Lexical Variation in Lowman's LAMSAS

John Nerbonne

Rijksuniversiteit Groningen

LSA Linguistics Institute

Summer, 2005

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

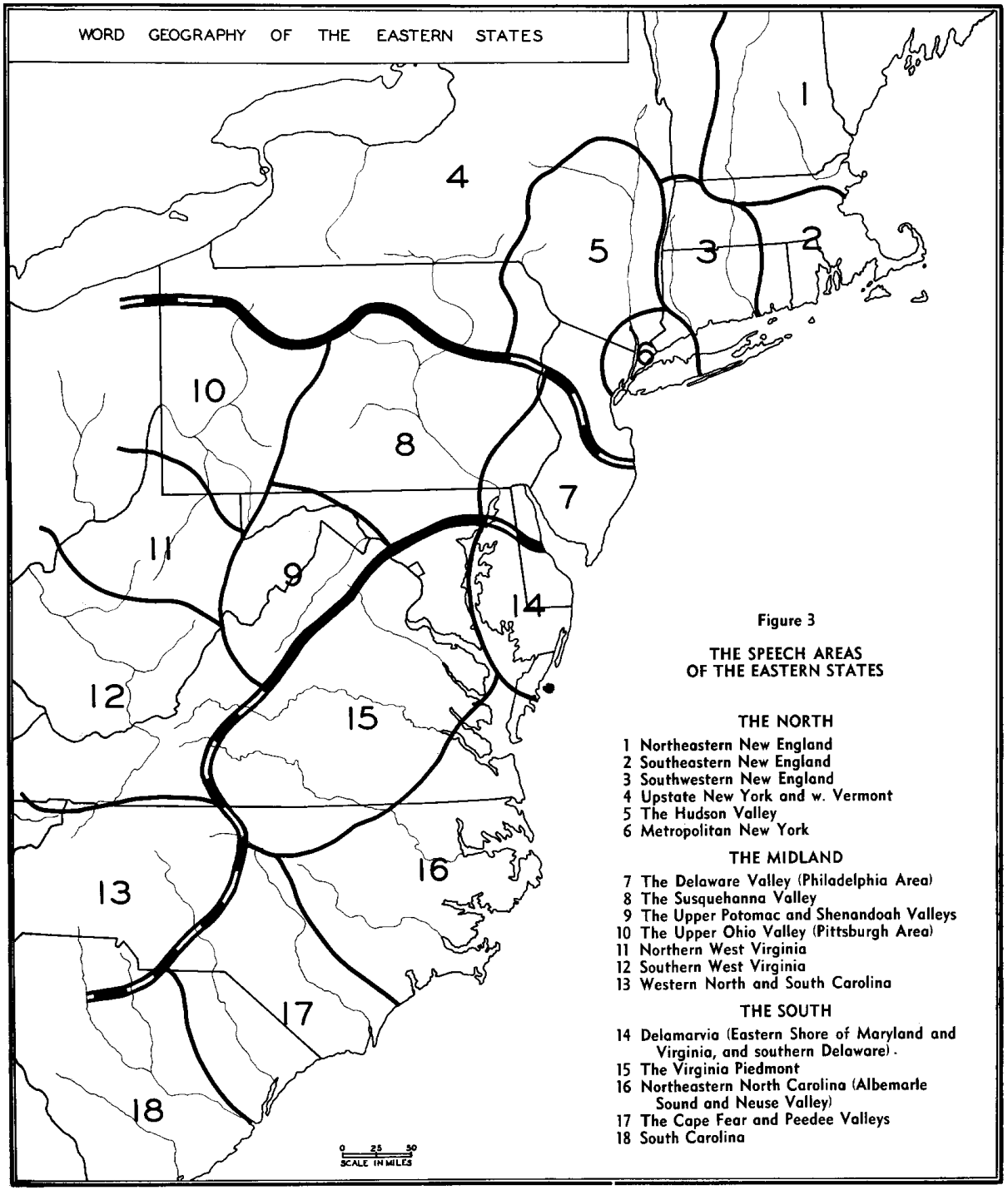


Figure 3

THE SPEECH AREAS
OF THE EASTERN STATES

THE NORTH

1 Northeastern New England
2 Southeastern New England
3 Southwestern New England
4 Upstate New York and w. Vermont
5 The Hudson Valley
6 Metropolitan New York

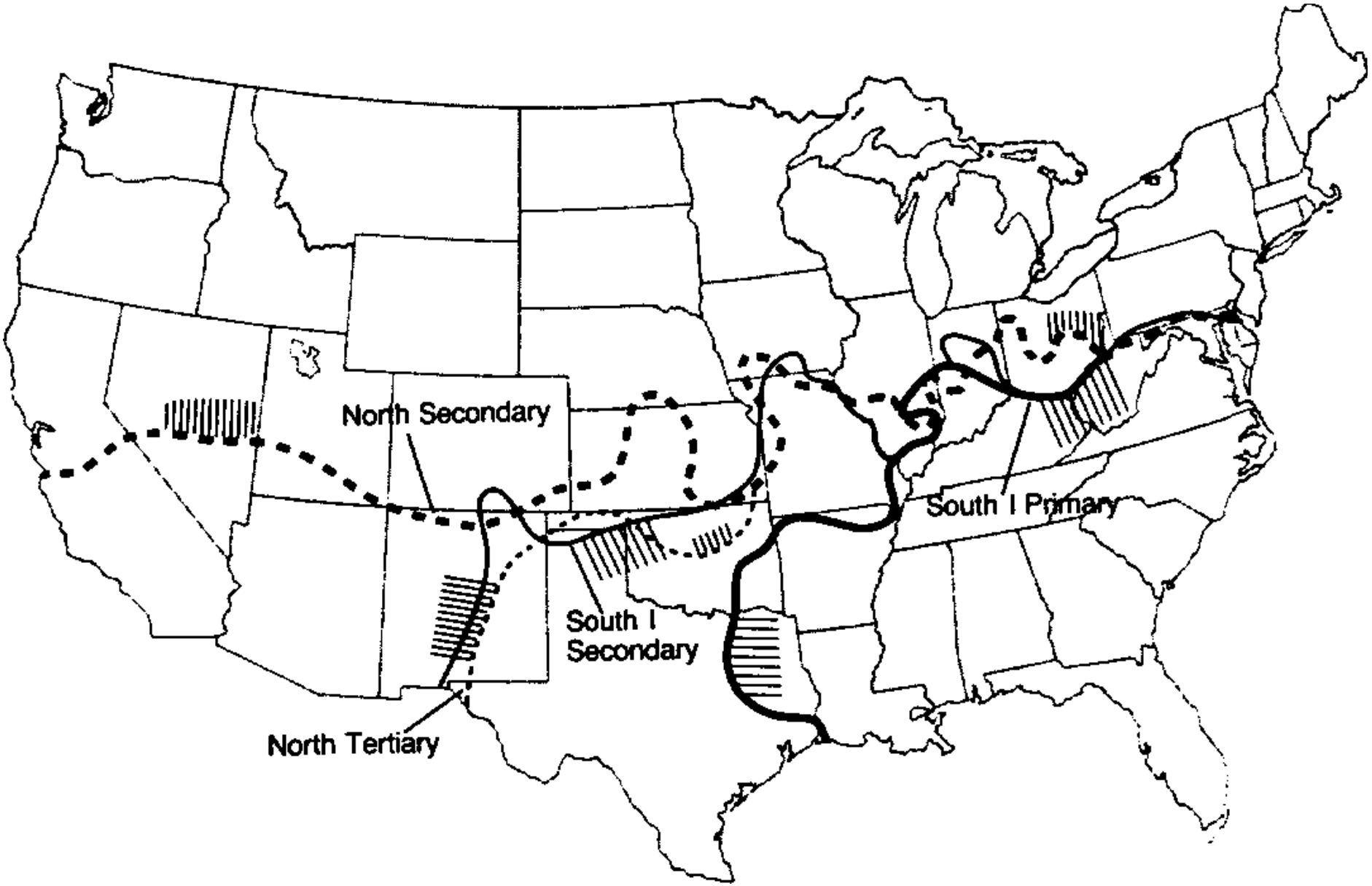
THE MIDLAND

7 The Delaware Valley (Philadelphia Area)
8 The Susquehanna Valley
9 The Upper Potomac and Shenandoah Valleys
10 The Upper Ohio Valley (Pittsburgh Area)
11 Northern West Virginia
12 Southern West Virginia
13 Western North and South Carolina

