Measuring Language Variation

Therese Leinonen

university of groningen

25 June 2008
Overview

- Background
- Measuring dialect variation with Levenshtein distance
- The phonetic puzzle
- Levenshtein distance and perceptual distance
- SweDia2000
- Measuring dialect variation acoustically
- Visualizing results: Multidimensional scaling
- Future work
Background

- dialectometry = measuring dialect. Term invented by Jean Séguy.
- aim: find dialect borders and explore dialect continua
- method: find a measure for measuring linguistic distance between dialects
Levenshtein distance

- edit distance, calculates the cost of changing one string to another
- applied for comparison of Irish dialects by Kessler 1995
- later applied to American English, Bantu languages, Bulgarian, Chinese, Dutch, German, Norwegian, Sardinian
- example Lyngby [\textipa{'eni]} vs. Helsinki [\textipa{e:ni\textalpha}] ‘agreed’

<table>
<thead>
<tr>
<th></th>
<th>\textipa{e:ni}</th>
<th>remove</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lyngby</td>
<td>\textipa{e:ni}</td>
<td>substitute i by i</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>\textipa{e:ni}</td>
<td>insert a</td>
<td>1</td>
</tr>
<tr>
<td>Helsinki</td>
<td>\textipa{e:ni\textalpha}</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>
Levenshtein distance

Length normalization

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lungby</td>
<td>?</td>
<td>e:</td>
<td>n</td>
<td>i</td>
<td></td>
</tr>
<tr>
<td>Helsinki</td>
<td>e:</td>
<td>n</td>
<td>i</td>
<td>a</td>
<td></td>
</tr>
<tr>
<td></td>
<td>del</td>
<td>sub</td>
<td>ins</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

non-normalized distance: 3
normalized distance: $\frac{3}{5} = 0.6$ or 60 %
Phonetic Puzzle

• theorem: given segment distances, Levenshtein algorithm finds optimal alignment

• what are good segment distances?

• various feature systems: Vieregge-Cucchiarini, Almeida-Braun

• "acoustic" distance

• stochastic learning procedure (Pair Hmms)

• very limited improvement over binary segmental table
Why is detailed phonetic information not helping?

- hypothesis 1: transcriptions are phonetically unreliable
- hypothesis 2: previous attempts were too ambitious, trying to characterize all distinctions
- hypothesis 3: we are past the size where fine discrimination matters
- others?
Predicting intelligibility and perceived linguistic distance
(Beijering, Gooskens and Heeringa 2008)

Research questions:

- How well can Levenshtein distance predict perceptive distance and intelligibility?
- How well can normalized Levenshtein distance predict perceptive distance in comparison to non-normalized Levenshtein distance?

Data:

- recordings of The North Wind and the Sun in 18 Scandinavian varieties
- phonetic transcriptions of cognates (on average 98 words)
Predicting intelligibility and perceived linguistic distance
(Beijering et al. 2008)

Perceptual distance:

- listeners: 3 groups 15-19-year-olds from Copenhagen
- stimulus data: the whole recording of the fable in 6 varieties
- task: judge distance to Standard Danish on a scale from 1 to 10

Intelligibility:

- listeners: 18 groups 15-19-year-olds from Copenhagen
- stimulus data: 6 sentences in 6 varieties
- task: translate into Standard Danish
Predicting intelligibility and perceived linguistic distance
(Beijering et al. 2008)

Correlation with Levenshtein distance:

<table>
<thead>
<tr>
<th></th>
<th>normalized</th>
<th>non-normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perceptual distance</td>
<td>0.52</td>
<td>0.62</td>
</tr>
<tr>
<td>Intellegibility</td>
<td>-0.86</td>
<td>-0.79</td>
</tr>
</tbody>
</table>

Differences between normalized and non-normalized Levenshtein distances are not significant.

Conclusion: Levenshtein distance a better predictor of intelligibility than of perceived linguistic distances
Swedish vowel data

- SweDia2000: project carried out by the universities of Lund, Stockholm and Umeå 1998-2001 (Bruce, Elert, Engstrand and Eriksson 1999)

- 105 sites in Sweden and Swedish-speaking Finland

- 12 speakers from each site: 3 elderly women, 3 elderly men, 3 young women, 3 young men

- Vowels elicited with existing one-syllable words with the target vowel in a coronal consonant context

- 19 words of which the vowels cover the standard Swedish vowel space
Acoustic method

- principal component analysis (PCA) on bandfiltered spectra (Jacobi, Pols and Stroop 2005, Pols, Tromp and Plomp 1973)

- vowel spectra filtered up to 18 Bark

- PCA built on 4 anchor vowels ([i], [æ], [a] and [u]) of equally many men and women from every site (in total 300 speakers from 83 sites)

- two first principal components (85.6 % of total variance explained) used as acoustic measure of vowel quality

- creaky voice is a problem for the method: F0 controll
Factor loadings

![Factor loadings plot](image-url)
Factor scores

pc1

100
120
140
160
180

80
100
120
140
160
180

pc2

-40
-20
0
20
40

i

u

æ

a
Dialect distances

- Linguistic distances measured for all pair of sites: Euclidean distances of pc1 and pc2 of all words (averages per site) \( \sqrt{n \sum_{i=1}^{n} (p_i - q_i)^2} \)

- Distances analyzed with multidimensional scaling (MDS): visualization of distances in a low dimensional space

- Visualizing three dimensions with RGB-colours gives maps that can show a dialect continuum (Heeringa 2004)
MDS: dimensions 1 and 2
MDS: dimensions 3, 4 and 5
MDS: dimensions 3-5
Future work

• work on the acoustic method (rotation)

• include more measuring points within a segment (diphthongization)

• extracting underlying linguistic structure (PCA)
References


