LAMSAS Pronunciation

John Nerbonne
LSA.107 Dialectology in the Aggregate
Rijksuniversiteit Groningen
j.nerbonne@rug.nl
http://www.let.rug.nl/alfa/

Summer, 2005

RuG

This Lecture: LAMSAS Pronunciation

- Levenshtein distance applied to pronunciations
- LAMSAS feature system
- Clustering, MDS, etc. applied to pronunciation distances

Aggregate Pronunciation Difference

Dialectometry (Séguy, Goebl) has focused on collecting lots of features, treated as qualitative variables.

Levenshtein distance (aka "edit distance", "string distance") provides a way to measure pronunciation difference numerically.

- numerical, therefore additive
- collect random sample of pronunciations (of the same word)
- use sum of differences as **aggregate pronunciation distance** between varieties

Experimental material

LAMSAS: Linguistic Atlas of the Middle and South Atlantic States

- data collection 1933–1974
- 71% of data collected by Guy Lowman 1933–1941
- most analyses based on lexical overlap
- now maintained by Bill Kretzschmar
- restriction to Lowman data: NC VA WV DC DE MD PA NY
  - omitting FL, GA, SC because geographically dispersed
  - apparent transcriber effects (see remarks on lexicon)

Data Preparation

- Parser recognizes a segment as a sequence of:
  1. zero or more pre-modifiers
  2. one head
  3. zero or more post-modifiers
- Results:
  - 57803 strings (99.95%) parsed correctly
  - 30 strings (0.05%) rejected

Some numbers

92,537 strings, 1.3 mil. phonetic tokens

363 locations, 797 informants

On average, 14.0 characters per string, parsed into 7.9 sound tokens per string

1,677 unique sounds (combinations of heads and modifiers)

1,132 unique vowel sounds

LAMSAS Feature System

Strategy: contrast analysis with LAMSAS's feature system (design by Kurath and Lowman) with the simple analysis in which segments are alike or different

One feature, DIRECTION, is not instantiated anywhere in Lowman's data. We ignore it, since it doesn't distinguish anything.

Feature: VOICE

p t k θ s ...	0
m b g ð z ...	0.1
vowels	0.1

RuG

Feature: NASAL

m ɱ n ɲ ŋ	0.1
all others	0

RuG

Feature: ORAL

m ɱ n ɲ ŋ	0
l ɭ ɬ ʎ ʟ	0.15
all others	0.05

RuG

Feature: CLOSURE

ɾ ɹ	0.025
n l ʎ p d g ...	0.05
f z ɹ x w ...	0
vowels	0

Feature: ORIFICE

p	t	g	...		0	i	y		0.23	ɘ			0.33	ɐ		0.44
f	z	ʒ	x	...	0.1	ɨ	ʉ		0.25	ə	ɚ	o	0.35	æ		0.45
m	l	ʎ	ɹ	...	0.2	ɪ	ʏ	ɯ	ʊ	0.28	ɵ			0.36	ɑ	ɒ	0.48
ʔ	h	ɦ			0.35	ɨ̞	ʊ̞			0.29	ɛ	ɞ		0.38	ɑ		0.5
						e	ø			0.3	ɜ	ɞ	ʌ	0.4	a		0.53
						ɣ	u			0.31	ɔ			0.43			

RuG

Feature: PLACE

m p b Φ ß	0
ɱ f v ꜰ ʋ	0.1
θ ð	0.2
n l ɭ ɬ ɾ ɻ t	0.3
d s z ɹ ɤ r ʀ	0.3
ʃ ʒ	0.45
ɲ ʎ c ɟ ç j ɥ	0.6
ʔ h ɦ	0.8
ŋ k g x ɾ w	0.9

i y	0.6
e ø	0.65
ɪ ʏ	0.68
ɛ ɘ	0.7
æ ɨ ʉ ɵ	0.75
ɨ ʉ	0.78
a ə ɚ	0.8
ɜ ɞ	0.83

ɐ	0.86
ɤ u ɵ	0.88
ɑ	0.89
ɯ ʊ ʌ	0.9
o	0.93
ɔ	0.95
ɑ ɒ	0.98

RuG

Feature: PLACE2

ʯ ʍ	0
ɭ̵	0.45
ɻ̵ ʁ R	0.9
all others	undefined

Feature: ROUNDING

i̤	ɪ	e̞	ɛ	æ	a̟	i̙	ɨ	ɘ	ə	ɚ	ɜ	ɐ	ɑ	ɯ	ɤ	ʌ	ɑ	0
y̟	ʏ	ø	ɞ	ʉ̞	ʊ̞	ɷ	ʊ	u	o	e̞	ɔ	ɒ						0.9
consonants																		undefined

