Methods for aggregate corpus-based dialectometry

Natalia Levshina
Mainz, May 10 2016
What is dialectometry?

- Introduced by Séguy (1971)
- Examines the differences and similarities between geographical locations with regard to a large number of linguistic variables
  - Strong agreement between neighbouring locations suggests a dialect area.
  - Weak agreement between neighbouring locations suggests a dialectal boundary.
Main inspiration
Outline

1. Data: different approaches
   • Szmrecsanyi (2013) dataset of BrE dialects
   • Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   • Preliminary univariate analyses
   • Aggregate distance matrix
   • Exploration of dialect continua (MDS)
   • Identification of dialect areas (cluster analysis)
Szmrecsanyi (2013) data

• Available from
  https://sites.google.com/site/bszmrecsanyi/datasets

• The Freiburg Corpus of British dialects (FRED): interviews (interviewers’ utterances excluded)

• 158 different locations in 34 counties (pre-1974) of Great Britain + the Isle of Man and the Hebrides

• 57 morphosyntactic features
Groups of features with examples

A. Pronouns and determiners.
   • [1] non-standard reflexives, e.g. *They didn’t do it *theirself.*

B. The noun phrase
   • [9] the ‘s-genitive, e.g. John’s book

C. Primary verbs
   • [15] the primary verb TO HAVE, e.g. *The time has passed.*

D. Tense and aspect
   • [20] used to as a marker of habitual past, e.g. *He used to go around killing pigs.*
Groups of features (cont.)

E. Modality
F. Verb morphology
G. Negation
H. Agreement
I. Relativization
J. Complementation
K. Word order and discourse phenomena
Selection of features

• Dialectological, variational and corpus-linguistic literature
• Non-standard and standard features
• Alternation variables, e.g. standard and non-standard reflexives (e.g. *theirself* – *themselves*), and non-alternating forms (e.g. preposition stranding, *The house I lived in*)
Frequency matrix

- The extracted instances of variables are summarized for each county (34 in total)
- Normalized (per 10K words in the corpus)
- $\log_{10}$-transformed (to reduce the effect of large frequency differences and emphasize small frequency differences)
- $N \times p$ matrix ($N =$ number of objects, i.e. 34 counties, $p$ is the number of features, 57)
A fragment of the data set

<table>
<thead>
<tr>
<th>County code</th>
<th>County</th>
<th>avg_longitude</th>
<th>avg_latitude</th>
<th>a1_nonstandard_refl</th>
<th>a2_standard_refl</th>
<th>a3_thee_thine</th>
<th>a4_ye</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANS</td>
<td>Angus</td>
<td>-2.6269</td>
<td>56.6594</td>
<td>0</td>
<td>0.845098</td>
<td>0</td>
<td>0.6020</td>
</tr>
<tr>
<td>BAN</td>
<td>Banffshire</td>
<td>-2.9495</td>
<td>57.5431</td>
<td>0.3010299</td>
<td>0.602059</td>
<td>0.3010299</td>
<td>-1</td>
</tr>
<tr>
<td>CON</td>
<td>Cornwall</td>
<td>-5.5021</td>
<td>50.1754</td>
<td>0</td>
<td>0.698970</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>DEN</td>
<td>Denbighshire</td>
<td>-3.7429</td>
<td>53.1463</td>
<td>-1</td>
<td>0.698970</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>DEV</td>
<td>Devon</td>
<td>-3.681</td>
<td>50.3777</td>
<td>0</td>
<td>0.778151</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>DFS</td>
<td>Dumfriesshire</td>
<td>-3.8395</td>
<td>55.0028</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>
Outline

1. Data: different approaches
   • Szmrecsanyi (2013) dataset of BrE dialects
   • Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   • Preliminary univariate analyses
   • Aggregate distance matrix
   • Exploration of dialect continua (MDS)
   • Identification of dialect areas (cluster analysis)
Grieve (2009) data set

- Regional variation in written American English
- 206 cities from across the U.S.
- Letters to editors in local newspapers
- 40 lexical alternation variables only, e.g. though/although, whilst/while, about/around + Num
- The values in the matrix are the proportions of one variant (e.g. though) divided by the sum frequency of both variants (e.g. though + although) (from 0 to 1)
- The approach is more Labov-style.
Figure 3  About/Around Alternation Raw Values

From Grieve, Speelman & Geeraerts (2011)
Outline

1. Data: different approaches
   • Szmrecsanyi (2013) dataset of BrE dialects
   • Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   • Preliminary analyses
   • Aggregate distance matrix
   • Exploration of dialect continua (MDS)
   • Identification of dialect areas (cluster analysis)
How reliable is the frequency matrix?

- Cronbach’s $\alpha$: often used as a measure of reliability of test scores in psychology.
- A high score means that the variables are sufficiently intercorrelated for performing further analyses.
- Conventional minimum: $\alpha = 0.7$
- Here: $\alpha = 0.77$
Boxplot by areas: Var20 used to
Correlation with geographic distances

• Step 1. For each variable separately, compute pairwise linguistic distances by using the Euclidean metric. In this case (1 variable), the distances are absolute frequency differences between the points.

