Two Approaches to Genre Analysis

Three Genres in Modern American English

ZHONGHUA XIAO
ANTHONY McENERY
Lancaster University, United Kingdom

This article compares two approaches to genre analysis: Biber's multidimensional analysis (MDA) and Tribble's use of the keyword function of WordSmith. The comparison is undertaken via a case study of conversation, speech, and academic prose in modern American English. The terms *conversation* and *speech* as used in this article correspond to the demographically sampled and context-governed spoken data in the British National Corpus. Conversation represents the type of communication we experience every day whereas speech is produced in situations in which there are few producers and many receivers (e.g., classroom lectures, sermons, and political speeches). Academic prose is a typical formal-written genre that differs markedly from the two spoken genres. The results of the MDA and keyword approaches both on similar genres (conversation vs. speech) and different genres (the two spoken genres vs. academic prose) show that a keyword analysis can capture important genre features revealed by MDA.

Keywords: *multidimensional analysis; keyword analysis; genre; conversation; speech; academic prose*

This article compares two approaches to genre analysis: Biber's (1988) multidimensional analysis (MDA) and Tribble's (1999) use of the keyword function of WordSmith (Scott 1999). The comparison is undertaken via a case study of conversation, speech, and academic prose in modern American English. The terms *conversation* and *speech* as used in this article correspond to the demographically sampled and context-governed spoken data in the British National Corpus (BNC; see Aston and Burnard 1998, 31). Conversation represents the type of communication we experience every day (Biber 1988, 10) whereas speech is produced in situations in which there are few producers and many receivers (e.g., classroom lectures, sermons, and political speeches). Academic prose is a typical formal-written genre that differs markedly from the two spoken genres. In this article, we will compare

AUTHORS' NOTE: We thank the UK ESRC for supporting us to undertake this pilot study (grant reference RES-000-23-0553). We are equally grateful to Charles F. Meyer, Anne Curzan, and two anonymous reviewers from the *Journal of English Linguistics* for their insightful and constructive comments on an earlier draft of this article.

Journal of English Linguistics, Vol. 33 / No. 1, March 2005 62-82
DOI: 10.1177/0075424204273957

the results of the MDA and keyword approaches on both similar genres (conversation vs. speech) and different genres (the two spoken genres vs. academic prose).

While MDA was originally developed to compare written and spoken registers in English (Biber 1988), the approach has been used extensively in (1) synchronic analyses of specific registers and genres (Biber 1991; Biber and Finegan 1994b; Conrad 1994; Reppen 1994; Tribble 1999) and author styles (Biber and Finegan 1994a; Connor-Linton 1988; Watson 1994), (2) diachronic studies describing the evolution of registers (Biber and Finegan 1989, 1992; Atkinson 1992, 1993), and (3) register studies of non-Western languages (Besnier 1988; Biber and Hared 1992, 1994; Kim and Biber 1994) and contrastive analyses of different languages (Biber 1995). In addition, MDA has also been applied in addressing corpus design issues (e.g., Biber 1993) and the definitional issues of register/genres and text types (e.g., Biber 1989). More recently, Biber et al. (2002) have considered the implications of MDA for the development of teaching materials. Two edited volumes published recently (Conrad and Biber 2001; Reppen, Fitzmaurice, and Biber 2002) demonstrate the ongoing development of the MDA approach.

MDA is undoubtedly a powerful tool in genre analysis. But associated with this power is complexity. The approach is very demanding both computationally and statistically in that it requires expertise not only in extracting a large number of linguistic features from corpora but also in undertaking sophisticated statistical analysis. In this article, we will demonstrate that using the keyword function of WordSmith can achieve approximately the same effect as Biber's MDA. This approach is less demanding as WordSmith can generate wordlists and extract keywords automatically.

The primary corpus data used in this case study were taken from the Santa Barbara Corpus of Spoken American English (SBCSAE) and the Corpus of Professional Spoken American English (CPSA). Based on hundreds of recordings of naturally spoken English from all over the United States, SBCSAE represents a wide variety of people of different regional origins, ages, occupations, and ethnic and social backgrounds and reflects the many ways that people use language in their lives: conversation, gossip, arguments, on-the-job talk, card games, city council meetings, sales pitches, classroom lectures, political speeches, bedtime stories, sermons, weddings, and so forth (cf. Dubois et al. 2000-2004). CPSA is a two-million-word corpus that has been constructed using a selection of transcripts of interactions of various types occurring in professional settings recorded from 1994 to 1998. It has two components. The first component is made up of transcripts of press conferences from the White House, while the second component consists of transcripts of faculty meetings and committee meetings related to national tests (Barlow 1998). After classifying the 43 corpus files from SBCSAE into the conversation and speech genres on the basis of the topic and number of participants described in the documentation,[1] we found that there were only twelve files for the

TABLE 1
Corpus Data

Genre	Corpus	Sampling Date	Number of Texts	Tokens	Tokens by Genre
Conversation	SBCSAE	1988-93	31	135834	135834
Speech	SBCSAE	1988-93	12	46312	203810
	CPSA	1997	2	157498	
Academic prose	FROWNJ	1991-92	80	166169	166169
Total			125	505813	

NOTE: SBCSAE = Santa Barbara Corpus of Spoken American English; CPSA = Corpus of Professional Spoken American English.

speech genre in the SBCSAE data available to us, considerably less than the data for the conversation genre. Consequently, we decided to include two files from CPSA (comm797.txt and comr797.txt) to improve the balance of conversation and speech in our study.

As we also wish to contrast the two spoken genres with a typical written genre, the section of academic prose (eighty text samples in category J) from the FROWN corpus (hereafter referred to as FROWNJ) is also included in this case study. FROWN is an update of the Brown corpus containing data from the early 1990s (see Hundt, Sand, and Skandera 1999). Table 1 shows the data we have used.

