Multi-dimensional register classification using bigrams

Scott A. Crossley and Max M. Louwerse
Mississippi State University / University of Memphis

A corpus linguistic analysis investigated register classification using frequency of bigrams in nine spoken and two written corpora. Four dimensions emerged from a factor analysis using bigram frequencies shared across corpora: (1) Scripted vs. Unscripted Discourse, (2) Deliberate vs. Unplanned Discourse, (3) Spatial vs. Non-Spatial Discourse, and (4) Directional vs. Non-Directional Discourse. These findings were replicated in a second analysis. Both analyses demonstrate the strength of bigrams for classifying spoken and written registers, especially in locating distinct collocations among spoken corpora, as well as revealing syntactic and discourse features through a data-driven approach.

Keywords: multi-dimensional analysis, register variation, collocations, bigrams

1. Introduction

Identification and categorization of registers and genres has a long history. Early studies on specialized languages centered on the role of registers following Firth's (1957) concept of lexical collocation. Firth's theory of text cohesiveness and register types is based on the idea that text register and text cohesiveness can be determined by the distribution of the words in a text and their combinations. Later work in register analysis followed Halliday (1978) in which linguists identified special registers on the basis of lexical aspects, which were considered sufficient in themselves in order to distinguish specific registers. While these early register studies generally focused on isolated words and their frequency within texts, they did not consider how registers compared to each other in their respective differences. Recently, however, many linguists have begun to focus on the study of registers from a comparative perspective

ISSN 1384–6655 / E-ISSN 1569–9811

known as register variation (Biber 1988; Conrad & Biber 2001). While much of this current work is located under the paradigm of multi-dimensional variation analysis (Biber 1988) and considers the co-occurrence of syntactic constructions, only some of the research considers how lexical items can be used to analyze register variation (e.g.. Biber, Conrad & Cortes 2004). Of specific importance to this study is the analysis of variation based on shared bigrams, or the co-occurrence of words across corpora, to categorize texts into different dimensions of register variation.

This paper aims to classify registers using the frequency of bigrams shared across corpora. It differs from past examinations of register variation in that it examines the ability to distinguish various registers (both spoken and written) through the use of shared lexical information, thereby capturing syntactic, semantic, and discourse features of registers. An analysis using shared bigrams across corpora follows the linguistic approach of studying systematic variation across registers through co-occurrence features which allows for the examination of registers based on associated features (Conrad & Biber 2001).

The paper is structured as follows. We first discuss related studies that have used an n-gram analysis approach. We next discuss Biber's (1988) work on register variation using factor analysis, since our work is heavily based on this research. We then describe the corpora used in the current study, as well as a description of the method used. This is followed by the results for two studies. The paper concludes with a discussion of the findings.

2. Related work on n-gram analysis

The current studies use bigram analyses as their method. Bigrams (or rather n-grams) have been used historically in computational linguistics to model language based on co-occurrences. For instance, they have been used in a number of language models such as determining the probability of a sequence of words, speech recognition models, spelling correction, machine translation systems, and optical character recognizers (Jurafsky & Martin 2000; Manning & Schutze 1999). In addition, n-grams have been used in genre analysis, specifically for information retrieval purposes. For instance, Peng et al. (2003) used simple n-gram language modeling analyses on well-defined corpora in various languages (Greek, Japanese, Chinese, and English) to classify texts based on languages, authorship, genres, and topics. Using overall accuracy, they were able to classify texts based on languages using bigrams (100% accuracy), attribute correct authorship with trigrams (90% accuracy), classify text genre using bigrams (86% accuracy), and detect topics in various languages at or above 80% accuracy. Furthermore, Stamatatos et al. (2001) used common word unigram frequencies taken from the British National Corpus and found that they were more reliable indicators of genre detection than lexial items taken from their testing corpus, which was the *Wall Street Journal*. In a stylometric approach, Burrows (1987, 1992) used the frequency of common words as markers of style to successfully train a corpus, but unlike Stamatatos et al. (2001), he only used the most common words from the training corpus and not words taken from the "entire" written English language. Recently, Biber et al. (2004) used 'lexical bundles' (quadgrams) to compare the language found in classroom teaching, conversation, textbooks, and academic prose. In their study, they were able to inductively distinguish between classroom teaching and conversation based on stance and discourse organizing bundles and classroom teaching and academic prose based on referential bundles. This research challenged the idea that language is compositional, and instead produces a view of language that often relies on prefabricated expressions (Biber et al. 2004).

Though this overview is far from complete, it does illustrate that n-gram analyses are a useful tool in corpus linguistic analyses. Thus, n-gram analysis was selected as an approach to register analysis for this study. However, in a similar fashion to Peng et al. (2003), and following the computational linguistic standards of Jurafsky and Martin (2000), the current study only considers bigrams. This is because unigrams do not capture enough of the syntactic and semantic context and larger n-grams, such as trigrams and quadgrams, create sparse data problems. Bigrams, on the other hand tap into both the paradigmatic and syntagmatic features of the text and, in addition to being extremely simple to compute, they have been found to be effective in many computational applications to include text categorization.

3. Multi-dimensional analysis and register variation

Multi-dimensional analysis was developed as a methodological approach to identify significant linguistic co-occurrence patterns in language and compare spoken and written genres (Biber 1988, 1993; Conrad & Biber 2001). Because of the wide disparity of linguistic and lexical features found across texts, it is generally difficult to reliably distinguish registers based on all language characteristics. It is, however, possible to analyze registers based on the co-occurrence and alternation patterns using the statistical technique of factor analysis that explains the variability among a number of observable variables in terms of a smaller number of unobservable variables (i.e., factors).

The standard for delimiting registers based on underlying factors of syntactic features is Biber (1988). Biber's corpus analysis consisted of 481 written and spoken texts comprising 17 written and six spoken registers. The study focused on word-level information (e.g., parts-of-speech) and looked at 67 linguistic features including tense and aspect markers, place and time adverbials, pronouns and pro-verbs, questions, nominal forms, passives, stative forms, subordination features, prepositional phrases, adjectives and adverbs, modals, specialized verb classes, reduced forms and dispreferred structures, and coordinations and negations. The normalized frequencies of these features in each of the registers were then entered in a factor analysis, from which six factors emerged. These factors can be seen as dimensions on which registers can be placed. Biber defined the sets of relations among texts as follows: (1) Involved versus Informational Production; (2) Narrative versus Non-Narrative Concerns; (3) Explicit versus Situation Dependent Reference; (4) Overt Expression of Persuasion; (5) Abstract versus Non-Abstract Information; (6) Online Informational Elaboration. For instance, registers such as romantic fiction, mystery fiction and science fiction were positioned high on the second dimension (Narrative); whereas registers such as academic prose, official documents, hobbies, and broadcasts scored low (Non-Narrative). Biber's research is also supported by the findings of Louwerse et al. (2004) who used 250 measures of linguistic cohesion to reconstruct Biber's original work with similar findings.

