A Dialectological Yardstick

John Nerbonne and Peter Kleiweg

Rijksuniversiteit Groningen

Quantitative Linguistics Conference 2003
The University of Georgia
May 29, 2003

Dialectometry: Measuring linguistic “distances”

- Seguy (1971), Goebl (1982, 1984, ...) alternative foundation for dialectology — inverse overlap in linguistic features, interpreted categorically.

- Kessler (1995), Nerbonne et al (1996, 1999, ...), Heeringa (2003) measure pronunciation distance using (metric) sequence distance measures.

- Palander et al. (2003), Speelman et al. (2003) frequency profiles of linguistic alternatives

Problem: Validity—When are measurements right?

- Many computational, mathematical alternatives

- Often no expert consensus, sometimes no opinion

Categorical 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*	—

- Ignore items for which data is missing (*smallest finger*)
- Distance is $(1 - o)$, where o is proportional overlap
  - distance(Brownsville, White Plain) $= 0.25$
- Number of different items or proportion?
- Treatment of multiple responses, close variants (*clear/clears*)
- Frequency weighting à la Goebl?

Porter Stemming

- Poor man's lemmatizer (used in Information Retrieval)
- Public Domain versions available

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

abat     abated
abat     abating

blew     blew
blew     blewed

ceas     cease
ceas     ceased
ceas     ceases
ceas     ceasing
```

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 response to query about i.

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

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

This *emphasizes* rather than ignores infrequent words. We try $1 - S(w_j^i, w_{j'}^i)$ as an alternative distance measure.

Nerbonne, Heeringa et al. on Pronunciation Differences

- Phonetics describes sounds using *features*, allowing distance measurement, e.g., as city-block distance

  Example: d([i],[e]) $<$ d([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