To facilitate the extraction of linguistic features, we decided to annotate our corpus grammatically. As we did not have access to the tagging system used by Biber, we tagged our data with the Lancaster CLAWS system, applying the BNC C7 tagset (see Garside, Leech, and McEnery 1997). As the search patterns Biber developed for his tagger could not readily be used on CLAWS-tagged data, we developed a set of search algorithms that can easily be used in combination with the advanced search functions of WordSmith (e.g., *file-based search* and *context search*) to extract the required linguistic features from corpora tagged using CLAWS. These search algorithms allow the extraction of features such as *THAT deletion*, which are difficult to extract from unannotated texts (for further discussion, see McEnery, Xiao, and Tono 2005). While some of the patterns we devised may extract only typical instances and some of them may even generate false matches due to tagging errors, the same patterns are applied to all corpus files. Hence, we consider the results for the different genres to be comparable and sufficiently reliable, despite the small margin of error associated with our pattern-matching procedure, as it is assumed that the errors are distributed evenly across the files.

Having presented our methodology and data, in the remainder of this article we will compare the results achieved using the MDA and keyword approaches. But before the comparison is introduced, a brief review of the MDA approach is appropriate.

An MDA of the Three Genres

Biber (1988) presents a full analysis of twenty-one genres of spoken and written British English on the basis of sixty-seven linguistic features in 481 texts from the Lancaster/Oslo-Bergen (LOB) and London-Lund (LLC) corpora. This study established the multidimensional approach to genre analysis. Biber (1988, 63, 79) used factor analysis in concert with frequency counts of linguistic features to identify the sets of features that co-occur in texts with a high frequency. These are referred to as dimensions or factors. As these dimensions underlie linguistic features, they are conceptually clearer than the many features considered individually.

There are seven dimensions in Biber's MDA. They are informational versus involved production (dimension 1), narrative versus nonnarrative concerns (dimension 2), explicit versus situation-dependent reference (dimension 3), overt expression of persuasion (dimension 4), abstract versus nonabstract information (dimension 5), online informational elaboration (dimension 6), and academic hedging (dimension 7). Some factorial structures (namely, dimensions 1 and 3) include linguistic features with negative loadings. Positive and negative loadings along a dimension are written with a plus or minus symbol, as in factor +1 and factor –3. Biber observes that features with positive loadings co-occur frequently, whereas features with negative loadings occur together on a dimension.

The linguistic features Biber selected for his MDA are all functionally related. The features with positive loadings on dimension 1, for example, first- and second-person pronouns, THAT deletion, contraction, discourse markers, and private verbs such as *believe* and *think*, are all "associated in one way or another with an involved, non-informational focus" (Biber 1988, 105). Conversely, high frequencies of features with negative weights on dimension 1 (e.g., word length, type/token ratio, attributive adjectives and prepositions) are typically associated with a high informational focus and a careful integration of information in a text. The features with salient positive weights on dimension 2 (e.g., past-tense verbs, third-person pronouns, perfect aspect verbs, present participial clauses, and public verbs such as *agree*, *report*, and *say*) can all be used for narrative purposes (Biber 1988, 92), although narrative discourse depends heavily on the past tense and verbs marked for the perfect aspect (Biber 1988, 109). Alongside dimension 3, which is related to explicit versus situation-dependent reference, features with positive loadings include WH relative clauses, phrasal coordination, and nominalization. As Biber (1988, 110) observes, relativization specifies "the identity of referents within a text in an explicit and elaborated manner, so that the addressee will have no doubt as to the intended referent" while "the co-occurrence of phrasal coordination and nominalizations with these relativization features indicates that referentially explicit discourse also tends to be integrated and informational." The two features with negative weights on this dimension, time and place adverbials, on the other hand,

depend crucially on the addressee for text-internal references. The features associated with dimension 4 (e.g., prediction modals such as *will* and *shall*; necessity modals such as *ought*, *should*, and *must*; conditional subordination; and suasive verbs such as *ask*, *beg*, and *propose*) function together to mark persuasion, whether that be the overt marking of the addresser's own viewpoint or an assessment of the advisability or likelihood of an event presented to persuade the addressee (cf. Biber 1988, 111). The features associated with dimension 5 (i.e., abstract vs. nonabstract) are conjuncts, main/subordinate passive constructions, and adverbial subordinators. Discourse with a high frequency of passives is typically abstract and technical in content, as well as formal in style. This type of discourse is generally characterized by complex logical relations, which are achieved by conjuncts and adverbial subordinators (cf. Biber 1988, 112). Features with salient positive weights on dimension 6 (e.g., demonstratives such as *this* and *that*, THAT relative clauses, and THAT clauses as verb and adjective complements) function to mark informational elaboration in discourse that is informational but produced under real-time conditions (Biber 1988, 113-14). Dimension 7 has only one salient positive feature, SEEM/APPEAR. SEEM and APPEAR mark perception rather than assertion of fact and thus mark an assertion as uncertain. They are typically used in academic discourse as a downtoner to qualify the extent to which an assertion is known (Biber 1988, 114). As the factorial structure of this last dimension was not strong enough for a firm interpretation, it was not discussed in detail in Biber (1988). Accordingly, this dimension will also be omitted in our analysis.

A full list of the linguistic features, together with illustrative corpus examples, is given in the appendix at the end of this article. Note that in this case study, WH relative clauses in object and subject positions are kept together for ease of presentation. For the same reason, THAT relative clauses refer to those in either object or subject positions. Combinations of similar features to these reduce the number of linguistic features under consideration to fifty-seven.

The dimension score of a text is the sum of the scores for all linguistic features on a dimension.[2] The dimension score of a genre is the mean of the factor scores of the texts within the genre. After the frequency of each of the linguistic features was extracted using our search algorithms, we followed Biber's procedures to compute the dimension scores of the three genres under consideration, which are shown in Table 2. The table shows that the three genres differ significantly alongside dimensions 1, 3, and 5.