The current study follows the methodology outlined by Biber (1988) and Louwerse et al., (2004) except for two important differences. First, the current study puts an emphasis on different registers of spoken texts. In addition to using the London-Lund Corpus that was included in Biber's (1988) study, we used task-based dialogs as well as telephone and face-to-face conversations on a diversity of topics. The reason for these additions is related to our current work on multimodal communication in humans and computer agents for which we use the HCRC MapTask scenario (Louwerse et al. 2006). This task-based corpus is a transcript of dialog partners navigating a route on a map. One of the questions that will be raised in this study is to what extent task-based dialogs differ from other spoken registers.

The second difference from Biber (1988) and Louwerse et al. (2004) stems from a recently conducted analysis by Louwerse and Crossley (2006) in which a speech act classification system was created to classify utterances from the Map Task Corpus. Using n-grams (unigrams, bigrams, and trigrams), they designed an algorithm that assigned one of twelve dialog acts to an utterance from the

Map Task Corpus. The system's performance was on par with human performance. Louwerse and Crossley raised the question to what extent n-grams play a specific role in the MapTask corpus. Do speakers in the MapTask scenario use specific lexical items that allow for content analysis more so than in other corpora? Moreover, does a content-based analysis using n-grams allow for classification of registers and, if so, what does this classification look like? There is a practical component to our approach as well. Biber's analysis of 67 linguistic features requires a computational tool with sophisticated syntactic taggers and parsers, bag-of-words algorithms as well as a manual verification of the data. The question in this study is whether categorizations can be obtained using a simple n-gram algorithm. With their ability to capture syntactic, semantic, and discourse features of language, an n-gram analysis might be well suited toward a multi-dimensional analysis of register variation. Two studies were conducted to answer the question of how task-oriented dialogs differ from other dialogs and how these differences come about in lexical and syntactic collocations.

4. Corpora and methodology of the present study

4.1 Corpora

Because this investigation was primarily concerned with distinguishing differences between spoken registers, the bulk of the corpora selected were instances of spoken discourse. The spoken corpora chosen for this analysis were the six spoken corpora used in the London Lund Corpus (LLC) (broadcast speeches, face to face conversations, telephone conversations, interview, spontaneous speeches, and prepared speeches) (ICAME 1999), the Map Task Corpus (HCRC 1993), the TRAINS Corpus (Allen & Heeman 1995), the Santa Barbara Corpus (Du Bois et al. 2000), and the Switchboard Corpus (Godfrey & Holliman 1993). These corpora were augmented by two written corpora: the Brown Corpus (ICAME 1999) and the Lancaster Oslo Bergen (LOB) Corpus (ICAME 1999). Each of these corpora is discussed next.

The Map Task Corpus is a tasked-based corpus that is the linguistic product of a cooperative task involving two participants. The Instruction Givers have a marked route on their map and give directions to the Instruction Followers who have no route. The maps are not identical, which elicits unscripted problem solving dialog. Because of the domain, we predicted a strong spatial predisposition in the lexical and syntactic collocations of this dialog. However, to ensure that any Map Task Corpus findings were based purely on their spatiality, another task-based corpus was selected to include in the analysis: the TRAINS Corpus. This corpus is based on the routing and scheduling of freight trains. The corpus shares with Map Task its basis as a task based corpus, but it is more temporal and directional in nature than the spatial Map Task Corpus. A selection of non-task based spoken dialogs were also included in the analysis. Following the methodology of Biber (1988), the spoken dialogs found in the London Lund Corpus were included in the analysis and broken up into six different speech situations: spontaneous speech, prepared speech, face to face conversations, telephone conversations, interviews, and broadcast speech. Our primary purpose in including the LLC was to use it as a means to compare natural dialogs and task-based dialogs. As such, we were also interested in the possibility that bigrams might be powerful enough to delimit natural dialogs from task-based dialogs in a factor analysis based on the idea that task-based dialogs were more instructional in nature and depended on a more controlled lexical domain. Because there has been ample attention given to dialectical differences between American and British spoken dialects (Biber 1987; Helt 2001), we also included American spoken dialogs. These included the Santa Barbara Corpus and the Switchboard Corpus. The Santa Barbara Corpus is a collection of natural speech recordings taken from people across the United States. The Switchboard Corpus, on the other hand, is a collection of about 2,400 two-sided random topic telephone conversations taken from 543 speakers from all areas of the United States. Since one of Biber's (1988) primary research questions was the delineation of spoken and written dimensions, we also included the LOB and the Brown corpora. The Brown corpus is a collection of written American texts published in 1961. It comprises 500 text samples of about 2,000 words each and totals about one million words. Each text sample is categorized into one of fifteen registers including religion, science, fiction, humor, and press reports. The LOB Corpus is a direct replication of the Brown Corpus, but is based on 1961 text samples taken from British written sources. Based on the work of Biber (1988) and Louwerse et al. (2004) it was thought that the written corpora would be distinguished from the spoken corpora as a result of the written texts being more integrated and less fragmented and involved. A brief description of the corpus used in this investigation is found in Table 1.

4.2 Methodology

A study reproducing the methods employed by Biber (1988) was used to identify the role that bigrams played in shaping spoken and written genres. In his original analysis, Biber examined 23 corpora based on 67 linguistic features

Table 1. Corpora Overview

Spoken Corpora	**Description**	**Dialect**	**Year**	**Size**
Map Task	Task-based scenario (directions)	British / Scottish	1991	150,000
TRAINS	Task-based scenario (transportation)	American	1995	55,000
LLC Broadcast	Radio and television broadcasts	British	1980	90,000
LLC Face to Face	Casual conversations natural settings	British	1980	220,000
LLC Telephone	Casual conversations over telephone	British	1980	135,000
LLC Interviews	Radio and telephone interviews	British	1980	110,000
LLC Spontaneous	Spontaneous speeches	British	1980	80,000
LLC Prepared	Prepared speeches	British	1980	70,000
Santa Barbara	Casual Conversations, natural settings, natural speech	American	2000	200,000
Switch Board Task-based map scenarios (spatial)	Casual Conversations over telephone, topic based	British	1997	3,044,734
Written Corpora	**Description**	**Dialect**	**Year**	**Size**
LOB Corpus	Collections of writings including religion, humor, fiction, biographies, and press texts	British	1978	1,000,000
Brown Corpus	Collections of writings including religion, humor, fiction, biographies, and press texts	American	1982	1,000,000

using frequency counts that were normalized for corpus size. Unlike Biber, though, this project was concerned with linguistic features that could present larger data structures above 67 variables. For this reason, the linguistic features observed were marked for sharedness. This was especially important for the bigram analysis, as the number of bigrams had the potential to run at over 50,000 variables. Therefore, to control for size, only those bigrams that occurred in all text samples were included. While this is a similar approach to that used by Biber (1988) it appears to be different from all other n-gram analyses that have been conducted in the field of multi-dimensional analysis (c.f. Biber et al. 2004).

