Biber Redux: Reconsidering Dimensions of Variation in American English

Rebecca J. Passonneau
Center for Computational Learning Systems
Columbia University
New York, New York USA
becky@ccls.columbia.edu

Nancy Ide
Department of Computer Science
Vassar College
Poughkeepsie, New York USA
ide@cs.vassar.edu

Songqiao Su
Department of Computer Science
Columbia University
New York, New York USA
ss4555@columbia.edu

Jesse Stuart
Department of Computer Science
Vassar College
Poughkeepsie, New York USA
jestuart@cs.vassar.edu

Abstract

Genre classification has been found to improve performance in many applications of statistical NLP, including language modeling for spoken language, domain adaptation of statistical parsers, and machine translation. It has also been found to benefit retrieval of spoken or written documents. At its base, however, classification assumes separability. This paper revisits an assumption that genre variation is continuous along multiple dimensions, and an early use of principal component analysis to find these dimensions. Results on a very heterogeneous corpus of post-1990s American English reveal four major dimensions, three of which echo those found in prior work and the fourth depending on features not used in the earlier study. The resulting model can provide a basis for more detailed analysis of sub-genres and the relation between genre and situations of language use, as well as a means to predict distributional properties of new genres.

1 Introduction

Although a precise definition of the term "genre" has traditionally proven to be elusive, it cannot be disputed that a genre represents a set of *shared regularities* among written or spoken documents that enables readers, writers, listeners and speakers to signal discourse function, and that conditions their expectations of linguistic form. Genre distinctions are therefore an important aspect of language use and understanding. They clearly have a role to play in statistical language processing, which relies on regularities of form as well as content. Indeed, with the advent of the Web, statistical methods for genre differentiation have been applied to information retrieval to limit search criteria and organize results (Karlgren and Cutting, 1994; Kessler et al., 1997; Mehler et al., 2010; Ward and Werner, 2013), and the study of genres on the web has become a sub-field in its own right (see for example (Mehler et al., 2010)). More recently, the development of genre-dependent models for a variety of natural language processing (NLP) tasks such as parsing (Ravi et al., 2008; McClosky et al., 2010; Roux et al., 2012), speech recognition (Iyer and Ostendorf, 1999), word sense disambiguation (Martinez and Agirre, 2000), and machine translation (Wang et al., 2012) has been found to significantly improve performance. The ability to match documents by genre has also become important for collecting data to train language models for spoken language understanding, given the difficulty of creating large repositories of transcribed spoken language corpora (Bulyko and Ostendorf, 2003; Sarikaya et al., 2005).

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 565–576, Dublin, Ireland, August 23-29 2014.

While the utility of document characterization by genre for empirical language analysis is widely acknowledged, there is relatively little agreement on methodology. In part, this stems from the difficulty of providing a comprehensive list of genres or even an operational definition of what constitutes a distinct genre, much less a definitive set of features to characterize genre differences. The earliest large-scale statistical study of genre is that of Biber (Biber, 1988), who applied principal component analysis (PCA) to a one-million word corpus consisting of heterogeneous varieties of spoken and written discourse in order to identify multiple dimensions of variation in language. Biber argued that linguistic variation was continuous along six dimensions: involved vs. informational, narrative vs. non-narrative, explicit vs. situation-dependent reference, overt expression of persuasion, abstract vs. non-abstract information, and on-line information elaboration; he identified features associated with each dimension, and characterized kinds of discourse by joint assessment of similarities and differences across these dimensions. Interestingly, since Biber's study, there has been comparatively little investigation of how genres vary using multivariate distributional methods (see, for example, the discussion in (Kilgarriff, 2001)).

Biber's work, which was completed in the mid-1980's, relied on a large number of features extracted using somewhat *ad hoc* methods and reported no reliability measures. Given the renewed interest in genre classification and the increasing interest in automatic techniques to adapt NLP tools across different kinds of corpora, we feel it is worth subjecting Biber's thesis to a new test, utilizing state-of-the-art methods for extracting features from a high quality, very heterogeneous corpus. In addition to replicating Biber's basic approach with more reliable features, we include newer genres (e.g., email, blogs, tweets) in an attempt to verify that these methods can generalize over different kinds of data. We use a smaller feature set that overlaps with Biber's for the most part, but which also includes features unavailable in the earlier work. In our set, each feature was identified using freely available NLP tools and was manually validated. In our use of different features, our experiment constitutes a strong test of Biber's claim that the dimensions of variation he identified arise from underlying constraints on usage. We find three components similar to his, and a new one he did not find, based on our use of Named Entity features. We find that genres that are separable on one component are often co-extensive on another. To quantify the distinctiveness of each of the genres relative to the others, we use a metric that has previously been used to measure separability of classes.

2 Related work and motivation

Our work builds on Biber's 1988 study, but differs in the corpus and features used. Biber's corpus and MASC (Ide et al., 2010), the corpus used in our study, differ in source language (British English versus American English), time coverage (skewed towards a single year versus three decades), and the situations of use. Biber's corpus was drawn from the Lancaster-Oslo-Bergen (LOB) Corpus of British English, consisting of works published in 1961, the London-Lund corpus of spoken English, consisting of 87 texts of British English from private conversation, public interviews and panel discussions, telephone conversations, radio broadcasts, spontaneous speeches and prepared speeches produced in the 1970s. To these Biber added a collection of his own professional and personal letters. MASC represents a larger time slice (1990s to present) and is more heterogeneous, including a wider range of traditional genres as well as new social media (email, blogs, twitter) and collectively generated fiction (ficlets). We take advantage of MASC's rich set of validated annotations to include features that would not have been (easily) available at the time of Biber's study, and reconsider the use of some features used in his work.