- Which feature system? Vieregge-Cucchiarini, Almeida-Braun, Ladefoged, Chomsky-Halle (SPE), ... ?

- City block distance or Euclidean distance? Information-Gain weighting on features?

- Ceiling on segment distance or logarithmic correction?

Sequence Distance

Idea: *lift* segment distance to sequence distance.

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

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

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

Which Measurements are Probative?

- Choice of linguistic probes (material)
- Frequency weightings (Goebl)
- Individual variation (multiple responses)
- Status of inflectional variants (stemming/lemmatizing)
- Choice of phonetic features, distance measures
- Phonetics vs. lexicon vs. other

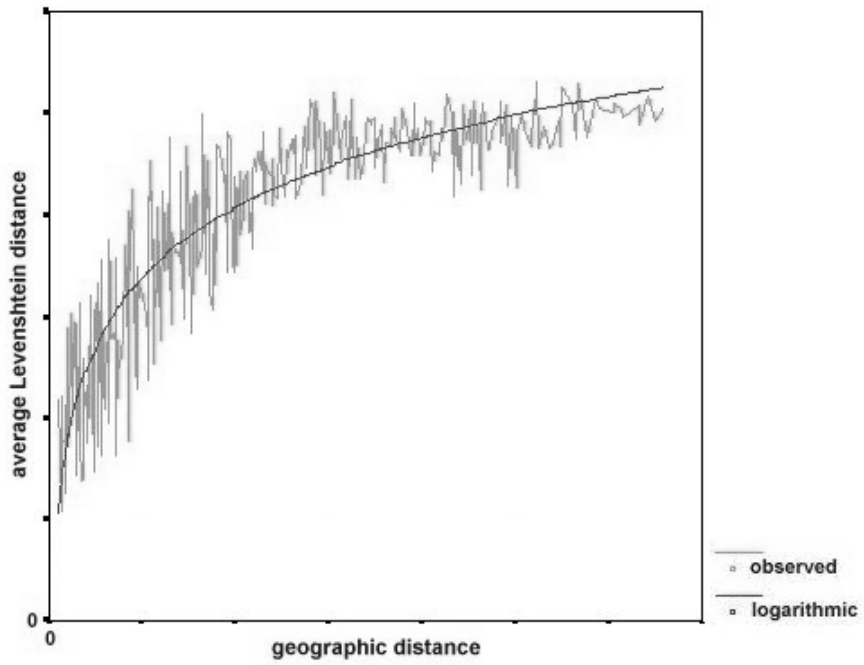
Proposal: prefer measures to maximize local linguistic coherence.

Fundamental Dialectological Postulate

- Neighboring varieties are linguistically similar
  - Exception: border areas