Following the methodology of Biber (1988), the data was then entered into a factor analysis.[1] The factor analysis (a principle component analysis) was used to cluster bigrams into groups that co-occurred frequently within the texts. At this point, the study departed slightly from the methods used by Biber.

While Biber was looking at 67 common linguistic features found within text, this study was searching for shared bigrams. Because of the widespread use of the Biber's linguistic features, he was able to divide registers into individual texts leading to a large number of available samples for his factor analysis (481 in his 1988 study). In this study, shared bigrams turned out to be infrequent and larger text samples were needed to provide better linguistic coverage leading to fewer samples available for the factor analysis. Such a small sample size could be considered problematic with some (Tabachnick & Fidell 2001) arguing for at least 300 samples. Others, however, have argued that factor loadings are more important than sample size. Guadagnoli and Velicer (1988) for instance contended that if a factor had four or more loadings greater than .6, it was reliable regardless of sample size. In a later study, MacCallum et al. (1999) found that if all commonalities had factor loadings above .6, then small sample sizes were acceptable. We acknowledge that larger sample size is always better, but in this analysis, we were limited by corpus constraints.

After the factor analyses were conducted, the factors were interpreted as register dimensions through the assessment of bigrams that clustered in each factor. This was done by computing a factor score for each register. The algorithm for these factor scores followed Biber's (1988) method in which all frequencies were standardized to a mean of 0.0 and a standard deviation of 1.0. A factor score was then computed by summing the number of occurrences of the bigrams in that factor if that bigram loaded highest in that factor as compared to other factors. This allowed each bigram to only be included in the computation of only one factor score. These frequency scores were then standardized using the mean and standard deviation to label registers as being marked or unmarked with respect to the register dimension. The resulting register dimensions were interpreted in consideration of the relations of the registers highlighted through the factor scores.

Two analyses were conducted. The first analysis considered all bigrams shared across corpora. The second analysis, meant as a complement to the first, broke down the larger corpora into smaller chunks to ensure that the results were not the result of size or scarcity issues. This involved dividing the Map Task Corpus and the TRAINS Corpus into three approximately equal parts and adding them to the existing corpora. The LOB Corpus and the Brown Corpus were also subdivided into smaller categories (press, recreation, fictional and informational) in order to ensure that any statistical differences were not the result of combining defined registers. This new collection of corpora was then analyzed for shared bigrams. After the shared bigrams were established, a new factor analysis was conducted.

5. Results and discussion: Study 1

In the first analysis, all 12 corpora were analyzed. The LOB and the Brown corpora were not broken down into their various registers and no attempt was made to divide any of the larger corpora into chunks. All the bigrams shared between these 12 corpora were extracted, but in order to control for spurious data, only those bigrams that had a mean occurrence per 1,000 words of .50 or greater across all corpora were analyzed using a factor analysis. The factor analysis using a Promax rotation with eigenvalues over .35 revealed four dimensions (see Table 2 for factor scores), which, like Biber's (1988) original dimensions, plotted registers into positive and negative loadings that were distributed across texts in complementary patterns. All factors in this study had four or more loadings greater than .6 and 11 of the bigrams loaded at below .6. Four dimensions emerged from the analysis: Scripted vs. Unscripted Discourse, Deliberate vs. Unplanned Discourse, Spatial vs. Non-spatial Discourse, and Directional vs. Non-directional Discourse. These dimensions are discussed next.

5.1. Dimension 1: Scripted vs. Unscripted Discourse

The first factor comprised 26 bigrams and explained 33% of the total variance. When the bigram frequency scores were computed for the registers in this dimension, the most appropriate label for this dimension was 'Scripted vs. Unscripted Discourse' because it separated natural dialogues from monologues, written texts, and tasked-based dialogues (see Figure 1). The dimension also distinguished between American dialects, which were more marked, and British dialects, which were less marked, but only those that were not task-based. The corpora that were most marked on this dimension included the Switchboard Corpus and the Santa Barbara Corpus (both American), but all natural dialogs were positively marked including the LLC registers of face to face dialogs, interviews, and telephone discussions. Those corpora which were negatively marked included monologs (LLC spontaneous and prepared speeches), written corpora (both Brown and LOB), and task based dialogs (Map Task and TRAINS Corpora).

The linguistic features that mark the Unscripted Discourse side of this dimension were the use of filler phrases such as *you know* and *I mean*, and the use of coordinating conjunctions collocated with first person pronouns such as *and I*, *so I*, *but I*, and *and we*, the general inclusion of first person pronouns,

Table 2. Factor loadings Study 1

Factor 1		Factor 2		Factor 3		Factor 4	
Bigrams	**Loading**	**Bigrams**	**Loading**	**Bigrams**	**Loading**	**Bigrams**	**Loading**
DON’T KNOW	.976	IN THE	.988	YOU GO	.993	NEED TO	.956
YOU KNOW	.969	OF A	.978	HAVE YOU	.985	GO TO	.925
I DON’T	.958	FOR THE	.977	UP TO	.984	BACK TO	.891
I KNOW	.957	TO A	.967	OF IT	.977	TO GET	.883
TO DO	.950	THE FIRST	.951	GOT A	.977	THE OTHER	.867
THAT I	.926	BY THE	.939	I’VE GOT	.967	HAVE TO	.865
I MEAN	.917	IT IS	.920	AT THE	.931	UP THE	.850
BUT I	.914	HAVE BEEN	.885	THE WAY	.930	THE SAME	.846
THAT WAS	.912	THAT THE	.878	TO THE	.923	AND THEN	.756
I WAS	.884	AND THE	.814	YOU HAVE	.867	SO THAT	.754
AND I	.881	THERE WAS	.810	DO YOU	.862	WE HAVE	.745
A LOT	.863	TO BE	.808	TO GO	.856	IS THAT	.732
AND IT	.861	FOR A	.808	IF YOU	.848	TO MAKE	.728
KIND OF	.850	IS A	.773	FROM THE	.812	I HAVE	.649
I THINK	.849	WITH A	.759	HAVE A	.759	WOULD BE	.536
AND YOU	.736	THERE IS	.724	AND A	.692	WHICH IS	-.575
YOU CAN	.723	WAS A	.711	WANT TO	.664	IN FACT	-.770
ALL THE	.718	IS THE	.704	OF THE	.654		
IT WAS	.714	INTO THE	.685	THAT IS	-.506		
I WOULD	.646	IN A	.658	AND THAT	-.620		
AND WE	.617	WILL BE	.572	ONE OF	-.635		
A LITTLE	.570	ON THE	.542				
TO HAVE	.556	OUT OF	.511				
GOING TO	.491	YOU SEE	-.479				
THIS IS	.442	SO I	-.666				
WITH THE	-.613	IS IT	-.709				

