Julian Brooke and Graeme Hirst
jbrooke, gh@cs.toronto.edu Department of Computer Science, University of Toronto

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?
- Will the direction of formality differences correspond to our intuitions?

Background: Quantificational Approaches to Stylistic Variation
- Multidimensional analysis of register (Biber 1988)
- Variationist sociolinguisitics (Labov 1972; Tagliamonte, 2006a)
- Lexicalized compututational stylistics (Argamon et al., 2007)
- Relevant tasks in NLP (Garera and Yarowsky, 2009; Peterson et al., 2011)

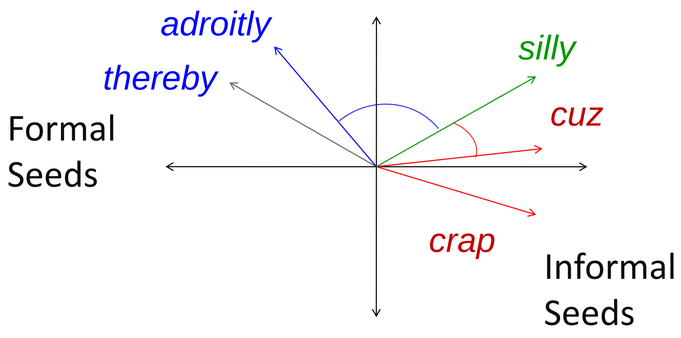
2. Building a Lexicon of Formality

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 Spinn3r 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

Normalization
- Normalize to –1 to 1 range
- –1 is most informal, 1 is most formal
- Core vocabulary generally near zero
  - Neutral word *and* is taken as absolute zero

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

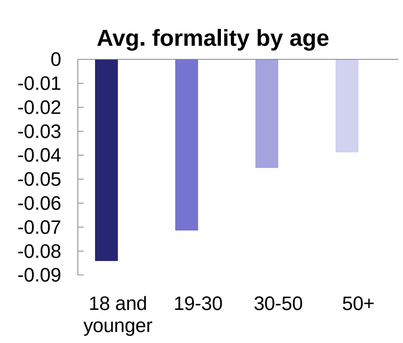
Brooke et al., 2011
- Lexicon applied to word choice (prediction of clipping, e.g. *doc/doctor)*
- Results similar to both word choice system and human performance

4. The Toronto Corpus

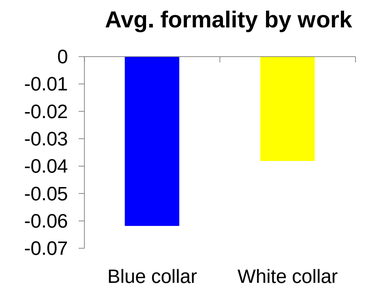
- 135 transcribed interviews with Toronto residents (Tagliamonte, 2006b)
- Collected by Sali Tagliamonte and colleagues between 2002 and 2004
- (Now) machine readable, automatically part-of-speech tagged
- Marked with social factors:
  - Age (9-85)
  - Work (blue collar, white collar, or student)
  -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

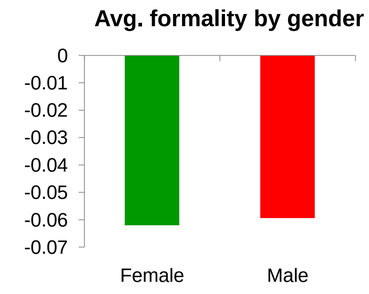


Results for Age
- Formality increases with age
- Young and old significantly different ($p < 0.001$)
- Children and young adults significantly different ($p < 0.01$)
- Key words: *like*, *yeah*, *just*, *stuff*, *okay, weird*



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

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 7:91–109.
Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press.
Brooke, Julian, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10).*
Brooke, Julian, Tong Wang, and Graeme Hirst. 2011. Clipping Prediction with latent semantic analysis. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP '11).*
Burton, Kevin, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. *In Proceedings of the Third Annual Conference on Weblogs and Social Media* (ICWSM '09)
Garera, Nikesh and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09)*.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia, University of Pennsylvania Press.
Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
Peterson, Kelly, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*.
Tagliamonte, Sali. 2006a. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Tagliamonte, Sali. 2006b. "So cool, right?": Canadian English entering the 21st century. *Canadian Journal of Linguistics*, 51:309–331.
Turney, Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

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

