Reconciling Conflicting Fieldworkers' Reports: Lowman vs. McDavid

John Nerbonne and Peter Kleiweg

University of Groningen

Oct. 9, 2003

New Ways of Analyzing Variation 2003

University of Pennsylvania

RuG

Structure of Talk

Thesis: Dialectometry can Reconcile Conflicting Reports

- Problem of conflicting reports
- “McDavid isoglosses”
- Normalizing distance measures
  - Focus on application to lexical differences
  - Application to pronunciation *better*, but more complicated to present.
- Evaluating results
- Conclusions and recommendations

RuG

LAMSAS

Linguistic Atlas of the Middle and South Atlantic States

- *"If the sun comes out after a rain, you say the weather is doing what?"*
  - *clearing up*
  - *fairing off* $[\ldots 40$ variants$]$
- 1162 interviews conducted 1933–1974
  - 71% of data collected by Guy Lowman 1933–1941
  - 25% of data collected by Raven McDavid 1939–1968
- Digitized data avail. from Bill Kretzschmar
- Records of lexical choice and pronunciation

How Fieldworkers' Reports Conflict

- Fieldworkers can confound data in subtle ways
- Inherent problem in analysis of historical data
  - Encouraging/Discouraging about eliciting alternatives
    - ∗ Infects lexical data as well
  - Transcription practices
    - ∗ Lowman/McDavid differed (*LAMSAS Handbook*, p.127)
    - ∗ Incl. frequent material (corrected relative freq. below)

diacritic	**example**	**L**	**McD**	**Tot. Token Freq.**
fronting	[ɔ˂]	0.30	0.70	33,206
raising	[əˆ] (IPA: [ə̝])	0.35	0.65	54,069

- Not eliminable in contemporary data
  - Instrumental analysis obviates some problems

RuG

Lexical Distance à la Seguy ’71

Site	**Vocabulary Item**				
	dog	*hat*	*horse*	*toilet*	*smallest finger*
Brownsville	*dog*	*hat*	*horse*	*bathroom*	*pinkie*
White Plain	*dog*	*cap*	*horse*	*bathroom*	—

1. Ignore items for which data is missing (*smallest finger*)
2. Distance is $(1 - o)$, where o is proportional overlap
   - distance(Brownsville, White Plain) $= 0.25$
3. Seguy used number of different items, we use proportion
4. Refinement for multiple responses (Nerbonne & Kleiweg, 2003)
5. Refinement weighing infrequent overlap (Goebl, 1982/1984)

