Groningen Work (1/2007 – 6/2007)

John Nerbonne

Alfa-informatica
University of Groningen

Linguistic Unity & Diversity Project
Tübingen, June, 2007

Outline

Introduction

- Jelena Prokić's work applying phylogenetic software
- Pair Hidden Markov Models inducing segment distances
- Experiments: bootstrap clustering vs. clustering with noise
- Conditional Entropy as Model of Comprehensibility
- Draft of Vision Paper on Aggregate Linguistic Analysis

Pair Hidden Markov Models: Background

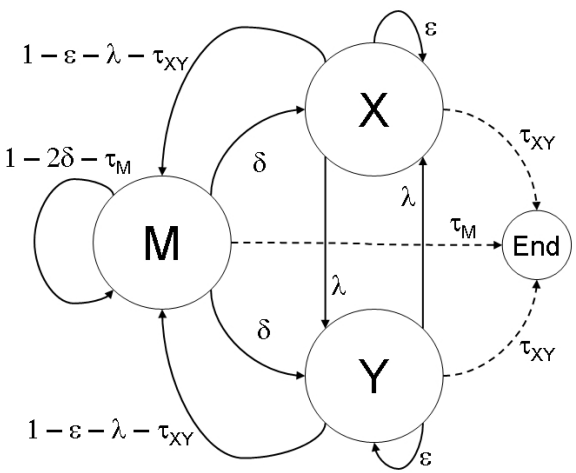
- Edit distance analyses dialect data well, even with minimal phonetic/phonological sensitivity
- C/V distinguished, only C/C, V/V aligned (w. exceptions for approximants, syllabic sonorants)
- Many attempts to include phonetic/phonological feature distances, acoustic differences (Heeringa '04, Kessler, '95)
  - Little difference—neither improvement nor degradation
- Solution Idea: Apply Machine Learning

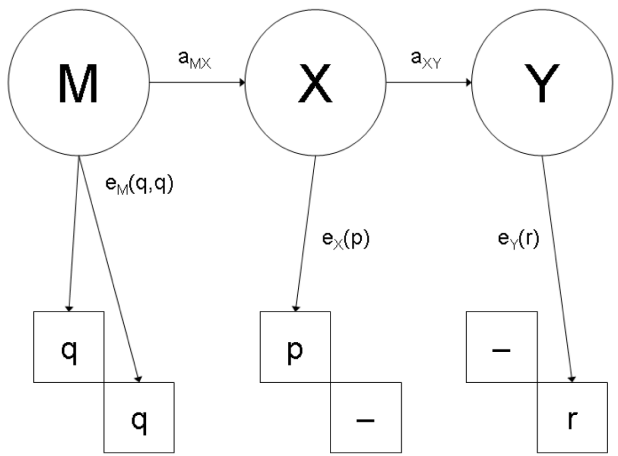
Pair Hidden Markov Models: Background

- Edit distance analyses dialect data well, even with minimal phonetic/phonological sensitivity
- C/V distinguished, only C/C, V/V aligned (w. exceptions for approximants, syllabic sonorants)
- Many attempts to include phonetic/phonological feature distances, acoustic differences (Heeringa '04, Kessler, '95)
  - Little difference—neither improvement nor degradation
- Solution Idea: Apply Machine Learning

Pair Hidden Markov Models (M. Wieling)

- Adapted Hidden Markov Model: 2 parallel output streams
  - Originally developed for aligning biological sequences
- Mackay and Kondrak (2005) applied PHMM's pronunciation
  - Thanks to Mackay and Kondrak for use of software!
- Three states: Match (M), Deletion (X), Insertion (Y)
- Transition, insertion, deletion and substitution probabilities determine probability of an alignment



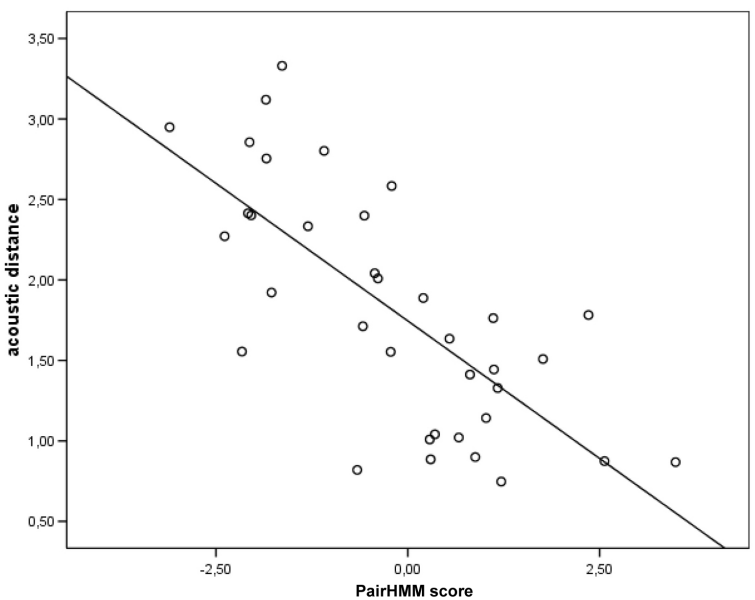


Material: the GTRP

- Data source: Goeman-Taeldeman-Van Reenen-Project (GTRP)
- IPA Transcriptions of 1876 items for 613 localities
- Most recent Dutch dialect data source: 1980 – 1995
- In our analysis a subset of the data is used
  - 424 Netherlandic varieties (transcription differences wrt Flemish)
  - 562 items (omitting items with morphological variation)