Feature: APEX

ɹ	0
r ɚ	0.1
all others	0.05

Feature: GLOTTIS

ʔ	0
h ɦ	0.1
all others	0.05

Feature: LENGTH

ɾ ɹ p b t d c ɟ k g ʔ	0
all others	0.05

Feature modifiers: consonants (1)

C^{h}	$glottis = 0.025$
C^{ɦ}	unknown, ignored
C$^{>}$ Ç C$^{>}$	backing, $place = place \times 1.08$
C$^{<}$ C$^{<}$	fronting, $place = place \times 0.92$
C$^{\cdot}$	half long (long??), $length = 0.1$
C$^{\frown}$	implosive ???, ignored
C$^{\circ}$	unknown, ignored
Ç	unknown, ignored
Ç$^{\cdot}$	retroflexion / r-color, $apex = 0.1$
Ç$^{\circ}$	unvoicing, $voice = 0.05$

Feature modifiers: consonants (2)

C◡	voicing, $voice = 0.05$
C◠	dentalization, $place2 = 0.25$
C̪	unknown, ignored
C̬	lenis, $voice = 0.05$
C̩	syllabic, ignored
Cʷ	labialization, ignored
Cʲ	palatalization, $place2 = 0.6$
Cˠ	velarization, $place2 = 0.9$
C̆	unknown, ignored
C̃	nasalization, $nasal = 0.1$

Feature modifiers: vowels

◌ː	long (overlong ??), $length = 0.15$
◌ˤ	pharyngealization ???, ignored
◌˞	r-color, $apex = 0.1$
◌̥	unvoicing, $voice = 0.05$
◌̲	spreading, $rounding = rounding - 0.3$
◌̹	rounding, $rounding = rounding + 0.3$
◌ˆ	raising, $orifice = orifice \times 0.9$
◌ˑ	half long (long?), $length = 0.1$
◌˂ ◌̘	fronting, $place = place \times 0.92$
◌ˇ	lowering, $orifice = orifice \times 1.1$
◌˃ ◌̙	backing, $place = place \times 1.08$
◌̆	short, $length = 0$
◌̃	nasalization, $nasal = 0.1$

differences between feature sets

- sum of all differences
- when comparing a normal character with a superscript character, the sum of feature differences is multiplied by 0.7
- when comparing two superscript characters, the sum of feature differences is multiplied by $0.7 \times 0.7 = 0.49$
- character ' ' ' is treated as superscript ' ʔ '
- character ' ` ' is treated as superscript ' ʰ '
- for $place2$:
  - both missing: 0
  - one missing: 0.3
  - weight: 0.5
- for $rounding$:
  - both missing: 0
  - one missing: 0.5

Feature Value Frequencies

feature	value	LAMSAS	LOWMAN
weight	0.7	67822	48817
weight	1	773834	586279
weight	1.1	17037	7839
weight	1.2	26028	10873
apex	0	3503	1762
apex	0.05	827237	610786
apex	0.1	53981	41260
closure	0	604687	445168
closure	0.025	6613	5078
closure	0.05	273421	203562
direction	0	884721	653808
glottis	0	1778	1510
glottis	0.025	818	806
glottis	0.05	870917	646219
glottis	0.1	11208	5273

Feature Value Frequencies

feature	value	LAMSAS	LOWMAN
length	0	164808	123095
length	0.05	664743	494588
length	0.1	53930	34917
length	0.15	1240	1208
nasal	0	795202	587980
nasal	0.1	89519	65828
oral	0	86961	64197
oral	0.05	760399	561630
oral	0.15	37361	27981
place2	0	23134	16961
place2	0.25	403	288
place2	0.45	1060	550
place2	0.6	1648	1191
place2	0.9	17632	15335

Feature Value Frequencies

feature	value	LAMSAS	LOWMAN		
rounding	-0.3	56	6		
rounding	0	315400	233565		
rounding	0.3	2566	1392		
rounding	0.6	1073	927		
rounding	0.9	74215	55347		
rounding	1.2	459	296		
voice	0	195753	142301		
voice	0.05	9670	4501		
voice	0.1	679298	507006		
orifice	51 values			min 0	max 0.58
place	53 values			min 0	max. 1.06

RuG

Example Computation: *bottom*

GA1H! Savannah --- PA29B Clinton Co.

