How much does Geography Influence Language Variation?

John Nerbonne

Center for Language and Cognition, University of Groningen

Language, Space and Geography III:
26-28 Nov. 2009, Freiburg
Overview

- Introduction to dialectometry
  - Levenshtein distance

- Examining the influence of geography
  - Understanding Séguy’s curve (distance)
  - Incorporating areas
Why dialectometry?

- Solve problems of earlier dialectology
  - Non-overlapping distributions
  - Selection of features too arbitrary
  - “Atomism” (Coseriu), idiosyncratic words (Bloomfield)
- Introducing replicable procedures
- Following Séguy, Goebel, Schiltz, Kretzschmar, Shackleton, ...
  - Aggregating over categorical variables (same vs. different)
- Seeking law-like relations in linguistic variation
Beyond categorical differences: Levenshtein distance

- Levenshtein distance enables analysis of phonetic transcriptions without manual alignment
  —move from categorical to numerical analysis of data.
- One of the most successful methods to determine sequence distance (Levenshtein, 1964)
  - biological molecules, software engineering, ...
- Levenshtein distance: the minimum number of insertions, deletions and substitutions to transform one string into the other
Example of the Levenshtein distance

<table>
<thead>
<tr>
<th></th>
<th>mɛlɛk</th>
<th>delete ə</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mɛlkə</td>
<td>subst. ɔ/ɛ</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>mɛlɛkə</td>
<td>delete ə</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>mɛlk</td>
<td>insert ə</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>mɛlək</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 1 1 1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

John Nerbonne 5/32
Calculating dialect distances

- To determine the aggregate distance between dialects:
  - We determine the distance between each dialect pair for every single word (in sample, e.g. dialect atlas)
  - We sum these distances for every word (hundreds of them) and compare them

- Besides dialect distances, this also yields interesting sound correspondences contained in the alignments (more on that elsewhere)
  - Note that a 100-word comparison already yields about 500 sound correspondences
Based on Dutch pronunciation data from the Goeman-Taeldeman-Van Reenen-Project data (GTRP; Goeman and Taeldeman, 1996)
- We use 562 words for 424 varieties in the Netherlands

Distribution of sites
Analytical steps

- Obtain the distances between each of the \( \approx 90,000 \) pairs of varieties
  - n.b. this involves \( 500 \times 5^2 \) segment comparisons
  - \( \approx 1.1 \times 10^9 \) segment comparisons in total
- Organize these in a \( 400 \times 400 \) table
- Seek groups (dialect areas) or continuum-like relations, e.g. by applying clustering or multi-dimensional scaling, respectively
MDS map
Large body of work using Levenshtein distance—positive aspects

- Dutch, German, American English, Norwegian, Swedish, Afrikaans, Sardinian, Tuscan, Sino-Tibetan, Chinese, Bulgarian, Bantu, Central Asian (Turkic & Indo-Iranian), ...
- Development of consistency measure (Cronbach’s $\alpha$) indicting whether data set is sufficiently large
- Novel reflection, work on validation aimed at assessing degree of detection of SIGNALS OF PROVENANCE
The Influence of Geography

- Regression design
- Dependent variable: varietal distance, as measured by aggregate categorical distance or Levenshtein distance
- Independent variable: geographical distance, regarded as an operationalization of the chance of social contact
- Statistical caution: correlations involving averages are much higher than correlations involving individual items
  - True, but we’re interested in the properties of entire varieties (dialects), not just in their individual features
Inspiration: Jean Séguy

Sublinear spread is general
According to Trudgill (1972) diffusion follows an inverse square law, with the consequence that linguistic distance should likewise increase with the square of the distance. Population size plays the role of mass.
Trudgill’s “Gravity hypothesis”

- Sublinear aggregate relation incompatible with a quadratic influence

How much does geography influence language variation?

<table>
<thead>
<tr>
<th>Area</th>
<th>Corr.(l,geo)</th>
<th>$r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gabon Bantu</td>
<td>0.47</td>
<td>0.22</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>0.49</td>
<td>0.24</td>
</tr>
<tr>
<td>Germany</td>
<td>0.57</td>
<td>0.32</td>
</tr>
<tr>
<td>Eastern U.S.</td>
<td>0.51</td>
<td>0.26</td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.62</td>
<td>0.38</td>
</tr>
<tr>
<td>Norway</td>
<td>0.41</td>
<td>0.16</td>
</tr>
</tbody>
</table>

But note that linguistic distance in Norway correlates more highly with 19th century travel times ($r = 0.54$, $r^2 = 0.29$).  
Gooskens (2005) Traveling time as a predictor of linguistic distance.  
_Dialectologia et Geolinguistica_ 13.
Geography and language variation?