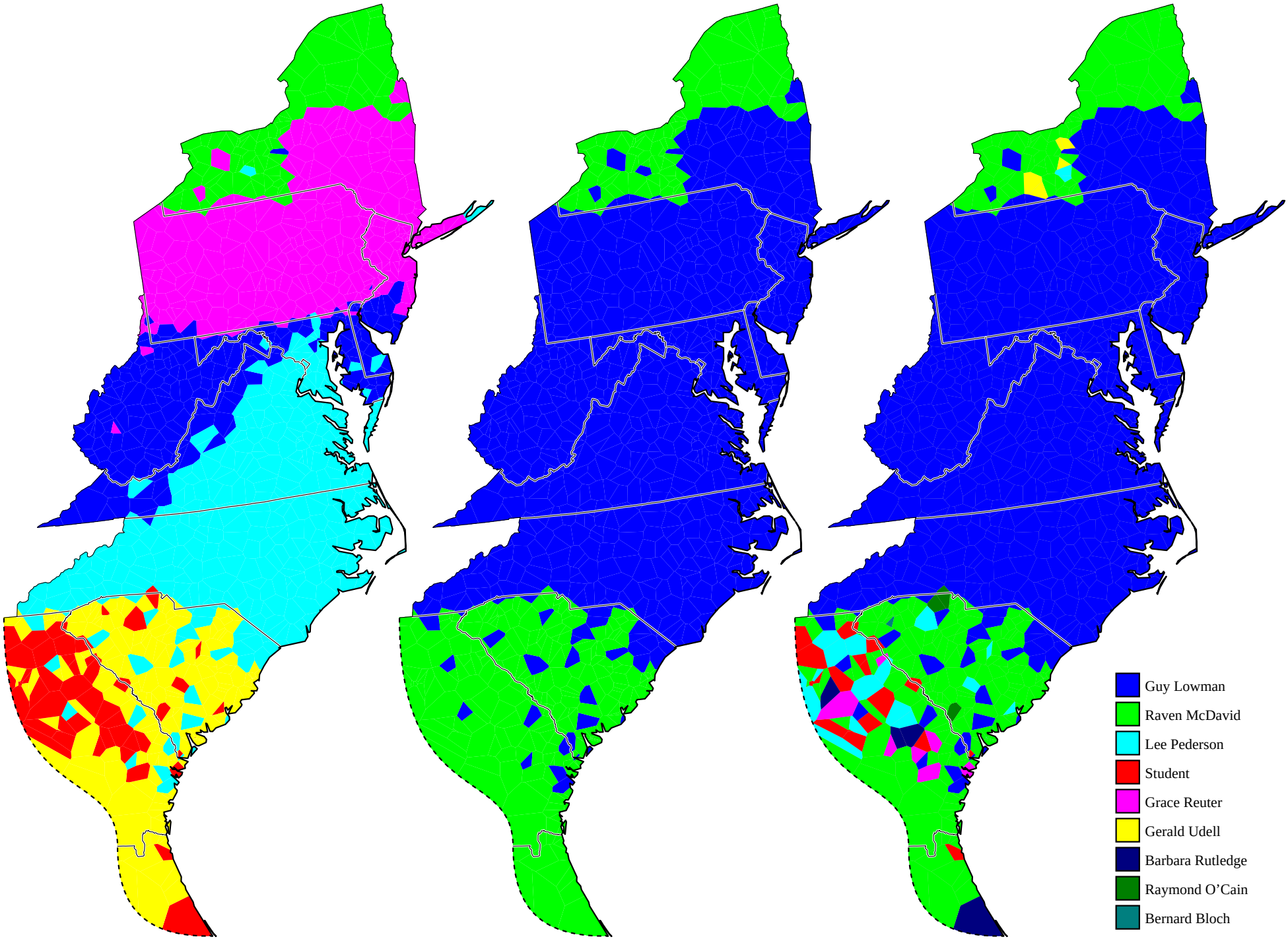
THE SOUTH

14 Delamarvia (Eastern Shore of Maryland and Virginia, and southern Delaware).
15 The Virginia Piedmont
16 Northeastern North Carolina (Albemarle Sound and Neuse Valley)
17 The Cape Fear and Peedee Valleys
18 South Carolina

Kurath's ideas on areas



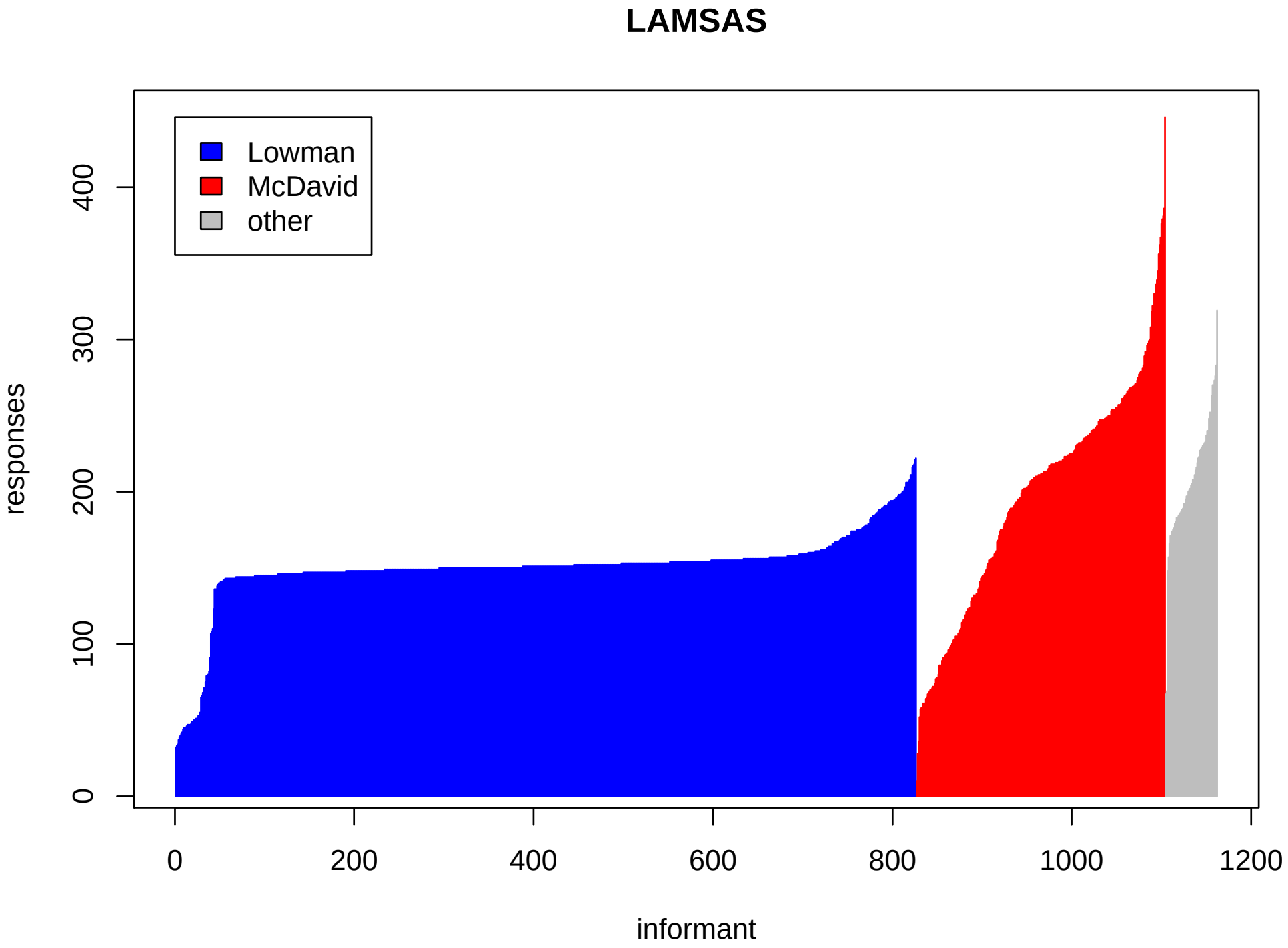
Carver's North/South Division



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

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 elicited fewer responses, but more consistently



responses per interview — Lowman, McDavid, other

Lexical Variation Needs Comparable Data

- Attempts to correct fieldworker bias
  - apparently no record of *order* of responses
  - restriction to most popular responses unsuccessful
- Therefore: concentrate on Lowman (71% of data)

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

Problem: close variants

- *fair off, fairing, fairing off, faired off, fairs off, ...*
- solution: use edit distance as measure of relatedness

Standard American	sɔəgIrl	delete r	1
	sɔəgIl	replace I/ɜ	2
	sɔəgɜl	insert r	1
Bostonian	sɔrəgɜl		
		Sum distance	4

- edit distance applied to *spelling*, not phonetics (in lexcial measurements)
  - lemmatizers would be most correct
    clear - clean - cleared

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

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

Problem: Infrequent Responses

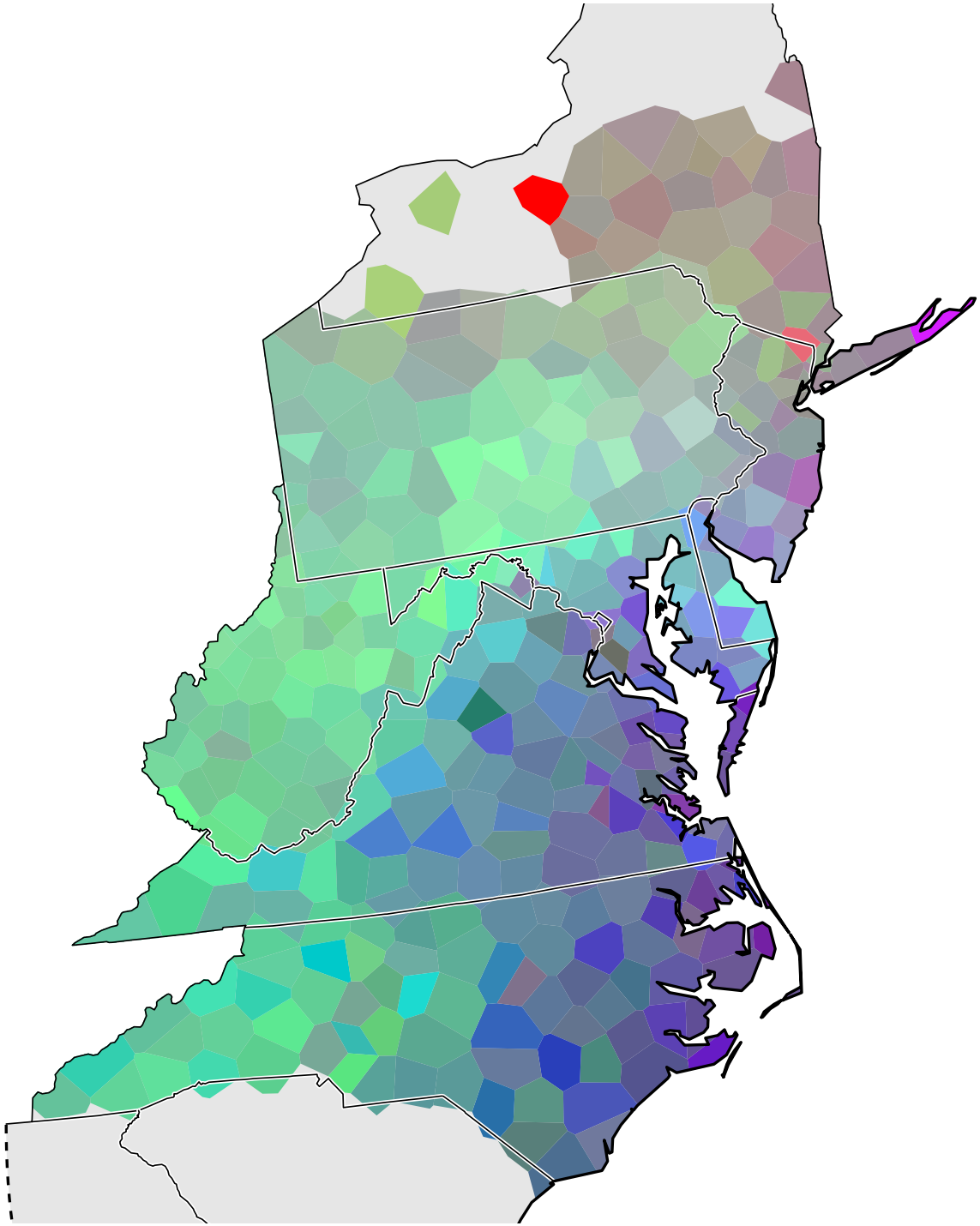
Two, diametrically opposed, views:

Goebl weight infrequent overlap most heavily (*Gewichteter Identitätswert*, frequently mentioned)

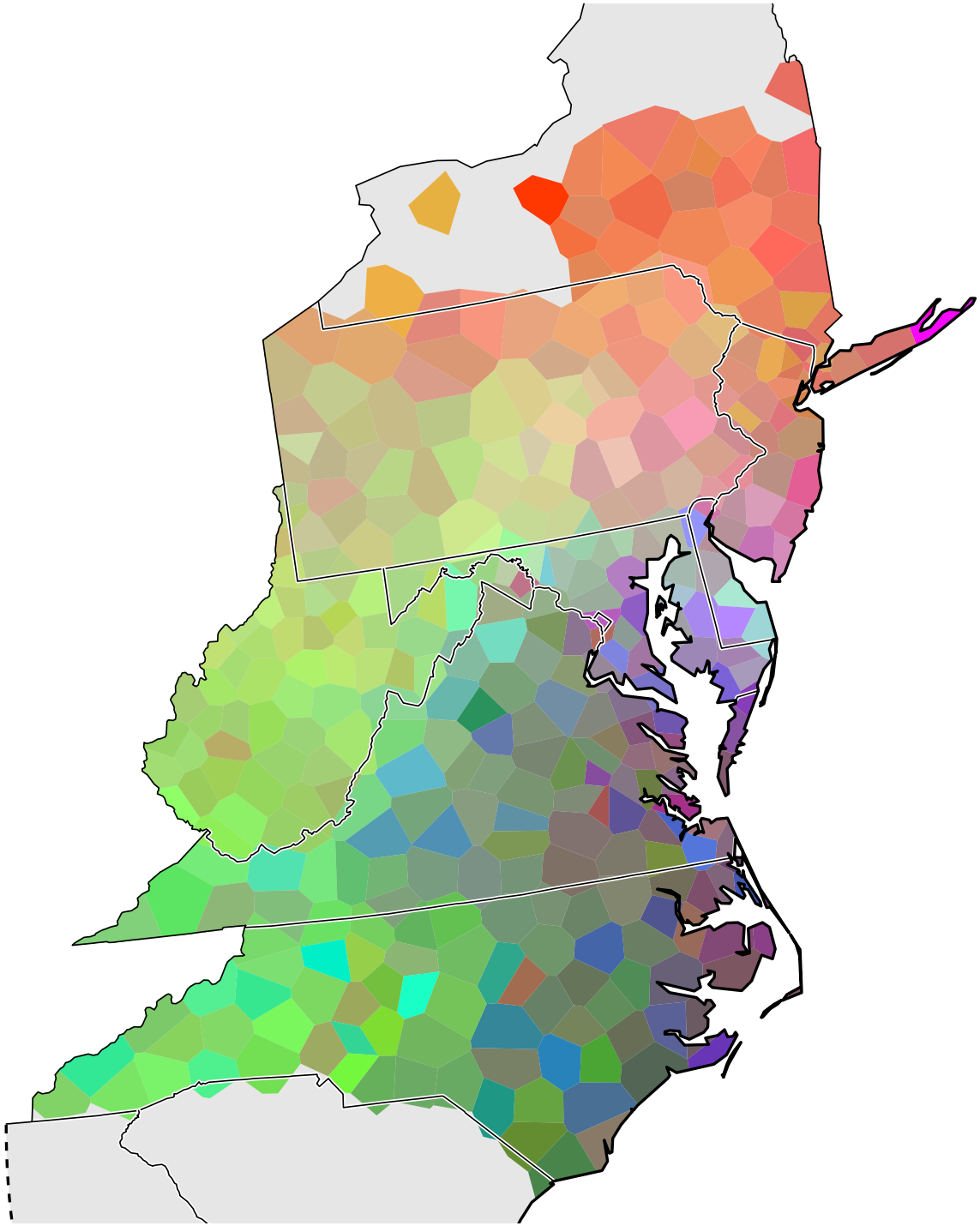
Carter discard least frequent items (*American Regional Dialects*, p.17)

Solution (here): discard responses which occur fewer than five times.

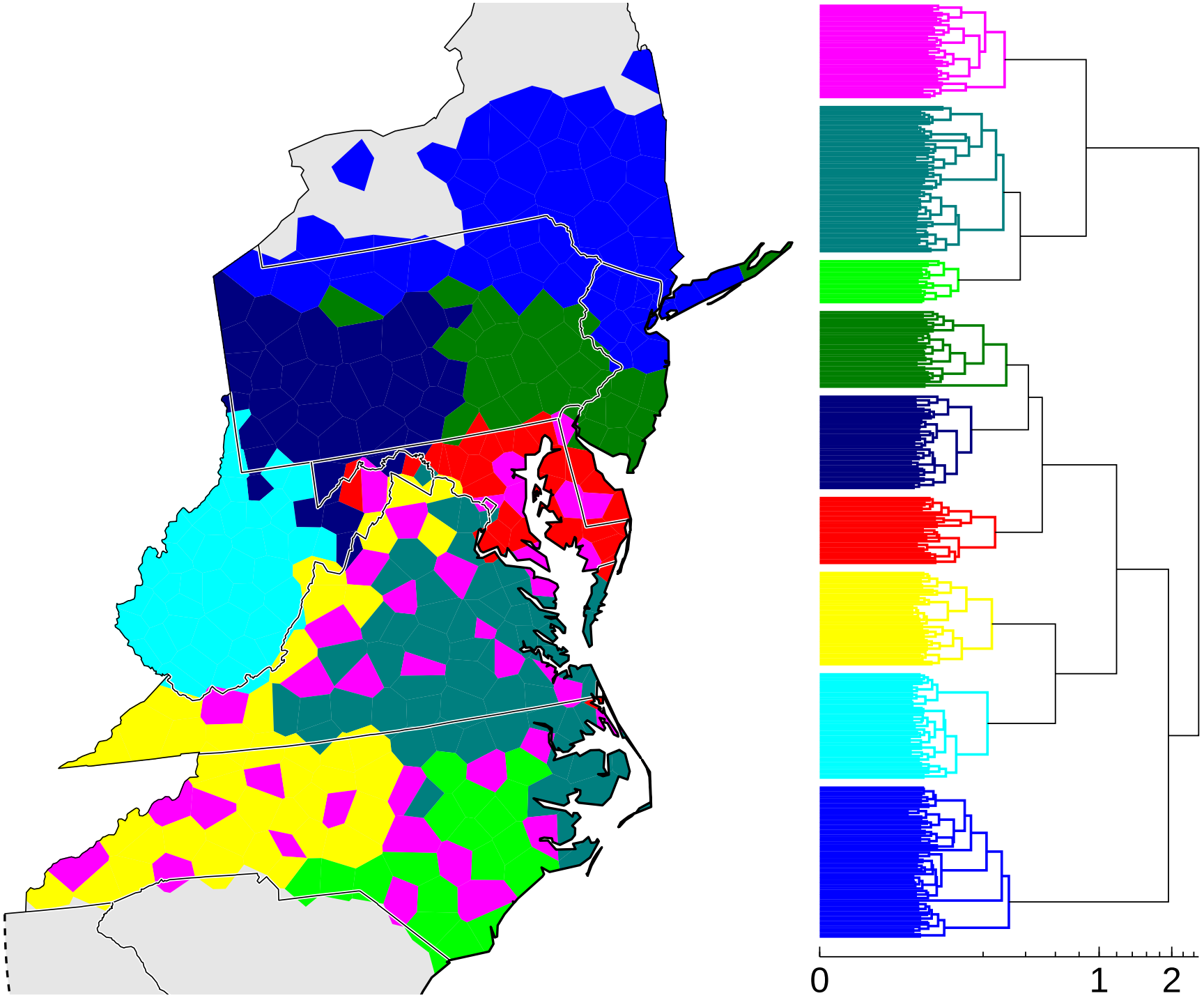
We examine this more systematically later in the course.