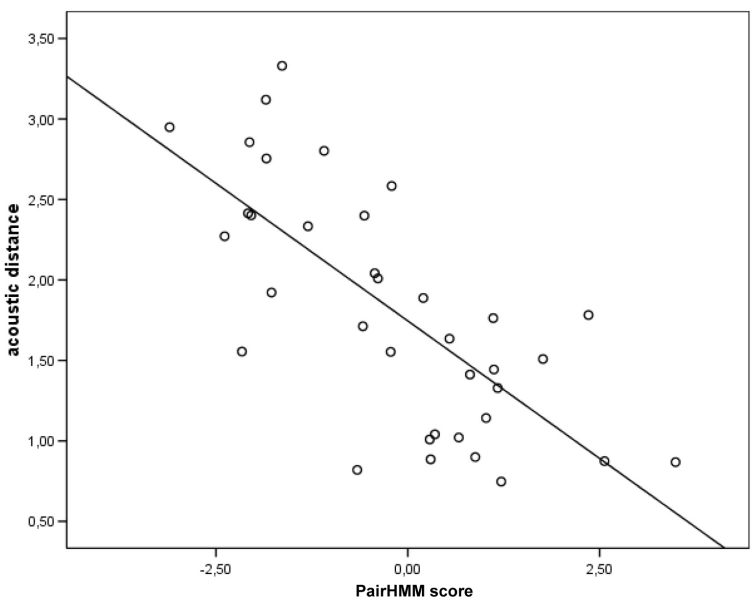
Training, calculating dialect distances

- Baum-Welch (EM) training; 100 CPU hours (200-node cluster)
- Viterbi (best alignment) best (with log-odds normalization for indels)
- Intuitive results: Prob[a/a] > Prob[V/V] > Prob[V/C]
- High correlation with acoustic vowel distances (Bark scale, z-score): $r = -0.72$
- Confirms that segment distances are learned!



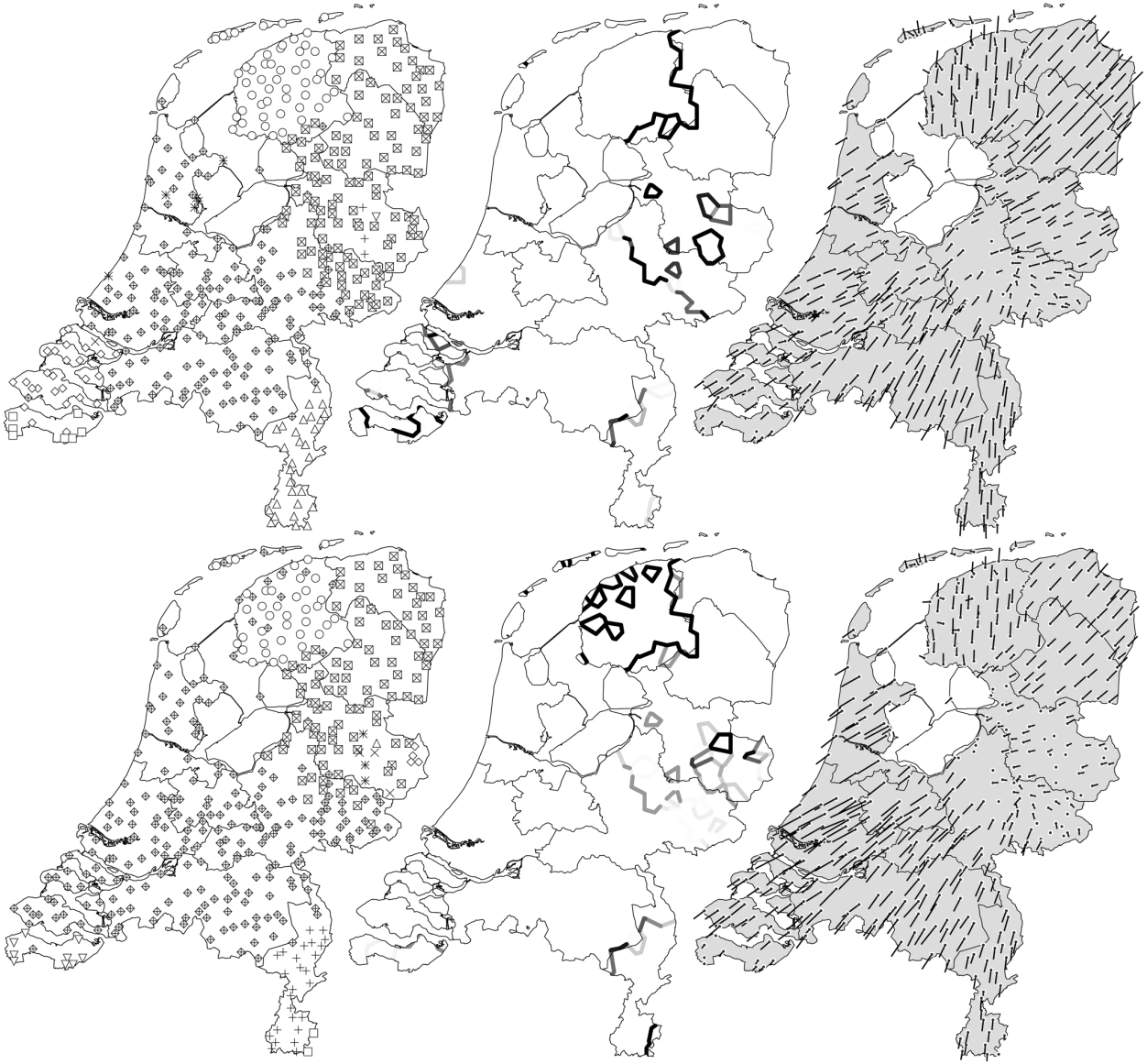
Training, calculating dialect distances

- Baum-Welch (EM) training; 100 CPU hours (200-node cluster)
- Viterbi (best alignment) best (with log-odds normalization for indels)
- Intuitive results: Prob[a/a] > Prob[V/V] > Prob[V/C]
- High correlation with acoustic vowel distances (Bark scale, z-score): $r = -0.72$
- Confirms that segment distances are learned!