Plotting the dimension scores of the three genres allows a clearer view of the differences, as shown in Figure 1. It can be seen from the figure that the most marked contrast between these genres lies in dimension 1 (informational vs. involved), though conversation/speech and academic prose also show noticeable differences in dimensions 3 (explicit vs. situation-dependent reference) and 5 (abstract vs. nonabstract). The relative "oralness" or "literateness" of a genre depends on the ex-

TABLE 2
Factor Scores of Three Genres

Dimension	Conversation	Speech	Academic Prose	*F* Score (two *df*)	Significance Level (*p*)
1. Involved	+28.96	+14.63	–9.24	51.358	<.001
2. Narrative	+1.20	+2.00	–0.61	0.976	.399
3. Reference	–6.63	–4.47	+2.45	17.155	<.001
4. Persuasion	+0.74	+1.33	+0.32	0.277	.762
5. Abstract	–1.76	–2.77	+2.90	82.782	<.001
6. Online	–1.19	+1.23	+0.84	3.468	.072

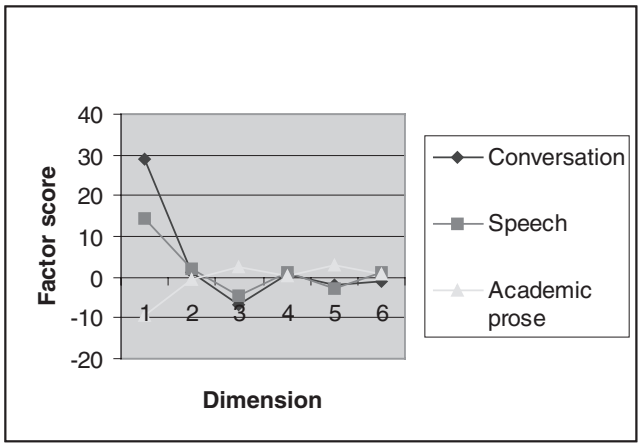

Figure 1: Multidimensional Analysis of the Three Genres.

tent to which texts within the genre are interactive or involved (Biber 1988, 108). Therefore, conversation, which has an involved, interactive purpose and is subject to real-time production constraints, is the most "oral" of the three genres in this study. In contrast, academic prose is the most "literate" of the three in that it is characterized by careful editing and informational density. Alongside dimension 3, conversation and speech are quite similar in that both genres make extensive reference to the physical and temporal situation of discourse, though to varying degrees (the difference is not statistically significant). In contrast, academic prose tends to make explicit text-internal reference. This feature is explicable in terms of the greater number of opportunities that writing affords one to undertake careful editing. Such opportunities are atypical of spontaneous speech. In terms of dimension 5, academic prose is much more technical and abstract than speech and conversation. Figure 1 also shows that speech has the highest score for dimension 6—the difference is marginally significant—which marks the degree of online informa-

tional elaboration under strict, real-time conditions. As conversation does not have an informational focus, it does not need stylistic elaboration. While academic prose is an informationally dense genre that needs elaboration, it is not subject to strict, real-time conditions. Rather, it is produced under circumstances that allow precise lexical choice and careful structural elaboration. Only the genre of speech, which is informationally dense but produced under real-time conditions, needs more online informational elaboration. This analysis is supported by the results of statistical tests in Table 3.

This section introduced the MDA approach to genre analysis and presented the result of an MDA of the three genres. While the process of extracting linguistic features and computing factor scores was not shown, it was very time-consuming and computationally/statistically demanding. In the section that follows, we will use WordSmith to analyze these three genres on the basis of the same data and compare the results of this keyword analysis with those of MDA.

A Keyword Analysis of the Three Genres

WordSmith is an advanced corpus exploration package that includes concordance as well as wordlist, keyword, and collocation functions. Tribble (1999) claims that the keyword function of WordSmith can be used to achieve an effect similar to that of Biber's MDA. Specifically, a keyword analysis not only indicates the "aboutness" (Scott 1999) of a particular genre but also reveals the salient features that are functionally related to that genre. As WordSmith can generate a wordlist and extract keywords automatically, the keyword approach to genre analysis does not require users to extract complicated structures from a corpus or undertake a sophisticated statistical analysis. If the approach can be shown to produce results comparable to the MDA approach, it could provide a "low effort" alternative to MDA. This section presents the results of the keyword analysis of the three genres, which are compared with the MDA results to explore the comparability of the results produced by the two techniques.

The first step in applying the keyword approach is to make a wordlist of the corpus files by genre, which is then compared with a reference corpus to extract those words whose frequency is unusually high (positive keywords) or low (negative keywords) in relation to the reference corpus. When this study was carried out, there were not many balanced corpora of American English available to us. As part of the FROWN corpus was already used as the target of our study, we could not use it as our reference corpus. We could have used the Brown corpus, but that corpus sampled texts in the early 1960s, well before the data used in this study. Consequently, we decided to use a corpus of British English as a reference. It is far from ideal to use a British English corpus to provide a reference wordlist for American English data. Using a British English corpus may give prominence to American-

Table 3
Paired Comparisons of Three Genres

Comparison	Dimension	*F* Score (one *df*)	Significance Level (*p*)
Conversation versus speech	1. Informational versus involved	21.516	<.001
	2. Narrative versus nonnarrative	0.131	.725
	3. Explicitness of reference	2.914	.114
	4. Overt expression of persuasion	0.148	.709
	5. Abstract versus nonabstract	3.277	.100
	6. Online informational elaboration	5.091	.065
Conversation versus academic prose	1 Informational versus involved	89.295	<.001
	2. Narrative versus nonnarrative	1.256	.289
	3. Explicitness of reference	38.504	<.001
	4. Overt expression of persuasion	0.104	.753
	5. Abstract versus nonabstract	86.064	<.001
	6. Online informational elaboration	4.569	.076
Speech versus academic prose	1. Informational versus involved	34.745	<.001
	2. Narrative versus nonnarrative	1.954	.192
	3. Explicitness of reference	34.788	<.001
	4 .Overt expression of persuasion	0.663	.434
	5. Abstract versus nonabstract	325.855	<.001
	6. Online informational elaboration	0.175	.690

isms in our data. However, given that the three genres are compared against the same reference corpus, using British English as a reference will not affect our observations notably if the assumption holds that the genres studied here use Americanisms with roughly similar frequency.[3]

One further issue related to the reference corpus is that it is clearly much larger than the corpora that are compared against it. Tribble (1999, 171) claims that the size of the corpus from which the reference wordlist is created is relatively unimportant. Before we undertake a keyword analysis of the three genres, we will first carry out a baseline test to verify this claim by comparing the keyword lists of the two spoken genres, which were created using a reference wordlist from the one-million-word Freiburg-LOB (FLOB) corpus (an update of LOB in the early 1990s; see Hundt, Sand, and Siemund 1998) and a reference wordlist from the 100-million-word BNC corpus. It is important to note that keywords were extracted from our American data, while FLOB or the BNC acted only as a reference corpus. Tables 4 and 5 show the top ten positive and negative keywords from the genres of conversation and speech.