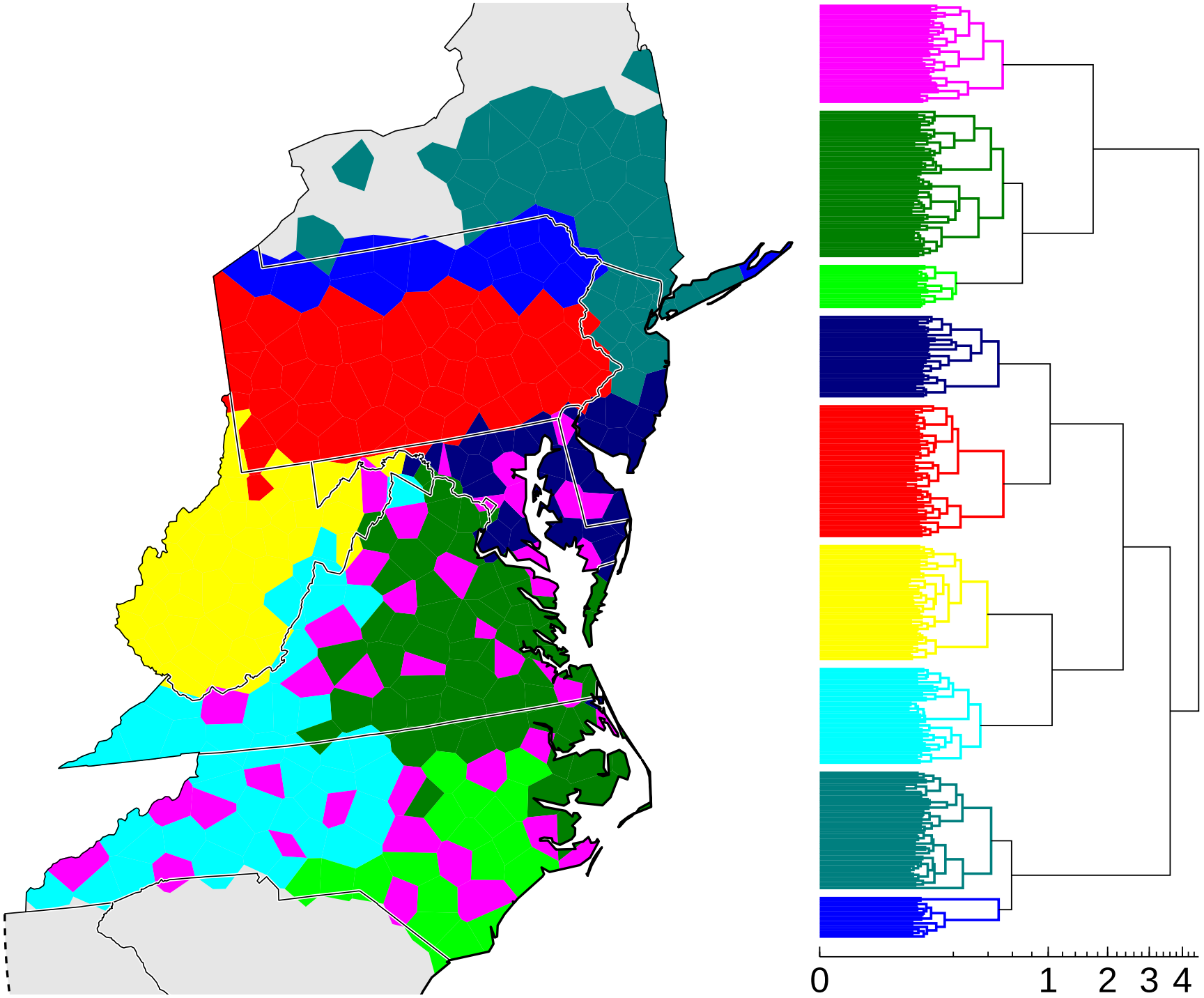
lexical, all words



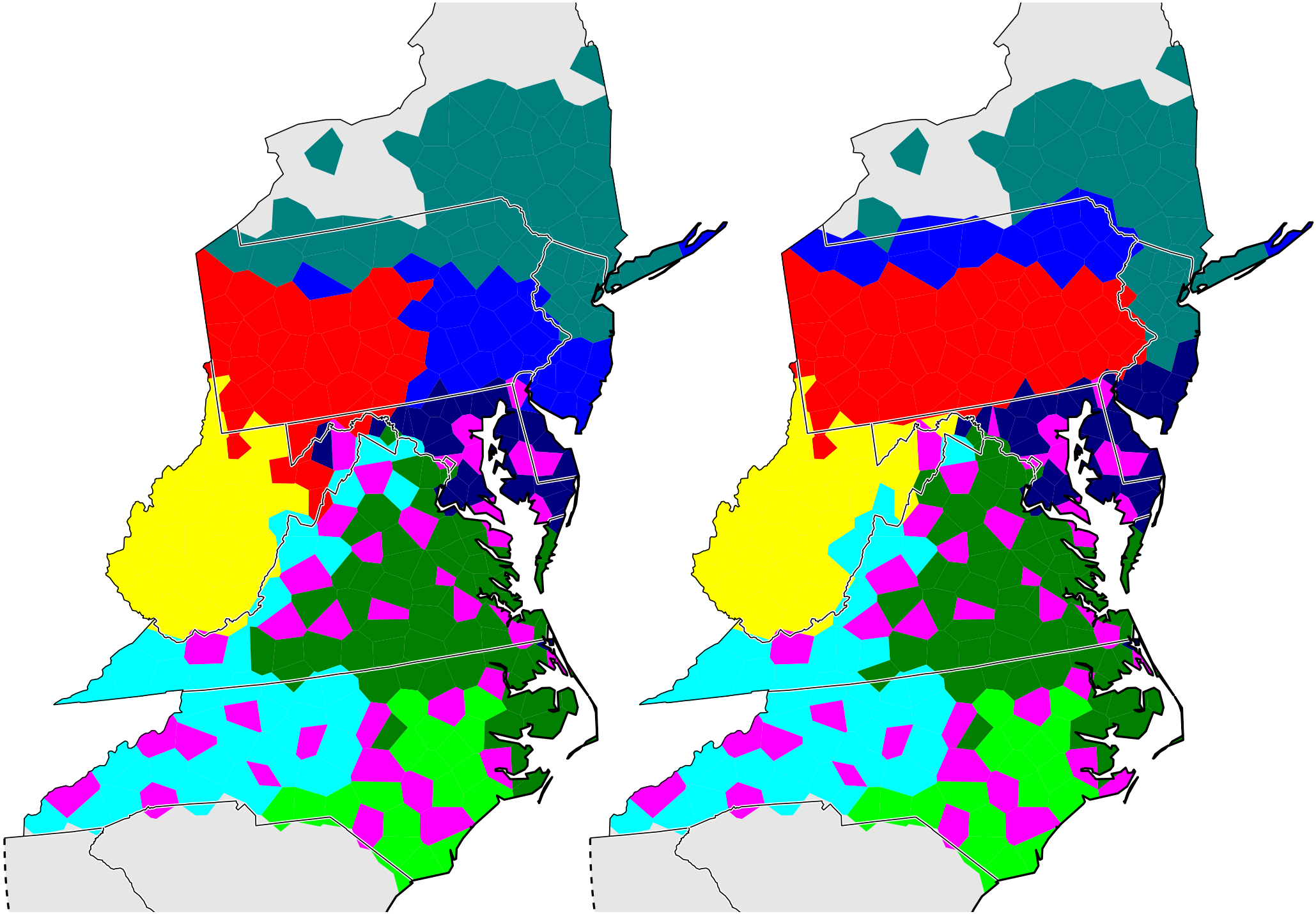
lexical, minimum 5 occurrences per word



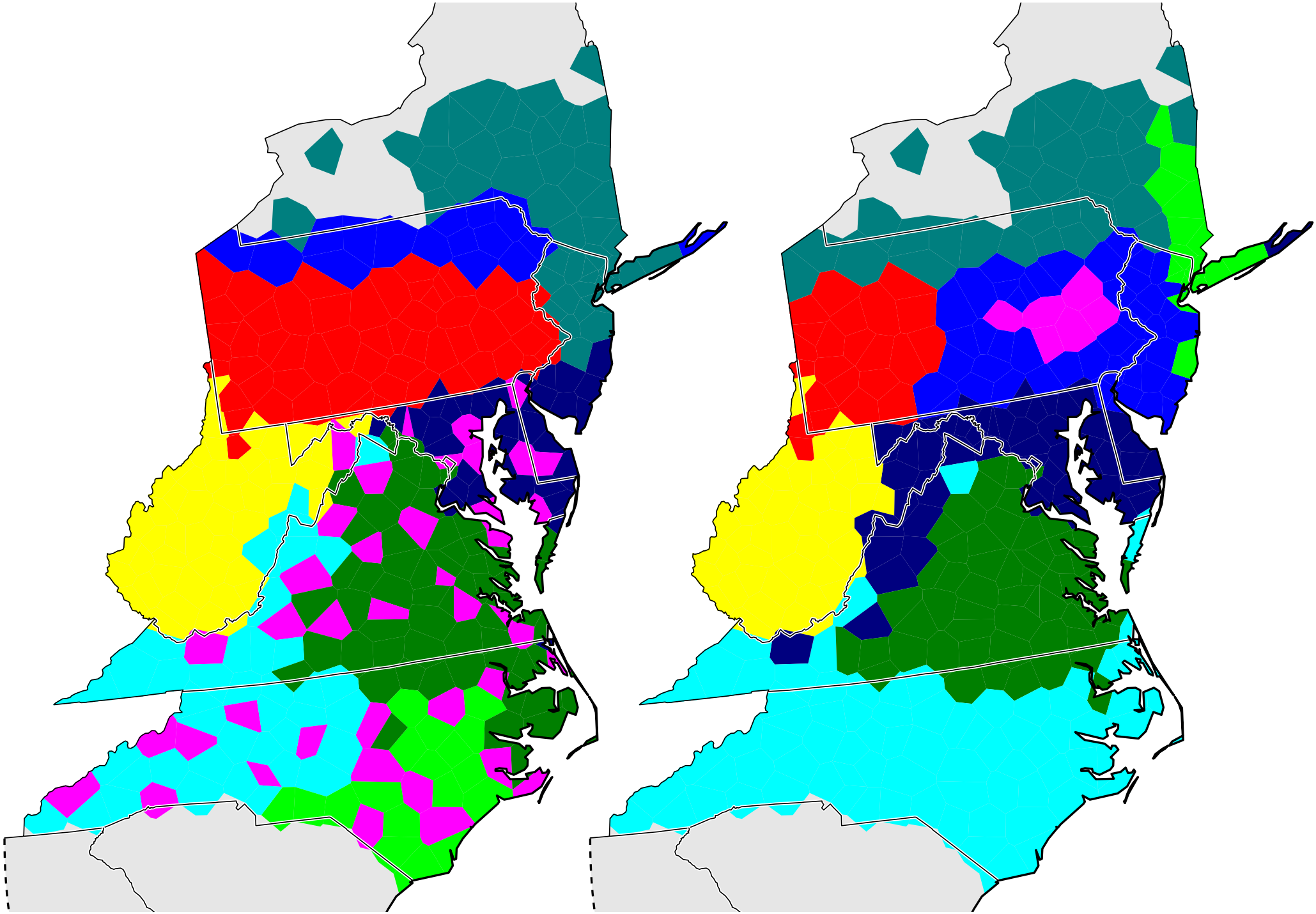
lexical, all words