ˈbɑt̬əm			---			bɒɾəm
b	ɑ	t̬		ə		m
b	ɒ	ɾ		ə		m
0	1.08	1.355		1.355		1.355
ˈ	b	ɑ	t	ˬ	ə	m
	b		ɒ	ɾ	ə	m
1	1	2	4	6	6	6

Example Computation: *ten*

GA1N Savannah --- PA29A Clinton Co.

t ˋ ɪ n	---		t ɛ̆ n
t	ˋ	ɪ̆	n
t		ɛ	n
0	0.07	0.24	0.24
t	ˋ ɪ	n	
t	ɛ ˘	n	
0	2	4	4

Example Computation: *hog pen*

GA1H! Savannah --- PA29A Clinton Co.

ˈhɔ˅˂•ᵊg			ˌp`ɛ˄ᵻn			---			hɒg			pɛn								
h		ɔ˅˂•ᵊ			g		p	`		ɛ˄		ᵻ		n						
h		ɒ			g		p			ɛ				n						
0		0.1956	0.6506		0.6506		0.6506	0.7206		0.7624		1.2734		1.2734						
ˈ	h	ɔ	˅	˂	•	{	ə	}	g	$	ˌ	p	`	ɛ	˄	{	ɪ	–	}	n	
	h							ɒ	g	$		p		ɛ						n	
1	1	2	3	4	5	6	7	9	9	9	10	10	11	11	12	13	14	15	16	16	

RuG

Example Computation: *four*

GA1B Savannah --- PA29B Clinton Co.

fo·ə		---		fɔɚ·	
f	o·	ə			
f	ɔ	ɚ·			
0	0.15	0.25			
f		o	·	ə	
f	ɔ	ɚ	·	·	
0	1	2	4	4	5

Example Computation: *forty*

GA1H! Savannah --- PA29B Clinton Co.

f ɔ ˂ • ə t̬ ɨ ˄ ˂ --- f ɔ ˄ ɚ ɾ ɨ ˄

f ɔ ˂ • ə t̬ ɨ ˄ ˂

f ɔ ˄ ɚ ɾ ɨ ˄

0 0.169 0.219 0.494 0.5564

f ɔ ˂ • ə t ˬ I – ˄ ˂

f ɔ ˄ ɚ ɾ I – ˄

0 0 1 2 4 6 8 8 8 8 9

RuG

Example Computation: *chair*

GA1H! Savannah --- PA29A Clinton Co.

tʃeːə			---		tʃɛɚː
t	ʃ	eː	ə		
t	ʃ	ɛ	ɚː		
0	0	0.18	0.28		
t	ʃ	e	ː	ə	
t	ʃ	ɛ	ɚ	ː	
0	0	1	3	3	4

Example Computation: *Wednesday*

GA1F! Savannah --- PA29B Clinton Co.

w	ɛ̃	ˬ	ə̃	n	z	d	ɨ	–	–	–		
w	ɛ̃	ˬ	ə̃	n	z	d	ɨ					
w	ɛ			n	z	d	ɨ					
0	0.138		0.593	0.593	0.593	0.593	0.593					
w	ɛ̃	ˬ	{	ə̃	}	n	z	d	ɪ	–		
w	ɛ					n	z	d	ɪ	–		
0	0	1	2	3	4	5	6	6	6	6	6	6

Example Computation: *thirty*

GA1E Savannah --- PA29A Clinton Co.

θ 3 ʌ ˂ ǂ t̬ ǂ ʌ ˂ — — — θ ə t̪ ǂ ʌ

θ 3 ʌ ˂ ǂ t̬ ǂ ʌ ˂

θ ə t ǂ ʌ

0 0.0964 0.6074 0.6574 0.7198

θ 3 ʌ ˂ { I – } t ˬ I – ʌ ˂

θ ə t ˽ I – ʌ

0 1 2 3 4 5 6 8 8 10 10 10 10 11

Example Computation: *dry spell*

GA1H! Savannah --- PA29B Clinton Co.

dɹaˀ·ə ˈspɛʌˀl --- draʌˁ·ɬ spɛᵊɬ

d	ɹ	aˀ·	ə	s	p	ɛʌˀ	l	
d	r	aʌˁ·ɬ	s	p	ɛ	ᵊ	ɬ	
0	0.2	0.381	0.461	0.461	0.461	0.5738	1.0288	1.1788

d	ɹ	a	ˀ	·	ə	$	ˈ	s	p	ɛ	ʌ	ˀ	l			
d	r	a	ʌ	ˁ	·	ɪ	–	$	s	p	ɛ	{	ə	}	ɬ	
0	2	2	3	5	5	6	8	8	9	9	9	9	10	12	14	16

RuG

Example Computation: *afternoon*

GA1G! Savannah --- PA7C Lancaster Co.

