Factors of Formality: A Dimension of Register in a Sociolinguistic Corpus

1. Introduction

Our Goals
• Evaluate a formality lexicon based on co-occurrence (Brooke et al., 2010)
• Investigate a sociolinguistic corpus (Tagliamonte, 2006b)

Questions
• Is our formality lexicon, derived from writing, applicable to speech?
• Is formality in speech indicative of underlying social factors?

Background: Quantification Approaches to Stylistic Variation
• Multidimensional analysis of register (Biber 1988)
• Variationist sociolinguistics (Labov 1972; Tagliamonte, 2006a)
• Lexicalized computational stylistics (Argamon et al., 2007)
• Relevant tasks in NLP (Garrera and Yarowsky, 2009; Peterson et al., 2011)

2. Building a Lexicon of Formality

Idea
• Assign every word a number indicating its level of formality
• Use corpus co-occurrence starting from a small set of seeds
• Inspired by methods for sentiment lexicons (Turney and Littman, 2003)

Seed sets
• 138 informal, slang (e.g. wuss) and interjections (e.g. yikes)
• 105 formal, discourse cues (e.g. hence) and adverbs (e.g. adroitly)

Corpus
• ICWSM Spin3r Dataset (Burton et al. 2009)
  • Mixed register
  • 7.5 million blogs
  • 1.3 billion word tokens
  • Filtering of rare words and short documents

Latent Semantic Analysis (Landauer and Dumais 1997)
• Similar to factor analysis as used for MD analysis (Biber 1988)
• Create word–document matrix
• Collapse word–document matrix to k dimensions
• For each word vector, calculate cosine similarity to seed words

Cosine similarity in two dimensions

3. Previous Experiments

Brooke et al., 2010
• Over 80% accuracy on near-synonym relative formality task
• Leave-one-out testing with seed words give nearly perfect accuracy
• LSA method better than word length and frequency-based metrics

Brooke et al., 2011
• Lexicon applied to word choice (prediction of clipping, e.g. doc/doctor)
• Results similar to both word choice system and human performance

4. The Toronto Corpus

• 135 transcribed interviews with Toronto residents (Tagliamonte, 2006b)
• Collected by Sali Tagliamonte and colleagues between 2002 and 2004
• (Now) machine readable, automatically part-of-speech tagged
• Marked with social factors:
  • Age (9–85)
  • Work (blue collar, white collar, or student)
  • Gender

5. Formality in the Toronto Corpus

Method
• Calculate a formality score for each text in Toronto Corpus
  • Average formality of all words in the text
  • Divide texts into groups by social factors
  • Calculate averages and significance (t-test)

Results for Age
• Formality increases with age
• Young and old significantly different (p < 0.001)
• Children and young adults significantly different (p < 0.01)
• Key words: like, yeah, just, stuff, okay, weird

Results for Work
• Students omitted
• White collar workers more formal (p < 0.001)
• Key words: gotta, stuff, guy, very, were

Results for Gender
• Women are slightly less formal
• Difference not significant
• Men say: gonna; women say: oh-my-god

Discussion
• Results correspond to our intuitions
• But which came first: the style, or the social group?

References and Acknowledgments


This work was supported by the Natural Sciences and Engineering Research Council of Canada. Many thanks to Sali Tagliamonte for the use of the Toronto Corpus.