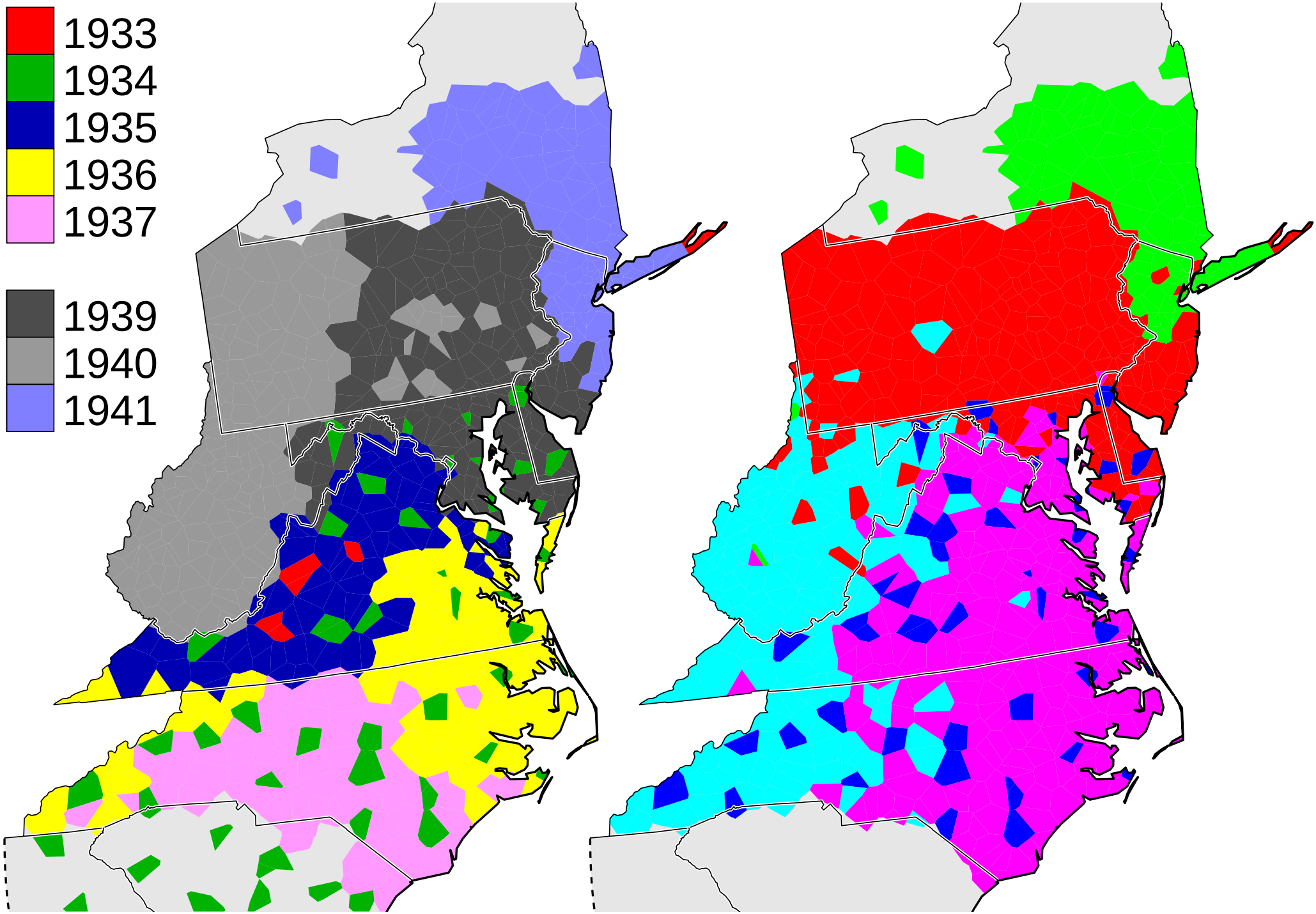
lexical, minimum 5 occurrences per word



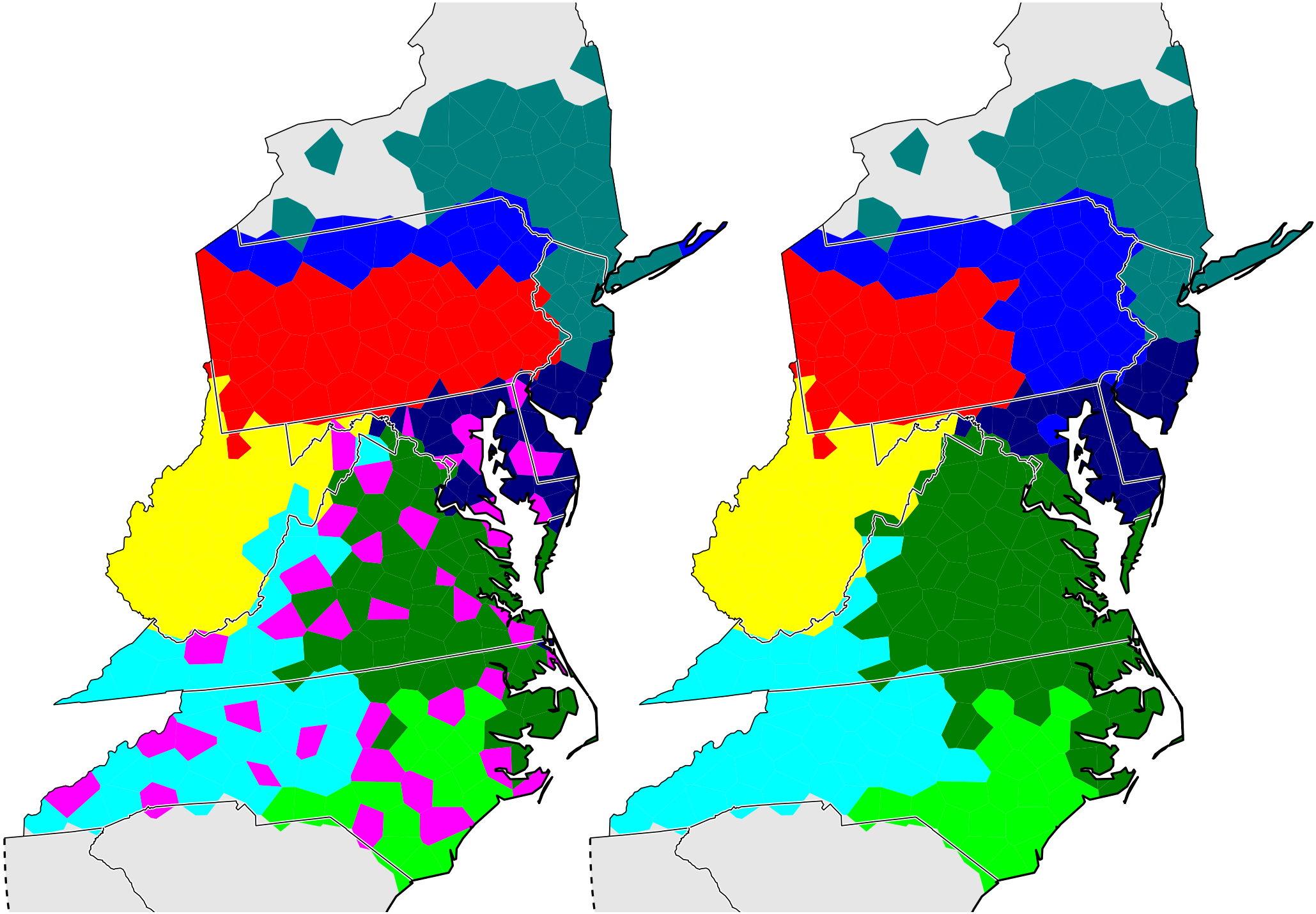
lexical, all words (left), minimum 5 occurrences per word (right)