Dialectologically: Levenshtein (top) vs. PairHMM?

$r = 0.89$



Discussion

- PairHMMs align linguistic material well and induce reasonable segment distances
- Confirming that frequent correspondences tend to be linguistically similar
  - Suggested by historical propagation and pressure toward convergence
- However, no marked improvement over the Levenshtein distance
- Possible cause: aggregate level of analysis

Discussion

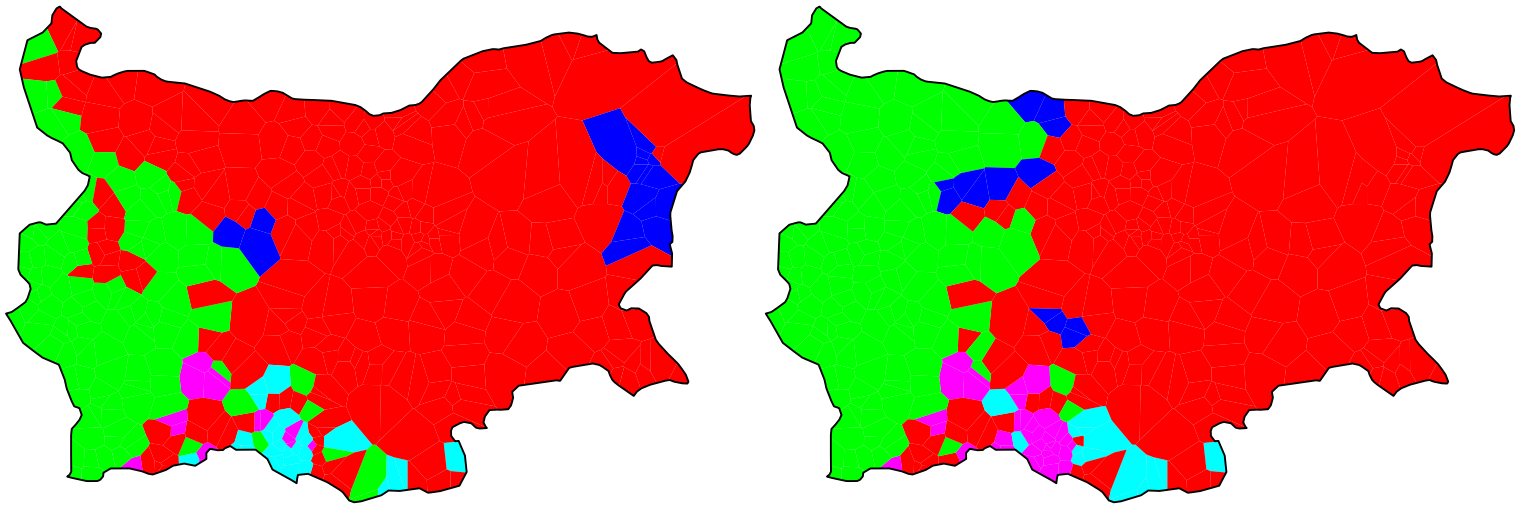
- PairHMMs align linguistic material well and induce reasonable segment distances
- Confirming that frequent correspondences tend to be linguistically similar
  - Suggested by historical propagation and pressure toward convergence
- However, no marked improvement over the Levenshtein distance
- Possible cause: aggregate level of analysis

Discussion

- PairHMMs align linguistic material well and induce reasonable segment distances
- Confirming that frequent correspondences tend to be linguistically similar
  - Suggested by historical propagation and pressure toward convergence
- However, no marked improvement over the Levenshtein distance
- Possible cause: aggregate level of analysis

Stable Clustering

Two Bulgarian Datasets ($r = 0.97$)

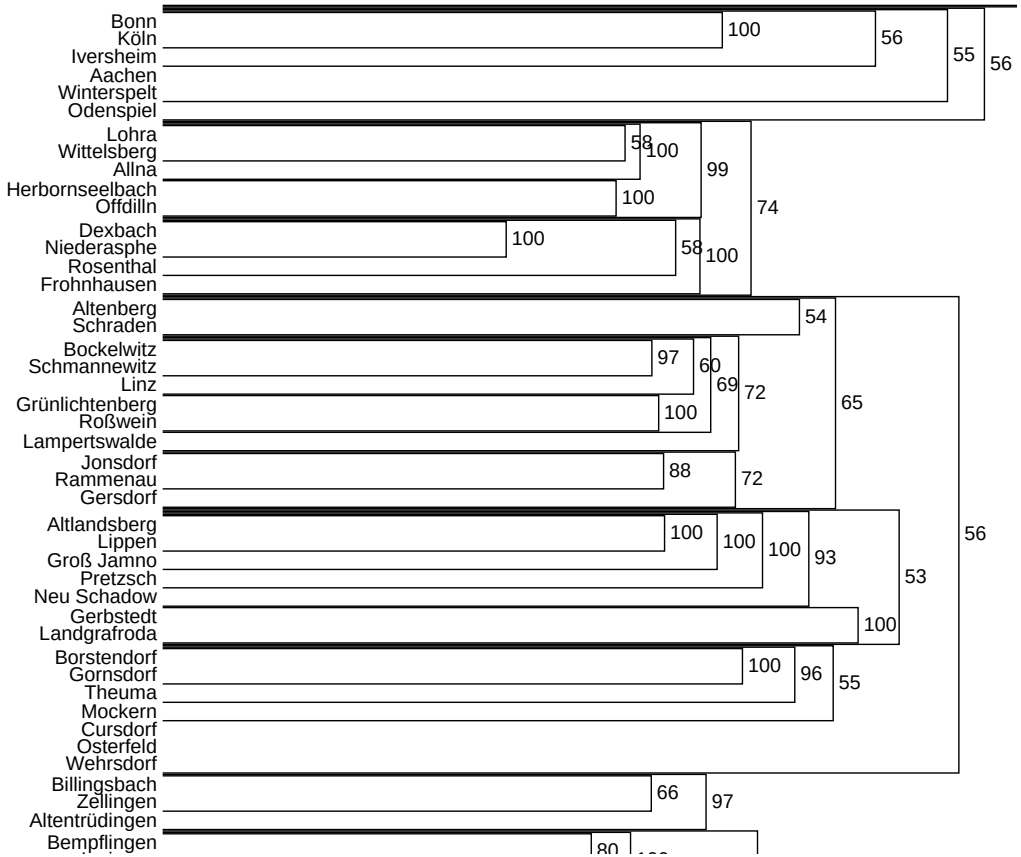


- Clustering isn't STABLE: small input differences can mean large differences in clusters.
- Stability is a real problem!
- Two solutions:
  - Bootstrap clustering
  - Adding small amounts of random noise

