Dialectology: Aggregate Dialectal Variation

John Nerbonne
University of Groningen, Humanities Computing
j.nerbonne@rug.nl

LSA Linguistics Institute
Harvard and MIT, Summer 2005
http://www.let.rug.nl/nerbonne/teach/dialectology/
Debts

We thank especially:

- Peter Kleiweg (software, visualization ideas)
- Wilbert Heeringa (collaboration on Dutch pronunciation, lexis)
- Bill Kretzschmar (making LAMSAS available)
- Marco Spruit (collaboration on Dutch syntax)

More about dialectometry in Groningen can be found at:
http://www.let.rug.nl/heeringa/dialectology/
http://www.let.rug.nl/kleiweg/lamsas/
Course Overview

- Introduction and Overview
- Lexical Analysis — Word Geography
- Pronunciation
- Validation and Meta-Analysis
- Explanation in Dialectology
- Linguistics in Dialectology
Goals

- Present dialectometry, emphasizing results
- Probe critical points.
  - Foundations
  - Relations to other approaches
  - Puzzle wrt sensitivity and quality
  - Applying explanatory models
  - Role of linguistics
    —so criticize!
What is dialectology?

• Study of language variation and how it’s used to mark geographic provenance.
• Geographic focus distinguishes dialectology from sociolinguistics.
• Variation that isn’t used (by language users) to mark geographic provenance ultimately falls outside dialectology.
  o Sub-perceptual variation (consider concomitants of fortis/lenis stop distinctions such as rise in $F_0$, $F_1$ of vowel following fortis, weaker bursts following lenis, longer duration of vowels preceding lenis, ...)
  o Noise, incl. individual variation, common variation
    But we don’t know a priori which variation does real work in marking provenance, and which doesn’t
• Therefore we need aggregating techniques.
Variation of *deur* ‘door’, *potten* ‘pots’ and *wijn* ‘wine’ as perceived by a traveler in Netherlands/Flanders. Extra-short sounds in superscript.
Consequences

- Dialectologists study how language variation is used to mark geographic provenance.
- This may exploit linguistic structure, but it may not.
  - **linguistic structure** syllable-final /r/’s in Eastern New England ([ba] ‘bar’, etc.);
    - Southern /ai/ ([ba] ‘buy’, etc.)
  - **haphazard** **tonic** ‘soda pop’ in Boston
- If we first examine all variation, we have a chance at saying how much is linguistically structured.
- Note that comprehensibility may be affected, but only when variation is extreme. There’s lots of variation which does not impede comprehensibility.
Consequences

- Dialectologists study how language variation is used to mark geographic provenance.
- Therefore, dialectology is ultimately grounded in how people perceive one another’s speech.
- Basis of Jo Daan’s “arrow method”
- Should we therefore turn over dialectology to psycholinguistics?
Dialect Speaker’s Perceptions

Daan
Perception

• If dialectology studies how language variation is used to mark geographic provenance, which is ultimately grounded in people’s perception of another’s speech, should we therefore turn over dialectology to psycholinguistics?
• There should be a role for perceptual studies in validating dialectological hypotheses and methods.
• But we want a linguistic perspective, not just a psycholinguistic one.
  o Perceptual studies are expensive, but dialect data is available for linguistic analysis.
  o We want to understand the linguistic basis of affinity judgements.
    —once we find that the Mason-Dixon line is important, perceptually, we want to know what linguistic features figure in the sensitivity
Bloomfield's Historical Perspective

Figure 6. Distribution of syllabic sounds in the words *mouse* and *house* in the Netherlands. — After Klocke.
Lessons from Bloomfield et al.

- Simple maps of forms are normally subject to exception, potentially contradicting dialectal trends

- Problem: the more detail one attends to, the less dialects appear to cohere. We need principled ways of aggregating.

- Seeking “isogloss bundles” is on the right track, because it looks to aggregate properties rather than identify small numbers of co-indicting diagnostics, but we need tools to analyse these aggregates.

- Coseriu (1956) also warned against “atomism” in dialectology
How to Progress in Dialectology

- Goal:
  Method for exploring dialect borders and dialect continua at any degree of detail.

