Regression Analysis on Levenshtein-Pointwise Mutual Information Segment Distances Across Languages and Acoustic Distances

Eliza Margaretha
Martijn Wieling
John Nerbonne

Rijksuniversiteit Groningen
Outline

- Overview
- Techniques
- Data
- Analysis
- Discussion
- Summary
Overview

- Compare phonetic segment distances
  - Dutch, German, Bulgarian
- Compare Levenshtein-Pointwise Mutual Information (PMI) distances to acoustic distances
- Regression analysis
  - Correlation
  - Prediction power
Techniques: Levenshtein-PMI (1/3)

- Segment Distance
  - How often segment $x$ is aligned with segment $y$

- Levenshtein
  - Insertion: a segment with a gap
  - Deletion: a gap with a segment
  - Substitution: 2 segment
Techniques: Levenshtein-PMI (2/3)

- Pointwise Mutual Information (Church and Hanks, 1995)
  \[
  PMI(x, y) = \log_2 \left( \frac{p(x, y)}{p(x)p(y)} \right)
  \]

- Wieling, et al. (2009)
  - \( p(x, y) \) is the number of the \( x \) and \( y \) occurrences at the same position in 2 aligned strings of \( X \) and \( y \), divided by the total number of aligned segments
  - \( p(x) \) or \( p(y) \) the number of the occurrences of \( x \) or \( y \) divided by the total number of segment occurrences
Techniques: Levenshtein-PMI (2/3)

- Training Wieling, et al. (2009)
  1. Align string with Levenshtein algorithm (w/o vocal-consonant)
  2. Compute PMI values and transform (subtract from 0 + max value)
  3. Apply Levenshtein to PMI-segment distances
  4. Repeat 2 and 3 till convergence is reached

<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th>ι</th>
<th>ε</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td>0</td>
<td>2331</td>
<td>1880</td>
</tr>
<tr>
<td>ι</td>
<td>2331</td>
<td>0</td>
<td>64905</td>
</tr>
<tr>
<td>ε</td>
<td>1880</td>
<td>64905</td>
<td>0</td>
</tr>
</tbody>
</table>
Techniques: Formant Measurements (1/3)

- **Vowel quality** (McArthur, 1998)
  - the property that makes vowels different, e.g. /i:/ as in sheep from /i/ as in ship
  - determined by the position of the vocal tracts during pronunciation

- **Formants**
  - measure vowel quality by means of acoustic signals
  - specify the energy concentration positions in the acoustic signals, i.e. the lowest resonance frequencies (Peterson & Barney, 1952)
Techniques: Formant Measurements (2/3)

- Formants: darker bands
  - 2 first formants are the most distinguishing
  - 3\textsuperscript{rd} formants and lip position
- /i/ and /u/ has similar first formants but the second formant of /i/ is much higher than that of /u/

![Formant Waveforms](Picture from (Leinonen, 2010))
Techniques: Formant Measurements (3/3)

- Acquiring acoustic distances
  - Compute Euclidean distances of formant values between vowel pairs (Wieling, et al., 2007)

- Normalizing non-linguistic speaker-dependent differences
  - Pitch, gender, shape & size of vocal tracts
  - Transforms Hertz frequency to the Bark or the Mel scales
Techniques: Mantel Test (1/2)

- Triangle inequality
  - Dependent: $D(a, \varepsilon)$ is dependent to $D(a, \theta)$ and $D(\theta, \varepsilon)$
    - $D(a, \varepsilon) < D(a, \theta) + D(\theta, \varepsilon)$
  - Acoustic distance
- Independent
  - Levenshtein PMI
- Mantel test
  - Significance Test of a Correlation Coefficient of Distance Matrices
Techniques: Mantel Test (2/2)

- Random permutation test
- H Null = No relation between 2 matrices
  - R should be equally likely to be larger or smaller
- Steps
  1. Permutate rows and columns of one of the matrices randomly
  2. Compute correlation between the permutated matrix and the other matrix
  3. Repeat 1 and 2
- Observation value:
  - Add 1 for every $r(\text{PD1}, \text{D2}) > r(\text{D1}, \text{D2})$
  - Divided by number of repetition
Dataset (1/5)