As can be seen in Table 4, nine out of the top ten positive keywords extracted from the American conversation corpus appear in both the FLOB- and BNC-based keyword lists. Only one item from the FLOB-/BNC-based list does not appear among the top ten in the BNC-/FLOB-based list. The contracted negation (*n't*) from the FLOB-based list ranks thirteenth in the BNC-based list, while *hm* from the

TABLE 4
Top Ten Positive and Negative Keywords from Conversation

	Number	FLOB as Reference Corpus	Number	BNC as Reference Corpus
Positive keywords				
	1	I	1	uh
	2	you	2	um
	3	yeah	3	I
	4	know	4	you
	5	uh	5	know
	6	oh	6	yeah
	7	mhm	7	mhm
	8	um	8	okay
	9	okay	9	oh
	10	n't	10	hm
Negative keywords				
	1	the	1	yes
	2	of	2	mm
	3	in	3	the
	4	as	4	as
	5	by	5	of
	6	his	6	've
	7	which	7	in
	8	its	8	quite
	9	for	9	terms
	10	their	10	very

NOTE: FLOB = Freiburg Lancaster/Oslo-Bergen corpus; BNC = British National Corpus.

BNC-based list ranks twenty-first in the FLOB-based list. The top ten negative keyword lists also show similarities, though not as marked as the positive keywords. A similar pattern is found for the American speech corpus. As can be seen in Table 5, eight positive keywords are the same whether the one-million-word FLOB or the 100-million-word BNC is used as a reference corpus. The two items of the top ten positive keywords from the FLOB-based list, *you* and *do*, appear as eleventh and thirteenth in the BNC-based list. The two items of the top ten positive keywords from the BNC-based list, *uh* and *NAEP* (the National Assessment for Education Progress), appear as eleventh and twentieth in the FLOB-based list. The top ten negative keywords from the two lists are exactly the same, though they appear in a slightly different order. The top ten positive and negative keywords created for academic prose (not shown in the tables) using a reference wordlist from FLOB and the BNC are also very similar. The above test provides evidence to show that the size of a reference corpus is not very important in making a keyword list. With Tribble's (1999) claim supported, we are now ready to compare the keyword lists of the three genres. We will examine positive keywords that were extracted from the American data using the BNC as a reference corpus.

TABLE 5
Top Ten Positive and Negative Keywords from Speech

	Number	FLOB as Reference Corpus	Number	BNC as Reference Corpus
Positive keywords				
	1	we	1	we
	2	I	2	that
	3	that	3	uh
	4	you	4	I
	5	think	5	test
	6	're	6	think
	7	okay	7	NAEP
	8	what	8	okay
	9	test	9	're
	10	do	10	what
Negative keywords				
	1	his	1	the
	2	the	2	his
	3	he	3	her
	4	her	4	by
	5	of	5	he
	6	by	6	of
	7	she	7	she
	8	had	8	had
	9	was	9	its
	10	its	10	was

NOTE: FLOB = Freiburg Lancaster/Oslo-Bergen corpus; BNC = British National Corpus.

In genre analysis, a key keyword list may prove more useful than a keyword list, because it excludes keywords that occur frequently in only a few texts of a genre. For example, with reference to the BNC, the keywords *test* and *NAEP* occur frequently in only two texts in our American speech corpus, namely, comm797.txt and comr797.txt, which were taken from the CPSA corpus. These files contain frequent uses of the two keywords simply because they are transcripts of a national meeting on reading tests and a national meeting on mathematics tests. As WordSmith can create a key keyword database automatically, key keywords are as simple to extract as keywords.

Table 6 lists the top ten key keywords from the three genres. It also shows as a percentage the frequency of the keyword in terms of how often it occurs and how widespread it is in the genre. It is clear that key keywords occur frequently in a wide range of texts in each genre. While there are both similarities and differences in the top ten key keywords for conversation and speech, the top ten key keywords for academic prose are totally different from the two spoken genres.

Let us first consider key keywords in the two spoken genres. Table 7 compares the top ten key keywords from conversation and speech. Note that in this table, as in

TABLE 6
Top Ten British National Corpus–Based Key Keywords from the Three Genres

Conversation				Speech				Academic Prose			
Number	Word	Frequency %	Cover %	Number	Word	Frequency %	Cover %	Number	Word	Frequency %	Cover %
1	I	4.00	100	1	uh	0.26	85.71	1	of	4.10	17.50
2	you	3.00	96.77	2	that	3.35	85.71	2	the	6.70	13.75
3	yeah	1.10	96.77	3	um	0.11	78.57	3	is	1.35	12.50
4	n't	1.29	93.55	4	I	2.67	64.29	4	formula	0.27	12.50
5	um	0.46	93.55	5	you	1.65	64.29	5	system	0.12	10.00
6	uh	0.67	90.32	6	n't	0.78	50.00	6	American	0.07	10.00
7	know	1.26	87.10	7	we	1.82	50.00	7	B	0.10	10.00
8	it	2.63	87.10	8	so	0.71	50.00	8	G	0.09	10.00
9	do	1.07	87.10	9	okay	0.30	50.00	9	C	0.09	8.75
10	oh	0.74	74.19	10	know	0.35	50.00	10	program	0.04	8.75