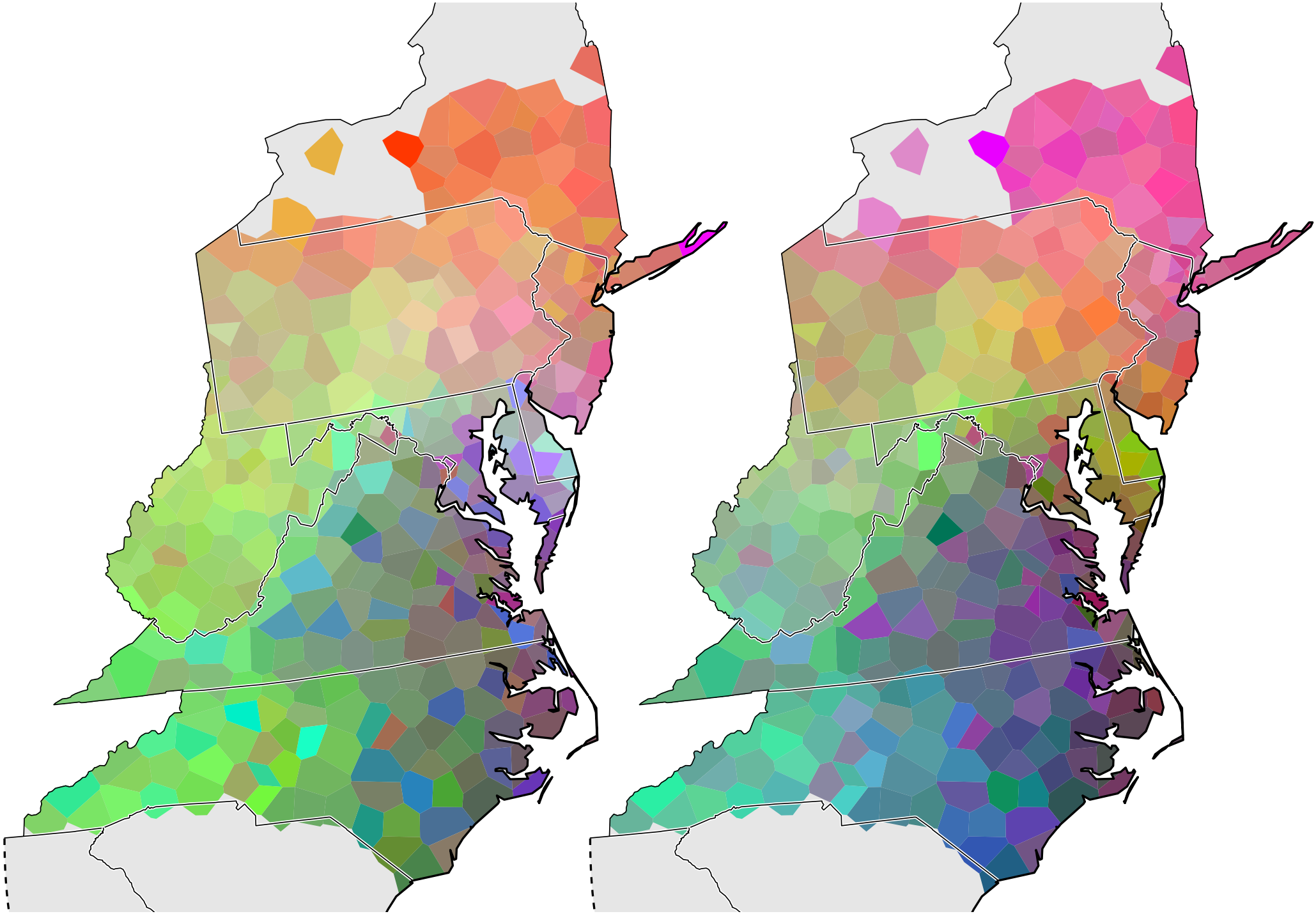
lexical, minimum 5 occurrences per word (left), phonetic (right)



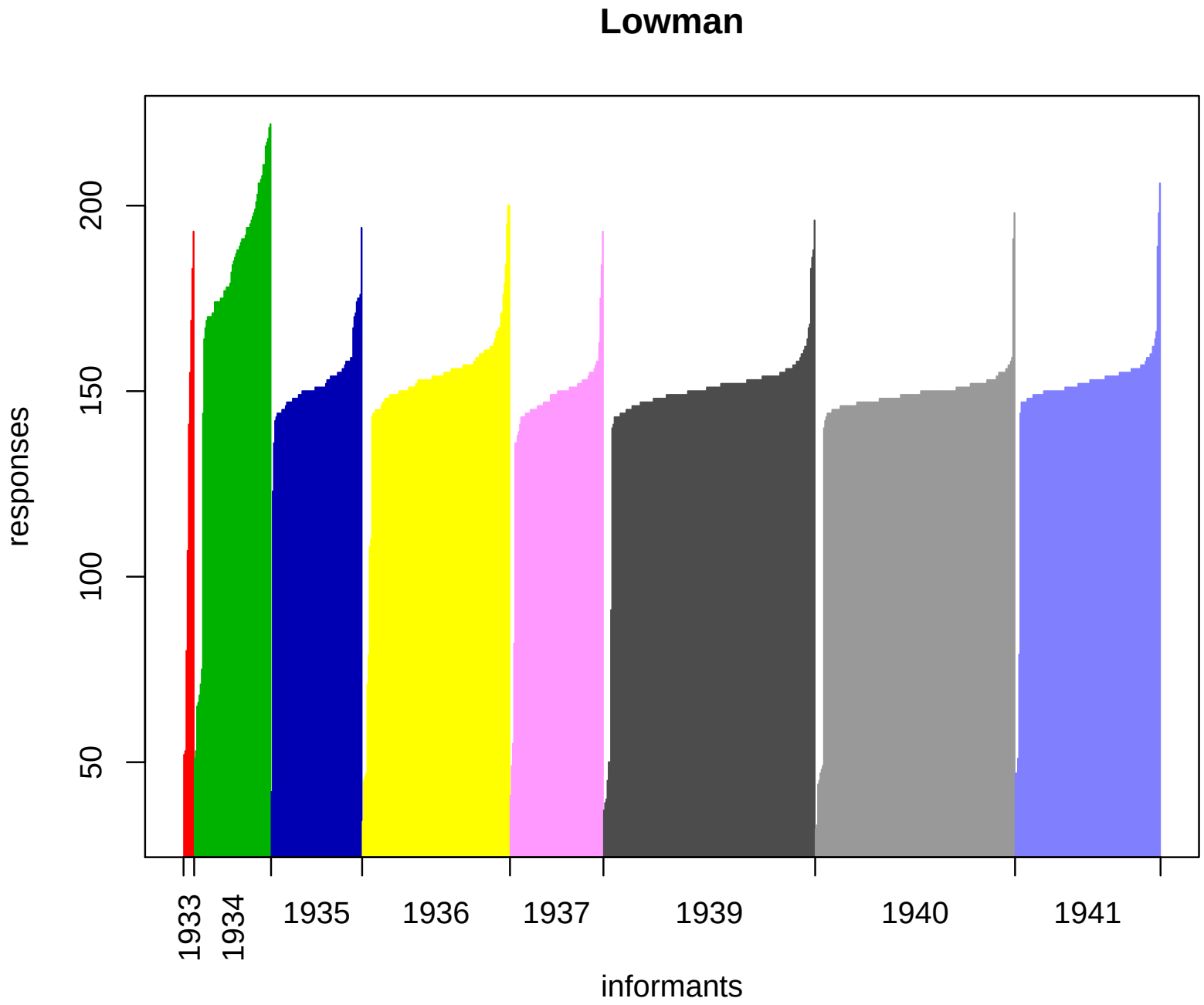
informants by year (left), lexical, minimum 5 occurrences per word (right)

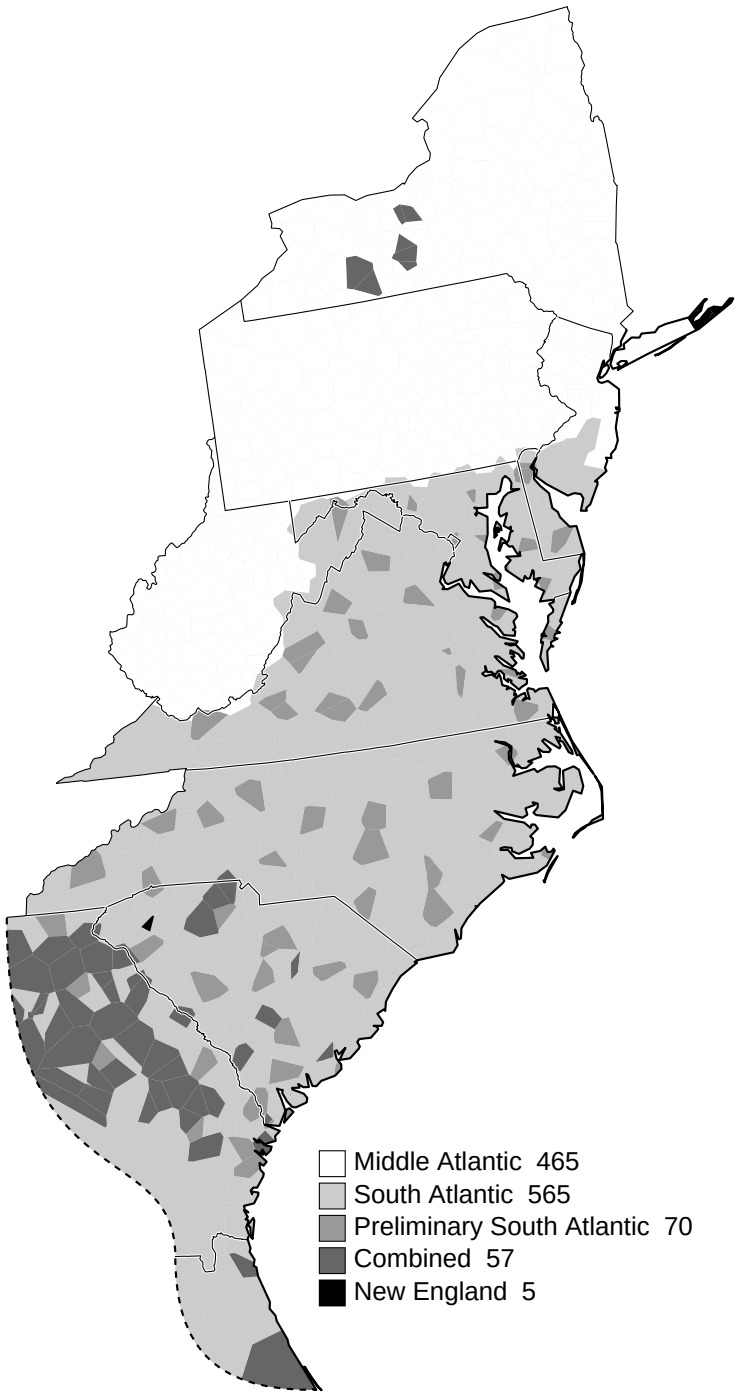


lexical, minimum 5 occurrences per word (left), without years 1933, 1934 (right)