Some work on genre classification contrasts with Biber's approach, which assumes that documents fall discretely into distinct classes or clusters. Genre classification has been treated as a standalone task (Karlgren and Cutting, 1994; Kessler et al., 1997; Feldman et al., 2009; Stamatatos et al., 2000a; Santini, 2004), or combined with topic classification (Rauber and Müller-Kögler, 2001; Lee and Myaeng, 2002). All of these studies assume that documents fall discretely into distinct classes or clusters. These studies vary in their approach to determining the genre of text, either by using corpora with pre-defined classes (Karlgren and Cutting, 1994), manually refining pre-existing classes (Kessler et al., 1997), creating genre classes using annotators, or locating *a priori* classifications (e.g., web product reviews). The feature sets in genre studies have remained rather stable over the past three decades, mostly utilizing word-based

features similar to many of Biber's such as individual lexical items and/or their orthographic characteristics (e.g., contractions), part-of-speech (POS), punctuation (Kessler et al., 1997; Stamatatos et al., 2000b), derivative statistics (e.g., average word/sentence length, ratios among lexical or POS classes), and POS-ngrams (Santini, 2004; Feldman et al., 2009).

Karlgren and Cutting (1994) apply discriminant analysis to pre-defined classes from the Brown corpus using easily identifiable information such as POS counts, type/token ratios, and sentence length. They achieve relatively low accuracy of 52%. Kessler et al. (1997) also use the Brown corpus and classify documents into three facets: brow, narrative, and genre. They extract 55 features, avoiding features at the syntactic level that are computationally expensive to identify, and characterize them as lexical, character-level, and derivative (log ratios and their sums). They achieve nearly 80% accuracy on their six *genre* classes (reportage, editorial, scitech, legal, non-fiction, fiction). Feldman et al. (2009) create a corpus of eight genres of speech and web text and test an approach to factor documents by genre, formality and number of speakers. They achieve accuracy of 55% using quadratic discriminant analysis on a representation consisting of features based on POS tags, words, and punctuation, reduced using PCA. Santini (2004) applies high-dimensional POS trigram vectors to ten BBC genres (four spoken, six written) with Naïve Bayes classification. A document representation using a length-835 vector achieves 82.6% accuracy for 10-fold cross-validation on all 10 genres, and a Kappa agreement of 0.80.

Rauber and Müller-Kögler (2001) apply self-organizing maps (Kohonen, 1995) for both topic and genre clustering, using features typical of readability measures (e.g., sentence and word lengths, punctuation frequency). Lee and Myaeng (2002) address classification of web text and also do simultaneous genre and subject (topic) classification, using a Naive Bayes learner. Tests on seven genres for both English and Korean achieve 0.80 micro-averaged f-measure or 0.87 cosine similarity.

More recent work finds good performance from the use of ngram features for words, characters and part-of-speech (Gries et al., 2009; Kanaris and Stamatatos, 2009; Sharoff et al., 2010). Gries et al. (2009) relies only on word ngrams of various lengths to produce clusters with high maximum average silhouette width, where higher widths represent more homogeneous clusters that are more distinct from one another. They find that trigrams do best. Kanaris and Stamatatos (2009) uses frequently occurring character ngrams without regard to their discriminatory power, and Sharoff et al. (2010) find that character ngrams outperform word and pos ngrams. On benchmark corpora with from 4 to 8 genres, the latter two works achieve accuracies of up to 96-97% on some corpora. They assume that genres can be taken as a given, although Sharoff et al. (2010) note that chance-corrected human agreement on the gold standard is only moderate.

Another strand of investigation addresses genre variation as a requirement for achieving better performance in new domains, as in language modeling for speech applications (Bulyko and Ostendorf, 2003; Sarikaya et al., 2005) or statistical parsers applied to text (Ravi et al., 2008; McClosky et al., 2010; Roux et al., 2012), where downstream applications can include assignment of semantic argument structure. Bulyko and Ostendorf (2003) select web text for class-based n-gram language modeling. They locate relevant documents using queries representative of conversational speech, rather than characterizing the documents as a whole in terms of statistical features, but demonstrate a significant reduction in Word Error Rate (WER) for their enhanced language models. Sarikaya et al. (2005) achieve even higher improvements using a similar query methodology, then use BLEU scores, a machine translation similarity method (Papineni et al., 2002), to find sentences that are closest to a domain sample. Ravi et al. (2008) propose a method to predict parser accuracy based on properties of the new domain of interest and properties of the domain on which the parser was trained. Lexical features for words other than the 500 most frequent were found to generalize less well than features such as POS and sentence length. Subsequent work models corpus differences using regression models to predict parser accuracy McClosky et al. (2010), or incorporates explicit genre classifiers Roux et al. (2012).