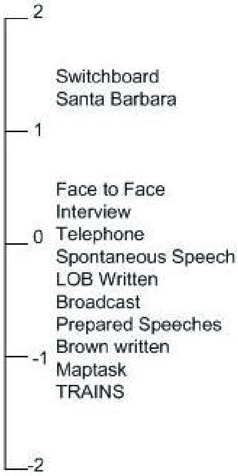

Figure 1. Dimension 1: Scripted vs. Unscripted Discourse Marked and Unmarked Registers

the expression of opinions such as *I think* and *I would*, the use of phrasal quantity adjectives such as *a little* and *a lot*, the use of pro-verbs such as *to do*, and finally the use of phrasal auxiliaries for the future aspect *going to* and *want to* as compared to the use of the modal *will*. The dimension was also constructed in opposition to prepositional phrases such as *with the*. Examples of these bigrams as found in the most marked corpora can be found in Table 3. The factor analysis demonstrated that natural, spoken dialogues showed a preference for using hedges, a high degree of coordinating conjunctions, especially when combined with first and second person pronouns, the expression of opinions, phrasal future aspects, phrasal adjectives for quantity, and a variety of prepositional phrases.

Many of these bigram correlate well with Biber et al.'s (2004) study of lexical bundles. That study reported that in spoken university registers and natural conversation, there was increase in the use of pronouns as compared to written registers such as textbooks and academic prose. The study also found that the use of first person and second person pronouns often occur with topic introducing bundles such as *going to talk about* and *want to talk about*. The bigrams in this study used to express opinion seem similar to the category 'epistemic stance bundles' used by Biber et al. (2004) to categorize bundles that commented on the knowledge status of information; in addition, the filler phrases in this study are also similar to the discourse organizing bundles labeled 'topic elaboration and clarification bundles' by Biber et al. (2004), which were analyzed in

Table 3. Bigram examples for Dimension 1

Bigrams	Corpus Example
Don't know	I don't know what her plans really are.
You know	And you know what we start out with.
I don't	I don't know what it's called either.
I know	All I know is you light a candle.
To do	Which is what I'm going to do.

their study as being useful when the speaker believes additional explanation is needed. The phrasal quantity bigrams found in this study were also very similar to Biber et al's. (2004) referential bundles used to specify attributes of following head nouns (i.e. *have a lot of* and *have a little of*).

In consideration of the possible dialectical differences computed through the factor analysis, a few multi-dimensional studies have addressed similar findings. These include Biber's (1987) study of American and British written registers in which he concluded that American registers were more colloquial and interactive, while the British registers were more situated and less abstract. Also, Helt (2001), using a similar methodology to Biber (1988), compared spoken British and American registers and found that American telephone conversations were more involved, less formal, and used more overt expressions of argumentation and persuasion than their British counterparts. Helt also found that British speakers were less abstract and more concrete (especially in reference to agents) than American speakers. Helt's (2001) findings have recently been supported by a series of studies conducted by Hall, McCarthy, and McNamara (Hall et al. 2006; McCarthy et al. 2007) in which British texts were found to have more concrete links at the discourse level, especially in reference to co-referential cohesion, causal connectives, intentional events, and logical operators. The findings of the current study seem to support many of the conclusions drawn from these past investigations.

5.2. Dimension 2: Deliberate vs. Unplanned Discourse

The second dimension comprised 26 bigrams and explained 27% of the total variance. The dimension has been labeled 'Deliberate vs. Unplanned Discourse' as it separated written and memorized texts from all other spoken texts (see Figure 2). The corpora that were most marked on this dimension included the LOB and Brown Written corpora. Prepared speeches, which are quite similar to written texts as they are usually memorizations of written texts, were marked positively on the scale, but they were noticeably lower than the other

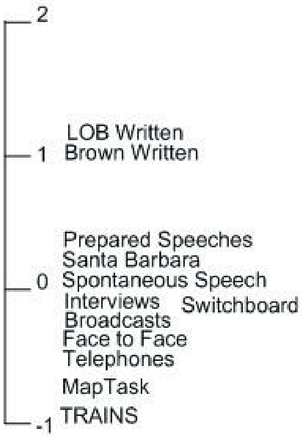

Figure 2. Dimension 2: Deliberate vs. Unplanned Discourse Marked and Unmarked Registers

written texts. The Santa Barbara Corpus and the LLC spontaneous speech texts were also marked positively, but only marginally so. Those registers which were marked negatively included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, or monologs.

The linguistic features that set apart dimension two from the other dimension were the high use of non-spatial prepositional phrases such as *in the*, *for the*, *in a*, *of a*, and *by the*, the use of declarative existential *there* and *it* clauses such as *there is*, *there was*, and *it is*, the use of *that* complementizers, the use of modals for future time such as *will be*, as compared to phrasal auxiliaries for expressing future time such as *going to* and *want to*, and the use of present perfects such as *have been*. The dimension was also constructed in opposition to bigrams with first person pronouns such as *you see* and *so I*, as well as question bigrams such as *is it*. Examples of these bigrams as found in the most marked corpora can be found in Table 4.

The results of the factor analysis demonstrate that in comparison to unplanned discourse, deliberate texts tend to favor non-spatial prepositional phrases that collocate with definite articles, the use of existential *there* and *it* clauses, the perfect tense, the use of *that* complementizers, a lack of first person pronouns, and the use of modals to express future time as compared to phrasal auxiliaries. These findings should be taken to complement the findings of Biber (1988), who, while analyzing the complexities of spoken and written genres, found that written genres were less likely to permit the use of reduced structures (that deletions, contractions, pro-verbs *do*, pronouns, and clausal coordinators),

Table 4. Bigram examples for Dimension 2.