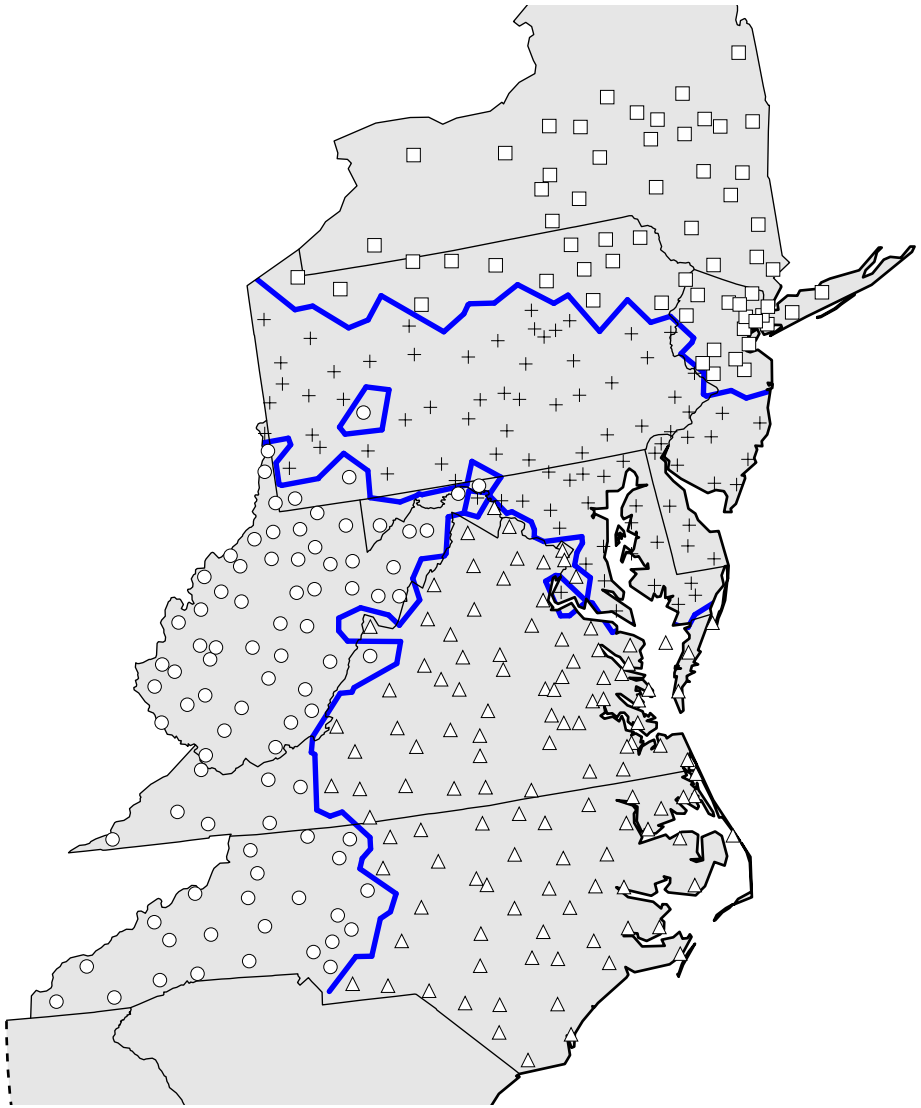


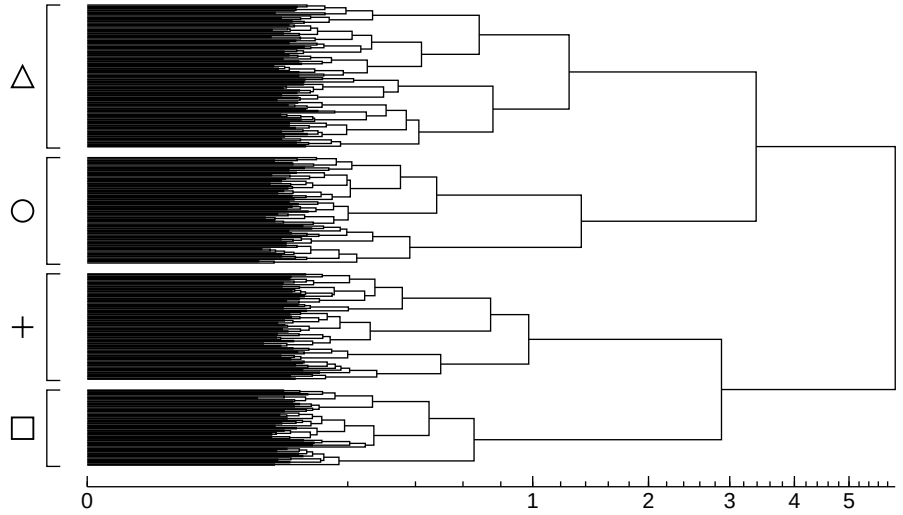
lexical, minimum 5 occurrences per word (left), without years 1933, 1934 (right)


Lowman
responses
50
100
150
200
1933
1934
1935
1936
1937
1939
1940
1941
informants



