbol is value of CLUSTER; NOTE: 194 obs hidden.)

Cluster analysis performs this grouping statistically, based on the scores for all three dimensions. Figure 1 shows the distribution across only two dimensions (1 and 3); these two dimensions were chosen because they provide a good visual display of how the conversations within each cluster are grouped based on their dimension scores. However, the actual cluster analysis uses all three dimension scores to identify the groupings of conversations that are maximally similar in their linguistic characteristics.

Cluster analysis is an exploratory statistical technique. The FASTCLUS procedure from SAS was used for the present analysis. Disjoint clusters were analyzed because there was no theoretical reason to expect a hierarchical structure. Peaks in the Cubic Clustering Criterion and the Pseudo-F Statistic (produced by FASTCLUS) were used to determine the number of clusters. These measures are heuristic devices that reflect goodness-of-fit: the extent to which the texts within a cluster are similar, while the clusters are maximally distinguished. In the present case, these measures had peaks for the 3-cluster solution and for the 6-cluster solution. The latter was chosen for subsequent analyses because it provided greater discrimination among the specialized clusters, facilitating the interpretation of those clusters as conversation text types.

Figure 1 shows the distribution of these six clusters in only a 2-dimensional space, whereas the cluster analysis is actually based on a 3-dimensional space. It turns out that the third dimension is also important in defining some clusters. For example, Cluster 4 is not sharply delimited in terms Dimensions 1 and 2, but all conversations in this cluster have large positive scores on Dimension 3 (‘narrative’).

Tables 4 and 5 provide a descriptive summary of the cluster analysis results. Table 4 shows the number of conversations grouped into each cluster, while Table 5 gives descriptive statistics for each dimension across the clusters. The clusters differ notably in their distinctiveness: the smaller clusters are more specialized and more sharply distinguished linguistically. For example, Cluster 1 has only 40 conversations; linguistically, the conversations grouped in Cluster 1 have extremely large positive scores on Dimension 1 (‘informational’); large negative scores on Dimension 2 (‘context-focused’); and scores near 0.0 on Dimension 3 (‘narrative’). At the other extreme, Cluster 5 is a ‘general’ text type: it is large (680 conversations) and relatively unmarked in its dimension scores.

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	40	4.6276	19.8319	2	18.2629
2	116	4.3710	16.9770	4	9.5460
3	496	3.2839	18.7622	5	8.6697
4	308	3.4828	18.2692	6	8.2551
5	680	3.2643	16.2853	4	8.3268
6	526	3.2447	17.4048	4	8.2551

Table 4. Summary of the Cluster Analysis

Cluster Means

Cluster	Dim. 1	Dim. 2	Dim. 3	
1	22.15	-5.08	-0.99	(Informational context-focused)
2	7.67	5.87	0.93	(Informational stance-focused)
3	-8.04	-5.19	-2.88	(Interactive context-focused)
4	2.12	-0.31	5.61	(Narrative)
5	-4.15	1.74	0.55	(Unmarked interactive)
6	2.63	-4.46	-1.50	(Unmarked context-focused)