TABLE 7
Comparison of Top Ten Key Keywords from Conversation and Speech

		Conversation		Speech	
Factor	Linguistic Feature	Keyword	Frequency %	Keyword	Frequency %
+1	Private verb	know	1.26	know	0.35
	Second-person pronoun	you	3.00	you	1.65
	DO as proverb Present-tense verb Questions	do	1.07	—	—
	Emphatic	—	—	so (9%*0.71%)	0.06
	Contraction Analytic negation	n't	1.29	n't	0.78
	First-person pronoun	I	4.00	I, we	4.49
	Pronoun IT	it	2.63	—	—
	Interjections	yeah, um, uh, oh	2.93	uh, um, okay	0.67
	Demonstrative pronoun	—	—	that (35.8%*3.35%)	1.20
Total of factor +1[a]			16.18		9.20
–3	Other adverbs	—	—	that (0.5%*3.35%), so	0.02
				(77%*0.71%)	0.55
Total of factor –3			—		0.57
+5	Conjuncts	—	—	that (0.2%*3.35%)	0.01
	Other adv sub	—	—	so (14%*0.71%)	0.10
Total of factor +5			—		0.11
+6	Demonstrative	—	—	that (10.1%*3.35%)	0.34
	THAT clause	—	—	that (53.4%*3.35%)	1.79
Total of factor +6			—		2.13

a. $F = 5.812$ (1 *df*); $p = .030$.

other similar tables in this section, we are talking about top ten keywords. As such, we cannot assume, for example, that there are no instances of a present-tense verb or of questions in speech simply because DO does not appear on the top ten key keyword list of the speech genre. As noted in the previous section, constructions such as private verbs, second-person pronouns, and the pronoun IT all carry an interactive and affective focus. While the two genres share eight key keywords, two key keywords, DO and IT, appear among the top ten of the conversation list but not among the top ten of the speech list. Concordances of DO from the thirty-one texts of the conversation genre show that DO appears in the following structures: (1) analytic negation (i.e., *do not*, *don't*), (2) special and general questions, and 3) proverb *do* in the present tense. As all of these are factor +1 features that have an interactive focus, we will not make a distinction between them; rather, the gross percentage will be used for all of these features. Similarly, *n't* is both a contraction and an analytic negation, so we will not draw a distinction between the two.

Another feature of note in Table 7 is interjections. Interjections are of note for two reasons. First, they are more common in conversation than speech. Second, in-

terjections were not considered as discourse markers by Biber (1988), though they are actually used in the same way as discourse particles to maintain conversational coherence (Schiffrin 1982) and are typical of spoken language.[4] Hence, while interjections are not included as a relevant linguistic feature in MDA, they are an important feature in a keyword analysis. Two key keywords that are found among the top ten of the speech list but not among the top ten of the conversation list are THAT and SO. As CLAWS makes a distinction between the different uses of these words, it is easy to determine their proportions. Concordances of THAT from the fourteen texts of the speech genre show that THAT is used in following contexts:

(1) THAT-clause [53.4%],[5]
(2) demonstrative pronoun [*this*, *that*, *these*, and *those* not followed by a noun; 35.8%],
(3) demonstrative [*this*, *that*, *these*, and *those* followed by a noun; 10.1%],
(4) emphatic [0.5%], and
(5) other adverbial subordination [0.2%].

The keyword SO is used in the following contexts: (1) other adverbs, 77%; (2) other adverbial subordination, 14%; and (3) emphatics, 9%. The overall percentages of the two keywords are allocated to appropriate features accordingly. It can be seen in Table 7 that the total of factor +1 for the conversation genre (16.18%) is significantly greater than the total for the speech genre (9.20%). Conversely, in relation to factors +3 and +5, the total of factor +6 for speech is much greater than that for conversation, suggesting a possibly significant difference between the two spoken genres along dimension 6, which indicates the level of online elaboration. These observations of the two spoken genres are in line with the MDA results in the earlier section (see Table 3).

Table 8 shows the top ten key keywords in academic prose. The first two key keywords are *of* and *the*. *Of* as a preposition adds a negative weight to the dimension of informational versus involved production. Tribble (1999, 175-77) observes that *of* and *the* are typically associated with nouns. In academic prose, for example, *of* is typically used as a postmodifier in the N1 + *of* + N2 structure (e.g., *center of mass*, *clusters of galaxies*). The definite article *the* is also associated with nouns. In MDA, nouns of the nominalization type are a feature with a positive loading for dimension 3 (explicit vs. situation-dependent reference), while nouns of other types are a feature with a negative loading for dimension 1 (informational vs. involved focus). The present tense verb *is* adds a positive weight to dimension 1; so does its use as a main verb. But as can be seen from Table 8, the positive weights of *is* are well offset by other features with a negative loading for dimension 1. The table also shows that academic prose has a high score for factor +3, that is, this written genre typically makes explicit text-internal reference. The low dimension 1 score and high dimen-

TABLE 8
Top Ten Key Keywords in Academic Prose

Factor	Linguistic Feature	Keyword	Frequency %
+1	Present-tense verb	is	1.35
	BE as a main verb	is	0.89
–1	Preposition	of	4.10
	Other nouns	the	1.05
Total of factor 1		–2.91	
+3	Nominalizations	the	5.65
Others	formula, system, American, B, G, C, program		

sion 3 scores are just what we found for academic prose using the MDA approach in the earlier section. Apart from *of*, *the*, and *is*, the four content words (*formula*, *system*, *American*, and *program*) indicate the "aboutness" of academic prose, while the three letters (*B*, *G*, *C*) are used mainly as part of a list or variable labels, typical of academic prose. *Is* as a main verb is typically used in academic prose to make a statement or claim (e.g., *it is an excellent example of . . .*, *it is the user's responsibility to . . .*). These key keywords enable us to get a general view of the content and style of academic prose.

Let us now consider the top ten negative keywords from the two spoken genres. Note that because negative keywords are omitted automatically from a key keyword list, we will compare negative keywords from the keyword lists. As negative keywords are relatively infrequent words in relation to a reference corpus, we cannot take the same approach as when studying positive keywords. We need to refer back to the reference corpus to find an explanation for the relatively low frequency of negative keywords in our American data. There is little advantage in using a relatively large reference corpus (Tribble 1999, 171). Furthermore, as WordSmith allows a maximum of only 16,368 concordances at a time,[6] it would be very inconvenient to use the BNC as the reference corpus to study negative keywords. As such, we will use the FLOB-based keyword lists to study negative keywords in conversation and speech. We will also include negative keywords from academic prose for a contrast.