Bigrams	Corpus Example
In the	The biological control of pests in the garden.
Of a	Used for the compiling of a dictionary.
For the	Signed to meet Marty Servo for the World's welter crown.
To a	Or to a different part of the town altogether.
The first	This was the first representative match.

non-complex integrated structures (nouns, prepositions, attributive adjectives, nominalizations, word lengths, and TTR (Type-Token Ratio) scores) and passive structures, moderately complex forms of referential elaboration (*wh-* and *that* relative clauses), and complex forms of framing elaboration (*wh-* and *that* clauses, sentence relatives, and subordinations). The findings of this analysis also support the research carried out by Rayson et al. (1997), who conducted a word frequency analysis of spoken and written language as found in the British National Corpus and found that written texts were more likely to contain prepositions and the article *the*, while spoken texts were more likely to contain first and second person pronouns, contractions and interjections. What this factor analysis demonstrates is that deliberate discourse can be distinguished based on its formality, existentialism, and use of non-spatial prepositional phrases.

5.3 Dimension 3: Spatial vs. Non-spatial Discourse

The third dimension comprised 21 bigrams and explained 15% of the total variance. The dimension has been labeled 'Spatial vs. Non-spatial Discourse' as it was controlled by the Map Task corpus, which was marked well above the other corpora (Figure 3). The Switchboard and Santa Barbara Corpora were also marked positively, but only marginally so. Those registers which were negatively marked included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, or monologs and all written texts.

The space constructing the dimension was controlled by linguistic features that provided spatial information or were interrogative in nature. These included spatial prepositions such as *of the*, *to the*, *up to*, *from the*, and *at the*, the use of question bigrams such as *have you* and *do you*, the use of confirmation bigrams such as *I've got*, and the use of conditional bigrams such as *if you*. The dimension was also constructed in opposition to the use of ambiguous demonstratives, which are counterproductive in spatial guidance tasks. This opposition can be seen in the negative loadings of phrases such as *and that* and *that is*.

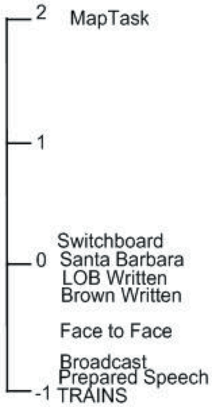

Figure 3. Dimension 3: Spatial vs. Non-spatial Discourse Marked and Unmarked Registers

Examples of these bigrams as found in the most marked corpora can be found in Table 5.

The factor analysis demonstrated that spatial discourse, and in this case, specifically, the Map Task Corpus, is defined by its spatiality (e.g. spatial prepositional phrases) and frozen expressions for coordinating a route on a map (e.g. *do you*, as in *Do you have a church on your map?*, *I want* as in *I want to go to the left side*, and *you have* as in *you have to go down here*). These frozen expressions come to form the Map Task dialog structure and have been classified into 12 different categories or conversational moves (Carletta 1996; Carletta et al. 1997). These conversational moves form the core of Map Task dialogs and correspond to linguistic features that bind a spatial task-based discourse. Many of these spatial bigrams are similar in nature to lexical referential bundles labeled by Biber et al. (2004) as 'time/place/text-deixis bundles' in that they refer to particular places or locations in the task. Biber et al., however, allocated these bundles to written registers only.

Table 5. Bigram examples for Dimension 3.

Bigrams	**Corpus Example**
You go	You go horizontal.
Have you	Have you got banana tree?
Up to	You're going up to the sort of left-hand.
Of it	The bottom edge of it.
Got a	You've got a camera shop.

5.4 Dimension 4: Directional vs. Non-directional Discourse

The fourth dimension comprised 17 bigrams and explained 11% of the total variance. The dimension was labeled 'Directional vs. Non-directional Discourse' as it was noticeably dominated by the TRAINS corpus and consisted of linguistic features that provided direction and managed time (Figure 4). Those registers which were marked negatively included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, monologs or written texts.

The space constructing the dimension was controlled by linguistic features that provided directional information or were directive in nature. These features included the use of temporal cohesion markers such as *and then*, directional bigrams such as *go to*, and *back to*, and the use of phrasal modals of volition such as *need to* and *have to*. Examples of these bigrams as found in the most marked corpora can be found in Table 6. These bigrams are unique to the TRAINS Corpus as they provide an indication of the both the task-based and directional nature of the discourse.

Like the Map Task corpus, the TRAINS corpus has also been categorized based on speech acts common to the discourse (Sikorski & Allen 1997). These speech acts are generally related to the corpus' purpose of building transportation plans for later use. What the factor analysis demonstrates is that these speech acts defined the directionality of the TRAINS corpus. In the problem solving discourse of TRAINS, many speech acts are strictly directional in nature. These include the speech acts 'establishing goal,' 'establishing solution,' and

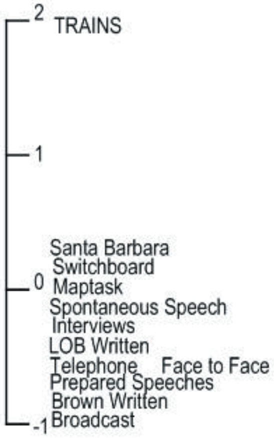

Figure 4. Dimension 4: Directional vs. Non-directional Discourse Marked and Unmarked Registers

Table 6. Bigram examples for Dimension 4.

Bigrams	Corpus Example
Need to	We need to get one boxcar of bananas to Bath.
Go to	Go to Bath and pick up two boxcars.
Back to	And then move back to Dansville.
To get	How long does it take to get from Avon to Dansville.
The other	But it could go the other direction.

'evaluating solution,' all of which are categorized as 'Action-Direction' speech acts. These 'Action-Direction' speech acts help inform the TRAINS dimension by providing the instructional nature of the dimension.

The findings of the first analysis provide a preliminary overview of how domain specific registers and wider genre dimensions can be better explained using the frequency of their shared bigrams. This initial investigation further demonstrates how bigrams, which are capable of capturing syntactic, semantic, and discourse features of language, can come to inform the field of multi-dimensional analysis and provide for the linguistic distinction of various spoken and written corpora.

6. Results and discussion: Study 2

Because a small number of corpora examples were used in the first analysis, a second analysis was undertaken with the purpose of ensuring that the findings from the initial investigation could be generalized and that the dimensions computed were not the result of heterogeneous datasets. In the second study, those large corpora that had constructed specific register types and controlled various dimensions were subdivided into smaller portions to ensure that it was neither the size, nor the scarcity of the corpora that were determining the findings. To this end, the Map Task corpus and the TRAINS corpus were divided into three parts of equal size and the LOB and Brown corpora were divided into four parts (fiction, press reportage, recreation, and informational). Equal parts rather than unique dialogs and texts were preferred to ensure a representative sample of bigrams across the corpora. This division gave the factor analysis 22 samples from which to compute factor loadings.