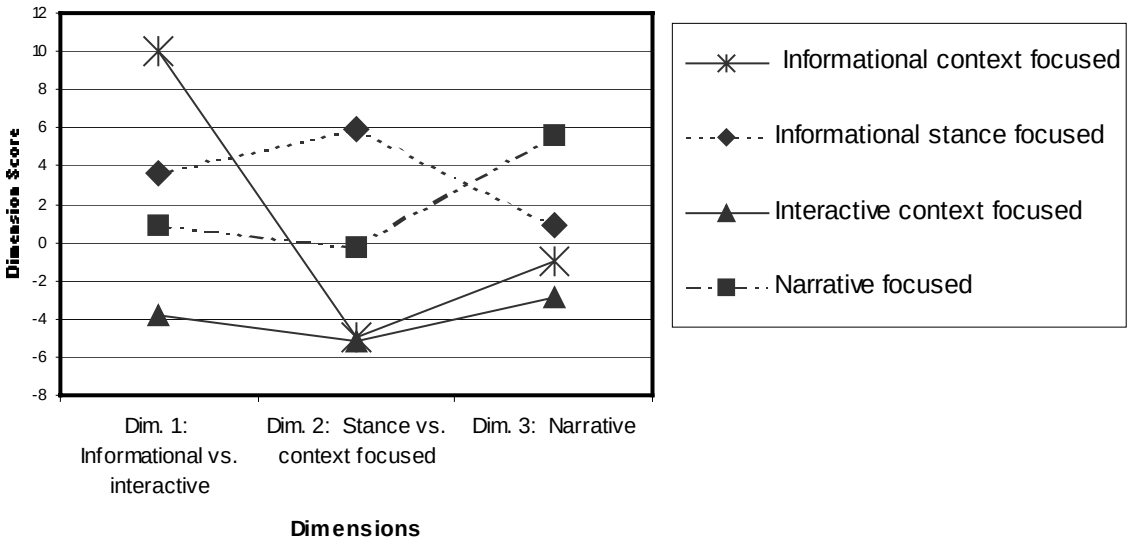
Cluster Standard Deviations

Cluster	Dim. 1	Dim. 2	Dim. 3
1	5.02	5.19	3.46
2	5.05	4.47	3.41
3	3.75	3.42	2.54
4	3.72	3.28	3.42
5	3.11	3.49	3.17
6	3.75	3.54	2.22

Table 5. Cluster descriptive statistics for each dimension

The clusters can be interpreted as Conversation Text Types, because each cluster represents a grouping of conversations with similar linguistic profiles. Figure 2 compares the linguistic characteristics of the four most distinctive of these conversation types, plotting their mean dimension scores. The 'general' conversation types — clusters 5 and 6 — are not plotted in Figure 2.

Figure 2: Multi-Dimensional profile for Conversational Text Types 1-4
(Note: Dimension 1 has been transformed to a scale of 10 for comparison)



Taken together, Table 4 and Figure 2 provide the basis for the interpretation of each conversation type. (These interpretations are refined by consideration of individual conversations from each type.)

Type 1 is the most specialized, with the fewest number of texts (only 40, or about 2% of the conversations in the corpus). Linguistically, these conversations are extremely informational (Dimension 1) and focused on the context (Dimension 2). Text Sample 1 provides an example of a conversation from this cluster. This text illustrates the dense use of 'informational' features, such as nominalizations (e.g., *conversation, sophistication, agreement, possibility, information*), other long words (e.g., *paperwork, computer-wise, computerized, consequently*), attributive adjectives (*modern, massive, great, preliminary, certified*), passives (*be/getting inundated*), prepositional phrases (*to you, about the conversation, with Alec, on a piece of paper*), and relative clauses (*things that you're liable to get asked*). Although texts from this cluster would be considered interactive and involved in comparison to written expository texts, they are highly informational in comparison to other conversational texts.

Text Sample 1 Conversation from Cluster 1: Informational, Context-focused

A: I, I want to talk to you about er the conversation I had with Alec <name> yesterday, he seems to be inundated with having to get details about <unclear> on his er, all his paperwork and so on, and he seems to be inundated and he sounded a bit low, quite frankly, to me yesterday on the phone that he was getting inundated with all this
B: Mm, mm
A: work. I said I'm quite sure there must be something that could be done computer-wise
B: Right
A: but he sort of pooh-poohed it and sort of said well you know, we're getting a bit too old for all this modern sophistication of computers and so on, well I said well quite frankly I am not totally in agreement with you, because as you probably know Clyde <name> was looking into a program which will could alleviate a lot
B: Yes I know, I know
A: of the work, that I do, but I
B: yes it's on the <unclear>
A: would tell you right here and now, er I'm still retaining my bible you know the book
B: Yeah, yes, yes
A: that I have downstairs, because it's, if it was to be computerized, it would be a massive great bloody great volume
B: Yes
A: and I would be carrying this around and it just wouldn't be feasible
B: Quite, right
A: so he said that apparently whenever he came back to B S H he was told by Neville roughly about eighteen hundred acres would be sort of his target
B: Target, right
A: and it's, it's multiplied by about three or four times that you see
B: Oh right, right, right
A: so consequently he's getting inundated, he really is apparently under pressure
B: Mm, mm, right
A: so this is why I raised the very conversation about it
B: Right, right
A: and er, I said well look I'll have a word with the erm, with <name> and see if he can think of anything that might
B: Yes alleviate the point
A: all things in mind that are possible on er, on er computer, and he said that he hadn't much time to think about it and said well look, maybe over Easter
B: Mm
A: put down on a piece of paper what essentials you want done
B: Right
A: and what things that you're liable to get asked
B: Right, mm, mm
A: so he's going to do that, so I said well look, do you mind if I had a wee sort of preliminary talk with him