RuG

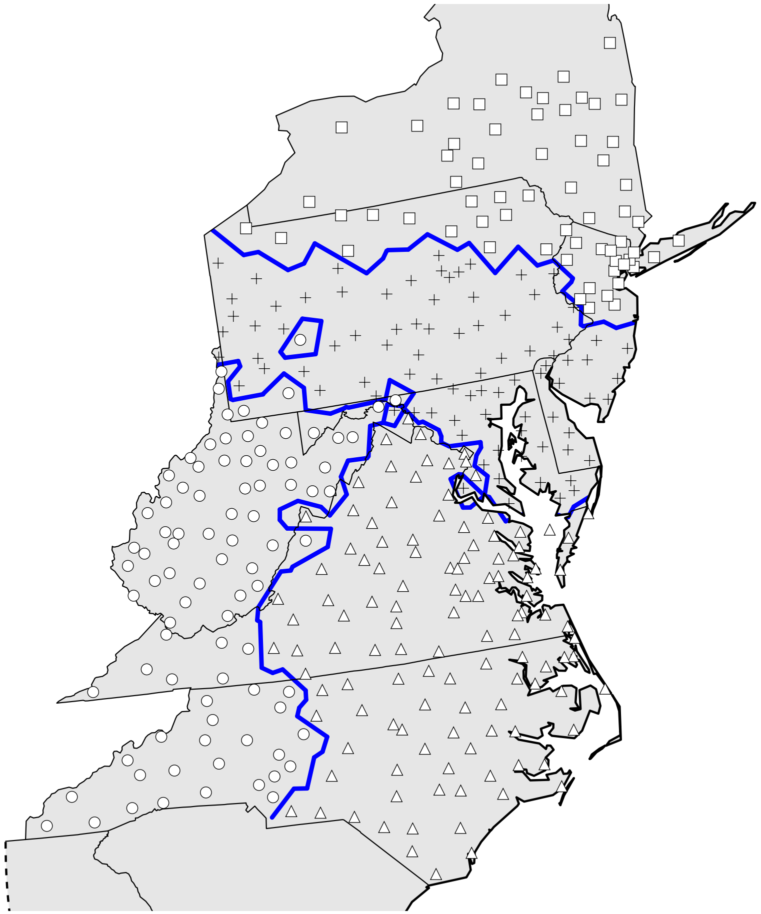
Problem: Close Variants

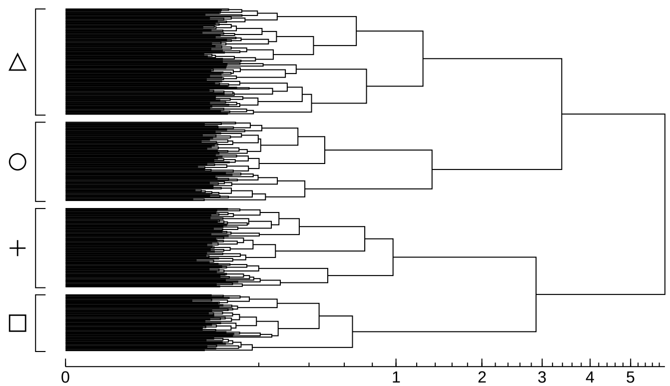
- *fair off, fairing, fairing off, faired off, fairs off, ...*
- Solution: use Porter stemming (poor man's lemmatizer)

```
a hundr year    a hundred year
a hundr year    a hundred years

blew    blew
blew    blewed

ceas    cease
ceas    ceased
ceas    ceases
ceas    ceasing
```