  - Exception: some distributed varieties (migration, trade)
- Campbell: “[...] neighboring languages often turn out to be related.”, referring to Dyen (1956), Sapir (1916)
- Experience in Dialectometry:
  - Very remote varieties show little correlation linguistic/geographic distance.
  - Therefore uninteresting for choice of measurement.
  - Emphasize closest varieties

Need to Ignore Distant Varieties

Phonetic distance as function of geography ($r \approx 0.75$)
—Heeringa & Nerbonne *LVC* 13, 2002



Toward a Measure of Incoherence

Idea: Measure linguistic distance in a number of varieties, then examine how far the closest varieties are (geographically).

$$D_i^L = \sum_{j=1}^{k} d_{i,j}^L$$

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

- Prefer measures which show linguistically closest varieties to be geographically closest, i.e., minimize D_i^L

Refinements

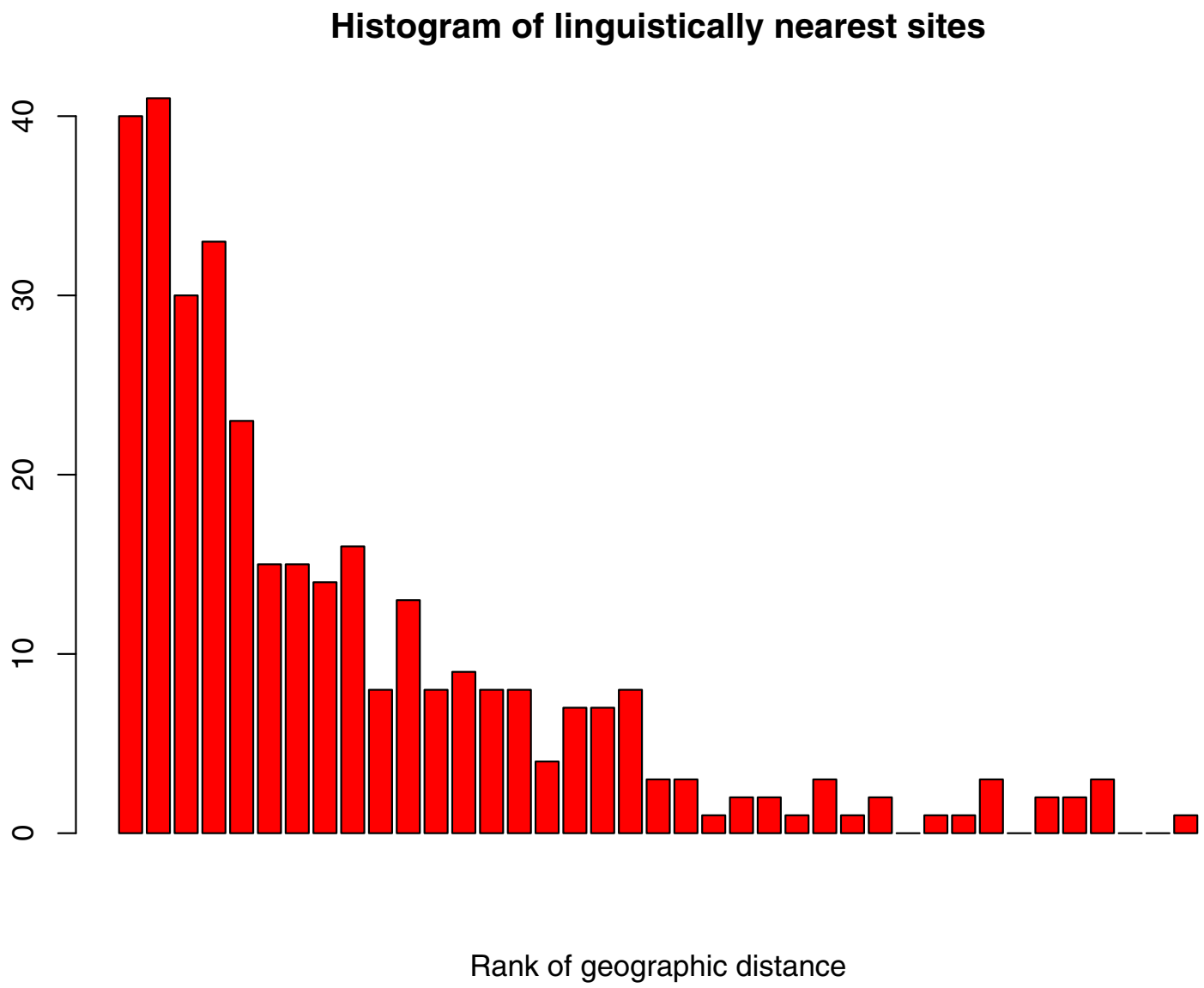
$$D_i^L = \sum_{j=1}^{k} d_{i,j}^L \cdot 2^{-0.5j}$$