B: Right
A: see if, if it's a possibility
B: Right
A: What he's looking for is certified numbers, field numbers
B: That sort of
A: all this sort of information

Type 2 is also relatively specialized (with only 116 conversations grouped into this cluster). Linguistically, this conversation type is relatively informational (Dimension 1) but especially marked for being highly stance-focused (Dimension 2). (This conversation type should be contrasted with Type 5: a much larger cluster that is stance-focused and highly interactive rather than informational.) Text Sample 2 illustrates the typical linguistic characteristics of Conversation Type 2. Notice especially the frequent mental verbs (e.g., *know, think, expect, want*), stance verbs controlling *that*-clauses, usually with the *that* omitted (e.g., *would have thought…, I think…, I suppose…*), and the frequent hedges and stance adverbs and adverbials (*surely, obviously, really, actually, probably, certainly, to be perfectly frank*).[2] Texts from this cluster are informational, in that they are focused on discussion of a particular topic rather than the immediate interpersonal interaction, but their primary purpose is the expression of personal stance in relation to that topic.

Text Sample 2. Conversation from Cluster 2: Informational, Stance-focused

A: No no no one person that's not right.
B: Oh, right.
A: There is no, statutory obligation for the person organizing it
C: Oh, I know.
B: Well not the organizer surely oh I know I would have thought you'd have to, <unclear> shoot it
A: I'm sure that the social services require psychiatric or
B: Mm, I would of thought so
A: obviously medical <unclear> what you're doing. Mhm but they're to be qualified people involved. But I would have expected that the whole thing would have to be operated by, somebody who was qualified.
B: I don't know, because like, you know like the doctors <unclear>
A: <unclear> I think it sort of depends how big that you want to get involved in. If you're just somebody who's on the outside providing services, to keep the smooth running of it then you don't really have to know anything about it.
C: Mm.
A: But if you're actually involved in it, and you want to be involved in the people, then I think you have to know something about it.
B: Well the other evening they were showing something on TV, one of these doctors', doctors' practices that are opting out or whatever. And they got a stockbroker, someone who used to be a stockbroker, actually managing the whole practice.
A: Yeah.
B: I mean he's obviously not qualified as a doctor.
A: Mhm.
B: So I mean I suppose they'll look at it in the same kind of way, somebody who's got managerial, management qualities rather than . — I suppose people who are interested in the other side of it, the medical side of it, probably, really be geared up to organizing the money side of it wouldn't they, usually one or the other.
C: So have you done any more calculations on it?
A: There's nothing really more <unclear> I mean the whole thing is a budget guesstimate. I've no idea yet, really what, I mean, you know, for instance I don't know how much ratio staff to patients they need, therefore you can't really, you know, follow that up because you've no idea what the costs themselves could be.
D: Well you don't know, have you, have you found the statutory requirement for space yet? Per person. —

[2] Note also the dense use of discourse markers (e.g., *I mean, you know*) supporting the expression of stance in this conversation.

A: I think the thing is going to come unstuck . — in the, I think the biggest thing is, I was thinking, is the fact that you've got to get <unclear> I wouldn't get a commitment from Social Services until they see a property actually ready for occupation. Now I'm not gonna be prepared to go through the whole business and then find them say oh sorry you're wrong.
C: Property is the biggest bugbear.
A: Yeah. Because I don't think
C: If you're actually sitting on <unclear>
A: I don't think the banks are gonna want to invest. To be perfectly frank. . — You see the only way we can get equity out and put money in ourselves is by selling this place.
C: Yes.
A: Therefore if we don't actually want to live in the same place as the residents, which I certainly wouldn't want to do, right. We'd have to buy two <unclear> adjoining -

In contrast, Type 3 conversations are much more common (496 conversations grouped into this cluster, or 23% of all conversations in the LSWE Corpus). These conversations are extremely interactive and focused on the immediate context, as illustrated by Text Sample 3. The turns in this conversation are short and highly interactive (notice the dense use of *I* and *you*), and there is a dense use of common nouns together with WH-questions to express context-dependent information.

