Some quantitative approaches to analyzing language dialects


- **Qualitative – Quantitative**
- **Single** feature – **Aggregate** of multiple features
- **In depth** analysis – **Overall** trends

**Quantitative analysis** of dialect data:
- coupled with qualitative methods
- cutting-edge **statistical and computational** methods
- exploratory/pattern discovery + hypothesis testing
- thorough
What is a “dialect”?

Oxford Dictionaries Online (http://oxforddictionaries.com/definition/english/dialect; Feb 2013)

Pronunciation: /ˈdɪəlɛkt/

noun

a particular form of a language which is peculiar to a specific region or social group

Origin

mid 16th century (denoting the art of investigating the truth of opinions): from French dialecte, or via Latin from Greek dialektos 'discourse, way of speaking', from dialegesthai 'converse with'

Grammar

A version of a language spoken in a particular geographical area or by a particular group of people. The English spoken in Newcastle is different from that spoken by natives of North Cornwall. Not only do speakers in these two areas have a different accent, they also use a number of different words. Different dialects also use slightly different grammar, too. For example, in Devon some people say ‘They do have …’ in preference to ‘They have …’ Such regional expressions are not ‘wrong’, they simply differ from standard English. They are sometime described as ‘non-standard’
What is a “dialect”? 

Accent: just the pronunciation

Dialect: everything (vocabulary, morpho-syntax, phonetics & phonology)

Language: linguistic, historical and political factors

http://www.youtube.com/watch?v=BOUTfUmI8vs
What is a “dialect”? 

Accent: just the pronunciation

Dialect: everything (vocabulary, morpho-syntax, phonetics & phonology)

Language: linguistic, historical and political factors
The data: sources and formats

- Various *features* & their *values* in different locations/for different speakers
- *Lexical, phonetic, morpho-syntactic*
Zeichenerklärung zu Karte 80

Simplizia oder Grundwörter in Komposita:

- >Kraut< n. /->Kräuter< Pl.  
  (z.B. khrůth, ɕʰɜɡrɑdɔ Pl., kʰtʰɛkerɔd)
- >Kräuterich< n. (z.B. grɛdriʃ)
- >Stengel< m. (z.B. šdɛpl, bɔdɛb̥rædɛpl)
- >Staude< f. (z.B. štɡo, ɕɛdɛpʃtɔudɔ je Pl.)
- >G'*staude/-g*stäude< n.  
  (206 Mmb ɡɛmbɛʃɔstɛtɛid, 229 Dkl ɡstɛuda)
- >Stock< m. (z.B. šdɛpl Pl.)
- >Pflichter< n. (z.B. ɻɛltɔ, 107 Aɪɛt ɻɛltɔd)
- >Rebe< f. (z.B. ɻɛba, ɕɛdɛpʃlɛbɔ, 98 Kaʊ ɻɛba)
- >Jaub< n. (263 Btz ɡrunbɛʃɔlɛb)

Zusatzzeichen über den Symbolen: Bestimmungswörter

- >Erdäpfel<  "  >Bumser<
- >Erdbirnen<  "  >Bodenbirnen<
- >Grundbirnen<  "  >Herbst<

Weitere Zeichen:

- unter dem Symbol: Beleg als „älter“, „richtiger“ u.a. qualifiziert
- unter dem Symbol: Beleg als E qualifiziert
- zwischen den Symbolen: semantische Differenzierung;  
  vgl. Belegliste
- nicht gefragt
- Hinweis auf Belegliste

The data: sources and formats

- Various **features** & their **values** in different locations/for different speakers
- Lexical, **phonetic**, morpho-syntactic
Feature 2 from the Dialect Atlas of Finnish (Kettunen, 1940; digitized by York University, Canada; Embleton & Wheeler 1997, 2000; checked and corrected by the BEDLAN project http://kielievolutio.uta.fi/doku.php?id=en:start)
The data: sources and formats

- Various **features** & their **values** in different locations/for different speakers
- Lexical, phonetic, **morpho-syntactic**
Table 1. Map 68a in SAND1 shows the five syntactic variables in the context of weak reflexive pronoun as object of inherently reflexive verb.

