Corpus-based dialectometry: why and how

Benedikt Szmrecsanyi (QLVL, KU Leuven)
Christoph Wolk (University of Giessen)
Introduction
Introduction & terminology

- **linguistic corpora**: principled & broadly representative collections of naturalistic texts or speech ⇔ usage data
- **corpus linguistics**: base claims about language on corpora ⇔ methodological outgrowth of the usage-based turn
  (see Bybee 2010; Tomasello 2003; papers in Szmrecsanyi and Walchli 2014)
- **classical dialectometry**: draws on atlas material to explore geolinguistic patterns using aggregation methodologies
- **corpus-based dialectometry (CBDM)**: draws on quantitative / distributional info derived from corpora
Why

• Goebl (2005: 499): “Extra atlantes linguisticos nulla salus dialectometrica” (because of comparability issues)
  ⇒ we respectfully disagree

• also (!) being able analyze naturalistic corpus data is central to the maturation of the dialectometry enterprise:

• CBDM not a second-best methodology – principled reasons for using usage data:
  • contextualization
  • usage versus knowledge
  • gradedness

• variationist (socio)linguists almost exclusively analyze usage/corpus data ⇒ methodological convergence
How

• challenge: corpus-derived datasets are noisier and dirtier (i.e. less balanced) than atlas- and survey-derived datasets

• 2 approaches:
  1. **top-down CBDM**: (1) define feature catalogue; (2) establish frequencies / probabilities associated with features; (3) aggregate
  2. **bottom-up CBDM**: (1) let features emerge in a data-driven fashion via identification of significant/distinctive POS n-grams; (2) aggregate
This presentation

- 2 case studies illustrating these approaches
- summarize work by Szmrecsanyi (2013) and Wolk (2014):
  (see also Szmrecsanyi 2008, 2011; Szmrecsanyi and Wolk 2011)
  - grammatical variation
  - traditional British English dialects
  - tapping into Freiburg Corpus of English Dialects (FRED)
The Freiburg Corpus of English Dialects (FRED)

- ca. 2.5 mio words of running text
  \(\approx 300h\) of recorded speech
- oral history interviews
- 431 dialect speakers, mainly NORMs
  - mean speaker age: 75 years
    (typically born around the beginning of 20\textsuperscript{th} century)
  - 64\% male
- see [www.helsinki.fi/varieng/CoRD/corpora/FRED/](http://www.helsinki.fi/varieng/CoRD/corpora/FRED/)
County Cornwall, Southwest of England (St. Ives)
speaker: male, born 1892 (interview recorded in 1978)

{<u IntRS> Well you’re a St. Ives man. Where were you born?}
<u CAVA_PV> Born Belyars Lane, eighteen ninety-two. Eighteenth of December. Worn sovereign in the cupper. Born sovereign. The poor times then, you know (gap ’indistinct’) boiling potatoes and tinkle mosses.
{<u IntRS> Did you, did you, how long did you live there?}
<u CAVA_PV> Oh we lived there about, oh about twelve years, I suppose. Then we went up to a rose wall terrace. Hmm. So everything’s altered now to what er was then, I mean.

audio
Top-down CBDM
Top-down CBDM: a cooking recipe

- **Step 1**: define feature catalogue  
  (motto: the more the merrier)  
- **Step 2**: identify features in the corpus texts  
  (automatically, semi-automatically, or manually)  
- **Step 3**: establish feature frequencies (per location);  
  normalize frequencies and/or model probabilistically  
- **Step 4**: aggregate: $N \times p$ feature matrix $\Rightarrow N \times N$  
  distance matrix  
- **Step 5**: project to geography, analyze & interpret
Our feature catalogue

• \( p = 57 \) features
• all major grammatical subdomains covered
• the usual suspects in the variationist & dialectological literature, e.g. . . .
  • non-standard past tense *done* (e.g., *you came home and done the home fishing*)
  • multiple negation (e.g., *don’t you make no damn mistake*)
  • *don’t* with third person singular subjects (e.g., *if this man don’t come up to it*)
Barebone frequencies: cluster maps

input: geographic distances

input: corpus-derived linguistic distances
Bare-bones frequencies and data availability

- corpora are constrained by the availability of suitable data
- measurements are imprecise and biased when little data is available

linguistic distance as a function of corpus size.