Not surprisingly, with the corpora broken into smaller chunks, there were less shared bigrams overall. In the second analysis there were only 82 shared bigrams available as compared to the 454 bigrams available in the original analysis. Considering the need to have as many variables as possible to provide a robust factor analysis, all bigrams regardless of frequency were used in the second factor analysis. All factors in this study had four or more loadings greater than .6 and 6 of the bigrams loaded at below .6. Because the shared bigrams in the second analysis differed from the first, some dimensions lost bigrams and, as a result, volume. Moreover, only 80 of the bigrams had loadings above .35, so only 80 bigrams were used to create the factor scores (see Table 3 for detail). These differences led to a rearrangement of the order of the dimensions, but not the significance of the dimensions. It also led to the inclusion of new, shared bigrams and the exclusion of bi-grams that occurred in the first study, but were no longer shared among the newly defined corpora. However, as in the first analysis, four dimensions emerged: Directional vs. Non-directional Discourse, Deliberate vs. Unplanned Discourse, Spatial vs. Non-spatial Discourse, and Scripted vs. Unscripted Discourse. These dimensions will be discussed next.

6.1 Dimension 1: Directional vs. Non-directional Discourse

The first dimension in the second analysis was comprised of 30 bigrams which explained 36% of the total variance. The dimension was dominated by all three parts of the TRAINS corpus and consisted of linguistic features that were both task and directionally based. A Pearson Correlation[2] between the dimension loadings of Dimension 4 in the first bigram analysis (labeled 'Directional vs. Non-directional Discourse' as well) and Dimension 1 of the second bigram analysis demonstrated a significant correlation ($r = .989$, $p < .001$, $N = 22$). As with the initial analysis, the Santa Barbara and Map Task Corpora were also marked positively, but only marginally so. Those registers which were marked negatively included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, monologs or written texts.

The linguistic features that were strongest in this dimension were similar to that of the fourth dimension of the first analysis. These features included the use of temporal markers such as *and then* and *then the*. Phrasal modals again helped to structure the dimension with both *have to* and *need to* being included as well as directional bigrams such as *go to*, *back to*, *up the*, *take the*, *take a*, *make it*, and *be at*, and, new to this analysis, numeric bigrams such as *the two*, *two and*, *the one*. Phrasal modals of volition such as *need to* and *have to* also shaped this factor. As in the initial analysis, these bigrams are unique to the TRAINS Corpus as they provide an indication of both the task-based and directional nature of the discourse.

Table 7. Factor loadings Study 2

Factor 1		Factor 2		Factor 3			
Bigrams	**Loading**	**Bigrams**	**Loading**	**Bigrams**	**Loading**	**Bigrams**	**Loadings**
GO TO	.973	BY THE	.966	UP TO	.939	TRYING TO	.871
NEED TO	.970	AS THE	.960	DO YOU	.913	AND I	.828
TO GET	.959	AND THE	.952	IF YOU	.905	BUT I	.763
TAKE THE	.955	IN THE	.940	YOU HAVE	.898	YOU CAN	.728
SO IT	.943	THE FIRST	.913	AT THE	.896	AND IT	.650
BACK TO	.935	IT IS	.904	UP AND	.882	TO DO	.646
TO TAKE	.927	THERE IS	.889	TO THE	.881	ALL THE	.600
BE AT	.915	IS THE	.868	TO GO	.870	AND SO	.578
IT TO	.902	WITH THE	.864	SHOULD BE	.841	AS WELL	-.712
IS THAT	.901	AFTER THE	.854	I WANT	.742		
HAVE TO	.900	WITH A	.830	AND YOU	.671		
THE TWO	.898	TO BE	.825	AND A	.667		
TWO AND	.875	IS TO	.795	WANT TO	.612		
UP THE	.873	WILL BE	.735	TO HAVE	-.726		
THAT WOULD	.865	ON THE	.690	AND THAT	-.746		
THE OTHER	.865	OF THE	.641				
UP A	.860	INTO THE	.410				
THEN THE	.805	KNOW WHAT	-.501				
TAKE A	.789	I DO	-.519				
AND THEN	.781	IF I	-.526				
TO MAKE	.716	WHAT I	-.655				
I HAVE	.715	I THINK	-.662				
THE ONE	.712	IS IT	-.667				
IT WOULD	.702	IT AND	-.780				
THE SAME	.690	GOING TO	-.810				
SO THAT	.679						
HAVE AN	.677						
MAKE IT	.644						
I CAN	.618						
WOULD BE	.591						

6.2 Dimension 2: Deliberate vs. Unplanned Discourse

The second dimension was comprised of 25 bigrams and explained 20% of the total variance. The dimension was dominated by written texts, which were in opposition to spoken texts and consisted of linguistic features that demonstrated more formality and less speaker centeredness. A Pearson Correlation between the dimension loadings of Dimension 2 in the first bigram analysis (labeled 'Deliberate vs. Unplanned Discourse' as well) and Dimension 2 of the second bigram analysis demonstrated significant correlation ($r = 0.850$, $p < 0.001$, $N = 22$). In this dimension, those texts that included more optimal varieties of written English such as press reportage, information, and recreation were marked highest on the dimension, while fiction writing, which presumably includes instances of reported speech and other spoken examples of speech, were marked lower, but still positively. One of the subdivided Map Task Corpora and the Santa Barbara and Switchboard Corpora were also marked positively, but only marginally so. Those registers which were marked negatively included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, or monologs.

The linguistic features that set apart dimension two from the other dimension were similar to the findings in the first study and included the high use of non-spatial prepositions that collocate with direct articles such as *of the*, *in the*, *on the*, *and the*, and *with the*, the use of existential *there* and *it* clauses such as *there is* and *it is*, the use of modals for future time such as *will be*, as compared to phrasal auxiliaries for expressing future time such as *going to*. Bigrams that collocated with first person singular pronouns also loaded negatively on this factor. These included the bigrams *what I*, *if I*, *I think*, and *I do*. As in the first analysis, these findings support the idea that deliberate and unplanned discourse can be distinguished based on its formality, existentialism, and use of non-spatial prepositional phrases.

6.3 Dimension 3: Spatial vs. Non-spatial Discourse

The third dimension comprised 15 bigrams and explained 17% of the total variance. This dimension, like the third dimension in the first analysis, was dominated by the Map Task subdivided corpora and has been labeled 'Spatial vs. Non-spatial Discourse'. A Pearson Correlation between the dimension loadings of dimension three in the first bigram analysis (labeled 'Spatial vs. Non-spatial Discourse' as well) and dimension three of the second bigram analysis demonstrated significant correlation ($r = 0.951$, $p < 0.001$, $N = 22$). In this dimension, the Switchboard, TRAINS, and Santa Barbara Corpora were also marked positively, but only marginally so. Those registers which were marked negatively included all other spoken texts regardless of whether they were natural dialogs, task-based dialogs, monologs or written texts.