1. Limit, e.g., $k = 8$ to avoid letting distant measurements confound local (in)coherence

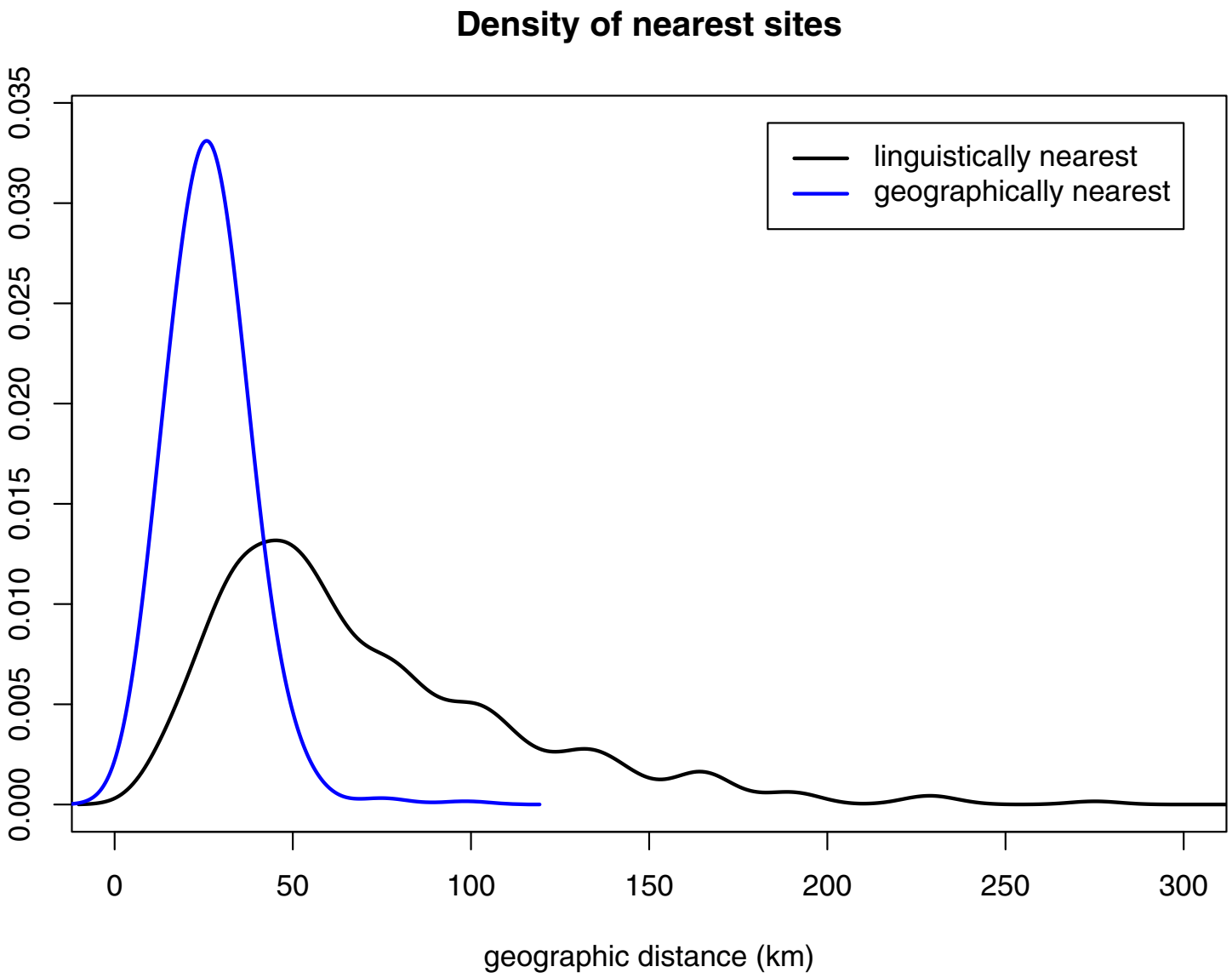
2. Let linguistically more distant measures weigh exponentially less ($\cdot 2^{-0.5j}$)

3. Compare to optimum (still not shown)

Why Limit to 8 Nearest Sites?



Distribution of Nearest Sites



Minimize Local Incoherence (I_L)

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

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

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

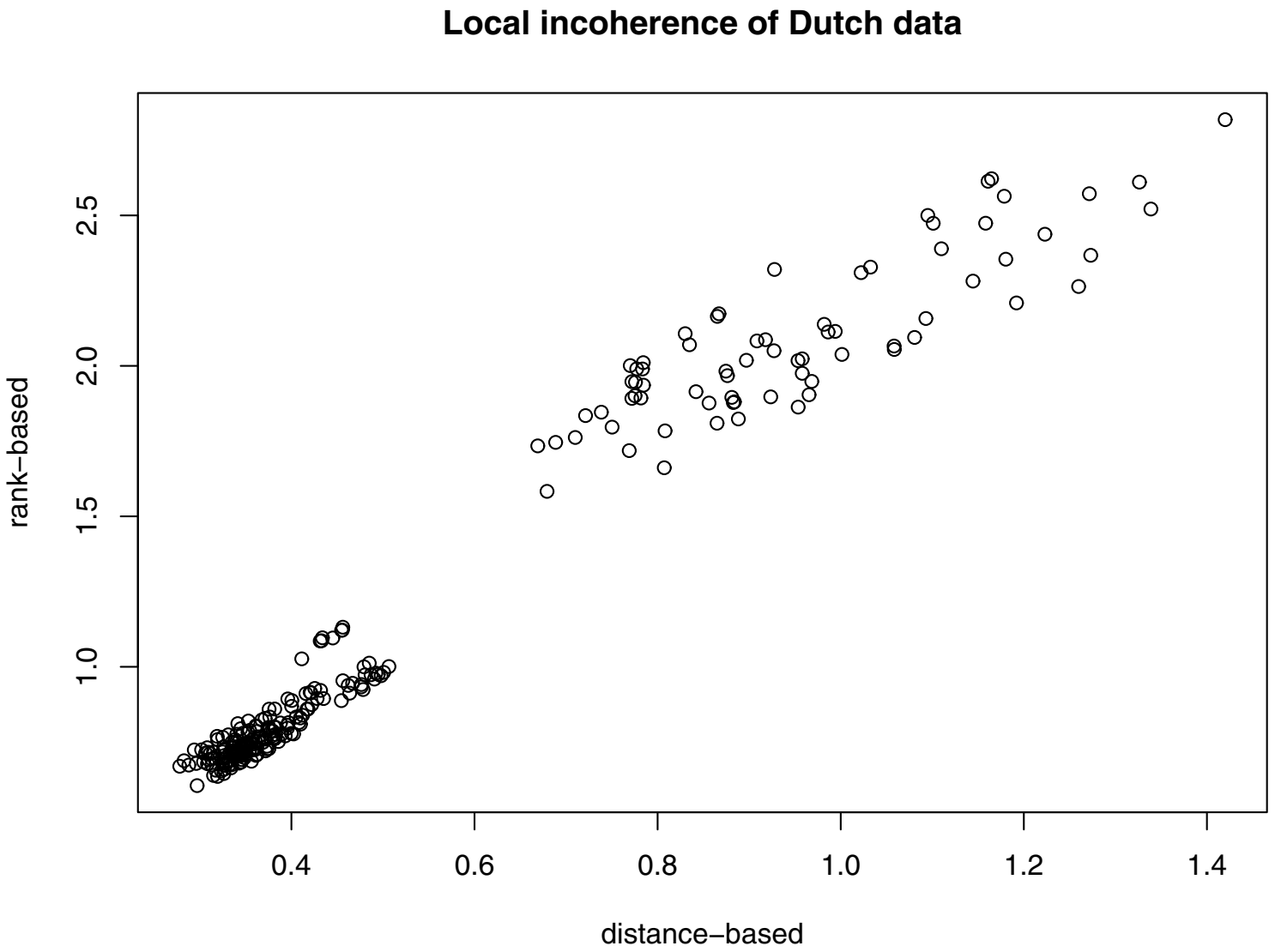
$d_{i,1\cdots n-1}^G$: sorted by increasing geographical distance

Local Incoherence (I_L)

$$\mathrm{I}_L \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{D_i^L}{D_i^G} - 1$$

- Dependent on geographical distribution of fieldwork sites
  - Density of site sampling
  - Informants at same site (dist$= 0$) — noise
- Simple notion of geographic distance used, others possible
- Using geographic distance is preferable to using geographic *ranks* because these vary in real distance