RuG

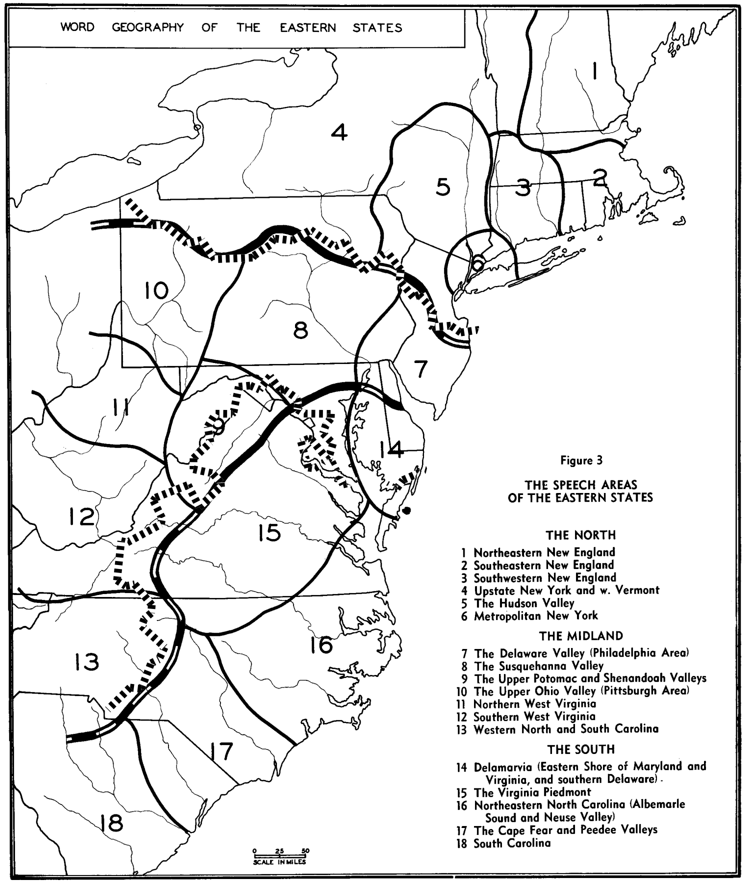
Results on Lowman's data



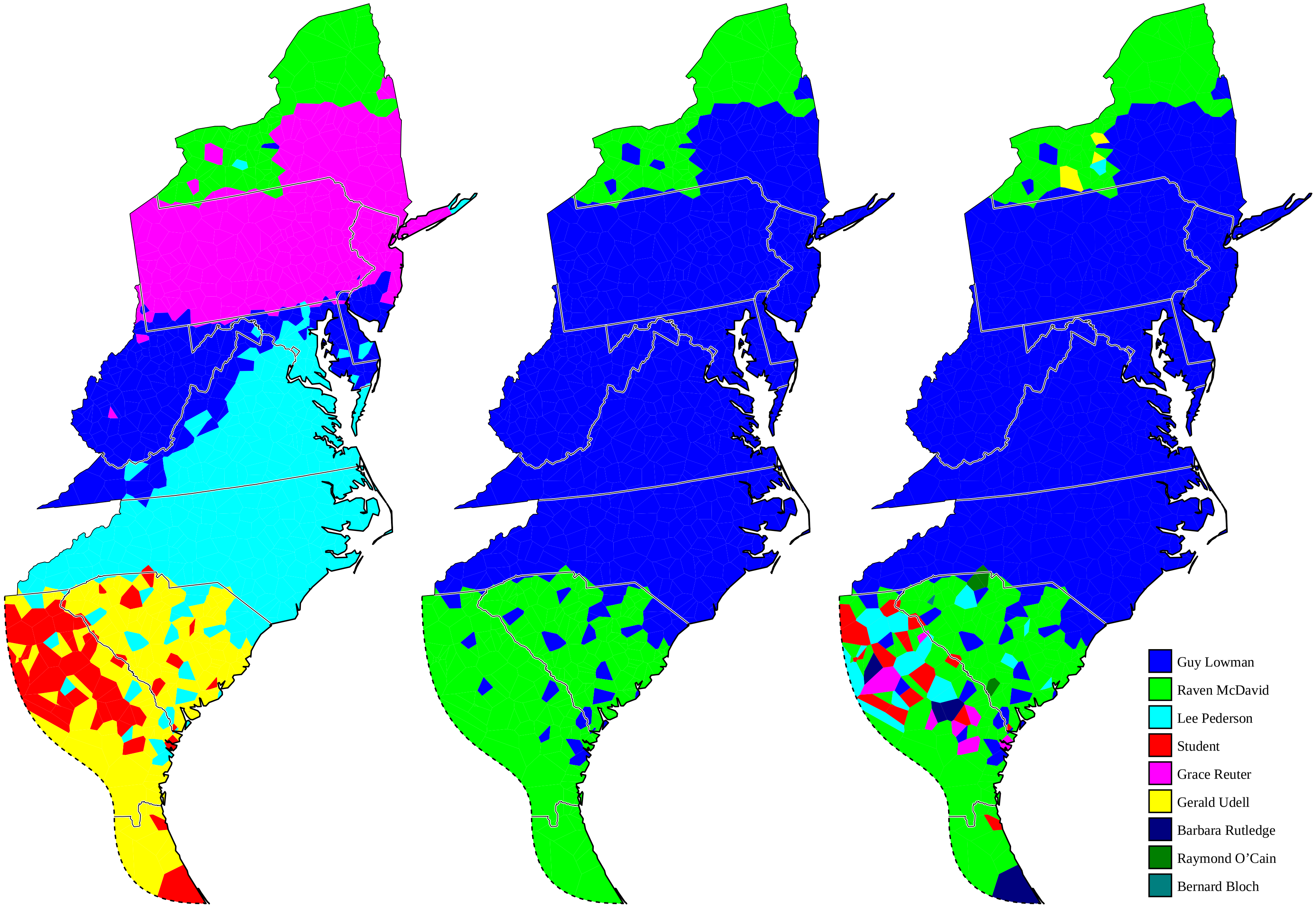


RuG

Fit to Kurath



So where's the conflict?



informants, 6 clusters (left), 2 clusters (middle), fieldworkers (right)

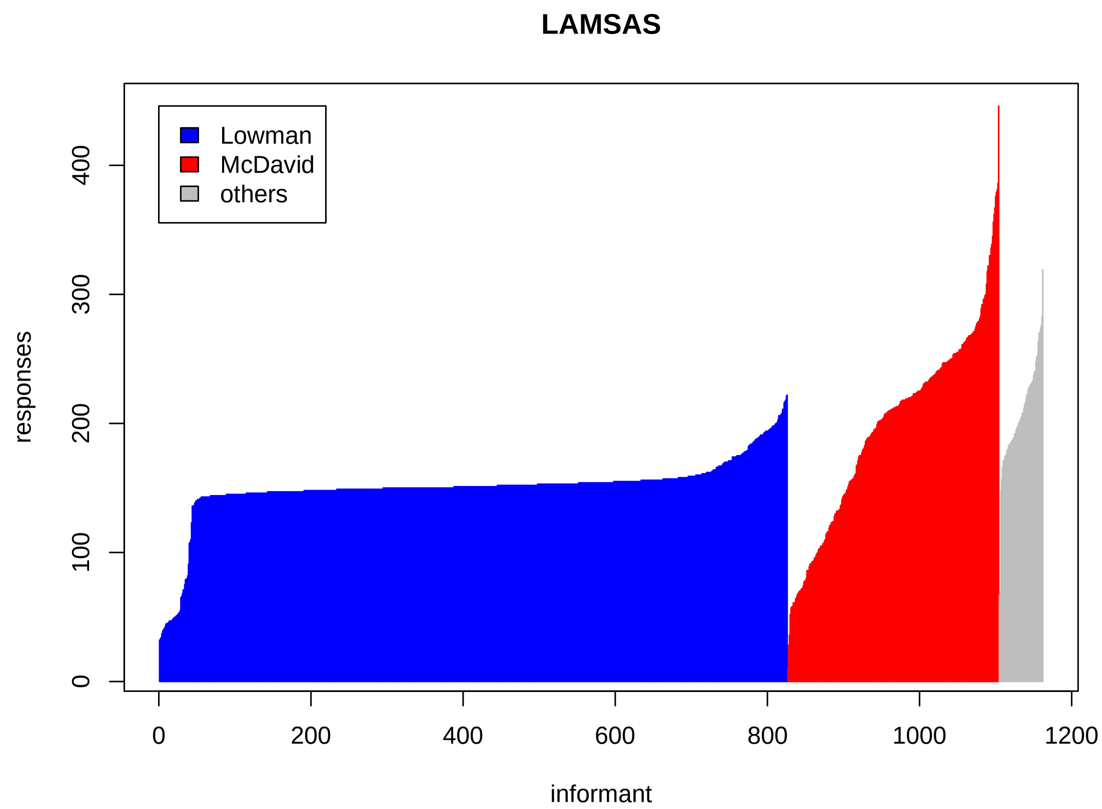
Lowman, McDavid et al.

