Lexical Variation in Lowman’s LAMSAS

John Nerbonne
Rijksuniversiteit Groningen
LSA Linguistics Institute
Summer, 2005
“If the sun comes out after a rain, you say the weather is doing what?”

- clearing up
- fairing off [... 40 variants]

1162 interviews conducted 1933–1974

71% of data collected by Guy Lowman 1933–1941

digitized data avail. from Bill Kretzschmar

focus on lexical overlap here, just as elsewhere (Kurath, ...)

- later goal: relation to pronunciation
Kurath’s ideas on areas
Carver's North/South Division
informants, phonetic, 6 clusters (left), 2 clusters (middle), fieldworkers (right)
<table>
<thead>
<tr>
<th>Fieldworker</th>
<th>Number of Interviews</th>
<th>Number of Responses</th>
<th>Mean Responses/Interview</th>
<th>SD Responses/Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowman</td>
<td>826</td>
<td>123990</td>
<td>150.1</td>
<td>25.3</td>
</tr>
<tr>
<td>McDavid</td>
<td>278</td>
<td>54855</td>
<td>197.3</td>
<td>76.8</td>
</tr>
<tr>
<td>others</td>
<td>58</td>
<td>12057</td>
<td>207.9</td>
<td>43.9</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>1162</strong></td>
<td><strong>190902</strong></td>
<td><strong>164.3</strong></td>
<td><strong>49.6</strong></td>
</tr>
</tbody>
</table>

Lowman elicited fewer responses, but more consistently.
responses per interview — Lowman, McDavid, other
Lexical Variation Needs Comparable Data

- Attempts to correct fieldworker bias
  - apparently no record of order of responses
  - restriction to most popular responses unsuccessful

- Therefore: concentrate on Lowman (71% of data)
Lexical Distance à la Seguy ’71

<table>
<thead>
<tr>
<th>Site</th>
<th>dog</th>
<th>hat</th>
<th>horse</th>
<th>toilet</th>
<th>smallest finger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brownsville</td>
<td>dog</td>
<td>hat</td>
<td>horse</td>
<td>bathroom</td>
<td>pinkie</td>
</tr>
<tr>
<td>White Plain</td>
<td>dog</td>
<td>cap</td>
<td>horse</td>
<td>bathroom</td>
<td>—</td>
</tr>
</tbody>
</table>

1. Ignore items for which data is missing (*smallest finger*)

2. Distance is $(1 - o)$, where $o$ is proportional overlap
   - distance(Brownsville, White Plain) = 0.25

3. Seguy used number of different items, we use proportion
Problem: close variants

- *fair off, fairing, fairing off, faired off, fairs off, ...*

- **solution**: use edit distance as measure of relatedness

<table>
<thead>
<tr>
<th>Standard American</th>
<th>◊œgIrI</th>
<th>delete r</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>◊œgII</td>
<td>replace l/ʒ</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>◊œgɔl</td>
<td>insert r</td>
<td>1</td>
</tr>
<tr>
<td>Bostonian</td>
<td>◊œgɔl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sum distance** 4

- edit distance applied to *spelling*, not phonetics (in lexical measurements)

  - lemmatizers would be most correct
    
    *clear - clean - cleared*
Problem: multiple responses

- *clear, fair off vs changing, clear, fair off*

- sol’n: lift distance measure from strings to string sets

\[
d(C) = \sum_{c \in C} d(c), \quad \text{where } C \text{ is a set of string pairs}
\]

Let \( C^1, C^2 \) be first, second projections of \( C \). \( C \) COVERS \( A \times B \) if, and only if \( C \subseteq A \times B \), and \( C^1 = A \) and \( C^2 = B \).

We shall seek the minimum cost COVER

\[
d(A, B) = \frac{1}{|C|} \min d(C'), \quad \text{where } C \text{ covers } A \times B
\]
Problem: multiple responses

Illustration: $A = \{a, b, c\}, B = \{a, c, d\}$

then $C = \{\langle a, a \rangle, \langle b, d \rangle, \langle c, c \rangle\}$ covers $A \times B$,
even though $|C| = 3$, while $|A \times B| = 9$.

Since $d(a, a) = d(c, c) = 0$, $d(A, B) = 1/3 \cdot d(b, d) = d(b, d)/3$

Likewise

\[
    d(\{a\}, \{b\}) = d(a, b) \\
    d(\{a\}, \{b, c\}) = \frac{1}{2} \cdot (d(a, b) + d(a, c))
\]
Problem: Infrequent Responses

Two, diametrically opposed, views:

**Goebel** weight infrequent overlap most heavily (*Gewichteter Identitätswert*, frequently mentioned)

**Carter** discard least frequent items (*American Regional Dialects*, p.17)

Solution (here): discard responses which occur fewer than five times.