- Geography accounts for 22 – 38% of aggregate linguistic variation.
- General — sublinear — characterization of relation between geographical distance and linguistic differences
- Like population geneticists’ “isolation by distance” (Wright, 1943; Malécot, 1955)
Wrede’s (1926-56) German Dialect Areas
Influence of Dialect Areas?

- We add to the regression design variables indicating whether two varieties belong to the same or different dialect areas.
- For example, $\text{var}_{56-dff}$ has the value 1 if the two sites compared differ in that one is in area 5 (Bavaria) and the other in area 6 (Southwest).
- Do dialect areas contribute to the explanation of linguistic differences?

(On going work.)
Influence of Dialect Areas?

Preliminary results: Dialect areas contribute substantially to the explanation of aggregate linguistic distance. $r^2$ increases from about 32% (based only on geographic distance) to about 45% (based on geographic distance and areal differences).

(Ongoing work.)
How much does geography influence language variation?

- Pure distance models explain 22% - 38% of aggregate linguistic variation.
- Areal distinctions are somewhat collinear, but nonetheless add substantially to simple models, perhaps as much as 50% (moving 30% to 45%, for example).
- Naturally, there is also subdialectal variation (social, sexual, individual), but few systematic data collections.
- Emerging questions:
  - What is the linguistic structure of the dialect differences we find?
  - Do typological constraints play a (confounding) role?
  - Can we tease apart geographical and historical explanations, and how?
Questions?

Thank You!
Criticisms of Levenshtein-based work

- Binary segment distances too rough
- Frequent concern in Groningen (Heeringa Diss., 2004)
  - Segment distances based on phonetic features, phonological features, canonical spectrograms
  - High correlation with rough measures when compared at aggregate (varietal) level
  - But no substantial improvement in aggregate distance measures (validation wrt dialect speakers’ judgments)
  - Compare height measurements in in., cm, mm, μm
- Difficult problem in general — due *inter alia* to fine detail in atlases, e.g., 1,300 different vowels in LAMSAS
- New procedure (Jelena Prokic) induces segment weights from data
Inducing segment distances

Sound correspondences were obtained using the Levenshtein algorithm using a Pointwise Mutual Information procedure (Wieling et al., 2009; included in RuG/L04)

- Levenshtein algorithm:

\[
\begin{array}{cccccc}
| & \varepsilon & | & k & \varepsilon & n \\
| & i & | & k & h & \varepsilon & n \\
\end{array}
\]

\[
\begin{array}{cccc}
1 & 1 & 1 & 1 \\
\end{array}
\]

- Segment distances based on Pointwise Mutual Information:

\[
\text{PMI}(x, y) = \log_2 \left( \frac{p(x, y)}{p(x) p(y)} \right)
\]
Evaluating segment weight induction

- Evaluation with respect to alignment correctness
  —more sensitive than aggregate correlations with judgments
- 50% less error using alignments with induced weights
- Competitive with sophisticated bio-informatic techniques from, (pair Hidden Markov Models)
- Future project: evaluate the segment weights against linguistic criteria, compare weights induced from different data sets

Area: Bantu
Data: 53 sites, 160 words
Source: Van der Veen, Lyon
Note: Late settlement
Area: Bulgaria
Data: 482 sites, 54 words
Source: Stoykov’s atlas
Note: Long Turkish domination
Area: Germany
Data: 186 sites, 201 words
Source: Kleiner Deutscher Lautatlas (Göschel)
Area: Eastern Seaboard, US
Data: 357 sites, 145 words
Source: Mid & South Atlantic, LAMSAS (Lowman)
Note: Settlement in last 400 yr.
Area: The Netherlands  
Data: 424 sites, 562 words  
Source: Goeman-Taeldeman-van Reenen Atlas
Norway

Area: Norway
Data: North Wind & Sun
15 sites, 58 words
Source:
www.ling.hf.ntnu.no/nos
References