- Objective:
  Find a ‘ruler’ or dialectometric method with which the linguistic distances between any pair of dialects can be measured in an objective way.
Alternative Methods

- (19th century) Tribes and intuition
- "isogloss method": no satisfying way to identify which isoglosses should be used
- “Structure geographic” method (Moulton, Goosens): phoneme inventories differences are sufficient, but not necessary indicators (e.g., degree of aspiration in Hiberno-English)
- Arrow method (e.g. Daan’s map, 1969): only dialects that border on each other can be compared.
- Perception experiment: many (groups of) subjects are necessary.
Jean Séguy
- Director of the *Atlas linguistique de la Gascogne*.
- He coined the term *dialectometry* (1973).
- Dialect distance: number of items of disagreement, expressed as a percentage.

Hans Goebl and Edgar Haimerl
- Strongly related to but independent of Séguy in development.
- Goebl explores connections with numerical taxonomy, geography and cartography.
- Goebl deepened and broadened dialectometry.

Brett Kessler (1995) introduced edit distance as a technique to measure pronunciation differences between individual words.

Heeringa (2004) deepened investigation into edit distance, developed several novel analytical techniques.
Analyzing & Illustrating Dialectometric Results

- The result of a dialectometric measurement is a table of place $\times$ place distances, which is often quite large (there are 484 sites in LAMSAS, and 1162 informants)
- We need ways of seeing and interpreting these results
- Today, some visualizations.
- Later, some analytical tools.
Distances

The averaged Levenstein distances between the dialects. Darker lines connect closer points, lighter lines more remote ones. Pearson’s r with geographic distances is 0.69.
Clustering

Apply Johnson’s algorithm to the upper half of the matrix (blue values):

- Iteratively,
  1. select shortest distance in matrix,
  2. fuse the two datapoints involved.
- To iterate, we have to assign a distance from the newly formed cluster to all other points (several alternatives, we used UPGMA).
- Repeat until one cluster is left over.

<table>
<thead>
<tr>
<th></th>
<th>Grouw</th>
<th>Haarlem</th>
<th>Delft</th>
<th>Hattem</th>
<th>Lochem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grouw</td>
<td>0</td>
<td>41</td>
<td>44</td>
<td>45</td>
<td>46</td>
</tr>
<tr>
<td>Haarlem</td>
<td>41</td>
<td>0</td>
<td>16</td>
<td>34</td>
<td>36</td>
</tr>
<tr>
<td>Delft</td>
<td>44</td>
<td>16</td>
<td>0</td>
<td>37</td>
<td>38</td>
</tr>
<tr>
<td>Hattem</td>
<td>45</td>
<td>34</td>
<td>37</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Lochem</td>
<td>46</td>
<td>36</td>
<td>38</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>
Using cluster analysis a dendrogram is derived from the $360 \times 360$ matrix. The scale distance shows percentages. Each of the 13 most significant groups is summed in one label.
Using cluster analysis a dendrogram is derived from the $360 \times 360$ matrix. From the dendrogram the 4 most significant groups can be identified. Diamonds represent dialect islands.
Using cluster analysis a dendrogram is derived from the 360 × 360 matrix. From the dendrogram the 13 most significant groups can be identified. Diamonds represent dialect islands.
Cluster composition maps were introduced by Peter Kleiweg. Idea: Start with drawing a line between the two main clusters; next draw line between the three main clusters by making the previous lines a little darker and adding a new one; repeat this for all levels (2 \ldots 360 clusters).
Multidimensional scaling

- Given a geographic map, distances between locations can be measured.
- Multidimensional scaling: given distances, locations on a map can be inferred.
- In our case: from \( n \times n \) distances we infer coordinates in 2- or 3-dimensional space.
  So \( n \) dimensions are reduced to two or three.
- We use Kruskal’s Non-metric Multidimensional Scaling.
Using MDS the 360 dimensions are reduced to 2. Y-coordinates represent the first and X-coordinates inversely represent the second dimension.
Using MDS the 360 dimensions are reduced to 3. Y-coordinates represent the first, X-coordinates inversely represent the second, and greytone represents the third dimension (distinct in the South).
Conclusions

- We need aggregation techniques to make sense of dialectological data.
- Dialectological reality is ultimately grounded in the indications dialect speaker communicate about geographic provenance.
- A wealth of techniques is available, leading to questions of how to choose.
Final remarks

Indebtedness (again):

- Wilbert Heeringa (general collaboration)
- Peter Kleiweg (software, visualization software)

More about dialectometry in Groningen:
http://www.let.rug.nl/~heeringa/dialectology/
http://www.let.rug.nl/kleiweg/lamsas/