ON THE IMPORTANCE OF STANDARDIZED WORD LIST DATA IN DIALECTOLOGY RESEARCH

Douglas S. Bigham
University of Texas at Austin
douglas.s.bigham@gmail.com

Methods In Dialectology XIII – Leeds, United Kingdom
4th August, 2008
OVERVIEW

- **Dialect Contact**
  - What happens when speakers of two different dialects interact with each other?
  - Trudgill, 1986; 2004

- **University Students**
  - Southern Illinois University – Carbondale (SIUC)
  - Close, persistent, intimate contact
  - Transient, dynamic, “anchored” populations

- **Emerging Adulthood**
  - Period between High School and a Career
  - Roughly age 18-26
  - Marked by exploration, self-discovery, and transience
  - J. Arnett, 2001

- **Vowels of Interest:**
  - LOT, THOUGHT, and TRAP
DIALECT GEOGRAPHY FOR THIS STUDY
DATA TO BE PRESENTED

- LOT, THOUGHT, and TRAP vowels
  - Wells’ Key Words (Wells, 1982)
- Eight speakers sampled (from 126 speaker corpus)
  - 4 Chicagoland / 4 Southern Illinois
- “Natural” speech data
  - Tokens taken from casual interviews
  - Interviews averaged 1.5 hours per speaker
- “Word List” recitation data
  - Peterson & Barney (1952) repetition (P&B-task)
  - 11 vowels of American English in h_d & b_t contexts
  - Each reading repeated 5 times
The “Stylistic Continuum” and Problems with “Natural” Speech

- The Stylistic Continuum
  - Capturing non-self-aware speech

```
Stylistic Continuum

Casual Speech   Careful Speech   Reading Task   Word List Task   Minimal Pairs   P&B Replication

less attention  more attention
```

- “Style” and “Audience Design” (Bell, 1984)
- “Awareness” and “Formality” (Schilling-Estes, 1998)
- Interviews as situated speech (Fuller, 2000)
- Variation as “Indexical” (Eckert, 2000)
VIRTUES OF P&B-TYPE DATA

- P&B-type data support interview-based data
  - Hagiwara (1997) – “Californian” features in California
  - Clopper, et al. (2005) – regional features confirmed for multiple American English dialects

- Practical Considerations
  - Easier to collect
  - Easier to analyze
  - Provide a uniform comparison against a known “benchmark”
  - Provide uniform comparability across tokens & speakers
COMPARABILITY OF DATA: OVERALL TOKEN FREQUENCY

Comparable token counts
- Often lacking in “interview speech”
- e.g., FOOT is especially uncommon in my data

<table>
<thead>
<tr>
<th></th>
<th>LOT</th>
<th>TRAP</th>
<th>DRESS</th>
<th>THOUGHT</th>
<th>GOAT</th>
<th>GOOSE</th>
<th>FOOT</th>
<th>STRUT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interview</td>
<td>17%</td>
<td>16%</td>
<td>14%</td>
<td>19%</td>
<td>7%</td>
<td>12%</td>
<td>9%</td>
<td>6%</td>
</tr>
<tr>
<td>WordList</td>
<td>12%</td>
<td>14%</td>
<td>14%</td>
<td>12%</td>
<td>12%</td>
<td>13%</td>
<td>13%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Word List Data, while not perfect, show a much greater comparability across tokens
This problem is exacerbated when consonantantal contexts is considered
Comparability of Speakers’ Data: Following Consonantal Context