Geographic Distance vs. Ranks (I_L)



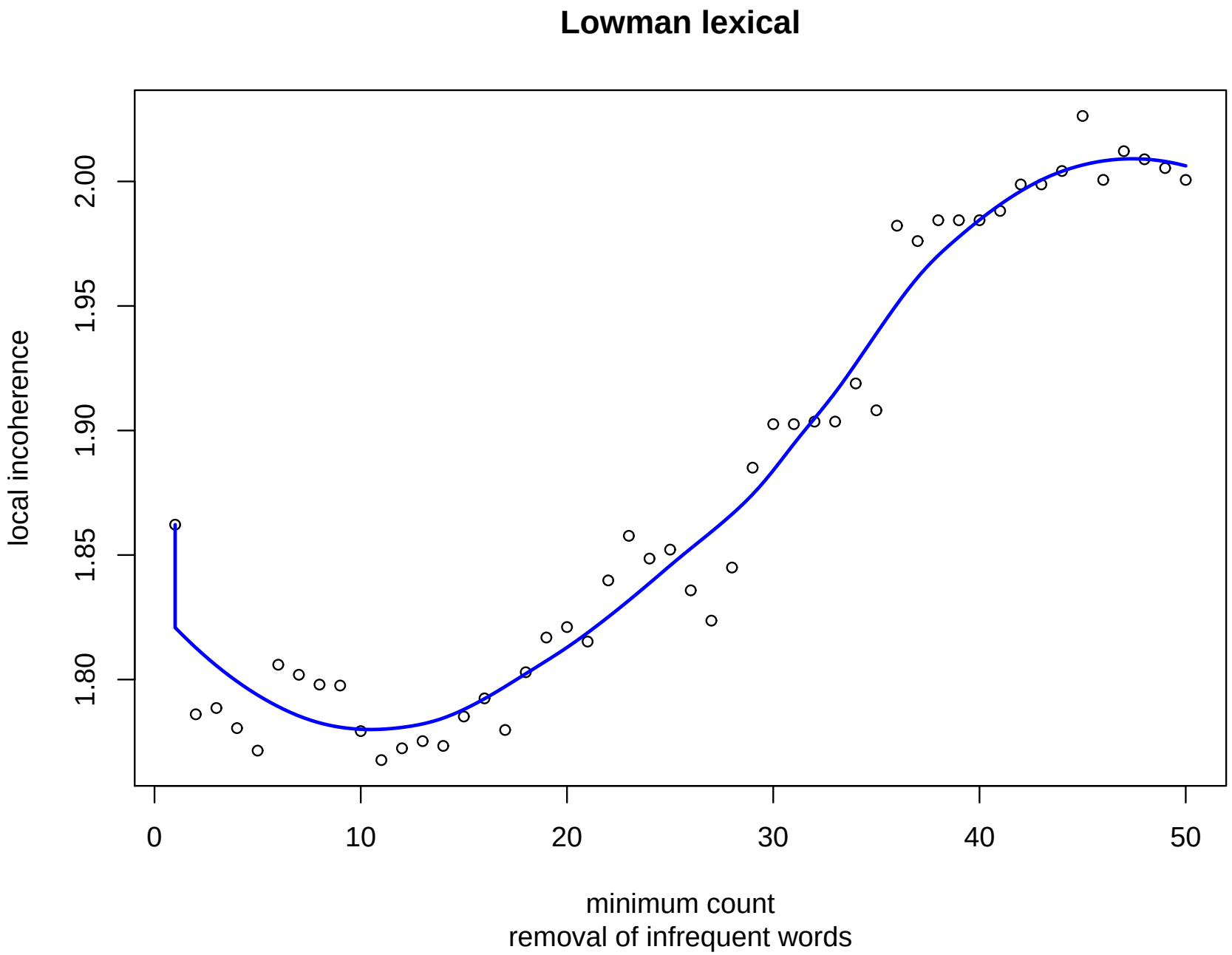
Data from LAMSAS: Linguistic Atlas of the Middle and South Atlantic States

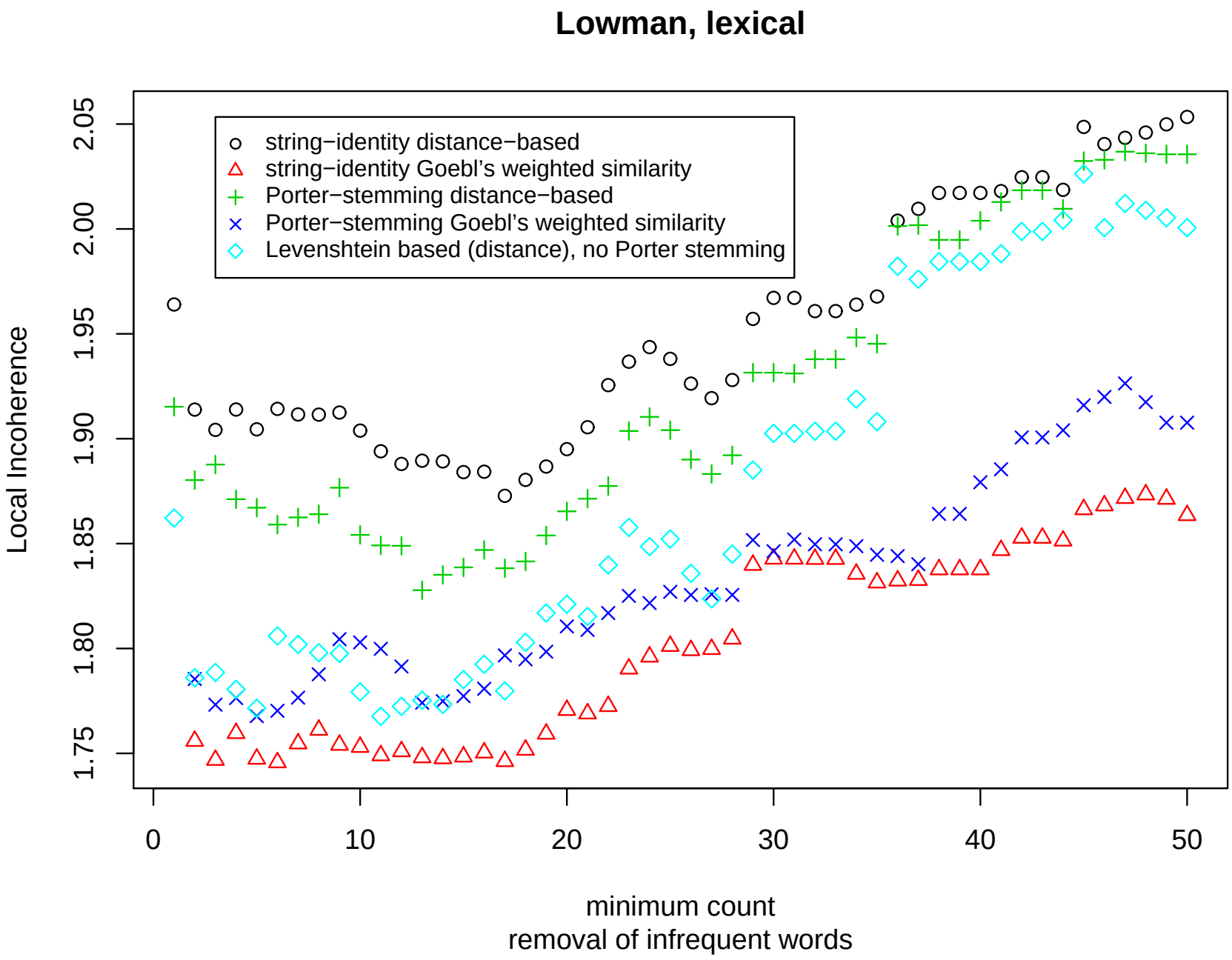
- *"If the sun comes out after a rain, you say the weather is doing what?"*
  - *clearing up*
  - *fairing off* $[\ldots 40$ variants$]$
- 1162 interviews conducted 1933–1974
- 71% of data collected by Guy Lowman 1933–1941
- digitized data avail. from Bill Kretzschmar
- focus on lexical overlap here, just as elsewhere (Kurath, ...)
  - later goal: relation to pronunciation

Focus: Infrequent Words

- Common remark: very infrequent words are *noise*, not evidence of linguistic coherence
  - Carver, *American Regional Dialects*, p.17
- But exactly where should the cut off be?
  - Words that occur twice, three times, ...
  - Words that occur with less than 1% of the frequency of the most frequent words
- Tension between this and Goebl's "Weighted Similarity"

Focus: Infrequent Words




Lowman, lexical
string−identity distance−based
string−identity Goebl's weighted similarity
Porter−stemming distance−based
Porter−stemming Goebl's weighted similarity
Levenshtein based (distance), no Porter stemming
Local Incoherence
1.75
1.80
1.85
1.90
1.95
2.00
2.05
0
10
20
30
40
50
minimum count
removal of infrequent words

LAMSAS Results

Local incoherence

measure	Lowman	LAMSAS
lexical	2.15	2.69
phonetic (symbols)	1.44	1.62
phonetic (features)	1.95	2.00

Conclusions

- Reanalyzing existing atlas materials is “data mining”— search for valuable ores in a huge area
- Wealth of computational techniques now really applicable
  - linguistic level, representation, detail, psychological fidelity, frequency, microvariation, ...
- Need “investigative” techniques
  - But also rigorous validation (see Heeringa, Nerbonne & Kleiweg in *Proc. of Gesellschaft für Klassifikation*, 2002)
- Leading “Dialectological Postulate”—which techniques expose geographic coherence?