Text Sample 3. Conversation from Cluster 3: Interactive, Context-focused

A: I'm coming home at lunchtime. There's milk on the step. Bye-bye.
B: But ... lunchtime
A: Right. We'll have to get cracking.
B: what d'ya mean lunchtime?
A: Well lunchtime I'm going to go in and pop back to get things. I've locked it, it was unlocked. Right.
C: Can I go in the front?
A: Tie your belt up please. Tie your belt up. — Okay, speedily . — now
B: Oh crash, bang, wallop you're a
A: but both doors were open, you know. Start the car.
C: <screaming>Ah! You happy
A: See the co=
C: now?
A: can you er zip your zips up please? Keira. Can you zip your zip up?
B: I can't.
A: What do you think you'll be doing at school today?
B: Recorder concert!
A: Oh! Have you got your recorder? In school?
B: No! Er, yes, yes
A: Yeah.
B: yes.
A: Now, what you gonna be playing?
B: Joe Joe stubbed his toe. Joe Joe stubbed his toe and ... Indian Warrior.
A: Oh!
B: <singing>Big chief, Indian warrior, warrior, warrior. Big chief, Indian warrior. High ... ho! High ... ho! High ... ho! oh! oh!
A: Right.
B: And erm ... the skateboard ride.
A: <crunches gears> Ooh! That gear. Keeps changing with the
B: Mummy. You know what I've
A: Skateboard ride?
B: you know what, that I
A: What's that one?
B: ca= just can play that, I couldn't do recorders that well?
A: Yes.
B: Well now erm, I'm really good at it.
A: Can you do all the musical notes?

B: Yeah.
— text omitted —
B: Guess what Kirsty was doing when
A: What?
B: we was just practising for recorders?
A: What?
B: She was going like this, and the music was on, she put her feet out and she put the music on her feet.
A: Oh well.

Finally, Type 4 conversations are also relatively common (308 texts). This cluster is relatively unmarked on Dimensions 1 and 2, but these conversations are extremely narrative in their Dimension 3 characterization. Text Sample 4 illustrates these characteristics. Note that these conversations are not necessarily extended stories (although some of them are). Rather, as in the present case, these conversations can be constructed out of extended discussion of past events (with frequent past tense verb phrases, 3rd person pronouns, and communication verbs — especially *said* in this conversation), often coupled with commentary on their immediate relevance.

Text Sample 4. Conversation from Cluster 4: Narrative

A: I've just explained that to him. And he said he didn't know that, that he would get hold of Sen and ring me first thing, thing in the morning . — er, to tell me why Sen hasn't paid. He's got the invoice and everything. I said well you've sent us twenty thou= . — I said there is no VAT on it which it should be! Deary me! He says. Has he got the invoice? I said yes. And I said, we've been having, having the invoice outstanding since October at two and half thousand pound! I said, you actually owe me six thousand, one hundred and odd! And I said, you must realize I've a small company, and that's, in one respect that I've had to send those conditions because you're failing to meet the agreed thirty days payment!

B: Yeah.

A: And I said it's not on! I said we couldn't survive like that. And he said, well would you like to carry on with the contract? I said we're too far committed now to, I says to back out. I said, you know, we can't back out at this stage. And I said, but I said if there isn't the payments of the invoice when they are sent . — then . — you know, we've go= you've gotta look at it. So that invoice wants -

B: Doing. Yeah.

A: it wants doing and sending, and put in i= put twenty eight days on.

B: Yeah.

A: Had to be paid, it can't be paid by the twenty eighth it's er . — you know — well if I could've got hold of David or er, Andrew <name>, I was gonna give Andrew <name> a right bollocking for just pushing it in and he should've sent it to er, Michael <name>, Michael <name>'s just got it shoved in front of his nose a= in Edinburgh. He's just gone in to see if everything's alright at Edinburgh . — and of course, that's why he's had to report for that. Which was fair play to him,

but bloody Andrew should have told him! It's agreed, the system of stage payments, it's all written to him.

B: And you've just spoken to him have you?

A: I've just spoken to Michael <name>. Michael's great!

B: And you, you . — so he understands after he's sent you this?

A: What?

B: What's going on.

A: Yeah. Because yo= di= I said I had to send that agreement because you're failing to meet the standard agreement, you're not paying within the twenty eight days or the thirty days!

B: Mm.

A: I said I've got an invoice outstanding for October, and I said I can't afford to do that! He said, I realize that. Then he said, we would want you to do that work he said, because you've got a good reputation. — It makes, you know, if we — we're not gonna go bust just to get twelve months bloody work out of him on a service contract! You know, but i= if we couldn't, if they . — we= . — as they said, if they wanted that money back tomorrow we could only give them half that money back because of what we've got.