In our initial exploration of genre variation in MASC, we exploited a set of features that subsume most of those discussed in the works reviewed above. We applied a variety of methods, including k-means clustering, discriminative classifiers such as Naïve Bayes, and PCA. Through comparison of results, we discovered that classification had variable performance, and that PCA provided an explanation: docu-

Genre	Code	No. words	Pct corpus
Court transcript	CT	30052	6%
Debate transcript	DT	32325	6%
Email	EM	27642	6%
Essay	ES	25590	5%
Fiction	FT	31518	6%
Gov't documents	GV	24578	5%
Journal	JO	25635	5%
Letters	LT	23325	5%
Newspaper	NP	23545	5%
Non-fiction	NF	25182	5%
Spoken	SP	25783	5%
Technical	TC	27895	6%
Travel guides	TG	26708	5%
Twitter	TW	24180	5%
Blog	BG	28199	6%
Ficlets	FC	26299	5%
Movie script	MS	28240	6%
Spam	SM	23490	5%
Jokes	JK	26582	5%
TOTAL		506768	

(a) Genre distribution in MASC

Annotation type	No. words
Logical	506659
Token	506659
Sentence	506659
POS/lemma (GATE)	506659
POS (Penn)	506659
Noun chunks	506659
Verb chunks	506659
Named Entities	506659
FrameNet	39160
Penn Treebank	506659
Coreference	506659
Discourse structure*	506659
Opinion	51243
TimeBank	*55599
PropBank	88530
Committed Belief	4614
Event	4614
Dependency treebank	5434

(b) Summary of MASC annotations

Figure 1: Composition of the Manually Annotated Sub-Corpus

ments from distinct classes often fell within an identifiable region on one or more dimensions discovered by PCA, but these regions overlapped one another along other dimensions. We concluded that whether or not a set of documents can be categorized into relatively distinct classes by their linguistic forms rather than content depends on how the documents are selected, how the classes are defined, and what features are used. Our goal here is to refine a method to learn key dimensions of variation relevant for the same types of applications referenced in work on genre identification, as discussed in Section 7.

3 Corpus and data preparation

MASC is a 500,000 word corpus of post 1990s American English comprised of texts from nineteen genres of spoken and written language data in roughly equal amounts, shown in Figure 1a). Roughly 15% of the corpus consists of spoken transcripts, both formal (court and debate) and informal (face-to-face, telephone conversation, etc.); the remaining 85% covers a wide range of written genres, including social media (tweets, blogs). The annotation types and coverage in MASC are given in Figure 1b); all MASC annotations are hand-validated or manually produced. The corpus is fully open and freely available.[1]

To prepare the data, we developed a framework in Groovy[2] (a dialect of Java) to extract linguistic features, using version 1.2.0 of the GrAF API[3] to access the MASC data and annotations. Most texts in MASC comprise complete discourse units, e.g. full conversations, letters, chapters from a book, etc., with the exception of tweets, jokes, and (to some extent) ficlets.[4] As shown in Figure 1a), although each MASC genre contains roughly 25,000 tokens, the number of texts in any given genre varies widely, from as few as two to over 100. To standardize the number of data points per genre, the texts in each genre were concatenated and then divided into samples of even length, rounded to the nearest sentence boundary. Portions of the texts containing email headers, bibliographic references, and computer code, which contain an excess of certain punctuation and other special characters, were eliminated prior to creating the samples.

Initially, we created sample sets consisting of 1,000 tokens per sample,[5] motivated by Biber's observation that even rare linguistic features are relatively stable across samples of this size (Biber, 1993). Our

[1] MASC is downloadable from http://www.anc.org/data/masc and available from the Linguistic Data Consortium (LDC).

[2] http://groovy.codehaus.org

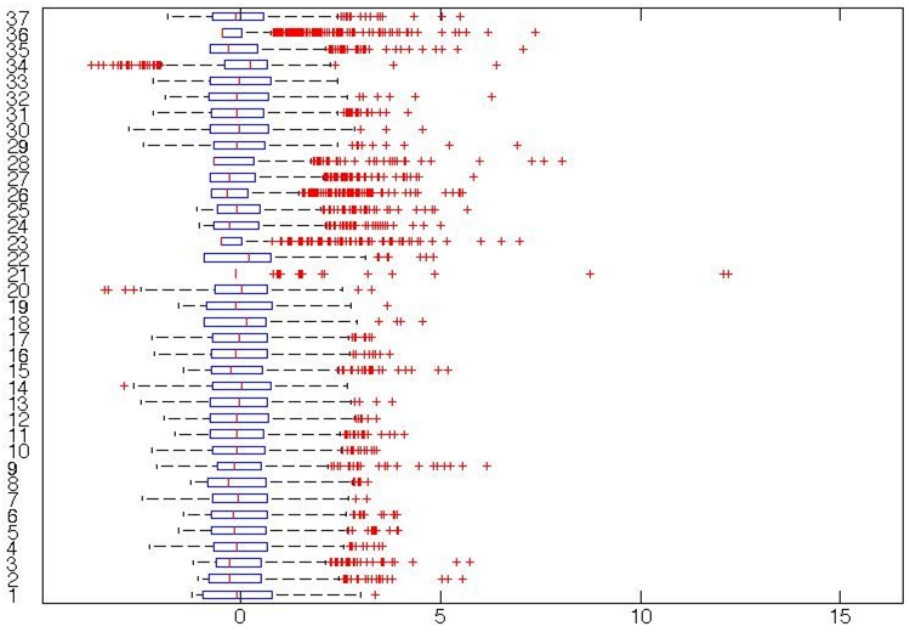
[3] http://sourceforge.net/projects/iso-graf/

[4] Ficlets are story fragments to which "prequels" or "sequels" are added by online participants.

[5] We use tokens as the unit of analysis rather than blank-separated words (strings), which, given the MASC tokenization strategy, means that hyphenated words such as "able-bodied" and possessive markers (*'s*) are treated as individual tokens.