Bootstrapping Clustering

- assume n (linguistic) distance matrices, $M_{1 \leq i \leq n}$, e.g. one matrix/word
- choose clustering technique, e.g. WPGMA
- repeat, e.g. 100 times
  - select $m \leq n$ matrices, allowing replacement
    - option 1: use repeated selection as weight (Mucha & Haimerl, GfKl 2006)
    - option 2: ignore repetition
  - cluster sum of matrices obtaining dendrogram, recording groups
  - "composite matrix" $M' \Leftarrow$ mean cophenetic distances
  - collect groups that appear a majority of times into a "composite dendrogram"
  - (new!) project dendrogram borders to map, reflecting cophenetic distance in darkness

Composite Dendrograms

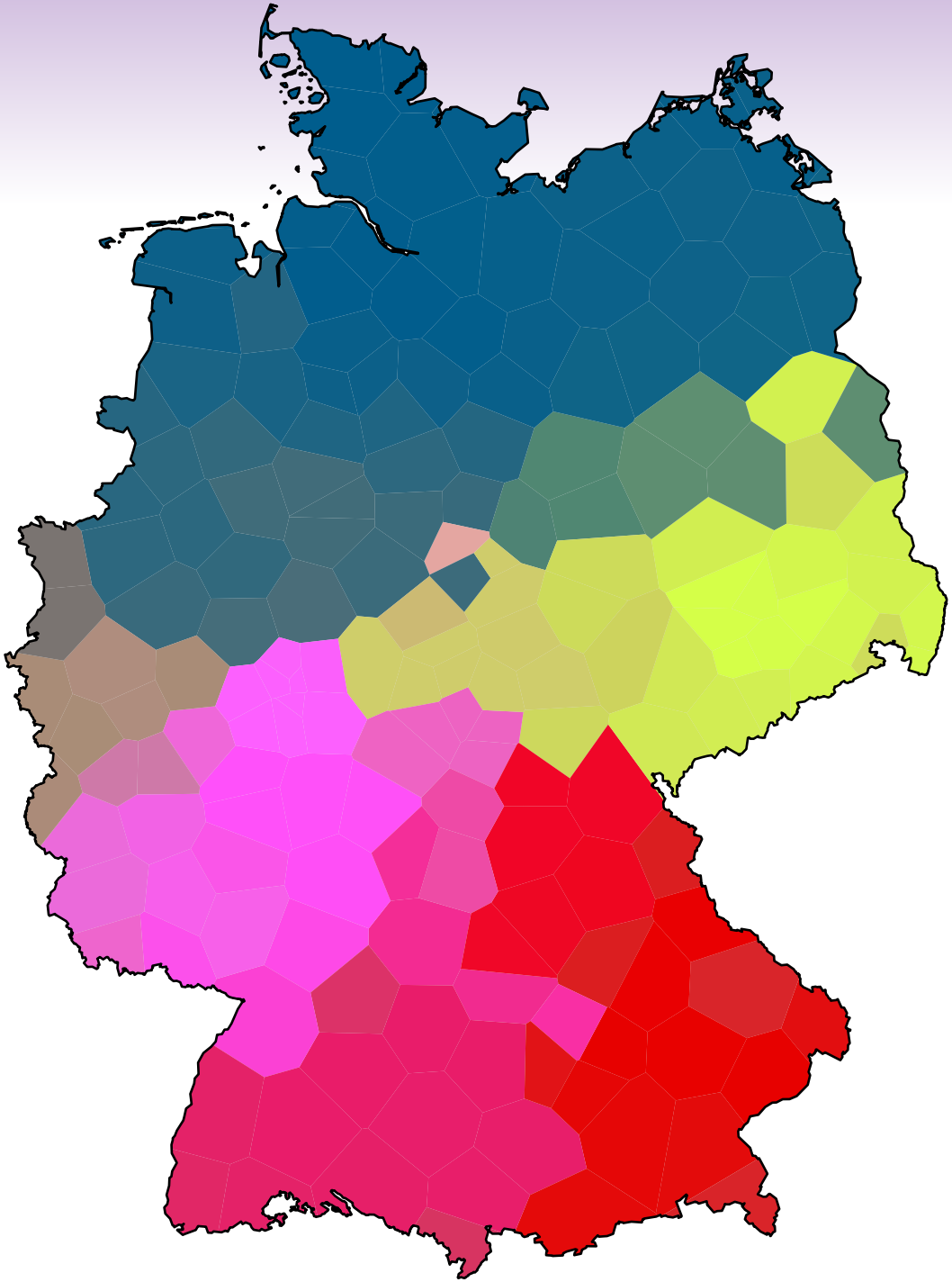


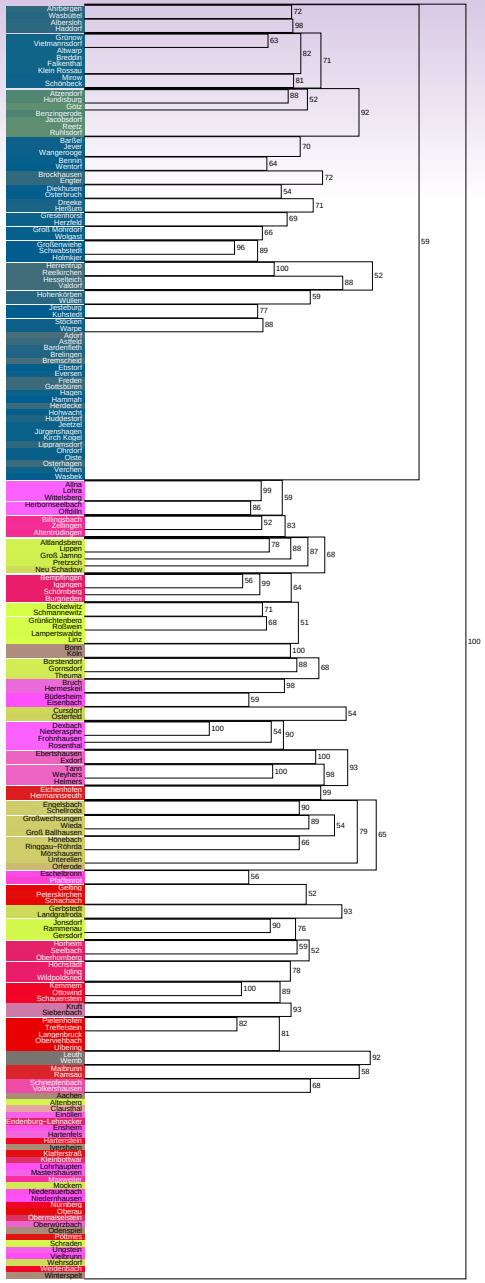
Composite dendrograms shows groups which appear in more than 50% of the repeated (bootstrapped) clusterings.

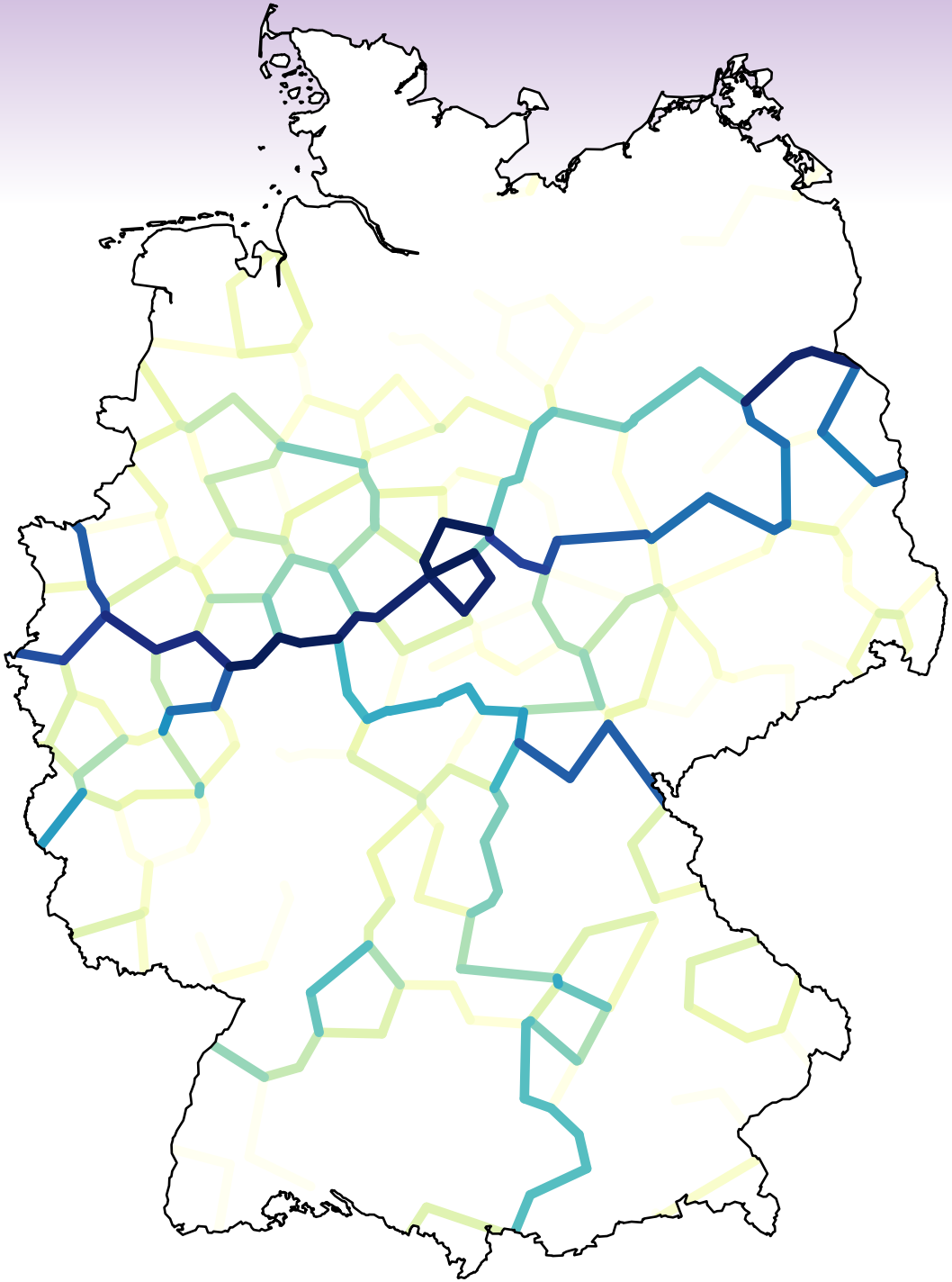
Cophenetic distance

Mean cophenetic distances M' obtained from bootstrapping $M_{1 \leq i \leq n}$:

- Apply (classical) multi-dimensional scaling to M', obtaining 3-dimensional solutions
  - remaining stress $\approx 10\%$
  - correlation with original M' very high, $r = 0.9$
- Interpret dimensions as red, green, blue intensities



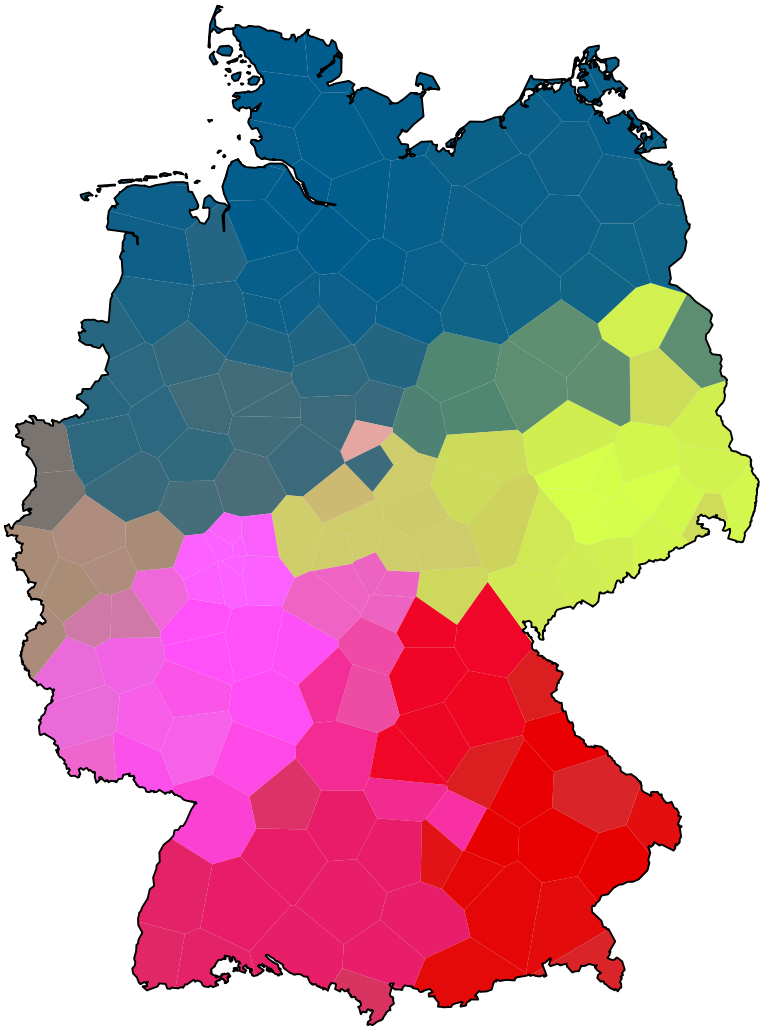




Correlation of Two Solutions

$r = 0.997$

clustering with noise



bootstrap clustering

Conclusions

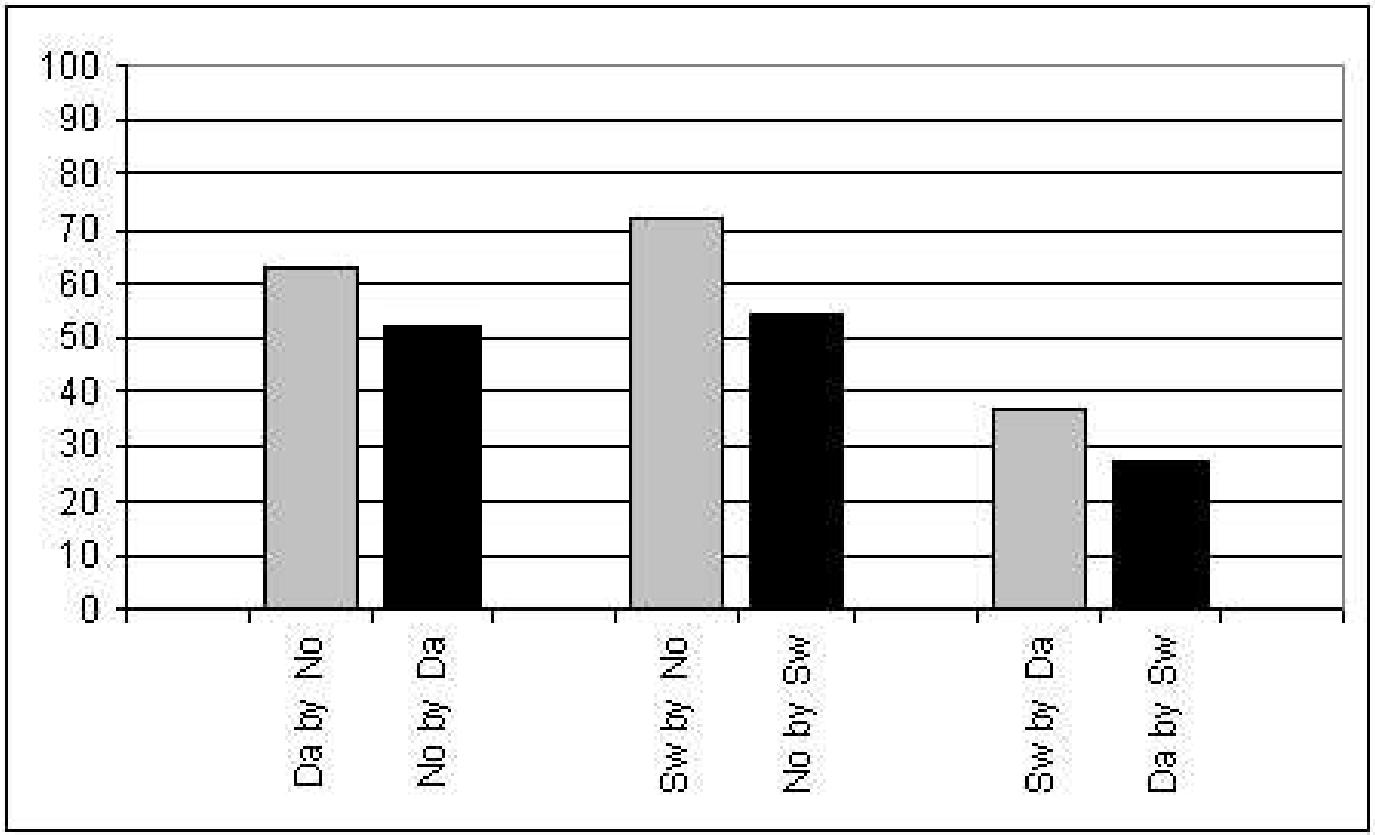
- Stability obtained—both through bootstrapping and through iteration with random small amounts of noise
- Noise-adding procedure needs a noise parameter, bootstrapping number of submatrices to use.
- Noise-adding procedure applicable to single matrices, bootstrapping requires that many be present.
- Choise of clustering technique still important
  But see: `http://www.let.rug.nl/~kleiweg/kaarten/MDS-clusters.html`