Different Questionnaires used in LAMSAS



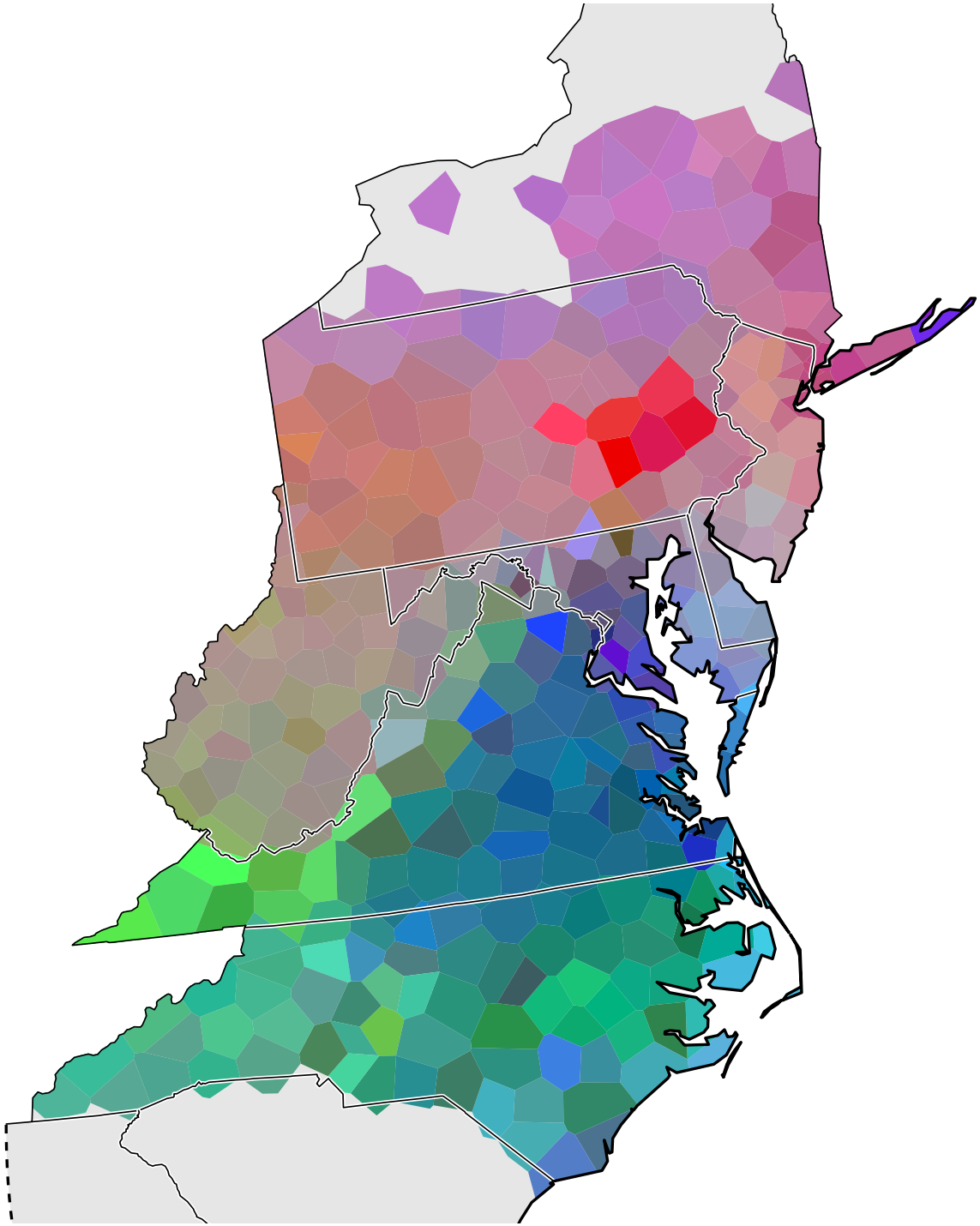


Final Division, Incorporating Questionnaire Intersection

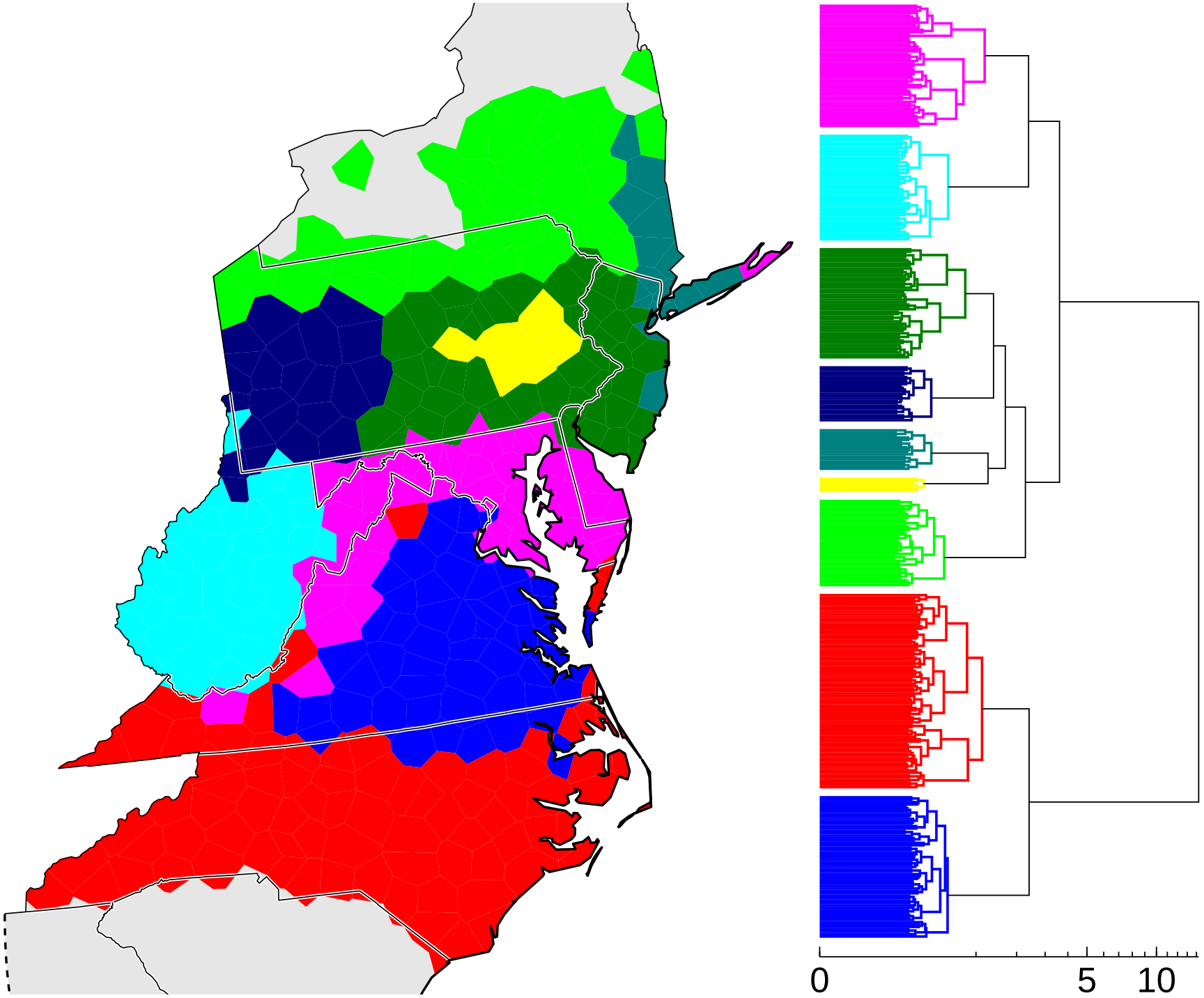
Concept	Response	North	Midland	S.Inland	S.Coast
dragonfly	darning needle	100%	13%	1%	0%
frost	dew	100%	2%	39%	0%
porch	stoop	92%	15%	0%	2%
quilt	comfort	2%	55%	84%	75%
night	evening	59%	74%	13%	8%
a little ways	a little piece	4%	64%	63%	17%
afternoon	evening	35%	21%	75%	82%
pallet	pallet	0%	6%	47%	59%
quilt	comfortable	61%	7%	0%	0%
northwest	northern	0%	0%	31%	58%
Sunday week	Sunday week	7%	25%	51%	67%
stairs	stairsteps	4%	28%	42%	66%
lightwood	lightwood	0%	5%	5%	54%
dragonfly	snake feeder	13%	44%	55%	2%
weatherboarding	clapboards	54%	14%	2%	2%
quarter to eleven	quarter till eleven	0%	20%	56%	19%
shades	shades	76%	26%	21%	53%
weatherboarding	weatherboarding	3%	41%	53%	50%
feet	feet	54%	5%	46%	52%
mantle	fireboard	0%	0%	48%	7%

Lexicon vs. Phonetics

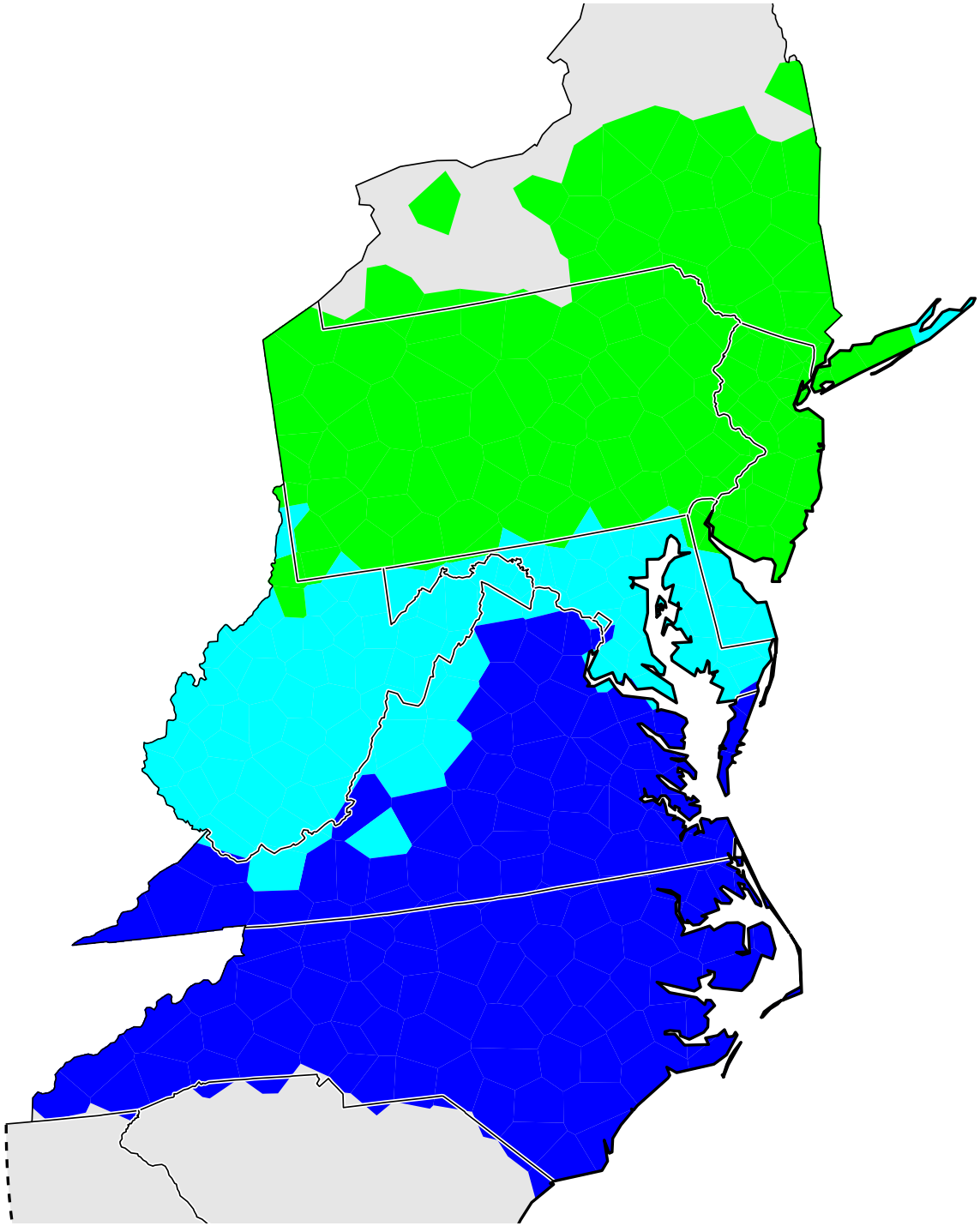
- Most attention has been paid to lexical overlap
  - Criteria clearer, simpler
- Phonetic proximity shows more coherence
  - Less volatile linguistically
  - Less likely to degenerate into “curiousity cabinet”
- Lexical-phonetic correlation $r = 0.65$
  - Kurath & McDavid (1961) claim that lexical and phonetic distributions “coincide fairly well”



phonetic



phonetic



phonetic

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)
- But are dialectometric techniques too sensitive to small differences in questionnaire size, interviewer technique, etc.?

Links

LAMSAS:

- `http://us.english.uga.edu/lamsas/`

Our research:

- `http://www.let.rug.nl/~kleiweg/lamsas/`

Our software:

- `http://www.let.rug.nl/~kleiweg/levenshtein/`