Table 9 lists the top ten negative keywords from the three genres. As can be seen, negative keywords are as revealing as positive keywords. The four linguistic features with positive weights on dimension 1, which are associated with interactive and affective discourse, are found among only the top ten negative keywords of academic prose. Conversely, the two features with negative weights on dimension 1 are found among only the top ten negative keywords of conversation and speech. It is also interesting to note the contrast between the two spoken genres: six out of ten negative keywords from conversation are associated with factor –1 features, while only three negative keywords from speech are associated with factor –1 features.

TABLE 9
Top Ten Freiburg Lancaster/Oslo-Bergen Corpus–Based Negative Keywords of Three Genres

Factor	Linguistic Feature	Conversation	Speech	Academic Prose
+1	Second-person pronouns	—	—	you
	First-person pronouns	—	—	I
	Contraction	—	—	n't
	Analytic negation	—	—	n't
–1	Nouns	the, of	the, of	—
	Prepositions	in, as, by, for	by	—
+2	Third-person pronouns	his, its, their	his, he, her, she, its	he, she, her, his
	Past-tense verbs	—	had, was	had, was, said
	Public verbs	—	—	said
+3	WH relative clauses	which	—	—
	Pied piping	which	—	—

Similarly, seven negative keywords and three linguistic features from academic prose are associated with dimension 2, which suggests that academic prose has a less narrative focus than the two spoken genres. While the difference between the three genres alongside dimension 2 is not statistically significant, academic prose has the lowest score for this dimension (see Table 2). It is more difficult to interpret *which*, which is found only in the top ten negative keywords from the conversation genre. Concordances of *which* from the reference corpus show that *which* is primarily used in WH relative clauses and pied piping constructions (67.7%),[7] which are salient features associated with dimension 3. The unusually low frequency of these features indicates that conversation relies heavily on context-dependent reference. The relatively low frequency of WH relative clauses in conversation is conformant with its lowest score for dimension 3 (see Figure 1).

The above analysis demonstrates that both positive and negative keywords can be good indicators of genre features. While it would seem that a keyword analysis can reflect only some MDA dimensions, the results obtained by both approaches are consistent across the three different genres under consideration; keywords can be used to achieve an approximation to an MDA analysis.

Conclusion

In this article, we compared Biber's MDA approach and Tribble's keyword approach to genre analysis via a case study of conversation, speech, and academic prose in modern American English. The results obtained by the two approaches are similar. The most significant difference between conversation and speech lies in dimension 1, a measure of the informational versus involved distinction. These genres also differ marginally alongside dimension 6, which indicates the level of online elaboration. This means that conversation is considerably more interactive and af-

fective than speech. While speech is informationally dense, it is subject to real-time production conditions, and thus speech needs online informational elaboration. The two spoken genres differ significantly from academic prose along dimensions 1, 3, and 5. This means that on one hand, academic prose is the most "literate," technical, and abstract of the three genres under consideration; on the other hand, it tends to make explicit in-text reference, whereas the two spoken genres make context-dependent references.

Methodologically, the MDA approach, while providing a powerful and comprehensive tool for genre analysis, requires considerable expertise in data extraction and statistical analysis. The keyword approach, in contrast, provides a less demanding approach to genre analysis. But since this approach provides a less comprehensive contrast of genres and may not work for more fine-grained types of genre analysis, it is not simply a substitute for MDA. Nevertheless, as the keyword approach requires little expertise to undertake and can be undertaken swiftly, it provides a quick and simple means of evaluating a genre against Biber's dimensions. The keyword approach to genre analysis provides linguists with a powerful and easily used tool.

APPENDIX
Linguistic Features in Multidimensional Analysis

Dimension 1: Informational versus Involved Production (twenty-three features with positive loadings and five with negative loadings)

Features with Positive Loadings

(1) Private verbs: all morphological forms of the following verbs: *anticipate*, *assume*, *believe*, *conclude*, *decide*, *demonstrate*, *determine*, *discover*, *doubt*, *estimate*, *fear*, *feel*, *find*, *forget*, *guess*, *hear*, *hope*, *imagine*, *imply*, *indicate*, *infer*, *know*, *learn*, *mean*, *notice*, *prove*, *realize*, *recognize*, *remember*, *reveal*, *see*, *show*, *suppose*, *think*, *understand*.
(2) THAT deletion: for example, *I think [that] it's so funny.*
(3) Contraction: *n't*, *'ll*, *'d*, *'m*, *'re*, *'ve*, *'s* [excluding possessive form].
(4) Present-tense verbs: all base forms and third-person singular present verb forms.
(5) Second-person pronouns: *you*, *your*, *yourself*, *yourselves*, *yours*.
(6) DO as a proverb, for example, *You did that?*
(7) Analytic negation: *not*, *n't*.
(8) Demonstrative pronouns: *this*, *that*, *these*, and *those* [not followed by a noun].
(9) General emphatics: *for sure*, *a lot*, *such a*, *real*, *so*, *just*, *really*, *most*, *more*, and DO + verb.
(10) First-person pronouns: *I*, *my*, *our*, *myself*, *ourselves*, *mine*, and *ours*.
(11) pronoun IT: *it*.