1	1st/2nd person pro.
2	3rd person pro.
3	Pronoun *it*
4	Copula verbs
5	All NEs
6	NEs w/o date
7	Verbs, base
8	Verbs, past
9	Gerunds/Pres. ptp.
10	Past ptp.
11	1st/2nd pres. sg. V
12	3rd pres. sg. V
13	Common nouns
14	All verbs
15	Proper nouns
16	Adjectives
17	Adverbs
18	Superlatives
19	All pers. pro.
20	Prepositions
21	Foreign words
22	Exist. *there*
23	Interjec.
24	NEs, person
25	NEs, date
26	NEs, location
27	NEs, org.
28	Suasive verbs
29	Stative verbs
30	Noun chunk length
31	Verb chunk length
32	Tokens/sentence
33	Characters/token
34	Periods
35	Questions
36	Exclamations
37	Commas

(a) Thirty-seven features



(b) Boxplots of the 37 features: the box shows the range of the 25th to 75th percentiles with the median value identified by the vertical red bar. The black whiskers show the extreme values not considered outliers, and the red are the outliers. The most extreme outliers of feature 21 were dropped to save space.

Figure 2: Feature names and boxplots

experiments showed, however, that for the features used here, results were comparable using 500-token chunks, which enabled us to work with a set of data points of the same size as Biber's. Our process generated 965 500-token chunks, with roughly 50 chunks per genre.

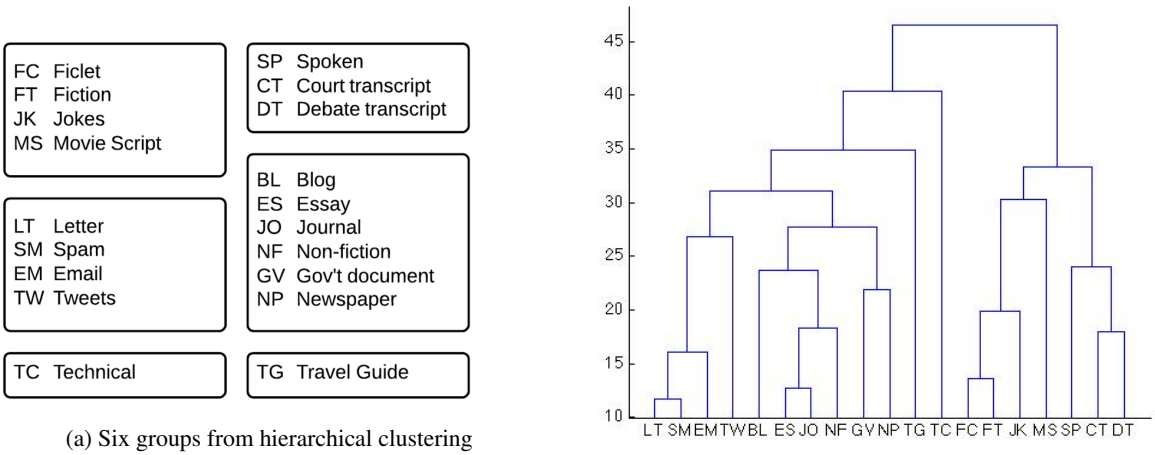
4 Features and feature analysis

Biber used sixty-seven features consisting primarily of lexical items and groups, parts of speech, and quasi-syntactic features such as coordination, negation, relative pronoun deletion, *that*-clauses, and so on. Many of the features in our set overlap with Biber's, but we also exploit annotations in MASC to provide additional features. All the MASC annotations have been manually validated, including those produced by automated tools such as POS-taggers, NE recognizers, and shallow parsers.

PCA is appropriate for data with normally distributed values and can be used to reduce the number of features to include only those that are the least correlated. It highlights features with the greatest variation. Figure 2b) shows boxplots of thirty-seven features we began with. These are mainly frequencies normalized by the total token count in the document samples we created. They also include the average characters per word, and average tokens per sentence, noun chunk, and verb chunk. Figure 2a) lists the features by number. Features 21, 23, 28 and 36, which are foreign words, interjections, suasive verbs and exclamations, have median values (red line within the box) near the 25th percentile, so are highly skewed. We therefore dropped these and carried out the PCA with the remaining thirty-three.[6]

Hierarchical clustering of the dataset by MASC genre yields the dendogram in Figure 3. We used the city block metric (also known as taxicab distance), which is similar to Euclidean distance but less sensitive to outliers. The legend identifies six major clusters for the 19 genres, with two singletons (Travel guides and Technical documents), a cluster with three spoken genres (Court and Debate transcripts, and transcripts of face-to-face and telephone conversations), two four-genre clusters, and one six-genre

[6]To insure comparability of feature influence, all our features were re-scaled in [-1,1] with mean 0.



(a) Six groups from hierarchical clustering

(b) Hierarchical clustering

Figure 3: Hierarchical clustering of 19 MASC genres

cluster. These larger clusters consist of "story-telling" genres (ficlets, fiction, jokes and movie scripts), offline-interactive genres (letters, spam, email and tweets), and discursive text (blog, essay, journal, non-fiction, government documents, and news). Thus the distribution of our features across the data predict groupings that correspond well with our intuitions about the genres defined in MASC, providing some justification for both our feature selection and the genre assignments in the corpus. The groupings also reflect several of Biber's dimensions of variation, as discussed in Section 7.

Here, we describe PCA in general terms to present four principal components identified in our analysis. We focus on features associated with the components, and on the six MASC document clusters.