*linear $r^2 = 0.61$*
Probabilistically enhanced CBDM

• we can combat this bias with some form of smoothing
• per the *Fundamental Dialectological Postulate* (Nerbonne and Kleiweg, 2007), geography-based smoothing seems most appropriate
• while several forms of geographic smoothing exist (e.g. Grieve, 2009; Pickl et al., 2014), we believe that generalized additive modeling (*GAM*, see also Wieling, 2012), a regression variant, provides a particularly adequate solution
• using *GAMs*, we can also take other information, such as sociolinguistic predictors like speaker age or gender, into account simultaneously
The process

- instead of normalizing the observed counts, build a regression model (GAM) per feature
- use the model to predict counts for the locations
- proceed as usual

Frequency of multiple negation (log scale)
Results

input: barebone CBDM

input: probabilistically enhanced CBDM
Top-down CBDM: interim summary

- the approach can uncover a geolinguistic signal in naturalistic usage data
- probabilistic modeling reduces noise
- correlation linguistic/geographic distances (least-cost travel time):
  - barebone: $R^2 = 7.6\%$ (mildly sublinear)
  - probabilistically enhanced: $R^2 = 44.3\%$ (sublinear)
Bottom-up CBDM
can we replace the manual feature selection and extraction with an automatic process?

idea: building on Nerbonne and Wiersma (2006) and Sanders (2010), use part-of-speech n-grams to measure syntactic distance and evaluate using permutation (see also Lijffijt, 2013)

the FRED Sampler (FRED-S) is available in a POS-annotated form
Bottom-up CBDM

- construct and count all part-of-speech n-grams (here: bigrams)
- create new corpora by resampling
  - pairwise, to detect differences between two dialects
  - globally, to identify reliable locations of high or low frequency
- compare original counts against large number of resampled counts: how often is it larger/smaller?
Example

- VDO + VVI, *do* + lexical verb (infinitive); includes unstressed periphrastic *do*

input: normalized frequency
input: reliability
input: non-significant differences
Aggregational results

input: top-down (bare-bone)

geographic $R^2 = 27.6\%$

input: bottom-up reliability

geographic $R^2 = 26.2\%$
Bottom-up CBDM: interim summary

- the approach works roughly as well as the manual feature selection process
- method detects known features of British dialect grammar (e.g. non-standard uses of was and were)
- the relation between bigram frequencies or related scores and dialectal features may be opaque - what do, for example, significant differences in article + noun sequences mean?
- the results seem to "correlate with syntactic differences as a whole, even if it does not measure them directly" (Nerbonne and Wiersma, 2006: 84)
Conclusion
Extensions and related work

- extension to other linguistic levels
  - phonetics & phonology (via aggregation of acoustic measurements or auditory classifications)
  - lexis (building on Speelman et al. 2003; Ruette et al. 2013)
- correlating aggregate variation on different linguistic levels (Spruit et al. 2009-style), based on measurements from the same corpus
- regional variation in corpora sampling written language (see Grieve 2009)
Thank you!

benszm@kuleuven.be
christoph.b.wolk@anglistik.uni-giessen.de


References II


References IV


Aggregation in the barebone frequency approach

1. the frequency matrix

<table>
<thead>
<tr>
<th></th>
<th>text frequencies</th>
<th>text frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>feature 1</td>
<td>feature 2</td>
</tr>
<tr>
<td>dialect a</td>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>dialect b</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>dialect c</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

\[ d(a, b) = \sqrt{(11 - 5)^2 + (8 - 2)^2} = 8.5 \]
\[ d(a, c) = \sqrt{(11 - 1)^2 + (8 - 7)^2} = 10.0 \]
\[ d(b, c) = \sqrt{(5 - 1)^2 + (2 - 7)^2} = 6.4 \]

2. aggregation via the Euclidean distance measure

3. the distance matrix

<table>
<thead>
<tr>
<th></th>
<th>dialect a</th>
<th>dialect b</th>
<th>dialect c</th>
</tr>
</thead>
<tbody>
<tr>
<td>dialect a</td>
<td></td>
<td>8.5</td>
<td></td>
</tr>
<tr>
<td>dialect b</td>
<td></td>
<td></td>
<td>10.0</td>
</tr>
<tr>
<td>dialect c</td>
<td>6.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The importance of data availability

- from ongoing work with Tobias Streck (Freiburg)
- pronunciation variation in southwest Germany, 189 lexemes in spontaneous speech, 354 locations
- distance only stabilizes at $\sim 100$ observations per location
- geographic $R^2 = 0.12$ total / 0.20 good support only