Fieldworker	Number of Interviews	Number of Responses	Mean Responses/ Interview	SD Responses/ Interview
Lowman	826	123990	150.1	25.3
McDavid	278	54855	197.3	76.8
others	58	12057	207.9	43.9
Totals	1162	190902	164.3	49.6

Lowman (& others) encountered “no-response” for 10% of items; McDavid for 15%. Significantly distinct ($p < 0.05$, binomial w. $n_1 = 826$, $n_2 = 278$).

Preliminary focus was therefore on Lowman data — 71%

Differing Practices



Elicitation confound

Idea: differences are error due to fieldworker

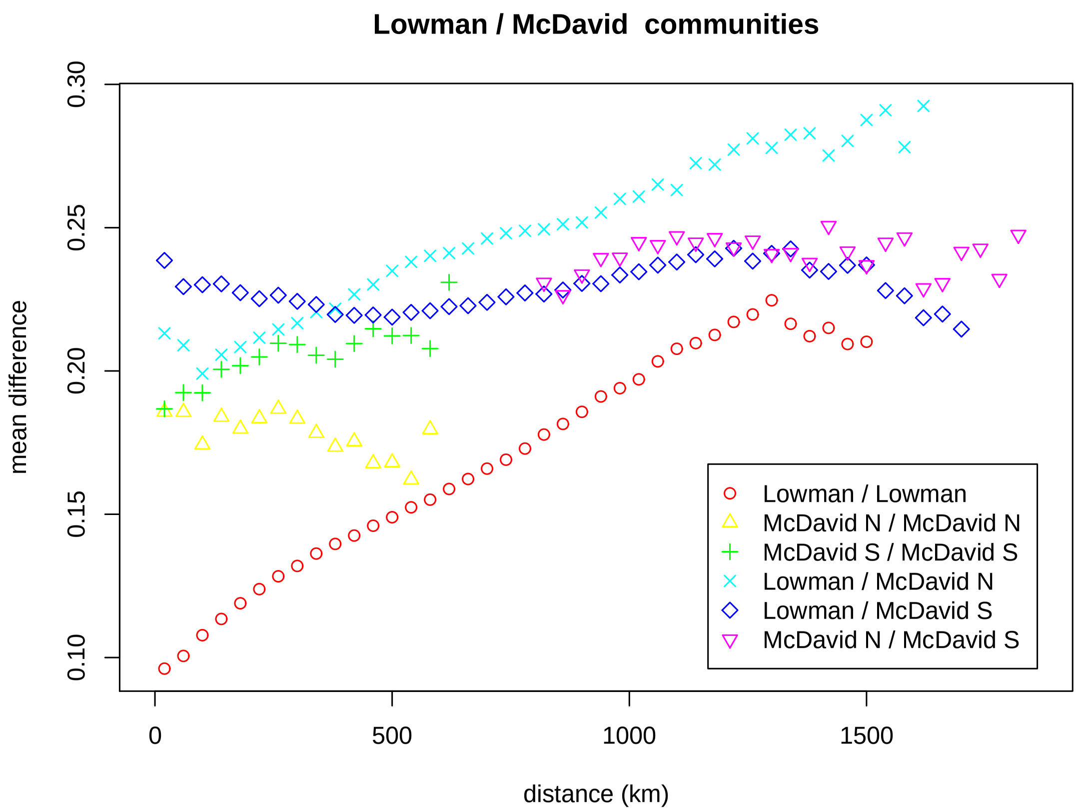
Normalize the measurements, i.e., to express distances as z-scores,

$$z_{\text{a,b}} = (x_{\text{a,b}} - m)/\text{SD}$$

where each linguistic distance is normalized according to the mean (m) and standard deviation (SD) of the respective fieldworkers/fieldworker-areas.

Complication: Linguistic distance is clearly dependent on geographic distance

Are there differences due to fieldworker?