(continued)

APPENDIX (continued)

(12) BE as a main verb [excluding BE as an auxiliary], for example, *You are right*.
(13) Causative subordination: *because*.
(14) Discourse markers: *well*, *anyway*, *anyways*, *anyhow.*
(15) Indefinite pronouns, for example, *none*, *one*, *anyone*, *someone*, *somebody*, *anybody*, *nobody*, *everything*, *nothing*.
(16) General hedges: *about* [not as a preposition], *something like*, *more or less*, *almost*, *maybe*, *sort of*, and *kind of* [excluding *sort* and *kind* as true nouns].
(17) Amplifiers: *absolutely*, *altogether*, *completely*, *enormously*, *entirely*, *extremely*, *fully*, *greatly*, *highly*, *intensely*, *perfectly*, *strongly*, *thoroughly*, *totally*, *utterly*, and *very.*
(18) Sentence relatives: for example, *The present book, which is the first . . .*
(19) WH questions: for example, *What is it?*
(20) Possibility modals: *can*, *could*, *may*, and *might* [including contracted forms].
(21) Nonphrasal coordination: for example, *Yeah, and it has*.
(22) WH clauses: for example, *You know what I mean*.
(23) Final prepositions: for example, *Where did you get it from?*

Features with Negative Loadings

(24) Other nouns: all noun forms excluding nominalizations [see 38 below].
(25) Word length: [WordSmith wordlist function: average word length].
(26) prepositions: All prepositions like *at*, *by*, *in* and *of*;
(27) Type/token ratio: [WordSmith wordlist function: standardized type/token ratio].
(28) Attributive adjectives: for example, *young girl* and *new regulatory requirements*.

Dimension 2: Narrative versus Nonnarrative Concerns (six linguistic features, all with positive loadings)

(29) Past-tense verbs: all past-tense verbs.
(30) Third-person pronouns: *she*, *he*, *they*, *her*, *him*, *them*, *his*, *its*, *hers*, *their*, *theirs*, *himself*, *herself*, and *themselves* [including contractions].
(31) Perfect-aspect verbs: for example, *That hasn't finished*.
(32) Public verbs: all morphological forms of the following verbs: *acknowledge*, *admit*, *agree*, *assert*, *claim*, *complain*, *declare*, *deny*, *explain*, *hint*, *insist*, *mention*, *proclaim*, *promise*, *protest*, *remark*, *reply*, *report*, *say*, *suggest*, *swear*, and *write*.
(33) Synthetic negation: *neither*, *nor*, and *no* [excluding *no* as a response].
(34) Present participial clauses: for example, *So, you got this Oscar there, swimming there in the tank*.

Dimension 3: Explicit versus Situation-Dependent Reference (four features with positive loadings and three with negative loadings)

Features with Positive Loadings

(35) WH relative clauses: for example, *You know the little folks who live above me*.
(36) Pied piping constructions: for example, *the problems with which he is concerned*.

APPENDIX (continued)

(37) Phrasal coordination: for example, *economic and social conditions*, *racism and sexism*, *pick and choose*.
(38) Nominalizations: all nouns ending in *-tion*, *-ment*, *-ness*, *-ity* [including plural forms].

Features with Negative Loadings

(39) Time adverbials: all adverbs of time.
(40) Place adverbials: all adverbs of place.
(41) Other adverbs: all adverbs minus all totals of hedges, amplifiers, downtoners, place adverbials, and time adverbials.

Dimension 4: Overt Expression of Persuasion (six linguistic features, all with positive loadings)

(42) Infinitives: *to* + base form of a verb (may be separated by one or two adverbs).
(43) Prediction modals: *will*, *shall*, and *would* (including contractions).
(44) Suasive verbs, including all morphological forms of the following verbs: *agree*, *arrange*, *ask*, *beg*, *command*, *decide*, *demand*, *grant*, *insist*, *instruct*, *ordain*, *pledge*, *pronounce*, *propose*, *recommend*, *request*, *stipulate*, *suggest*, and *urge*.
(45) Conditional subordination: *if* and *unless*.
(46) Necessity modals: *ought*, *should*, and *must*.
(47) Split auxiliaries: for example, *You're just saying that*.

Dimension 5: Abstract versus Nonabstract Information (six linguistic features, all with positive loadings)

(48) Conjuncts: *alternatively*, *altogether*, *consequently*, *conversely*, *eg*, *e.g.*, *else*, *furthermore*, *hence*, *however*, *i.e.*, *instead*, *likewise*, *moreover*, *namely*, *nevertheless*, *nonetheless*, *notwithstanding*, *otherwise*, *rather*, *similarly*, *that is*, *therefore*, *thus*, *viz*, *in (comparison, contrast, particular, addition, conclusion, consequence, sum, summary, any event, any case, other words)*, *for example (instance)*, *by contrast (comparison)*, *as a result (consequence)*, *on the contrary (other hand)*.
(49) Agentless passives: for example, *And this book was written in nineteen ten*.
(50) Past participial clauses: for example, *This problem, combined with administrative failure to meet . . .*
(51) BY-passives: for example, *It is shared by preacher and audience*.
(52) Past participial WHIZ deletions: for example, *tests designed for old age groups*.
(53) Other adverbial subordinators: *since*, *while*, *whilst*, *whereupon*, *whereas*, *whereby*, *such that*, *so that*, *inasmuch as*, *forasmuch as*, *insofar as*, *insomuch as*, *as long as*, and *as soon as*.

(continued)

APPENDIX (continued)

Dimension 6: Online Informational Elaboration
(four linguistic features, all with positive loadings)

(54) THAT clauses as verb complements: for example, *So he knew that the oil was leaking?*
(55) Demonstratives: *this*, *that*, *these*, and *those* followed by a noun.
(56) THAT relative clauses: for example, *In fact, I eat stuff that he doesn't eat*.
(57) THAT clauses as adjective complements: for example, *I'm just happy that I beat you*.

Notes

1. The Linguistic Data Consortium has published forty-six corpus files of the Santa Barbara Corpus of Spoken American English (SBCSAE) in three parts. As part 3 was not published when this study was undertaken, three SBCSAE files contained in that release are missing from this study. We studied only the forty-three files available for download at the TalkBank site (http://talkbank.org/data/Conversation/) in December 2002.

2. The dimension score of a text is computed by adding together the factor score of each feature with a positive loading and then subtracting the factor score of each feature, if any, with a negative loading. For example, suppose for the genre of academic prose the mean factor scores of the four features with positive weights on dimension 3 are –0.57, +0.53, +0.51, and +0.60, while those for features with negative weights are –0.44, –0.43, and –0.51. The dimension score of dimension 3 for academic prose would therefore be +2.45:

$$-0.57 + 0.53 + 0.51 + 0.60 - (-0.44) - (-0.43) - (-0.51) = 2.45.$$

3. See Mike Scott's comments about using reference corpora to create wordlists in the Corpora Archive dated 13 June 2003 (http://nora.hd.uib.no/corpora/2003-1/0545.html).