Background

Scandinavian “Semicommunication”

- Scandinavians—e.g., Swedes and Danes—hold conversations
  - where each speaks his own language, the Swede Swedish, and the Dane Danish
  - comprehend one anothers’ languages, but imperfectly and asymmetrically
    Danes understand Swedes better than *vice versa*
- Haugen 1966: “semicommunication”, Braunmüller, ca. 2004 “receptive multilingualism”
- Research has focused on attitudes and experiences als explanatory factors
- What about linguistic structure?

Comprehension



Sources: Maurud (1976), Bø (1978), Delsing & Åkesson (2005)

Idea

Danish	j	ɑ	i	
Swedish	j	ɑː	g	

Danish	l	ɑ	ŋ	ʔ
Swedish	l	ɔ	ŋ	#

Swedish problem: map Danish to Swedish
Danish problem: map Swedish to Danish

Map Foreign to Native

Conditional Entropy

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log_2 p(x|y)$$

$$H(\text{Native}|\text{Foreign}) = -\sum_{n \in N, f \in f} p(n,f) \log_2 p(n|f)$$

Given foreign words, how hard is it to map to native words?

See Unity & Diversity project proposal (Nathan Vaillette's section)

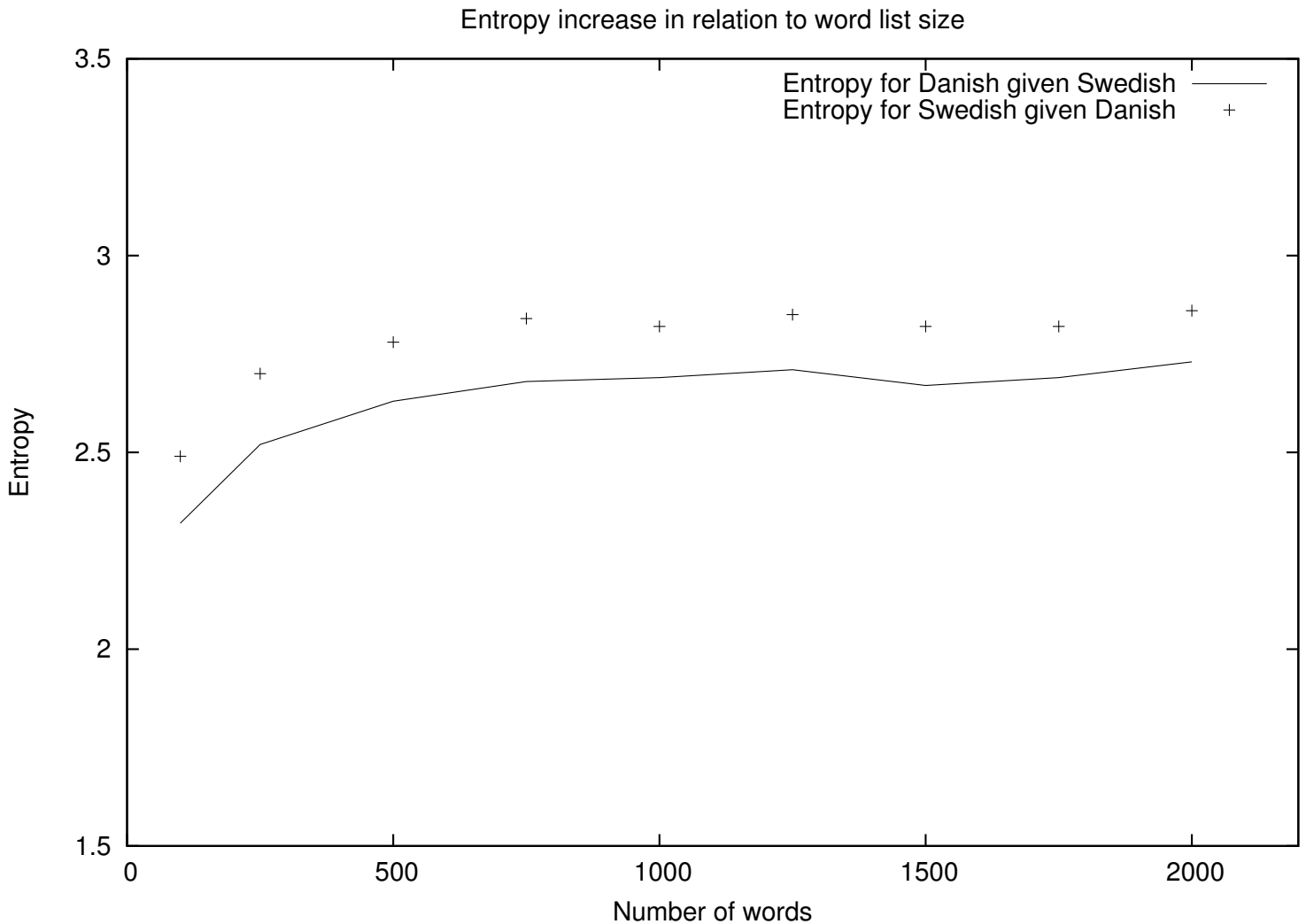
Calculations

Danish	j	ɑ	i	
Swedish	j	ɑː	g	

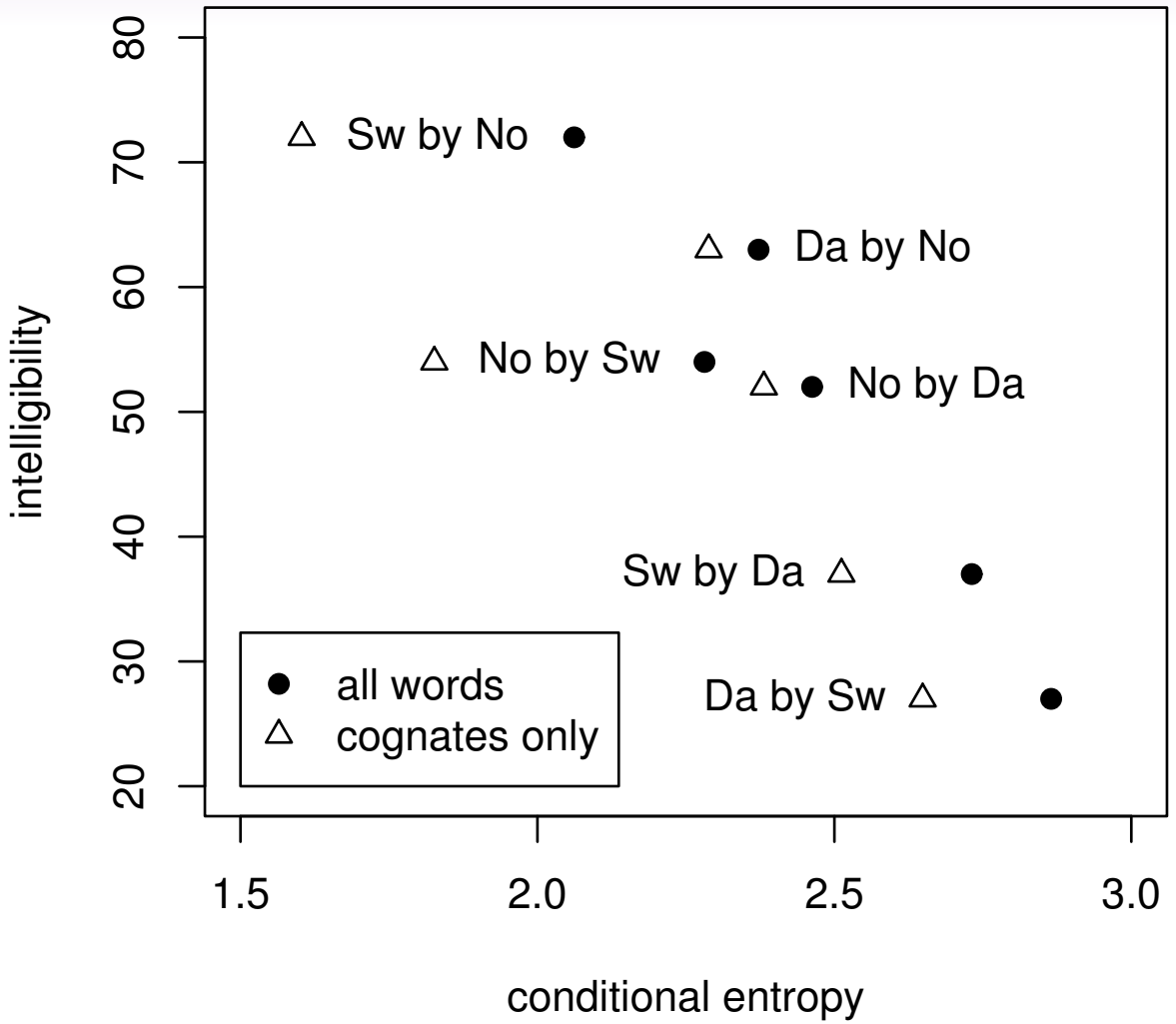
Danish	l	ɑ	ŋ	ʔ
Swedish	l	ɔ	ŋ	#

Except for $p(S|\text{ɑ}_{\text{Dane}})$, all conditional probabilities (mappings) are certain, $p(n|f) = 1, log_2(p(n|f)) = 0$, contributing nothing to entropy.

How Much Data for Estimation?



Intelligibility



Conclusions and Discussion

- Intelligibility correlates negatively, and nearly perfectly with the conditional entropy of the phoneme mapping.

- Only six data points—pairs of Scandinavian languages —More needed!
- Technical refinements possible, but difficult: contextual sensitivity, special status of identity mapping, phonetic detail
- Dutch varieties in sights!

Conclusions and Discussion

- Intelligibility correlates negatively, and nearly perfectly with the conditional entropy of the phoneme mapping.

- Only six data points—pairs of Scandinavian languages
  —More needed!
- Technical refinements possible, but difficult: contextual sensitivity, special status of identity mapping, phonetic detail
- Dutch varieties in sights!

Vision: Variation in the Aggregate

- Analyse entire varieties rather than single features!
- In fact most variationist linguistics—dialectology and sociolinguistics—aggregate but less aggressively.
- Large-scale aggregation protects against small data sets, missing data, contradictory features.
- Theory is tighter because range of hypotheses is more limited.
- Generalizations available at aggregate level, e.g. that variation distances correlates positively but sublinearly over geography.

See `http://www.let.rug.nl/nerbonne/papers/`

Comments welcome!

Thanks for your attention!

Questions?