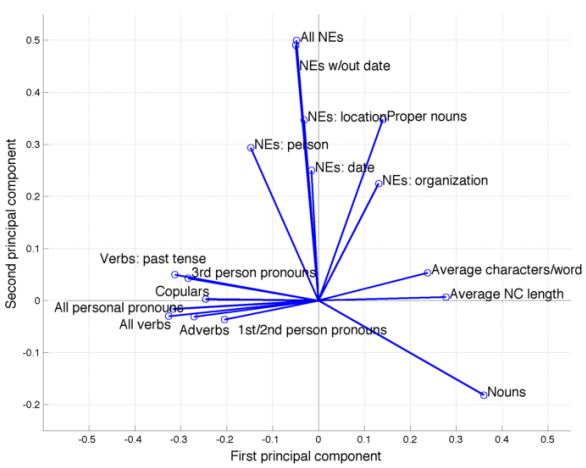
PCA starts with a covariance matrix of all features: a square matrix where each cell value is the covariance of feature x_i with feature x_j for $i, j \in M$. Covariance of x_i, x_j is analogous to variance: for all datapoints $n \in [1 : N]$, you subtract x_{i_n} from $\overline{x_i}$, x_{j_n} from $\overline{x_j}$, sum the products of these differences, and normalize by n-1.[7] A common explanatory visualization will show a scatterplot of hypothetical data values in a sausage shape at a diagonal to the x-axis. A line along the maximum width of the sausage represents the dimension of greatest variation. A second axis can be placed orthogonal to this first component; it will account for less of the variance in the data, and in a different direction. PCA consists of computation of these axes (eigenvectors) from a covariance matrix.

5 PCA results

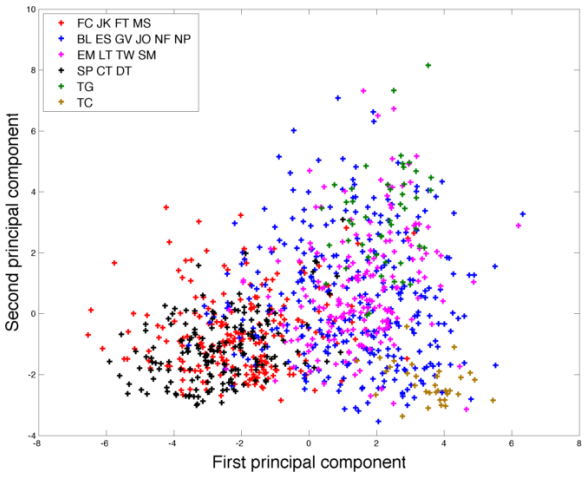
Figure 4a) shows a plot of our first principal component by the second component and the features that contribute most to each, based on the features' loadings (weights) on the new components. The components are rotated to become the new x,y axes and centered at zero. Projection of the individual features onto the rotated axes shows which features contribute most directly to each dimension. Figure 5a) shows a similar plot for the third and fourth components. Twenty-seven features have loadings of at least 0.2 on any component. Many have similar loadings (e.g., commas and prepositions on the fourth component), indicating the data could be represented with fewer, uncorrelated features.

Past tense verbs, copula verbs, personal pronouns, and adverbs load heavily on one pole of the first principal component, while characters per word, noun chunk length and nouns load higher on the opposite pole. This component corresponds rather well to Biber's first component, which had similar loadings for personal pronouns, adverbs, nouns and word length, and which he interpreted as *involved versus informational*–i.e., interactive, unplanned, primarily spoken data vs. polished written documents conveying (sometimes dense) information about a given topic.

[7] See any text on covariance for an explanation of why n-1 is a better normalization term than n.

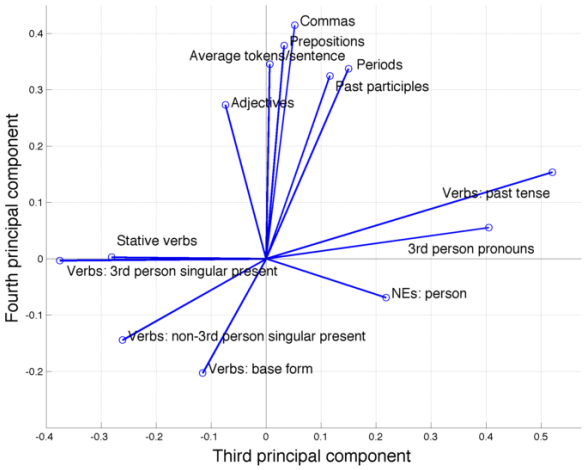


(a) First and second principal components

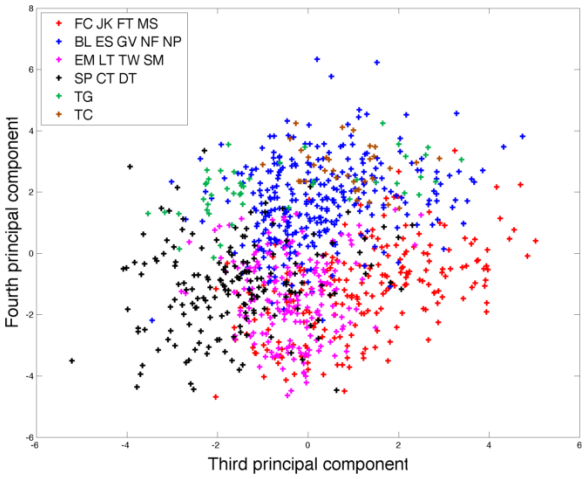


(b) Document regions for components one and two

Figure 4: First and Second Principal Components



(a) Third and fourth principal components



(b) Document regions for components three and four

Figure 5: Third and Fourth Principal Components

Our second principal component is defined almost entirely by the contrast between NEs and common nouns. It corresponds to none of Biber's components; he had no NE features. Our third component has loadings from 3rd person present tense verbs (and other verb forms) at one end, and past tense verbs, third person pronouns, and person NEs at the other. It corresponds to Biber's second component, which had similar loadings for past tense verbs and third person pronouns, and somewhat less for present tense verbs. He interpreted this dimension as representing the variation from non-narrative to narrative.