<table>
<thead>
<tr>
<th>Context:</th>
<th>Weak reflexive pronoun as object of inherent reflexive verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables:</td>
<td>{ zich, hem, zijn eigen, zichzelf, hemzelf }</td>
</tr>
<tr>
<td>Example:</td>
<td>Jan herinnert zich dat verhaal wel.</td>
</tr>
<tr>
<td></td>
<td>Jan remembers himself that story AFFIRM</td>
</tr>
<tr>
<td></td>
<td>&quot;John certainly remembers that story.&quot;</td>
</tr>
</tbody>
</table>
Wordlists

- A set of fixed words collected from several locations/social groups

<table>
<thead>
<tr>
<th>TABLE 1. The word list</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
</tr>
<tr>
<td>ash</td>
</tr>
<tr>
<td>bath</td>
</tr>
<tr>
<td>better</td>
</tr>
<tr>
<td>bite</td>
</tr>
<tr>
<td>blood</td>
</tr>
<tr>
<td>bone</td>
</tr>
<tr>
<td>brother</td>
</tr>
<tr>
<td>calf</td>
</tr>
<tr>
<td>cold</td>
</tr>
<tr>
<td>corn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Historical</th>
<th>Engl. &amp; Wales: S/M</th>
<th>England: North</th>
<th>Scotland &amp; Ireland</th>
<th>North America</th>
<th>Rest of the World</th>
<th>Other Germanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proto-Germanic</td>
<td>RP</td>
<td>Buxton</td>
<td>Standard Scottish</td>
<td>Standard Canadian</td>
<td>S.Africa: Jo'burg</td>
<td>West Frisian: Grou</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>OE: Mercian</td>
<td>Norwich TYP</td>
<td>Rosendale TYP</td>
<td>Coldstream TYP</td>
<td>Ohio TYP</td>
<td>NZ: Auckland TYP</td>
<td>German: Standard</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>ME: Kentish</td>
<td>Somerset EMGT</td>
<td>Morley TYP</td>
<td>Edinburgh TRAD</td>
<td>Chicago TYP</td>
<td>Nigeria (Igbo) TYP</td>
<td>'Plattdeutsch'</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>ME: E. Midlands</td>
<td>North Devon TRAD</td>
<td>Middlesbrough TYP</td>
<td>Edinburgh TYP</td>
<td>Chicago AAVE</td>
<td>India: Delhi STD</td>
<td>Luxembourg</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>ME: SW Midlands</td>
<td>North Devon TYP</td>
<td>Tyneside TRAD</td>
<td>Edinburgh EMGT</td>
<td>Boston TRAD</td>
<td>Singapore STD</td>
<td>Alsace</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>ME: Northern</td>
<td>North Devon EMGT</td>
<td>Tyneside TYP</td>
<td>Buckie TRAD</td>
<td>New York TRAD</td>
<td>-</td>
<td>South Tyrol</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>-</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>ME: Older Scots</td>
<td>S.Wales TYP</td>
<td>Tyneside EMGT</td>
<td>Lewis TYP</td>
<td>N.Carolina TRAD</td>
<td>Yiddish</td>
<td>-</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['tʃartə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>Shakespearean</td>
<td>Cornhill TYP</td>
<td>Antrim TRAD</td>
<td>N.Carolina AAVE</td>
<td>Icelandic STD</td>
<td>Faroese</td>
<td>Norwegian</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>Middle Scots</td>
<td>Longtown TYP</td>
<td>Antrim TYP</td>
<td>Alabama TRAD</td>
<td>-</td>
<td>-</td>
<td>Danish</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>-</td>
<td>-</td>
<td>['dɔktə:]</td>
</tr>
<tr>
<td>L.Mod: South-East</td>
<td>Berwick TYP</td>
<td>Belfast TYP</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>['dɔktə:]</td>
<td>['dɔktə:]</td>
<td>['dɔctə:]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Holy Island TRAD</td>
<td>Tyrone TRAD</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>['dɔktə:]</td>
<td>['dɔctə:]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Example from several English dialects (and other Germanic languages) for *daughter*; [http://soundcomparisons.com/](http://soundcomparisons.com/)
Wordlists