6. Conclusion

The three dimensions identified by this factor analysis of a conversational corpus are surprisingly similar to the dimensions of variation found in the earlier MD analysis of general spoken and written registers (Biber, 1988). Both analyses have a dimension that reflects the distinction between involved/interactive versus informational discourse; both analyses have a narrative dimension; and both have dimensions related to the expression of stance. The large-scale MD analyses of spoken and written registers in Somali and Korean similarly identified dimensions associated with these functions; composed of similar kinds of linguistic features.

Even more surprisingly, several MD analyses of restricted discourse domains have identified dimensions with similar formal and functional correlates (compare, for example, Reppen's (1994, 2001) analysis of elementary school registers with White's (1994) analysis of job interview registers). The fact that similar dimensions are found to be basic even in a corpus restricted to conversation suggests that these might be candidates for universal parameters of variation.

Comparing the present analysis to previous MD studies provides two complementary perspectives on the characteristics of conversation. In comparison to the full range of spoken and written registers, conversation is distinctive in being extremely interactive, involved, focused on the immediate context and personal stance, and constrained by real-time production circumstances. However, when conversation is considered on its own terms, we discover systematic patterns of variation among conversational texts (see also Carter and McCarthy, 1997; McCarthy, 1998; Quaglio, 2004; Quaglio and Biber, to appear). Interestingly, the present analysis indicates that the major parameters of variation internal to conversation are a mirror image to the dimensions of variation that distinguish among spoken and written registers.

References

Biber D. (1985). Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics*, vol. (23): 337-60.

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

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber D. (1989). A typology of English texts. *Linguistics*, vol. (27): 3-43.

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

Biber D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, vol. (15) : 133-163. (Reprinted in Conrad and Biber (Eds) (2001): 215-240.)

Biber D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, vol. (19): 219-241.

Biber D. (1994). An analytical framework for register studies. In Biber D. and Finegan E. (Eds), *Sociolinguistic perspectives on register*. Oxford University Press: 31-56.

Biber D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Biber D. (2001). Dimensions of variation among 18th century speech-based and written registers. In Conrad S. and Biber D. (Eds): 200-214.

Biber D. (2003), Variation among university spoken and written registers: A new multi-dimensional analysis. In Meyer C. and Leistyna P. (Eds), *Corpus analysis: Language structure and language use*. Rodopi.

Biber D. and Finegan E. (2001). Diachronic relations among speech-based and written registers in English. In Conrad S. and Biber D. (Eds): 66-83.

Biber D., Johansson S., Leech G., Conrad S. and Finegan E. (1999). *The Longman grammar of spoken and written English*. Longman.

Brown P. and Fraser C. (1979). Speech as a marker of situation. In Scherer K.R. and Giles H. (Eds), *Social markers in speech*. Cambridge University Press: 33-62.

Carter R. and McCarthy M. (1997). *Exploring Spoken English*. Cambridge University Press.

Connor-Linton J. (1989). *Crosstalk: A multi-feature analysis of Soviet-American spacebridges*. Ph.D. Dissertation: University of Southern California.

Conrad S. and Biber D. (Eds) (2001). *Variation in English: Multi-Dimensional Studies*. Longman.

Ervin-Tripp S. (1972). On sociolinguistic rules: Alternation and co-occurrence. In Gumperz J.J. and Hymes D. (Eds), *Directions in sociolinguistics*. Holt: 213-250

Halliday M.A.K. (1988). On the language of physical science. In Ghadessy M. (Ed.), *Registers of written English: Situational factors and linguistic features*. Pinter: 162-178

Hymes D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. University of Pennsylvania Press.

Jang S.-Ch. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study*. Ph.D. Dissertation. University of Hawaii.

Grabe W. (1987). Contrastive rhetoric and text-type research. In Connor U. and Kaplan R.B. (Eds.), *Writing across languages: Analysis of L2 text*. Addison-Wesley: 115-138.

Kanoksilapatham B. (2003). *A Corpus-based Investigation of Biochemistry Research Articles: Linking Move Analysis with Multidimensional Anal*ysis. Ph.D. Dissertation. Georgetown University.

McCarthy M. (1998). *Spoken Language and Applied Linguistics*. Cambridge University Press.