Our fourth component corresponds to Biber's fifth, which he characterized as abstract versus non-abstract. At one extreme we have commas, prepositions, sentence length (in tokens) and past participles, with base verbs loading to some degree on the other extreme. The features loaded on Biber's fifth component were conjuncts, which might correlate with longer sentence length, past participles, and agentless passives. In the corresponding scatterplots (Figures 4b and 5b), each datapoint (document chunk) has been color-coded according to the six clusters found in the preceding section. There are clearly distinct regions along the first component for spoken interactions (black), story telling (red), offline interaction (pink) and discursive (blue), but with a great deal of overlap. Travel guides (green) and technical (gold) are at the blue extreme, but at different locations along the second dimension. Moving from left to right in Figure 4b), each next color has greater dispersion along the second component, apart from green and gold, which have clearly separate locations from each other, at the top and bottom,

	Story telling	Discursive	Offline Interaction	Spoken Interaction	Travel Guide	Technical
Story Telling	0.00	0.23	0.13	<u>0.06</u>	0.63	0.91
Discursive	0.23	0.00	0.21	0.24	0.15	0.35
Offline Interaction	0.13	0.21	0.00	0.07	0.57	**1.07**
Spoken Interaction	<u>0.06</u>	0.24	0.07	0.00	0.68	0.88
Travel Guide	0.63	0.15	0.57	0.68	0.00	0.78
Technical	0.91	0.35	**1.07**	0.88	0.78	0.00

Table 1: Mean Bhattacharyya Distance of all Genre Pairs using PCA Scores

respectively. In Figure 5b), the overall dispersion is more even across both dimensions, with separate centers for each of the four major colors (black, pink, red and blue), but again without sharp separation.

6 Genre Distance Measurement

A metric that summarizes how separable a pair of genres are in the defined PCA space would be more convenient than the visualizations in Figures 4b and 5b. Bhattacharyya distance, which measures the similarity of two discrete or continuous probability distributions, has been used in image segmentation and signal selection, to minimize the probability of misclustering for segmentations (Coleman and Andrews, 1979), or the probability of misclassifying different signals (Kailath, 1967). Here we illustrate its use in summarizing the separability of a pair of genres across the four principal components.

In statistics, the Bhattacharyya distance measures the similarity of two discrete or continuous probability distributions. It is closely related to the Bhattacharyya coefficient, which measures the amount of overlap between two statistical samples or populations.

The Bhattacharyya coefficient for two continuous probability distributions $p(x)$ and $q(x)$ is:

$$\text{Bhattacharyya coefficient} = \rho = \int_C \sqrt{q(x)p(x)}dx$$

Where C is the domain of probability density $p(x)$ and $q(x)$. The Bhattacharyya coefficient takes on values in [0,1]. Bhattacharyya distance maps the Bhattacharyya coefficient to [0,∞]:

$$\text{Bhattacharyya distance} = B = -\ln \rho$$

We take the mean Bhattacharyya Distance of a pair of genres across all four components as a summary measure of seprability. As an illustration, consider the two clusters of offline interaction (pink) and discursive text (blue) from Figures 4b) and 5b). Their Bhattacharyya Distances on the first through fourth components, using the PCA scores, are: 0.05, 0.01, 0.14, 0.63. They have the largest distance on the fourth component, the axis of abstract vs non-abstract, which is consistent with the visualizations. The summary statistic is then the mean of the four individual distances: 0.21.

Table 1 gives the mean Bhattacharyya Distance of each pair of genres for the four components. The pair of genres that is the closest on all four components is story telling and spoken interaction (0.06; underlined). The pair that is the most distant on all four components is technical and offline interaction (1.07; in bold). Bhattacharyya Distance can also be computed for each pair of genres using the original normalized feature values. In three cases the Bhattacharyya Distance in the PCA space is the same as in the original feature space, but in all other cases the Bhattacharyya Distance is much greater.

7 Discussion

Strong patterns of similarity in dimensions of variation across many genres of English emerge from our comparison with Biber's study, despite differences in the features used, the contrast between American and British English, and the use of new media types. The results support the view that relatively stable dimensions of variation arise from properties of the situations of use across varieties of English. This applies as well to genres that did not exist in Biber's time (email, twitter, spam), which group with the interactive genre included in Biber's corpus (letters) and are similar to other offline discourse despite representing an interactive form–albeit an "offline interactive" form–of discourse.

A significant departure from Biber's results concerns the component defined primarily by Named Entities (NEs), which emerges as the second strongest dimension of variation in our study. This demonstrates that additional features–in particular, features beyond those based on orthographic and morpho-syntactic properties that have figured in most genre studies to date–can dramatically impact Biber's original model and extend the range of properties that can characterize particular text types. It also suggests that higher-level linguistic properties and other more complex features can contribute substantially to genre characterization and discrimination, a topic we plan to pursue in the future.