- A set of fixed words collected from several locations/social groups
- One idea: Compute distances between corresponding words at different locations
- Similar to sequence alignments
- Variants of Levenshtein (1965) distance:

\[
\begin{array}{c}
j & a & s \\ \hline 1 & 1 & 1 \\
\end{array} \begin{array}{c}
\alpha & z & i \\ \hline 1 & 1 & 1 \\
\end{array} = 3
\]

- Phonetic/articulatory features (e.g. V ≠ C), swap operation (Damerau distance; 1964), pairwise mutual information (frequencies of alignment; Wieling, 2012), ...
- Distances between dialects = average across word distances →

\[
M = \begin{bmatrix}
66 & 66 & 52 & 122 & 77 & 47 \\
52 & 56 & 56 & 134 & 81 & 51 \\
122 & 134 & 116 & 116 & 63 & 59 \\
77 & 81 & 63 & 115 & 115 & 72 \\
47 & 51 & 59 & 111 & 72 & \\
\end{bmatrix}
\]
Wordlists

- A set of fixed words collected from several locations/social groups
- One idea: Compute distances between corresponding words at different locations
- Similar to sequence alignments
- Slot-matching algorithm (Heggarty et al., 2005)

Averaging across items → $M$
Distance-based methods

- From wordlists or features → inter-dialects distances matrix

\[
M = \begin{bmatrix}
66 & 66 & 52 & 122 & 77 & 47 \\
52 & 56 & 134 & 81 & 51 \\
122 & 134 & 116 & 63 & 59 \\
77 & 81 & 63 & 115 & 111 \\
47 & 51 & 59 & 111 & 72
\end{bmatrix}
\]

- From \( M \) we can produce maps, do clustering, dimensionality reduction, phylogenetics, networks, statistical tests...

- Many can be done in general-purpose statistical packages such as \texttt{R}

- Specialized software e.g. GabMap (http://www.gabmap.nl/; Nerbonne \textit{et al.}, 2011), DiaTech (Aurrekoetxea \textit{et al.}, 2013)
Distance-based methods

- Distance matrix as heatmap
Distance-based methods

- Distance matrix as beam and network maps

Small distance (high similarity)  Large distance (low similarity)
Distance-based methods

- Dimensionality reduction with Multidimensional Scaling (MDS)
- French dialect data: 2D MDS scatterplot vs geographic locations:

![2D MDS scatterplot vs geographic locations](image)
Distance-based methods

- Dimensionality reduction with Multidimensional Scaling (MDS)
- [French dialect data] 2D MDS scatterplot vs geographic locations
- [Bulgarian dialect data] 3D MDS → RGB color components → plotted on map
Distance-based methods

- Clustering and mapping using different colors/symbols for each cluster
- Hierarchical clustering [Bulgarian data]
Distance-based methods

- Clustering and mapping using different colors/symbols for each cluster
- Hierarchical clustering [Bulgarian data]
- K-means clustering (requires $k$; can search for optimal $k$ given criterion)
- Quantify uncertainty: bootstrap, noisy/composite clustering (Nerbonne et al., 2008)
Distance-based methods

- Phylogenetic trees (Neighbour-Joining, UPGMA)
Distance-based methods

- Phylogenetic trees (Neighbour-Joining, UPGMA)
- Big issues with horizontal processes → phylogenetic networks (Neighbour-Net in SplitsTree4; Bryant & Moulton, 2004)
Distance-based methods


- 3D interactive plot:
Distance-based methods

- Correlation between geographic and linguistic distances: Mantel correlation (1967)

\[ r_{(D,M)} \]

\( p \) - permutations
Distance-based methods

- Correlation between geographic and linguistic distances: **Mantel correlation** (1967)
- Regression of **distances to a fixed variety** on location (and other co-variates)
- **Generalized Additive Models (GAM) and Mixed-Effects Models:**
  \[ d = \text{linguistic distance to Standard Dutch} \]
  \[ d \sim \text{longitude} + \text{latitude} + (1|\text{location} + \text{transcriber} + \text{word}) \]
Distance-based methods

- Correlation between geographic and linguistic distances: Mantel correlation (1967)
- Regression of distances to a fixed variety on location (and other co-variates)
- Various tests (t-test, ANOVA – permutations!) on distances between subsets of varieties
Character-based methods

- **Phylogenetic trees** using likelihood and Bayesian models:
  - “usual approach”: *cognate judgments* between dialects