Refinement: normalize separately per 50-km “bin” (and per fieldworker pair)

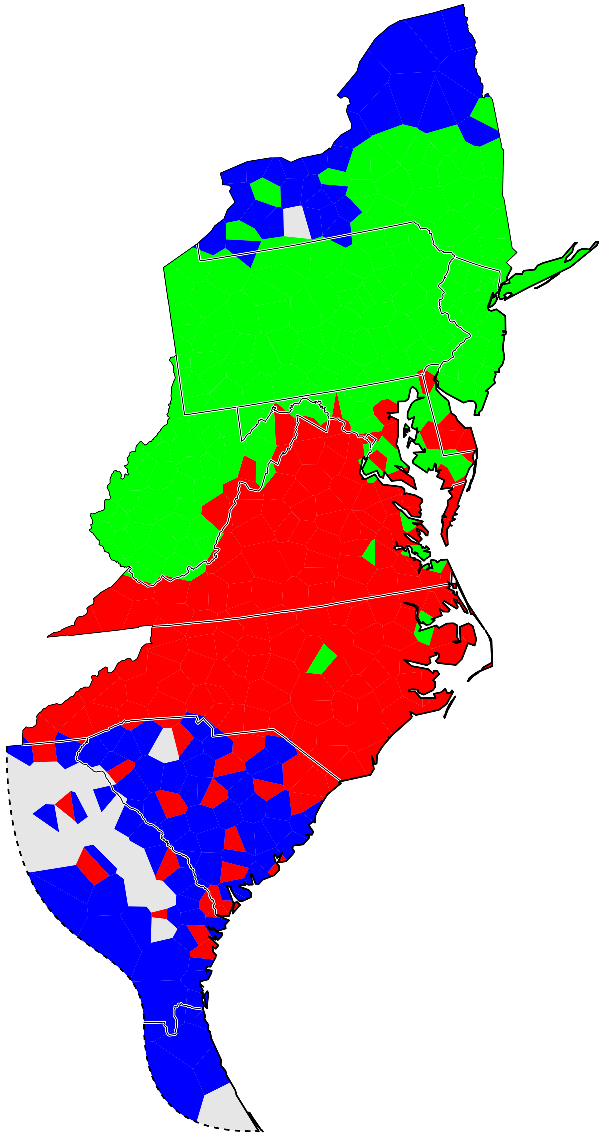
Reintroducing Geography

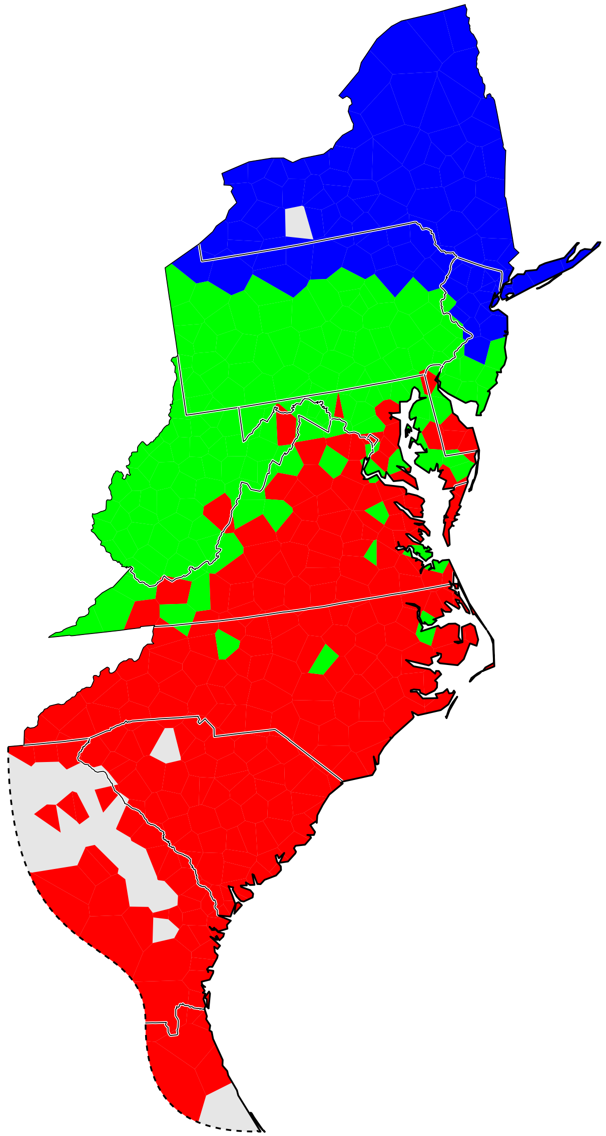
After we normalize in 50-km bins, the effect of geography is eliminated—even though geographic and linguistic distance are highly correlated.

We reintroduce it, now aggregated over all fieldworker pairs.

$$
\begin{aligned}
z_{\text{a,b}} &= (x_{\text{a,b}} - m_{\text{fw}_1,\text{fw}_2})/\text{SD}_{\text{fw}_1,\text{fw}_2} \\
z'_{\text{a,b}} &= (z_{\text{a,b}} \times \text{SD}_{\text{agg}}) + m_{\text{agg}}
\end{aligned}
$$

We now cluster using these corrected, normalized distances.