We examine this more systematically later in the course.
lexical, all words
lexical, all words
lexical, minimum 5 occurrences per word
lexical, all words (left), minimum 5 occurrences per word (right)
lexical, minimum 5 occurrences per word (left), phonetic (right)
informants by year (left), lexical, minimum 5 occurrences per word (right)
lexical, minimum 5 occurrences per word (left), without years 1933, 1934 (right)
lexical, minimum 5 occurrences per word (left), without years 1933, 1934 (right)
Different Questionnaires used in LAMSAS
Final Division, Incorporating Questionnaire Intersection
<table>
<thead>
<tr>
<th>Concept</th>
<th>Response</th>
<th>North</th>
<th>Midland</th>
<th>S.Inland</th>
<th>S.Coast</th>
</tr>
</thead>
<tbody>
<tr>
<td>dragonfly</td>
<td>darning needle</td>
<td>100%</td>
<td>13%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>frost</td>
<td>dew</td>
<td>100%</td>
<td>2%</td>
<td>39%</td>
<td>0%</td>
</tr>
<tr>
<td>porch</td>
<td>stoop</td>
<td>92%</td>
<td>15%</td>
<td>0%</td>
<td>2%</td>
</tr>
<tr>
<td>quilt</td>
<td>comfort</td>
<td>2%</td>
<td>55%</td>
<td>84%</td>
<td>75%</td>
</tr>
<tr>
<td>night</td>
<td>evening</td>
<td>59%</td>
<td>74%</td>
<td>13%</td>
<td>8%</td>
</tr>
<tr>
<td>a little ways</td>
<td>a little piece</td>
<td>4%</td>
<td>64%</td>
<td>63%</td>
<td>17%</td>
</tr>
<tr>
<td>afternoon</td>
<td>evening</td>
<td>35%</td>
<td>21%</td>
<td>75%</td>
<td>82%</td>
</tr>
<tr>
<td>pallet</td>
<td>pallet</td>
<td>0%</td>
<td>6%</td>
<td>47%</td>
<td>59%</td>
</tr>
<tr>
<td>quilt</td>
<td>comfortable</td>
<td>61%</td>
<td>7%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>northwest</td>
<td>northern</td>
<td>0%</td>
<td>0%</td>
<td>31%</td>
<td>58%</td>
</tr>
<tr>
<td>Sunday week</td>
<td>Sunday week</td>
<td>7%</td>
<td>25%</td>
<td>51%</td>
<td>67%</td>
</tr>
<tr>
<td>stairs</td>
<td>stairsteps</td>
<td>4%</td>
<td>28%</td>
<td>42%</td>
<td>66%</td>
</tr>
<tr>
<td>lightwood</td>
<td>lightwood</td>
<td>0%</td>
<td>5%</td>
<td>5%</td>
<td>54%</td>
</tr>
<tr>
<td>dragonfly</td>
<td>snake feeder</td>
<td>13%</td>
<td>44%</td>
<td>55%</td>
<td>2%</td>
</tr>
<tr>
<td>weatherboarding</td>
<td>clapboards</td>
<td>54%</td>
<td>14%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>quarter to eleven</td>
<td>quarter till eleven</td>
<td>0%</td>
<td>20%</td>
<td>56%</td>
<td>19%</td>
</tr>
<tr>
<td>shades</td>
<td>shades</td>
<td>76%</td>
<td>26%</td>
<td>21%</td>
<td>53%</td>
</tr>
<tr>
<td>weatherboarding</td>
<td>weatherboarding</td>
<td>3%</td>
<td>41%</td>
<td>53%</td>
<td>50%</td>
</tr>
<tr>
<td>feet</td>
<td>feet</td>
<td>54%</td>
<td>5%</td>
<td>46%</td>
<td>52%</td>
</tr>
<tr>
<td>mantle</td>
<td>fireboard</td>
<td>0%</td>
<td>0%</td>
<td>48%</td>
<td>7%</td>
</tr>
</tbody>
</table>
Lexicon vs. Phonetics

- Most attention has been paid to lexical overlap
  - Criteria clearer, simpler

- Phonetic proximity shows more coherence
  - Less volatile linguistically
  - Less likely to degenerate into “curiousity cabinet”

- Lexical-phonetic correlation $r = 0.65$
  - Kurath & McDavid (1961) claim that lexical and phonetic distributions “coincide fairly well”
phonetic
Conclusions

- Reanalyzing existing atlas materials is “data mining”—search for valuable ores in a huge area

- Wealth of computational techniques now really applicable
  - linguistic level, representation, detail, psychological fidelity, frequency, microvariation, ...

- Need “investigative” techniques
  - But also rigorous validation (see Heeringa, Nerbonne & Kleiweg in *Proc. of Gesellschaft für Klassifikation*, 2002)

- But are dialectometric techniques too sensitive to small differences in questionnaire size, interviewer technique, etc.?
Links

LAMSAS:

• http://us.english.uga.edu/lamsas/

Our research:

• http://www.let.rug.nl/~kleiweg/lamsas/

Our software:

• http://www.let.rug.nl/~kleiweg/levenshtein/