Words that come from the same proto-word in the ancestor language
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
- “usual approach”: cognate judgments between dialects
Character-based methods

- **Phylogenetic trees** using likelihood and Bayesian models:
  
  - “usual approach”: **cognate judgments** between dialects
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
Character-based methods

- **Phylogenetic trees** using likelihood and Bayesian models:
  - “usual approach”: **cognate judgments** between dialects

- **(Hierarchical) Bipartite graph partitioning** → cluster dialects **and** features

- **Linguistic features**

<table>
<thead>
<tr>
<th></th>
<th>HOE?</th>
<th>EEL</th>
<th>night</th>
<th>hog</th>
<th>pʰ</th>
<th>tʰ</th>
<th>asp.</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appleton</td>
<td>+</td>
<td>A</td>
<td>[natt]</td>
<td>[hɔg]</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Brownsville</td>
<td>+</td>
<td>A</td>
<td>[nat]</td>
<td>[hɔg]</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Charleston</td>
<td>+</td>
<td>B</td>
<td>[nat]</td>
<td>[hʊg, hɔg]</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Downe</td>
<td>-</td>
<td>B</td>
<td>[nat]</td>
<td>[hɔg]</td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Evanston</td>
<td>?</td>
<td>B</td>
<td>[natt]</td>
<td>[hɔg]</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Character-based methods

- **Phylogenetic trees** using likelihood and Bayesian models:
  - “usual approach”: **cognate judgments** between dialects

- (**Hierarchical**) Bipartite graph partitioning → cluster dialects *and* features

- **Linguistic features**
  - → **simult.** cluster varieties and their most distinctive feats. (here, sound changes)

![Bipartite Graph](image)

*Fig. 2. Example of a bipartite graph of four varieties and three sound correspondences.*
Character-based methods

- **Phylogenetic trees** using likelihood and Bayesian models:
  - “usual approach”: *cognate judgments* between dialects
- (Hierarchical) **Bipartite graph partitioning** → cluster dialects *and* features

- **Linguistic features**
  - → *simult.* cluster varieties and their most distinctive feats.
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis:
  - Test/estimate spatial patterning of feature values: e.g., Ripley's K, Moran's I, (semi-)variogram/correlogram, kriegering, 2D continuous kernel density estimation....
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis:
  - Test/estimate spatial patterning of feature values: e.g., Ripley's K, Moran's I, (semi-)variogram/correlogram, kriging, 2D continuous kernel density estimation....
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning $\rightarrow$ cluster dialects and features
- Spatial statistics/Point Pattern Analysis:
  - Test/estimate spatial patterning of feature values: e.g., Ripley's $K$, Moran's $I$, (semi-)variogram/correlogram, kriging, 2D continuous kernel density estimation....
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis:
  - Test/estimate spatial patterning of feature values: e.g., Ripley's $K$, Moran's $I$, (semi-)variogram/correlogram, kriging, 2D continuous kernel density estimation....
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis:
  - Test/estimate spatial patterning of feature values: e.g., Ripley's K, Moran's I, (semi-)variogram/correlogram, kriging, 2D continuous kernel density estimation....
Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis
- STRUCTURE/DESTRUCT & friends:
  - Population genetics → admixture

Character-based methods

- Phylogenetic trees using likelihood and Bayesian models:
  - “usual approach”: cognate judgments between dialects
- (Hierarchical) Bipartite graph partitioning → cluster dialects and features
- Spatial statistics/Point Pattern Analysis
- STRUCTURE/DESTRUCT & friends:
  - Population genetics → admixture

Some of the stuff \textbf{not} covered here

- Computer models of dialect formation, maintenance and evolution
- Network analysis of social & communicative networks
- Dialect change through time (stability, etc.)
- ....
Conclusions and discussion

- Dialect data is *multi-faceted, multi-dimensional*, and has complex *geographic* and *social* structure
- Several quantitative approaches, *continuously developed* and made *user-friendlier*
- Some techniques inspired from various branches of *biology*
- Some could be used to analyze *animal dialect* structure (geography?, bipartite graph?)
- What are the *fruitful parallels* and how to *capitalize* on them?

**Thank you!**

Special thanks to **Didier Demolin, Outi Vesakoski, and Jelena Prokić**

Genetic biasing in Speech and Language