<table>
<thead>
<tr>
<th></th>
<th>LOT</th>
<th>TRAP</th>
<th>DRESS</th>
<th>THOUGHT</th>
<th>GOAT</th>
<th>GOOSE</th>
<th>FOOT</th>
<th>STRUT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TOTAL FREQUENCY</strong></td>
<td>17%</td>
<td>16%</td>
<td>14%</td>
<td>19%</td>
<td>7%</td>
<td>12%</td>
<td>9%</td>
<td>6%</td>
</tr>
<tr>
<td><strong>LABIAL</strong></td>
<td>40%</td>
<td>8%</td>
<td>15%</td>
<td>26%</td>
<td>12%</td>
<td>30%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td><strong>ALVEOLAR</strong></td>
<td>36%</td>
<td>71%</td>
<td>79%</td>
<td>57%</td>
<td>35%</td>
<td>43%</td>
<td>76%</td>
<td>53%</td>
</tr>
<tr>
<td><strong>VELAR</strong></td>
<td>24%</td>
<td>21%</td>
<td>6%</td>
<td>11%</td>
<td>6%</td>
<td></td>
<td>24%</td>
<td>7%</td>
</tr>
<tr>
<td><strong>WORD FINAL</strong></td>
<td></td>
<td></td>
<td>6%</td>
<td>47%</td>
<td>27%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VOICED</strong></td>
<td>64%</td>
<td>39%</td>
<td>21%</td>
<td>52%</td>
<td>44%</td>
<td>91%</td>
<td>57%</td>
<td>47%</td>
</tr>
<tr>
<td><strong>VOICELESS</strong></td>
<td>36%</td>
<td>61%</td>
<td>79%</td>
<td>48%</td>
<td>63%</td>
<td>9%</td>
<td>43%</td>
<td>53%</td>
</tr>
<tr>
<td><strong>PLOSIVE</strong></td>
<td>67%</td>
<td>45%</td>
<td>64%</td>
<td>15%</td>
<td>29%</td>
<td>20%</td>
<td>95%</td>
<td>40%</td>
</tr>
<tr>
<td><strong>FRICATIVE</strong></td>
<td>2%</td>
<td>34%</td>
<td>36%</td>
<td>32%</td>
<td>24%</td>
<td>13%</td>
<td>33%</td>
<td></td>
</tr>
<tr>
<td>**LIQUID/GLIDE/</td>
<td>17%</td>
<td></td>
<td></td>
<td>23%</td>
<td>23%</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>APPROXIMANT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NASAL</strong></td>
<td>14%</td>
<td>21%</td>
<td>23%</td>
<td>17%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>AFFRICATE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>27%</td>
</tr>
</tbody>
</table>
PHONEMIC VS. PHONETIC FINDINGS

- LOT, THOUGHT, the Low-Back Vowel Merger
  - LOT~THOUGHT merged for most U.S. Midlands and Western speakers
  - Unmerged among Inland North and Core South speakers
  - Areas near the Midlands east of the Mississippi River show a variety of transitional and near-merged forms
  - The low-back vowel merger is a progressive merger... younger speakers are more likely to merge than older speakers in the same region
LOT, THOUGHT, & TRAP: 
UNMERGED CHICAGOLAND SPEAKER

Chicagoland Speaker, Interview and P&B data

Normalized First Formant (F1/s1) vs. Normalized Second Formant (F2/s2)

- α - LOT
- θ - THOUGHT
- æ - TRAP
LOT, THOUGHT, & TRAP (INTERVIEW DATA): Merged Southern Illinois Speakers
LOT, THOUGHT, & TRAP (P&B DATA):
Merged Southern Illinois Speakers
LOT, THOUGHT, & TRAP: MERGED OR NOT?

- Cartesian Distances
  - Baranowski, 2006
- Problem
  - Where is the merger?

<table>
<thead>
<tr>
<th></th>
<th>Distance in Normalized “Units”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOT ~ THOUGHT</td>
</tr>
<tr>
<td>Females, Interview Data</td>
<td>23</td>
</tr>
<tr>
<td>Females, P&amp;B Data</td>
<td>4</td>
</tr>
<tr>
<td>Males, Interview Data</td>
<td>24</td>
</tr>
<tr>
<td>Males, P&amp;B Data</td>
<td>1</td>
</tr>
</tbody>
</table>

- Target Undershoot (Lindblom, 1963)
  - Rapid speech produced more centralized vowels
  - Interview speech is more rapid than word list reading
- Word Lists = More Attentive Speech
  - “More attentive” speech more accurately reflects phonological categories
PHONETIC DETAIL: VOWEL TRAJECTORIES

- P&B tokens are longer in duration
  - Longer duration = better phonetic detail
  - ... such as comparing vowel trajectories

- MERGED
  - most speakers
  - all regions
**PHONETIC DETAIL: VOWEL TRAJECTORIES**

- **Southern Illinois**
  - Un-merged
  - THOUGHT = stable
  - LOT = moving

- **Chicagoland**
  - Un-merged
  - THOUGHT = moving
  - LOT = stable
CONCLUSION:  
IN DEFENSE OF WORD LIST DATA

- Comparability
  - Across disciplines / researchers / subjects
  - Across token types
  - Across consonantal & linguistic contexts

- Higher Level Phonological Access
  - More attentive speech provides more accurate phonological targets
  - Word list speech not subject to “supra-phonetic interference”

- Finer-grained Phonetic Detail
  - tokens are longer in duration
  - vowel trajectories provide compelling information
PROBLEMS, ISSUES, ASSUMPTIONS

- Is the “attention paid to speech” model accurate?
- Does “more attentive” speech fall closer to the “phonological target”?
- Where does attentive (“unnatural”) speech fit in a model of language change?
- How can these two kinds of data be reconciled?
- Why are some speakers’ interview data and P&B data more closely “matched” than others?
THANK YOU!

Contact Info for References, Further Questions... job offers...

Douglas S. Bigham, Ph.D.
University of Texas at Austin

douglas.s.bigham@gmail.com
REFERENCES (PARTIAL)