The space constructing the dimension was controlled by linguistic features similar to that of the first analysis and were based on spatial preposition such as *to the*, *at the*, and *up to*, the use of questions tags such as *do you*, and the use of conditional bigrams such as *if you*. This dimension was also constructed in opposition to the use of ambiguous demonstratives which can be seen in the negative loadings of phrases such as *and that*. This second analysis gives further support for the idea that spatial, task-based registers can be distinguished based on their use of spatial prepositional phrases, question types, and conditional markers.

6.4 Dimension 4: Scripted vs. Unscripted Discourse

The fourth dimension was comprised of 9 bigrams and explained 6% of the total variance. This dimension, like the first dimension in the initial analysis, favored natural dialogues over monologues, written texts, and tasked-based dialogues and, like the first analysis, was subsequently labeled 'Scripted vs. Unscripted Discourse'. Like the findings of the first analysis, it also loaded American dialects higher than British dialects. A Pearson Correlation between the first dimension in the initial analysis and this dimension demonstrated a significant correlation between the two ($r = 0.944$, $p < 0.001$, $N = 22$). The corpora that were marked highest on this dimension included the Switchboard Corpus and the Santa Barbara Corpus (both American), but all natural dialogs were marked positively including the LLC registers of face to face dialogs, interviews, and telephone discussions. Those registers which were marked negatively included monologs (LLC spontaneous and prepared speeches), written corpora (both Brown and LOB), and task-based dialogs (Map Task and TRAINS Corpora).

With the division of the corpora into smaller groups, many of the original shared bigrams that constructed this dimension were lost. No filler phrases such *You know* and *I mean*, were shared between the corpora. However, the linguistic features that did come to construct this dimension were similar to the first study and included the use of common coordinating conjunctions such as *and it*, *and so*, and *and now*, and the use of coordinating conjunctions that collocated with first and second person pronouns such as *and I*, and *but I*. Additionally, the logical connector *as well* scored negatively on this dimension. The factor analysis, therefore, supports the findings of the initial analysis and shows that natural, spoken dialogues, demonstrate a preference for using hedges, a high degree of coordinating conjunctions, especially when combined with first and second person pronouns, and a dearth of prepositional phrases.

The findings of the second analysis support the findings of the first analysis. The dividing of the larger corpora and the subsequent analysis of their shared bigrams provides additional, supporting evidence that domain specific registers and wider genre dimensions can be better explained using the frequency of their shared bigrams. The strong correlations between the dimension loadings of the first analysis and the second analysis provide support for the idea that it is the linguistic properties of the bigrams, which are capable of capturing syntactic, semantic, and discourse features of language, that provide evidence for the linguistic distinction between spoken and written corpora. Findings such as these are also important as they can be used to inform the field of multi-dimensional analysis.

7. Discussion and conclusions

Previous research (e.g. Biber 1988) has shown that registers can be classified using linguistic features operating at the word level. These features particularly focused on syntactic characteristics producing dimensions like 'Involved versus Informational Production'; 'Narrative versus Non-Narrative Concerns'; 'Explicit versus Situation Dependent Reference'; 'Overt Expression of Persuasion'; 'Abstract versus Non-Abstract Information'; and 'Online Informational Elaboration' in a factor analysis. At the same time, studies have been conducted using bigrams for dialog act classification (Louwerse & Crossley 2006) and speech recognition (Jelinek 1997). The current study is the first that uses bigrams for register classification and it is noteworthy that the frequencies in this study were taken from bigrams shared across corpora to avoid idiosyncrasies of a particular corpus. The advantage of a bigram approach is that it does not assume syntactic information, but rather lexical information by taking frequent collocations in different corpora. At the same time, however, the bigrams isolated here are not only based on lexical differences, but also on more latent, syntactic and discourse features. So, while many of the dimensions realized in this analysis are ostensibly based on lexical collocations, when scrutinized in their total relation to other bigrams found within the same dimension, it becomes clear that bigrams inform register variation studies based on more than just simple word collocations and word meanings. For instance, natural, spoken dialogues show a preference for using hedges, a high degree of coordinating conjunctions, especially when combined with first and second person pronouns, the expression of opinions, and a lack of prepositional phrases. In comparison to spoken texts, written texts tend to favor non-spatial prepositional phrases, the use of existential *there* and *it* clauses, the perfect tense, the use of *that* complementizers, a lack of first person pronouns, and the use of modals to express future time as compared to phrasal auxiliaries. In reference to more specific corpora, such as the Map Task and TRAINS Corpora, this analysis has demonstrated that they are easily defined by their intentions. The Map Task corpus, for instance, is delineated by its use of spatial prepositions, question tags, and its negative use of ambiguous demonstratives. These linguistic features that build the Map Task corpora correspond to the preconceived notions of what linguistic features bind a spatial task-based information gap conversation. In addition to the Map Task corpus, the TRAINS corpus is defined by its use of temporal cohesion markers, directional bigrams, and its use of phrasal modals of volition. These linguistic features are likely key to the shaping of directional task-based activities.

In consideration of Biber's (1988) seminal study on register variation, we find that this study is compatible in its findings. For instance, like Biber's study, a bigram approach was also able to distinguish between spoken and written genres. In addition this approach allowed for a distinction between seemingly comparable spoken genres such scripted and unscripted discourse, spatial discourse, and directional discourse. Perhaps the most salient finding of this research is the simplicity of the methodology used and the strength of the findings in relation to the approach. The bigram approach discussed in this paper demonstrates that examining the relatively straightforward frequency of word pairings can function as a powerful approach to multi-dimensional analysis. An approach such as this is computationally less complex and time consuming than approaches that depend on syntactic features. Additionally, while this approach depends explicitly on simple word collocations, the collocations themselves are far from simple and can display not only lexical and semantic information, but also syntactic and discursive information.

A bigram approach to register classification has limitations. While this analysis works well at distinguishing disparate registers, it does not seem to discriminate between similar registers (e.g. all the LLC registers were positively marked in one register dimension only). Additional studies should also be conducted that expand the number of registers analyzed and allow for larger observation sizes. Finally, while an approach based on shared bigrams seems successful, it is not an ultimate solution for register classification, but rather should be used in conjunction with other computational methods such as part of speech tagging, syntactic parsing, para-linguistic information, and multimodal behavior (Louwerse et al. 2006). Furthermore, as Biber et al. (2004) state, no single approach to multi-word units in discourse can provide the whole story.

