**Biber Redux: Reconsidering Dimensions of Variation in American English**

Rebecca J. Passonneau  
Songqiao Su  
Columbia University

Nancy Ide  
Jesse Stuart  
Vassar College

### Genre
- A genre consists of shared regularities of form among written or spoken documents
  - Narrative
  - Personal email
  - Personal letters
  - Encyclopedia articles
  - . . .
- Genre is orthogonal to topic
- Genres are an artifact of the situation of language use

### Genre Variation
- Genre classification: accuracy varies widely
- Biber’s (1988) claim: Genre variation is continuous along multiple dimensions
- Based on Principle Component Analysis of a heterogeneous 1M word corpus of British English, from a narrow time slice

### Features with High Loadings on 1st/2nd Components

#### Plot of PCA scores using 1st/2nd Components as Axes

- categorically
- Continuous
- 

### Features with High Loadings on 3rd/4th Components

#### Plot of PCA scores using 3rd/4th Components as Axes

- categorically
- Continuous
- 

**Biber’s 6 components**
1. Involved vs. Informational  
2. Narrative vs. Non-Narrative  
3. Explicit vs. Situation-dependent reference  
4. Overt persuasion  
5. Abstract vs. non-abstract information  
6. On-line information elaboration

**Our 4 components**
1. Involved vs. Informational (= Biber’s #1)  
2. Naming (Not in Biber’s features or analysis)  
3. Narrative vs. Non-Narrative (= Biber’s #2)  
4. Abstract vs. non-abstract information (=Biber’s #5)

**Strong Test of Biber’s claim**
- There are underlying dimensions of variation based on situations of use
- We had a different corpus with a different variety of English and different genres
- We used a different (overlapping) set of features among the many possible that co-vary

**Comparison**
- Similarities: 3 of the same components, same rank

**Differences**
- We found an additional rank 2 component due to Named Entity features Biber did not have
- We had few suasive verbs or other indicators of persuasion, so no analog to Biber’s 4th component

**Contributions**
- Verification of Biber’s basic approach on a different corpus, more reliable features
- Discovery of new, second ranked component based on Named Entities
- Visualization
- Conclusion that no single set of features can separate (clearly) more than a few genres
- Use of Bhattacharyya Distance to measure genre distinctness

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