<table>
<thead>
<tr>
<th>Chapman_code</th>
<th>county</th>
<th>a20_USED_TO_habitual_past</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ANS</td>
<td>Angus</td>
</tr>
<tr>
<td>2</td>
<td>BAN</td>
<td>Banffshire</td>
</tr>
<tr>
<td>3</td>
<td>CON</td>
<td>Cornwall</td>
</tr>
<tr>
<td>4</td>
<td>DEN</td>
<td>Denbighshire</td>
</tr>
<tr>
<td>5</td>
<td>DEV</td>
<td>Devon</td>
</tr>
<tr>
<td>6</td>
<td>DFS</td>
<td>Dumfriesshire</td>
</tr>
<tr>
<td>7</td>
<td>DUR</td>
<td>Durham</td>
</tr>
</tbody>
</table>
# Distance matrix (Var20)

<table>
<thead>
<tr>
<th></th>
<th>ANS</th>
<th>BAN</th>
<th>CON</th>
<th>DEN</th>
<th>DEV</th>
<th>DFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAN</td>
<td>0.632</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CON</td>
<td>0.368</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEN</td>
<td>0.883</td>
<td>1.515</td>
<td>0.515</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV</td>
<td>0.447</td>
<td>1.079</td>
<td>0.079</td>
<td>0.436</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>0.851</td>
<td>1.483</td>
<td>0.483</td>
<td>0.031</td>
<td>0.404</td>
<td></td>
</tr>
<tr>
<td>DUR</td>
<td>0.437</td>
<td>1.069</td>
<td>0.069</td>
<td>0.446</td>
<td>0.010</td>
<td>0.415</td>
</tr>
</tbody>
</table>
Step 2. Compute geographical distances from the coordinates

<table>
<thead>
<tr>
<th></th>
<th>ANS</th>
<th>BAN</th>
<th>CON</th>
<th>DEN</th>
<th>DEV</th>
<th>DFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAN</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CON</td>
<td>464</td>
<td>520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEN</td>
<td>247</td>
<td>306</td>
<td>219</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEV</td>
<td>437</td>
<td>497</td>
<td>82</td>
<td>192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFS</td>
<td>124</td>
<td>179</td>
<td>341</td>
<td>128</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>DUR</td>
<td>128</td>
<td>190</td>
<td>363</td>
<td>146</td>
<td>323</td>
<td>85</td>
</tr>
</tbody>
</table>

(the distances are in rounded miles, taking into account the Earth curvature)
Step 3. Correlation between linguistic and geographic distances

- Mantel’s permutation test for distance matrices:
  \[ r = 0.43, \ p = 0.001 \]
- Therefore, the closer are two locations, the more similar frequencies of the feature they have.
Exercise. Task 1

- Var32, the use of *ain’t*, is distributed across the regions as shown below. What does the boxplot say?
Exercise. Task 2

• What is the linguistic distance between Kent and Suffolk with regard to *ain’t*?

<table>
<thead>
<tr>
<th>Chapman_code</th>
<th>apriori_area</th>
<th>a32_aint</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>KEN</td>
<td>SE</td>
</tr>
<tr>
<td>27</td>
<td>SFK</td>
<td>SE</td>
</tr>
</tbody>
</table>

• Is there a significant correlation between the linguistic and geographic distances?
  • Mantel \( r = 0.3, p = 0.001 \)
Outline

1. Data: different approaches
   - Szmrecsanyi (2013) dataset of BrE dialects
   - Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   - Preliminary univariate analyses
   - Aggregate distance matrix
   - Exploration of dialect continua (MDS)
   - Identification of dialect areas (cluster analysis)
Euclidean distance

• Computed for every pair of locations
• The distance represents:
  • the square root...
  • of the sum...
  • of the squared differences between two locations for every variable
An example with two variables

<table>
<thead>
<tr>
<th>code</th>
<th>county</th>
<th>a20_USED_TO_hab.past</th>
<th>a32_aint</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEN</td>
<td>Kent</td>
<td>2.176091</td>
<td>0.4771213</td>
</tr>
<tr>
<td>SFK</td>
<td>Suffolk</td>
<td>2.075547</td>
<td>0.3010300</td>
</tr>
</tbody>
</table>

\[ \sqrt{((2.176091 - 2.075547)^2 + (0.4771213 - 0.3010300)^2)} \]

\[ [1] \quad 0.2027739 \]
Outline

1. Data: different approaches
   • Szmrecsanyi (2013) dataset of BrE dialects
   • Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   • Preliminary univariate analyses
   • Aggregate distance matrix
   • Exploration of dialect continua (MDS)
   • Identification of dialect areas (cluster analysis)
MDS (non-metric, 2 dimensions)

Stress = 0.18
[England + Wales] vs. [Scotland]

Stress = 0.18
Outline

1. Data: different approaches
   • Szmrecsanyi (2013) dataset of BrE dialects
   • Grieve (2009) dataset of AmE dialects

2. Statistical analyses
   • Preliminary univariate analyses
   • Aggregate distance matrix
   • Exploration of dialect continua (MDS)
   • Identification of dialect areas (cluster analysis)
Hierarchical cluster analysis (Ward method)

- requires a distance matrix
- first merges two closest points, then finds the next closest, and repeats until all points are merged
- different methods of choosing the nearest point/cluster to another cluster (complete, average, single, Ward’s, etc.)
Divisive non-hierarchical clustering

• The algorithm divides the points into $n$ clusters such that the distances between the clusters are maximized and the distances between the cluster members are minimized.

• PAM: Partitioning Around Medoids
  • can be performed either on a distance matrix, or on the frequency matrix
  • robust to outliers
Additional tools

• Groningen software for making fancy maps:
  • http://www.let.rug.nl/~kleiweg/L04/
References