These research findings do, however, fill a much needed gap for creating a register classification system based on lexical and syntactic n-grams as called for by Santini (2004). As much of the work done in register classification does not attempt to explain why registers differ linguistically, but rather how linguistic differences in registers can be used to categorize, an approach to register classification that includes bigram distinctions not only provides essential information about the linguistic framework of both written and spoken genres, but would also allow for more accurate register classification algorithms, information retrieval systems, and assist in automated language processing that depends on probabilistic parsing and input possibilities. This research also furthers work done on lexical bundles and provides additional evidence that multi-word sequences are important linguistic features in the building of discourse and communicative functions.

Notes

[1] For an introduction to factor analysis, see Gorsuch (1983) and Morrison (1990).

[2] A Pearson Correlation is used to measure how closely two variables correlate with one another. They are used in this study to demonstrate the close correlations between the dimensions in the first and second analyses.

References

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

Biber, D. (1992). On the complexity of discourse complexity: a multidimensional analysis. *Discourse Processes, 15,* 133–163.

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

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

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371–405.

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

Burrows, J. (1992). Not unless you ask nicely: The interpretative nexus between analysis and information. *Literary and Linguistic Computing*, 7, 91–109.

Burrows, J. (1987). Word patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing*, *2*, 61–70.

Carletta, J. (1996). Assessing the reliability of subjective coding. *Computational Linguistics*, 22, 249–254.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. & Anderson, A. (1997).

HCRC Dialogue Structure Coding Manual. Technical Report HCRC/TR–82, Human Communication Research Centre, University of Edinburgh.

Conrad, S. & Biber, D. (2001). *Variation in English: Multi- Dimensional Studies*. London: Longman.

Firth, J. R. (1957). Modes of Meaning. In J. R. Firth (Ed.), *Papers in Linguistics* (pp. 190–215). Oxford: Oxford University Press.

Gorsuch, R. L. (1983) *Factor Analysis*. Hillsdale, NJ: Erlbaum

Guadagnoli, E. & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265–275.

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language And Meaning*. London: Edward Arnold.

Hall, C., McCarthy, P. M., Lewis, G. A., Lee, D. S. & McNamara, D. S. (2006). Using Coh-Metrix to Assess Differences between English Language Varieties. *Coyote Papers: Psycholinguistic and Computational Perspectives. University of Arizona Working Papers in Linguistics*, 15, 40–54.

Helt, M. (2001). A multi-dimensional comparison of British and American Spoken English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-Dimensional Studies* (pp. 173–181). London: Longman.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Johansson, S., Leech, G. & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo.

Jurafsky, D. & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Louwerse, M. M. & Crossley, S. A. (2006). Dialog act classification using n-gram algorithms. In G. Sutcliffe & R. Goebel (Eds.), *Proceedings of the 19th International Florida Artificial Intelligence Research Society (FLAIRS)* (pp. 758–763). Menlo Park, CA: AAAI Press.

Louwerse, M. M., Jeuniaux, P., Hoque, M. E., Wu, J. & Lewis, G. (2006). Multimodal communication in computer-mediated map task scenarios. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1717–1722). Mahwah, NJ: Erlbaum.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S. & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner &

T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science* Society (pp. 843–848). Mahwah, NJ: Erlbaum.

MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84–99.

McCarthy, P. M., Lehenbauer, B. M., Hall, C., Duran, N. D., Fujiwara, Y. & McNamara, D. S. (2007). A Coh-Metrix analysis of discourse variation in the texts of Japanese, American, and British Scientists. *Foreign Languages for Specific Purposes*, *6*, 46–77.

Manning, C. D. & Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Morrison, D. F. (1990) *Multivariate Statistical Methods*. New York: McGraw-Hill.

Peng, F., Schuurmans, D. & Wang, S. (2003). Language and task independent text categorization with simple language models. In M. Hearst & M. Ostendorf (Eds.), *HLT-NAACL 2003: Main Proceedings* (pp. 189–196). Edmonton, Alberta, Canada, May 27 — June 1 2003. Association for Computational Linguistics.

Rayson, P., Leech, G. & Hodges, M. (1997). Social differentiation in the use English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, *2*, 133–152.

Santini, M. (2004), *State-of-the-art on Automatic Genre Identification*, Technical Report ITRI–04–03, 2004, ITRI, University of Brighton (UK).

Sikorski, T. & Allen, J. F. (1997) A Task-Based Evaluation of the TRAINS–95 Dialogue System. In E. Maier, M. Mast & S. LuperFoy (Eds.), Dialogue Processing in Spoken Language Systems (pp. 207–220). Berlin: Springer.

Sikorski, T. & Allen, J. F. (1996). A task-based evaluation of the TRAINS–95 dialogue system. In E. Maier, M. Mast & S. Luperfoy (Eds.), *Proceedings of the Workshop on Dialog Processing in Spoken Language Systems, ECAI–96* (pp. 207–220). New York: Springer.

Stamatatos, E., Fakotakis, N. & Kokkinakis, G. (2001). Automatic text categorisation in terms of genre and author. *Computational Linguistics*, 26, 471–495.

Svartvik, J. & Quirk, R. (1980). A Corpus of English Conversation. Lund: CWK Group.

Tabachnick, B. G. & Fidell, L. S. (2001). *Using Multivariate statistics* (4th Edition). Boston: Allyn Bacon.

Corpora

Allen, J. & Heeman, P. A. (1995). *TRAINS Spoken Dialog Corpus* [CD-ROM]. Linguistic Data Consortium.

Du Bois, J. W., Chafe, W. L., Meyer, C. & Thompson, S. A. (2000). *Santa Barbara Corpus of Spoken American English* [CD-ROM]. Linguistic Data Consortium.

Godfrey, J. J. & Holliman, E. (1993). *Switchboard–1* [CD-ROM]. Linguistic Data Consortium.

Human Communication Research Centre (1993). *HCRC Map Task Corpus* [CD-ROM]. Linguistic Data Consortium.

ICAME (1999). *The ICAME Corpus Collection* [CD-ROM]. International Computer Archive of Modern and Medieval English.

Authors' addresses:

Scott Crossley
Department of English
Mississippi State University
P.O. Box E.
Starkville, MS 39759
Phone: (662) 325–2369

Email: scrossley@mail.psyc.memphis.edu

Max Louwerse
Department of Psychology
University of Memphis
Psychology Building
Memphis, TN 38152
Phone: (901) 678–5017

Email: mlouwers@memphis.edu