4. Aijmer (1987, 61-86) gives an interesting description of the functions of *oh* and *ah* in the London-Lund corpus.

5. We will not differentiate between the three types of THAT clauses because they are all features with positive weights on dimension 6 (online elaboration).

6. This problem is solved in WordSmith version 4.

7. The remaining 32.3% are mainly sentence relatives and WH clauses.

References

Aijmer, Karin. 1987. *Oh* and *Ah* in English Conversation. In *Corpus Linguistics and Beyond*, edited by Willem Meijs, 61-86. Amsterdam: Rodopi.

Aston, Guy, and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh, UK: Edinburgh University Press.

Atkinson, Dwight. 1992. The Evolution of Medical Research Writing from 1735 to 1985: The Case of the Edinburgh Medical Journal. *Applied Linguistics* 13:337-74.

———. 1993. A Historical Discourse Analysis of Scientific Research Writing from 1675 to 1975: The Case of the Philosophical Transactions of the Royal Society of London. Ph.D. diss., University of Southern California.

Barlow, Michael. 1998. *A Corpus of Spoken Professional American English*. Houston, TX: Athelstan.

Besnier, Niko. 1988. The Linguistic Relationships of Spoken and Written Nukulaelae Registers. *Language* 64:707-36.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge, UK: Cambridge University Press.

———. 1989. A Typology of English Texts. *Linguistics* 27:3-43.

———. 1991. Oral and Literate Characteristics of Selected Primary School Reading Materials. *Text* 11:73-96.

———. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8 (4): 243-57.

———. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge, UK: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Maria Helt. 2002. Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly* 36 (1): 9-48.

Biber, Douglas, and Edward Finegan. 1989. Drift and Evolution of English Style: A History of Three Genres. *Language* 65:487-517.

———. 1992. The Linguistic Evolution of Five Written and Speech-based English Genres from the 17th to the 20th Centuries. In *History of Englishes: New Methods and Interpretations in Historical Linguistics*, edited by Matti Rissanen, Ossi Ihalainen, and Terttu Nevalainen, 688-704. Berlin: Mouton.

———. 1994a. Multi-dimensional Analyses of Authors' Style: Some Case Studies from the Eighteen Century. In *Research in Humanities Computing 3*, edited by Don Ross and Dan Brink, 3-17. Oxford, UK: Oxford University Press.

———, eds. 1994b. *Sociolinguistic Perspectives on Register.* New York: Oxford University Press.

Biber, Douglas, and Mohamed Hared. 1992. Dimensions of Register Variation in Somali. *Language Variation and Change* 4:41-75.

———. 1994. Linguistic Correlates of the Transition to Literary in Somali: Language Adaptation in Six Press Registers. In *Sociolinguistic Perspectives on Register*, edited by Douglas Biber and Edward Finegan, 182-216. New York: Oxford University Press.

Connor-Linton, Jeff. 1988. Author's Style and World-view in Nuclear Discourse: A Quantitative Analysis. *Multilingual* 7:95-132.

Conrad, Susan. 1994. Variation in Academic Writing: Textbooks and Research Articles across Disciplines. Paper presented at the annual conference of the American Association of Applied Linguistics, Baltimore.

Conrad, Susan, and Douglas Biber, eds. 2001. *Variation in English: Multi-dimensional Studies*. Cambridge, UK: Cambridge University Press.

Dubois, John, Wallace Chafe, Charles Meyer, and Sandra Thompson. 2000-2004. Santa Barbara Corpus of Spoken American English Parts 1-3. Linguistic Data Consortium.

Garside, Roger, Geoffrey Leech, and Anthony McEnery, eds. 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Addison Wesley Longman Ltd.

Hundt, Marianne, Andria Sand, and Rainer Siemund. 1998. *Manual of Information to Accompany the Freiburg-LOB Corpus of British English ('FLOB')*. Freiburg, Germany: Freiburg University.

Hundt, Marianne, Andria Sand, and Paul Skandera. 1999. *Manual of Information to Accompany the Freiburg-Brown Corpus of American English ('Frown')*. Freiburg, Germany: Freiburg University.

Kim, Yong-Jin, and Douglas Biber. 1994. A Corpus-Based Analysis of Register Variation in Korean. In *Sociolinguistic Perspectives on Register*, edited by Douglas Biber and Edward Finegan, 157-81. New York: Oxford University Press.

McEnery, Anthony, Zhonghua Xiao, and Yukio Tono. 2005. *Corpus-based Language Studies: An Advanced Resource Book*. London: Routledge.

Reppen, Randi. 1994. Variation in Elementary Student Writing. Ph.D. diss., Northern Arizona University.

Reppen, Randi, Susan Fitzmaurice, and Douglas Biber, eds. 2002. *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Schiffrin, Deborah. 1982. Discourse Markers: Semantic Resource for the Construction of Conversation. Ph.D. diss., University of Pennsylvania.

Scott, Mike. 1999. *WordSmith Tools*. Oxford, UK: Oxford University Press.

Tribble, Christopher. 1999. Writing Difficult Texts. Ph.D. diss., Lancaster University.

Watson, Greg. 1994. A Multidimensional Analysis of Style in Mudrooroo Nyoongah's Prose Works. *Text* 14 (2): 239-85.

Zhonghua Xiao is a research fellow in Linguistics and Modern English language at Lancaster University. He is the author of Aspect in Mandarin Chinese: A Corpus-based Study *(John Benjamins),* Corpus-based Language Studies: An Advanced Resource Book *(Routledge), both coauthored with Anthony McEnery, as well as a number of research articles in the* Journal of Linguistics*,* Literary and Linguistic Computing*,* Language and Literature*,* Languages in Contrast*,* English Studies*, and* Journal of Universal Language.

Anthony McEnery is a professor of English language and linguistics at Lancaster University. He has published widely in the area of corpus linguistics, though within the area, his major interests are currently the contrastive study of aspect, epistemic modality, and corpus-aided discourse analysis.