Quaglio P. (in preparation). *Conversation and TV Dialogue: A Corpus-based Study of NBC's Friends*. Ph.D. Dissertation. Northern Arizona University.

Quaglio P. and Biber D. (to appear). The grammar of conversation. In McMahon A. and Aarts B. (Eds), *The Handbook of English Linguistics*. Blackwell.

Reppen R. (1994). *Variation in elementary student writing*. Ph.D. Dissertation. Northern Arizona University.

Reppen R. (2001). Register variation in student and adult speech and writing. In Conrad S. and Biber D. (Eds): 187-199.

White M. (1994*). Language in job interviews: Differences relating to success and socioeconomic variables*. Ph.D. Dissertation. Northern Arizona University.

Appendix A.
List of grammatical, syntactic, lexico-grammatical, and semantic features identified by the Biber Tagger

1. Pronouns and pro-verbs

first person pronouns
second person pronouns
third person pronouns (excluding it)
pronoun it
demonstrative pronouns (this, that, these, those as pronouns)
indefinite pronouns (e.g., anybody, nothing, someone)
pro-verb do

2. Reduced forms and dispreferred structures

contractions
subordinator that deletion (e.g., I think [that/0] he went)
stranded prepositions (e.g., the candidate that I was thinking of)
split auxiliaries (e.g., they were apparently shown to …)

3. Prepositional phrases

4. Coordination

8b. Passives

agentless passives
by passives

8c. Modals

possibility modals (can, may, might, could)
necessity modals (ought, must, should)
predictive modals (will, would, shall)

8d. Semantic categories of verbs

be as main verb
activity verb (e.g., smile, bring, open)
communication verb (e.g., suggest, declare, tell)
mental verb (e.g., know, think, believe)
causative verb (e.g., let, assist, permit)
occurrence verb (e.g., increase, grow, become)
existence verb (e.g., possess, reveal, include)
aspectual verb (e.g., keep, begin, continue)

8e. Phrasal verbs

intransitive activity phrasal verb (e.g., come on, sit down)
transitive activity phrasal verb (e.g., carry out, set up)
transitive mental phrasal verb (e.g., find out, give up)
transitive communication phrasal verb (e.g., point out)
intransitive occurrence phrasal verb (e.g., come off, run out)
copular phrasal verb (e.g., turn out)
aspectual phrasal verb (e.g., go on)

9. Adjectives

attributive adjectives
predicative adjectives

9a. Semantic categories of adjectives

size attributive adjectives (e.g., big, high, long)
time attributive adjectives (e.g., new, young, old)
color attributive adjectives (e.g., white, red, dark)
evaluative attributive adjectives (e.g., important, best, simple)
relational attributive adjectives (e.g., general, total, various)
topical attributive adjectives (e.g., political, economic, physical)

10. Adverbs and adverbials

place adverbials
time adverbials

10a. Adverb classes

conjuncts (e.g., consequently, furthermore, however)
downtoners (e.g., barely, nearly, slightly)
hedges (e.g., at about, something like, almost)
amplifiers (e.g., absolutely, extremely, perfectly)
emphatics (e.g., a lot, for sure, really)
discourse particles (e.g., sentence initial well, now, anyway)
other adverbs

10b. Semantic categories of stance adverbs

non-factual/manner-of-speaking adverbs (e.g., frankly, mainly, truthfully)

attitudinal adverbs (e.g., surprisingly, hopefully, wisely)
factual adverbs (e.g., undoubtedly, obviously, certainly)
likelihood adverbs (e.g., evidently, predictably, roughly)

11. Adverbial subordination

causative adverbial subordinator (because)
conditional adverbial subordinator (if, unless)
other adverbial subordinator (e.g., since, while, whereas)

12. Nominal post-modifying clauses

that relatives (e.g., the dog that bit me, the dog that I saw)
WH relatives on object position (e.g., the man who Sally likes)
WH relatives on subject position (e.g., the man who likes popcorn)
WH relatives with fronted preposition (e.g., the manner in which he was told)
past participial postnominal (reduced relative) clauses (e.g., the solution produced by this process)

13. *That* complement clauses

13a. That clauses controlled by a verb (e.g., we predict that the water is here)

non-factual/communication verb (e.g., imply, report, say, suggest)
attitudinal verb (e.g., anticipate, expect, prefer)