<table>
<thead>
<tr>
<th>Language</th>
<th>Locations</th>
<th>Words</th>
<th>Segment Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td>424</td>
<td>562</td>
<td>82</td>
</tr>
<tr>
<td>German</td>
<td>186</td>
<td>196</td>
<td>78</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>197</td>
<td>152</td>
<td>67</td>
</tr>
</tbody>
</table>

- **Dutch**: Goeman-Taeldean-Van Reenen-Project
- **German**: Kleiner Deutscher Lautatlas project
- **Bulgarian**: students’ theses at the University of Sofia, published monographs, dictionaries, and the archive of the Ideographic Dictionary of Bulgarian Dialects (Prokić, et al., 2009)
Dataset (2/5)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Shared Types</th>
<th>Segment Alignments</th>
<th>Vowel Alignments</th>
<th>Consonant Alignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch and Bulgarian</td>
<td>43</td>
<td>235</td>
<td>92</td>
<td>143</td>
</tr>
<tr>
<td>Dutch and German</td>
<td>71</td>
<td>870</td>
<td>261</td>
<td>609</td>
</tr>
</tbody>
</table>
Dataset (3/5)
Dataset (4/5)
Dataset (5/5)

- Acoustic data was obtained from Pols, et al. (1973) and Van Nierop, et al. (1973),
- three first formants
- 50 male and 25 female Dutch speakers
- 36 acoustic vowel alignments
- All alignments appear in Levenshtein-PMI Dutch matrix
Analysis: Lev-PMI Distance Across Languages (1/5)

- Regression analysis setup
  - Variables
    - Dutch (independent/explanatory) and Bulgarian (dependent/response)
    - Dutch and German
  - Cases
    - Segment alignments
  - Values
    - Levenshtein-PMI distance
Analysis: Lev-PMI Across Languages (2/5)
Analysis: Lev-PMI Across Languages (3/5)
## Analysis: Lev-PMI Across Languages (4/5)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Alignment Sets</th>
<th>Pearson Correlation</th>
<th>Effect size ($r^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch and Bulgarian</td>
<td>All</td>
<td>0.336</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>Vowel</td>
<td>0.418</td>
<td>0.178</td>
</tr>
<tr>
<td></td>
<td>Consonant</td>
<td>0.339</td>
<td>0.115</td>
</tr>
<tr>
<td>Dutch and German</td>
<td>All</td>
<td>0.630</td>
<td>0.397</td>
</tr>
<tr>
<td></td>
<td>Vowel</td>
<td>0.620</td>
<td>0.384</td>
</tr>
<tr>
<td></td>
<td>Consonant</td>
<td>0.587</td>
<td>0.345</td>
</tr>
</tbody>
</table>
Analysis: Lev-PMI Across Languages (5/5)

- Computing regression line
  - $y = 1586.562 + 0.300x$
  - $T$ ratio 5.454 ($p < 0.001$)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1568.562</td>
<td>128.322</td>
<td>12.224</td>
<td>0.000</td>
</tr>
<tr>
<td>NL</td>
<td>0.300</td>
<td>0.055</td>
<td>5.454</td>
<td>0.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: BUL
Analysis: Example (1/2)

- How does the prediction work?
- Lev-PMI distance between \(a\) and \(e\) in Dutch, \(x = 1556\)
- Predicted \(a-e\) distance in Bulgarian:
  \[\hat{y} = 1586.562 + 0.300 \times 1556 = 2053.362\]

<table>
<thead>
<tr>
<th>Model Summary(^b)</th>
<th>Descriptive Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 1</td>
<td>R</td>
</tr>
<tr>
<td>1</td>
<td>.336(^a)</td>
</tr>
<tr>
<td>NL</td>
<td>2207.7872</td>
</tr>
</tbody>
</table>

\(^a\) Predictors: (Constant), NL
\(^b\) Dependent Variable: BUL
Analysis: Example (2/2)

\[ SE_{\hat{y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum^n_{i=1}(x_i - \bar{x})^2}} \]