3-area normalizing, uncorrected lexical (left), corrected lexical (right)

Summary of Procedure

- Three areas Lowman's, McDavid's North, McDavid's South
- Six sorts of distance: $\{$L,McD-N,McD-S$\} \times \{$L,McD-N,McD-S$\}$
- Each class of distance normalized w.r.t. its own mean, sd, considered in separate 50-km. "bins"
- Final correction in normalization reintroduces aggregate geographic effect (again per bin, aggregated over **all** fieldworker pairs.
- Lexical measurements as in Nerbonne/Kleiweg CHUM article:
  - all concepts common to worksheets examined
  - elimination of 11 least occurring tokens
  - (at first) no inverse-frequency weighting à la Goebl
- Distances clustered via Ward's method, which tends to create large, evenly sized clusters

Goebl's "Weighted Similarity"

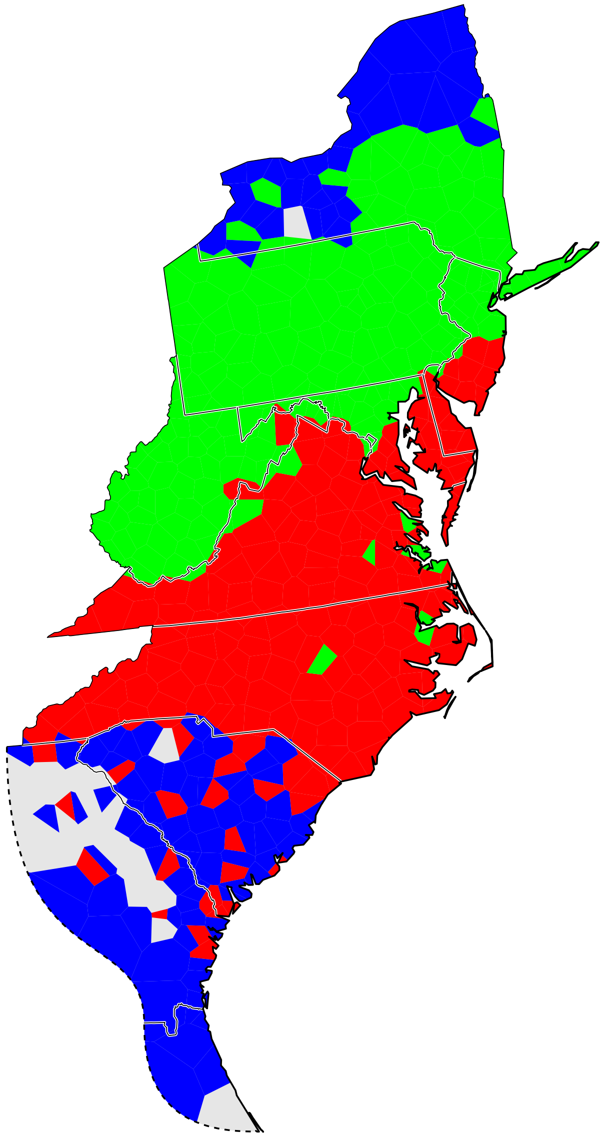
Goebl (1983) introduced *gewichteter Identitätswert*, a weighted similarity, counting overlap in infrequent words more heavily.

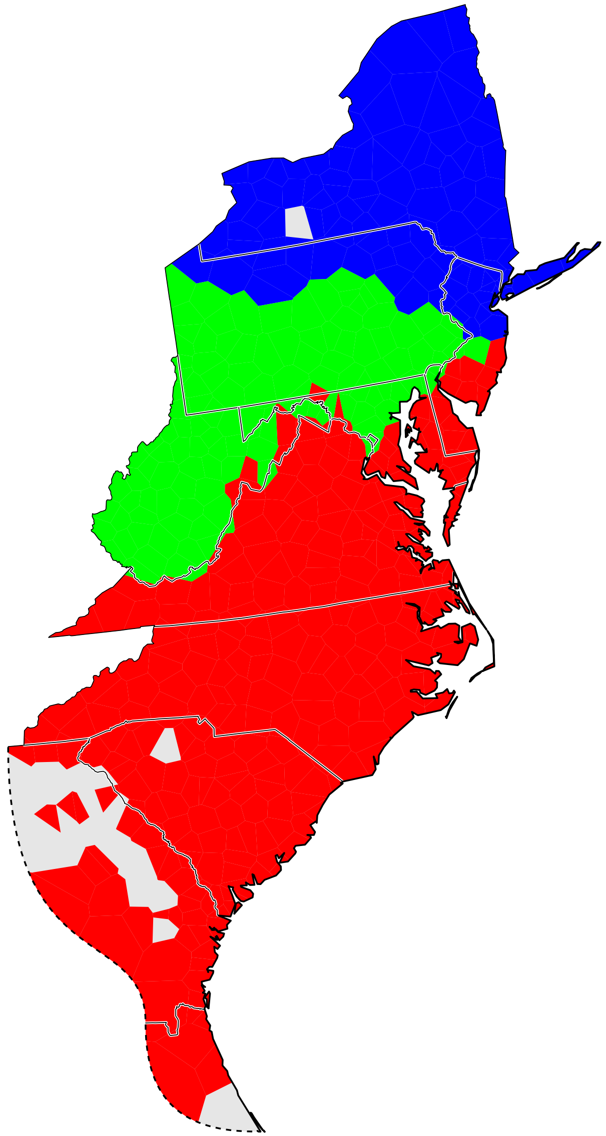
For concept i with n responses $w_1^i, w_2^i, \ldots, w_n^i$, we let $f(w_j^i)$ be the frequency of w_j as repsonse to query about i.