In what follows, we discuss similarities and differences in the two PCA analyses, the conclusions this leads to regarding the feasibility of genre classification, and ways in which the analysis can support retrieval, language modeling, and domain adaptation.

Our first principal component is very similar to Biber's first factor, which he interpreted as differentiating situations of use with more of an informational focus from those with an interactive or affective function. In addition, he noted a contrast between *online* and *offline* production–i.e., spoken vs. written production modes. The heavily loaded features the two analyses have in common are consistent with the interpretation: 1st/2nd person pronouns, many verb features, and adverbs are at one pole, with word length and nouns at the other. He claimed that this distinction *is obviously a very powerful factor . . . not an artifact of the factor extraction technique*, meaning that it arises from differences between the demands of face-to-face, online interaction and those of offline, expository discourse. Having found a very similar dimension using different (correlated) features, we agree with this claim. Figure 4b) shows that the spoken interaction documents in MASC fall on the "involved" side of this dimension, while expository texts fall on the "informational" side.

Interestingly, the genres that did not exist in Biber's time (email, twitter, spam) group with the interactive genre included in Biber's corpus (letters), and they are similar to other offline discourse despite representing an interactive form–albeit an "offline interactive" form–of discourse. This provides a strong argument for the validity of the first component and its link to underlying situational factors of language use. In Figure 4b), the hypothetical centroid of the pink (offline interactive) region seems somewhat less to the right on the x-axis than a corresponding centroid for the blue (expository) set, but the pink and blue are relatively co-extensive, and in particular, are clearly separated from both the black (face-to-face online interaction) and red (storytelling) genres. This makes intuitive sense, as storytelling genres often depict face-to-face interaction ("so the elephant says to the camel"), and therefore mimic its immediacy.

Our second principal component is defined primarily by Named Entities (NEs), which has no correlate in Biber's study; his features included proper nouns but not NEs. Person NEs load with past tense verbs and third person pronouns on our third component, which resembles Biber's narrative dimension. Most of the MASC genres seem to be dispersed all along our second dimension, suggesting that NE frequency varies across texts in these genres; the exception is travel guides, which consistently include larger numbers of NEs. The explanation here is less on production constraints than on function, as travel guides survey geographical points of interest, historical monuments and persons, hotels and restaurants, and so on.

As noted in Section 5, our third component is very similar to Biber's second (narrative versus non-narrative), and our fourth is somewhat similar to Biber's fifth (abstract versus non-abstract). Note that the fourth dimension shows a greater separation of expository (blue) and offline-interactive (pink) genres, which substantially overlap on the first dimension. This provides a good example of how the 4-dimensional visualization provided by the scatterplots reveals potentially very different relations among genres across the components, which in turn explains why fixed definitions of genre are difficult, if not impossible, and why genre classification can be hard to achieve. We observe that the genre classes can be more or less separable on one dimension but not another. As another example, travel guides and technical documents are at distinct locations on the second component, but span the same locations on the first.

This lack of separability on one or more dimensions is true for nearly all pairs of our six genre classes, as well as for any pair of dimensions. This suggests that an application that requires genre classification could use PCA to find dimensions of variation that lead to the best separation, and summarize the separability using the mean Bhattacharyya distance. As the number of genres one needs to classify increases,

it could be that the number of orthogonal dimensions required to lead to the best separation might also increase. In Table 1, for example, with the exception of the row for Discursive Text, all rows have at least one cell with a value close to or above 0.80, indicating that each of the six genres can be clearly separated from at least one other genre. We would predict that Discursive Text would be the most difficult to classify using genre features alone.

The strong similarities among the major components in Biber's study and ours support the view that genre variation is continuous along multiple dimensions due to contextual properties such as cognitive constraints, interactivity, and function. As such, we view the dimensions as arising from observable properties of discourse situations. Given a new genre, it should be possible to predict where it would be located in the PCA space defined here. We would predict that chats, for example, would pattern more closely with face-to-face interaction than with offline interactive genres. The same methodology could be applied to a sub-genre, such as the discursive texts, to discover more specific dimensions to differentiate among them.

Because language use changes over time, and new genres arise, we do not view the 4-dimensions as a definitive representation of genre space. We do, however, envision a concrete application of this particular representation, namely to measure corpus similarity in a multivariate fashion. Because our PCA analysis makes it possible to locate new documents in the defined space, it would be possible to identify which MASC documents a new set of documents is most similar to. PCA scores could be computed on the four dimensions for corresponding features in the new documents. This approach could be used in any application where it is desirable to find similar documents, such as retrieval, language modeling, or domain adaptation. For example, in recent work on domain adaptation of parsers, McClosky et al. (2010) present a confusion matrix with six corpora to demonstrate how performance of a Charniak parser (Charniak, 2000) varies depending on which corpus it is trained on. They assume that a new target domain will be a mixture of their six source domains and build a simple regression (three features) to predict which of the six parsers will perform best on a new corpus. They subsequently state that an alternative approach could use a high-dimensional vector space to compare corpora. Inspired by this suggestion, we are currently developing a web service that will allow researchers to locate their corpora in the 4-dimensional space identified in this study, and to compute the values of their PCA scores. This would make it possible to use Bhattacharrya distance as described in Section 6 to measure the similarity of corpora in genre space, which could be quite relevant for adapting parsers or other NLP tools. This contrasts with the similarity measures used in Ravi and Knight (Ravi et al., 2008) and McClosky (McClosky et al., 2010), which are based on lexical features.