- \[ SE_{\hat{y}} = 639.4 \times \sqrt{\frac{1}{235} + \frac{(1556 - 2207.8)^2}{578270.9}} = 549.6 \]
- \[ t \text{ for } (df = 200) = 1.97 \ (\alpha = 0.05) \]
- Confidence Interval 95% = \[ \hat{y} \pm t \times SE_{\hat{y}} = 2053.362 \pm 1.97 \times 549.6 = 2053.362 \pm 1083 \]
- With 95% certainty, mean of \( a-e \) distance in Bulgarian given the distance in Dutch = 1556, lies in the interval (970.3136).
- Real distance = 1675
Analysis: Lev-PMI and Acoustic Distances (1/3)

- **Response Variable**
  - Lev-PMI distance for Dutch segments

- **Explanatory variables (acoustic distance variations)**
  - Hertz: raw hertz measurements of formants
  - Bark: hertz values transformed to Bark scale
  - Mel: hertz values transformed to Mel scale
  - Z-score
    - hertz values transformed to Z-scores per speaker, normalizing over all the vowels for each speaker
    - average the Z-scores per vowel of all speakers
Analysis: Lev-PMI and Acoustic Distances (2/3)

<table>
<thead>
<tr>
<th>Acoustic variation</th>
<th>Number of first formants</th>
<th>Pearson Correlation</th>
<th>Effect Size ($r^2$)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hertz</td>
<td>2</td>
<td>0.481</td>
<td>23 %</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.426</td>
<td>18 %</td>
<td>0.010</td>
</tr>
<tr>
<td>Z-score</td>
<td>2</td>
<td>0.720</td>
<td>52 %</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.640</td>
<td>41 %</td>
<td>0.000</td>
</tr>
<tr>
<td>Bark Scale</td>
<td>2</td>
<td>0.616</td>
<td>38 %</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.517</td>
<td>27 %</td>
<td>0.001</td>
</tr>
<tr>
<td>Mel Scale</td>
<td>2</td>
<td>0.603</td>
<td>36 %</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.507</td>
<td>26 %</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Analysis: Lev-PMI and Acoustic Distances (3/3)

- Mantel test with 9999 replicates
- H0 = No relation between Lev-PMI distance with Acoustic distance
- Positive observations shows positive relationships

<table>
<thead>
<tr>
<th>Acoustic variation</th>
<th>Observation value</th>
<th>Significance (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hertz 2</td>
<td>0.168</td>
<td>0.0134</td>
</tr>
<tr>
<td>Hertz 3</td>
<td>0.132</td>
<td>0.035</td>
</tr>
<tr>
<td>Z2</td>
<td>0.410</td>
<td>1e-04</td>
</tr>
<tr>
<td>Z3</td>
<td>0.317</td>
<td>3e-04</td>
</tr>
<tr>
<td>Bark 2</td>
<td>0.303</td>
<td>2e-04</td>
</tr>
<tr>
<td>Bark 3</td>
<td>0.206</td>
<td>0.0027</td>
</tr>
<tr>
<td>Mel 2</td>
<td>0.286</td>
<td>2e-04</td>
</tr>
<tr>
<td>Mel 3</td>
<td>0.195</td>
<td>0.0036</td>
</tr>
</tbody>
</table>
Discussion

- Why is the correlation between Dutch and Bulgarian smaller than that between Dutch and German?
- Why do Z-scores yield better results than other variations (Hertz, Bark, Mel)?
- How are the relationships between Levenshtein-PMI distances and acoustic distances of other languages?
Summary

- Our results show that Levenshtein-PMI distances of Dutch are able to predict those of Bulgarian and German.
- Prediction of languages in the same category / with similar characteristics (Dutch-German) is better than those with different characteristics (Dutch-Bulgarian).
- Vowel quality as represented by acoustic distances correlate reasonably highly with Levenshtein-PMI distances, particularly in our Dutch case, the former can predict up to 52% of the latter.