$$S(w, w') = 1 - \frac{f(w_j^i) - 1}{n \cdot w}$$

where Goebl forsees experimentation with w, always $= 1$ here

This *emphasizes* rather than ignores infrequent words. We try $1 - S(w, w')$ as an alternative distance measure.





3-area normalizing, Goebl-weighting, uncorr. lex. (left), corr. lex. (right)

Local Incoherence

Measures how well analyses reflect tendency local varieties to be similar.

- Fundamental dialectological postulate: geographical proximity correlates with linguistic similarity.
- Basic Idea: sum of geographic distances to linguistically closest sites.
  - Summed over all collection sites.
  - Closest varieties weighted as more important.
  - Variables are too highly collinear (geographic and linguistic distance)
- Scale in $\Re+$, 0 is optimal
- Depends on area, distribution & density of sites.
- Varying geography not reflected

Correction results in small drop in LI for unweighted lexical measurements, large drop in weighted measurements.

Cautions

- Results shown here were selected for the ability to show proof of concept.
- Others less convincing, e.g., at finer levels of clustering
- Graph of linguistic distance vs. geographic distance suggests a finer correction (also correcting for contribution of geography)
  - Several experiments attempted
  - Very poor results
  - Variables are too highly collinear (geographic and linguistic distance)

Conclusions

Thesis: Dialectometry can Reconcile Conflicting Reports

- Genuine problem of conflicting reports
- “McDavid isoglosses”
- Normalizing distance measures by classes of area-pairs
  - Application to lexical differences demonstrated
  - Fieldworker isoglosses resolved
- Still exploratory

Phonetic Segment Distance

- Phonetics shows how to measure differences in segments, e.g. as city-block distance in *features*

  Example: difference between [i] & [e] much smaller than difference between [i] & [u].

	i	e	u	i-e	i-u
advancement	2(front)	2(front)	6(back)	0	4
high	4(high)	3(mid high)	4(high)	1	0
long	3(short)	3(short)	3(short)	0	0
rounded	0(not rounded)	0(not rounded)	1(rounded)	0	1
				1	5

- Diacritics [ĩ,e:,ə˞] can also be taken into account
- Vieregge-Cucchiarini system used, also Almeida-Braun
- Chomsky-Halle (SPE) system less useful (clever features for making rules compact)

Levenshtein Distance

Idea: *lift* segment distance to sequence distance.

Standard American	sɔəgIrl	delete r	0.5
	sɔəgIl	replace I/ɜ	0.1
	sɔəgɜl	insert r	0.8
Bostonian	sɔrəgɜl		
		Sum distance	1.4

- L-distance $=^{df}$ *minimal cost* of operation to rewrite one string to another.
- Insertions and deletions compare segment to silence

Levenshtein Distance aka edit distance, string distance also used in CL (bilinguial alignment), bioinformatics, software engineering.

`http://www.let.rug.nl/~kleiweg/lev/`

Problem: multiple responses

- *clear, fair off* vs *changing, clear, fair off*
- Sol'n: lift distance measure from strings to string *sets*

$$d(C) \doteq \sum_{c \in C} d(c), \quad \text{where } C \text{ is a set of string pairs}$$

Let C^1, C^2 be first, second projections of C. C COVERS $A \times B$ if, and only if $C \subseteq A \times B$, and $C^1 = A$ and $C^2 = B$.

We shall seek the minimum cost COVER

$$d(A, B) \doteq \frac{1}{|C|} \text{Min } d(C), \quad \text{where } C \text{ covers } A \times B$$

Problem: Multiple Responses

Illustration: $A = \{a, b, c\}$, $B = \{a, c, d\}$

then $C = \{\langle a, a\rangle, \langle b, d\rangle, \langle c, c\rangle\}$ covers $A \times B$,