ˈæˬᵊftəˌnʉ·n					---			ˌæftərˈnu·n							
æˬ		ᵊ		f		t		ə			n		ʉ·		n
æ				f		t		ə		r	n		u·		n
0.0594	0.5144		0.5144		0.5144		0.5144		1.3644		1.3644		1.6152		1.6152
ˈ	æ	ˬ	{	ə	}	f	t	ə		ˌ	n	u	–	·	n
ˌ	æ					f	t	ə	r	ˈ	n	u		·	n
2	2	3	4	5	6	6	6	6	7	9	9	9	10	10	10

RuG

Example Computation *my wife*

GA1H! Savannah --- PA29A Clinton Co.

mɐ˅		wɐ˅	ɪ˃	f		---		maɨ˄		wa ɨ˄	f	
m		ɐ˅			w	ɐ˅	ɪ˃	f				
m		a	ɨ˄		w	a	ɨ˄	f				
0		0.106	0.865		0.865	0.971	1.0356	1.0356				
m			ɐ	˅	$	w	ɐ	˅	ɪ		˃	f
m	a	ɪ	-		˄	$	w		a	ɪ	-	˄	f
0	1	2	4		6	6	6	7	9	9	10	12	12

RuG

Example Computation: *Georgia*

GA1H! Savannah --- PA7B Lancaster Co.

dʒɔʌˑˈədʒə --- dʒɔʌɚdʒə

d		ʒ		ɔʌˑˈə			d		ʒ		ə	
d		ʒ		ɔʌ		ɚ	d		ʒ		ə	
0		0		0.126		0.161	0.161		0.161		0.161	
d	ʒ	ɔ	ʌ	ˑ	ˈ	{	ə	}	d	ʒ	ə	
d	ʒ	ɔ	ʌ					ɚ	d	ʒ	ə	
0	0	0	0	1	2	3	4	6	6	6	6	

RuG

Example Computation: *Georgia* (again)

GA1H! Savannah --- PA7C Lancaster Co.

dʒɔˆ˂•ᵊdʒə --- t̪ʃɔrt̪ʃə

d	ʒ	ɔˆ˂•ᵊ		d	ʒ	ə
t̪	ʒ	ɔ	r	t̪	ʒ	ə
0.05	0.1	0.269	1.109	1.159	1.209	1.209

		d	ʒ	ɔ	ˆ	˂	•	{	ə	}	d	ʒ	ə				
t	_	ʃ	_	ɔ				r	t	_	ʃ	_	ə				
1	2	4	6	6	7	8	9	11	13	15	17	19	19				

RuG

Example Computation: *fog*

GA1A Savannah --- PA7C Lancaster Co.

fɒˆ·ᵊg --- faˀg

f ɒˆ· ᵊ g

f aˀ gˀ

0 0.962 1.1798 1.2298

f ɒ ˆ · { ə } g

f a ˀ g ₒ

0 1 2 3 4 6 8 8 9

Example Computation: *feet*

GA1H! Savannah --- PA5D Chester Co.

fi·ᵊt			---			fɪiˬt			haɨ			
f		i·		ᵊ		t						
f		ɪ		iˬ		t		h		a		ɨ
0		0.18		0.3879		0.3879		0.4879		1.3179		2.0479
f		i		·	{	ə	}	t				
f	ɪ	i					ˬ	t	$	h	a	ɪ –
0	1	1	2	3	4	6	6	7	8	9	10	11

