A data-driven approach to stylistic identification

LSA 2015

Christopher Ahern    Meredith Tamminga

University of Pennsylvania

January 8, 2015
Outline

1. Introduction
2. Data & methods
3. Results
4. Conclusions
Quantitative approaches to style

Top-down approaches
- Look for correlation of variation with pre-determined external factors
- Emphasis on topic, context, interlocutor, etc. as determining style

Bottom-up approaches
- Look for clustering of variables
- Emphasis on ordered sequences of observations
"As you listen to speech, set aside the first utterance of every response to yourself; then take every personal narrative and put it into the Casual speech bin; otherwise, exclude any discussion of language. Any group discussion not about language is Casual. Look for extended, long-winded general pronouncements and exclude them as Careful. Mark any discussion of kids’ affairs, from their own point of view as Casual, and include any sizeable excursion of the speaker into a different topic. Otherwise, interview speech is classed as Careful speech."
A top-down approach: Labov 2001

Image from Labov 2001: 104
New perspectives on style

- Increasing interest in how stylistic shifts take place in real time
- Attention to ordered sequences of observations
- See e.g. Podesva 2007, Sharma & Rampton 2011, Tamminga 2014
A bottom-up approach: Podesva 2007

Image excerpted from Podesva 2007 Fig. 2
"It should be noted that clusters present themselves visually, and that as yet I have not developed a mathematical method for isolating clusters; doing so would be a useful direction for future work."
New bottom-up approaches to style

- Begin from the data itself
- Make style "fall out" of the data
- Automatically infer underlying states
New bottom-up approaches to style

Potential Advantages

- Automation: less labor-intensive, can apply to large datasets
- Statistical validity: not reliant on subjective judgment
- Temporal sensitivity: more adaptable to dynamic questions
The data

/dh/-stopping

- Philadelphia Neighborhood Corpus (Labov & Rosenfelder 2011)
- Sociolinguistic interviews with 42 white English speakers from the PNC
- 18,022 observations of /dh/-stopping coded auditorily in Praat
The data

/dh/-stopping

(‘them’) ~ (‘dem’)
/ð/, /dð/ ~ /d/, /ɾ/
The data

"So I went with them to the park..."
The data

"So I went with 0 to 1 park..."
The data

1101000110010111011001101100110111110000011101111101111111...

Ahern, Tamminga (UPENN)
The data

Image excerpted from Podesva 2007 Fig. 2
The data

110100011001011101100110110011011111000011101111101111111...
Hidden Markov Model
Hidden Markov Model

**Parameters**

- **Starting States**: Probability of starting in a particular state
- **Transition Matrix**: Probability of transitioning from one state to another
- **Emission Matrix**: Probability of emitting a signal in a given state
Hidden Markov Model

**Procedure (R: HMM)**

- Initialize HMM: Specify starting transition and emission matrices
- Baum-Welch Algorithm: Fit parameters given the data
- Viterbi Decoding: Find most likely path through hidden states
- Model Comparison: Tradeoff between fit and complexity of model
Comparison
Differences
Main results

- Fit HMMs, found most likely path, compared models
- Two-state model fits most speakers best
Main results

![Graph showing stylistic identification results](Image from Labov 2001: 104)
Conclusions

- Pursue a bottom-up approach to identifying stylistic shifts
- Automatically extracted stylistic clusters from observed data
- Inferred state sequences can be used as input to a qualitative investigation
Future directions

Image excerpted from Podesva 2007 Fig. 2
Thanks!
Thanks!

- Students and Faculty at Penn, particularly Aaron Ecay
- Conference organizers
- Audience
References I

- Eckert (2003) The meaning of style
- Labov (2001) The anatomy of style shifting
- Podesva (2007) Three sources of stylistic meaning
- Sharma & Rampton (2011) Lectal focusing in interaction: A new methodology for the study of superdiverse speech
- Tamminga (2014) Persistence in the production of linguistic variation