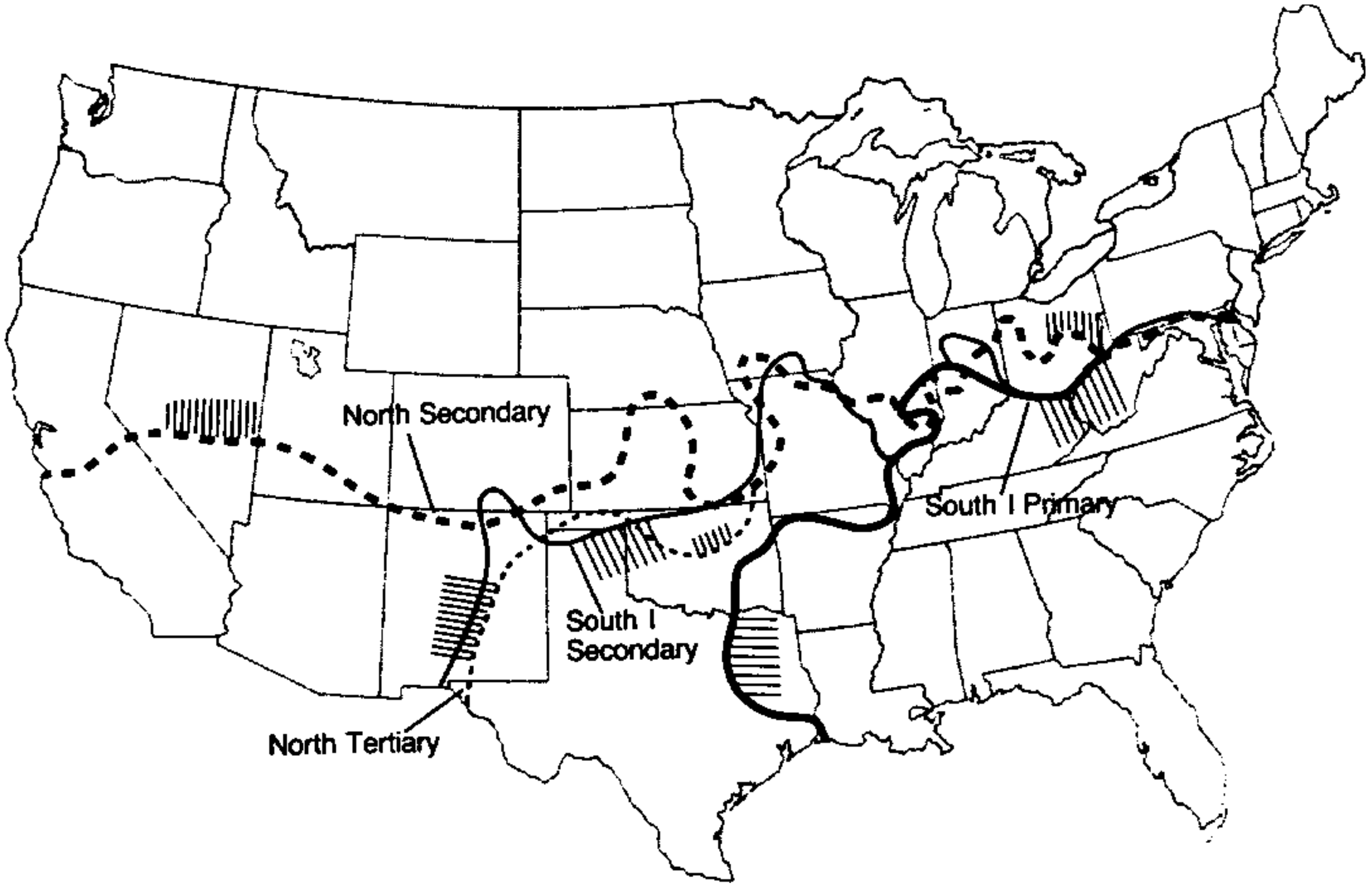
even though $|C| = 3$, while $|A \times B| = 9$.

Since $d(a, a) = d(c, c) = 0$, $d(A, B) = 1/3 \cdot d(b, d) = d(b, d)/3$

Likewise

$$d(\{a\}, \{b\}) = d(a, b)$$
$$d(\{a\}, \{b, c\}) = \frac{1}{2} \cdot (d(a, b) + d(a, c))$$

Carver’s North/South Division



Fundamental Dialectological Postulate

- Neighbouring varieties are linguistically similar
  - Exception: border areas
  - Exception: some distributed varieties (migration, trade)
- Experience in Dialectometry:
  - Very remote varieties show little correlation linguistic/geographic distance.
  - Therefore uninteresting for choice of measurement.
  - Emphasize closest varieties

Local Incoherence

$$
\begin{aligned}
\mathrm{I}_L &= \frac{1}{n}\sum_{i=1}^{n}\frac{D_i^L - D_i^G}{D_i^G} \\
D_i^L &= \sum_{j=1}^{k} d_{i,j}^L \cdot 2^{-0.5j} \\
D_i^G &= \sum_{j=1}^{k} d_{i,j}^G \cdot 2^{-0.5j}
\end{aligned}
$$

$d_{i,j}^L$, $d_{i,j}^G$: geographical distance between locations i en j

$d_{i,1\cdots n-1}^L$: sorted by increasing linguistic difference

$d_{i,1\cdots n-1}^G$: sorted by increasing geographical distance