8 Conclusion

Using a relatively small set of under three dozen features to represent the linguistic forms in discourse, PCA reveals four principal components of variation in a very heterogeneous corpus of post 1990s American English that are comparable to those identified in Biber's work, as well as additional dimensions based on features not included in that earlier study. Six genres derived from the MASC corpus using hierarchical clustering are separable on some but not all components. These differences in separability potentially explain the variations in performance across different works that do genre classification. The resulting 4-dimensional genre space provides a basis for more detailed analysis of sub-genres, for a better understanding of the relation between genre and situations of language use, and for predicting the distributional properties of new genres. In future work, we plan to build on this basis to develop an increasingly detailed and, at the same time, generalizable characterization of genre.

Our results depict a *big picture* for how discourse in English varies with respect to style or form, and how different genres are conditioned by aspects of the situations of language use. We believe that exploration of genre in these terms can provide a more viable approach to measuring distinctions among texts than the approach used in most recent work, and can provide a more informed basis to incorporate genre distinctions in information retrieval, language modeling, and domain adaptation for statistical NLP.

Acknowledgements

This work was supported in part by NSF CRI-1059312.

References

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

Douglas Biber. 1993. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, 26:331–345.

Ivan Bulyko and Mari Ostendorf. 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Proc. HLT-NAACL 2003*, pages 7–9.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, NAACL 2000, pages 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guy Barrett Coleman and Harry C Andrews. 1979. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5):773–785.

Sergey Feldman, Marius Marin, Julie Medero, and Mari Ostendorf. 2009. Classifying factored genres with part-of-speech histograms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 173–176, Boulder, Colorado, June. Association for Computational Linguistics.

Stefan Th. Gries, John Newman, Cyrus Shaoul, and Philip Dilts. 2009. N-grams and the clustering of genres. Paper presented at the workshop on Corpus, Colligation, Register Variation at the 31st Annual Meeting of the Deutsche Gesellschaft fr Sprachwissenschaft.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Rebecca Passonneau. 2010. The Manually Annotated Sub-Corpus: A Community Resource for and by the People. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 68–73, Uppsala, Sweden, July. Association for Computational Linguistics.

Rukmini Iyer and Mari Ostendorf. 1999. Relevance weighting for combining multi-domain data for n-gram language modeling. *Computer Speech & Language*, 13(3):267–282.

Thomas Kailath. 1967. The divergence and Bhattacharyya distance measures in signal selection. *Communication Technology, IEEE Transactions on*, 15(1):52–60.

Ioannis Kanaris and Efstathios Stamatatos. 2009. Learning to recognize webpage genres. *Information Processing and Management*, 45(5):499–512, September.

Jussi Karlgren and Douglass Cutting. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94, pages 1071–1075, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, pages 32–38, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):1–37.

Teuvo Kohonen. 1995. *Self-organizing Maps*. Springer-Verlag, Berlin.

Yong-Bae Lee and Sung Hyon Myaeng. 2002. Text genre classification with genre-revealing and subject-revealing features. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150, New York, NY, USA. ACM Press.

David Martinez and Eneko Agirre. 2000. One sense per collocation and genre/topic variations. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 207–215, Stroudsburg, PA, USA. Association for Computational Linguistics.

David McClosky, Eugene Charniak, and Mark Johnson. 2010. Automatic domain adaptation for parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Mehler, S. Sharoff, and M. Santini. 2010. *Genres on the Web: Computational Models and Empirical Studies*. Text, Speech and Language Technology. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Andreas Rauber and Alexander Müller-Kögler. 2001. Integrating automatic genre analysis into digital libraries. In *First ACM-IEEE Joint Conference on Digital Libraries*, pages 1–10.

Sujith Ravi, Kevin Knight, and Radu Soricut. 2008. Automatic prediction of parser accuracy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Honolulu, Hawaii, October. Association for Computational Linguistics.

Joseph Le Roux, Jennifer Foster, Joachim Wagner, Rasul Samad, Zadeh Kaljahi, and Anton Bryl. 2012. DUC-Paris13 systems for the SANCL 2012 shared task.

Marina Santini. 2004. A shallow approach to syntactic feature extraction for genre classification. Technical Report ITRI-04-02, Information Technology Research Institute, University of Brighton. Also published in Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, Birmingham, UK.

Ruhi Sarikaya, Agustín Gravano, and Yuqing Gao. 2005. Rapid language model development using external resources for new spoken dialog domains. In *International Congress of Acoustics, Speech, and Signal Processing (ICASSP)*, pages 573–576, Philadelphia, PA, USA. IEEE, Signal Processing Society.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of Babel: evaluating genre collections. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2000a. Text genre detection using common word frequencies. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, pages 808–814, Stroudsburg, PA, USA. Association for Computational Linguistics.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000b. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495, December.

Wei Wang, Klaus Macherey, Wolfgang Macherey, Franz Och, and Peng Xu. 2012. Improved domain adaptation for statistical machine translation. In *AMTA-2012*.

Nigel G. Ward and Steven D. Werner. 2013. Using dialog-activity similarity for spoken information retrieval. In Frédéric Bimbot, Christophe Cerisara, Cécile Fougeron, Guillaume Gravier, Lori Lamel, François Pellegrino, and Pascal Perrier, editors, *14th Annual Conference of the International Speech Communication Association, Interspeech*, pages 1569–1573. ISCA.