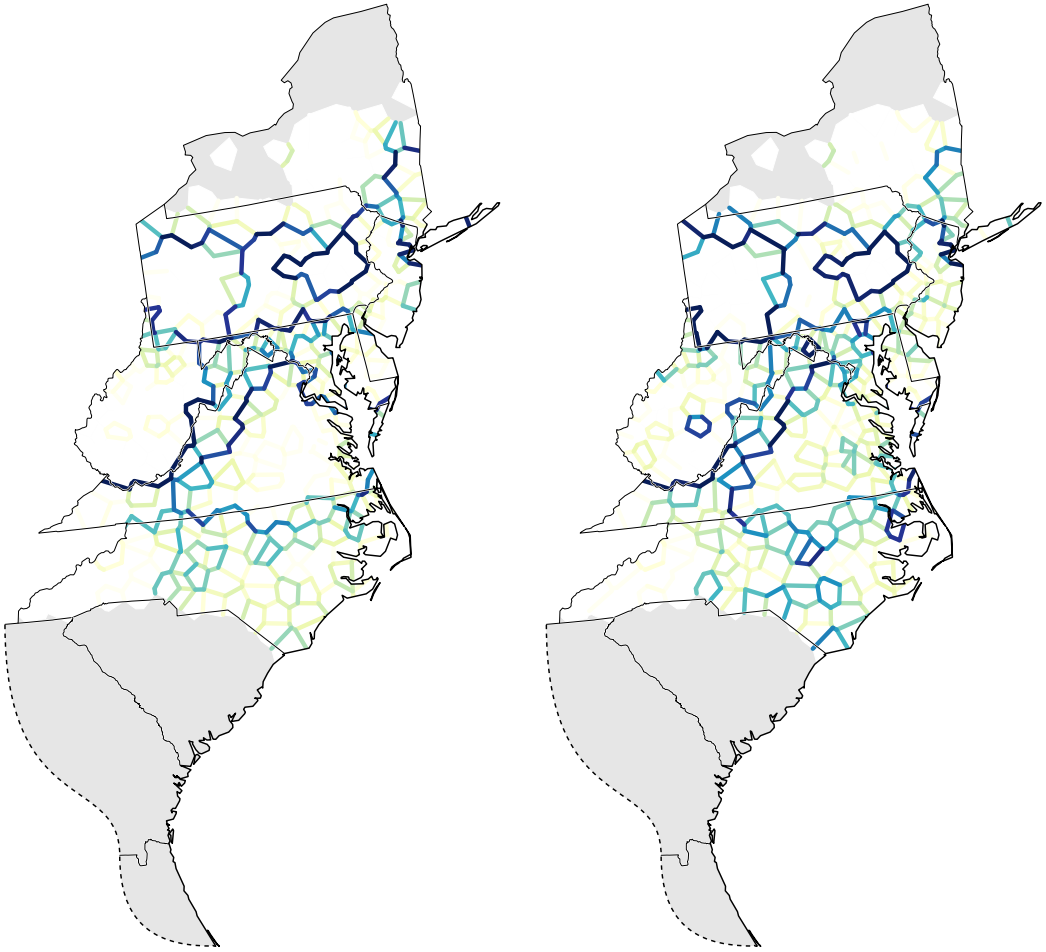
Analyzing Results: Clustering

Composite Clustering.

Repeat (50 times)

- cluster using weighted average and group average, include noise (0.5 standard deviations)
- draw each division in each dendrogram, redrawing when division is re-used (making lines darker through redrawing)

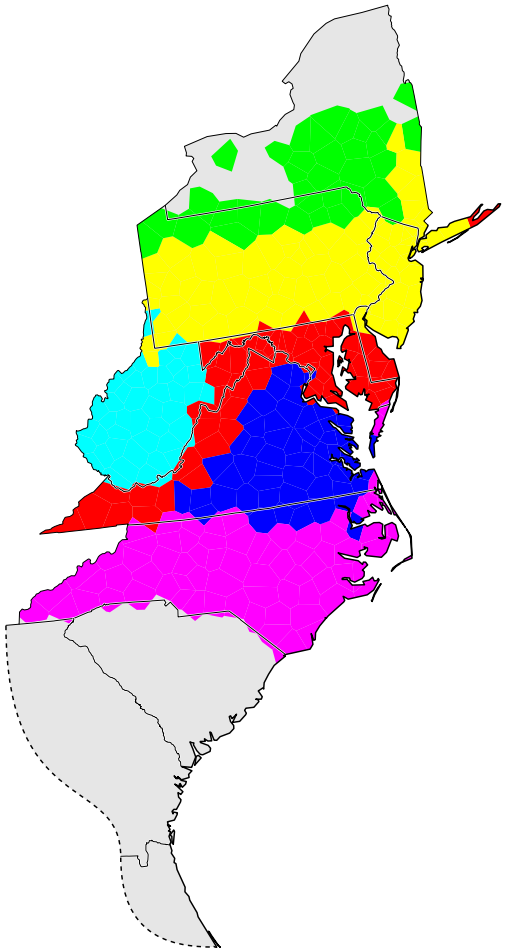
Symbols vs. Features

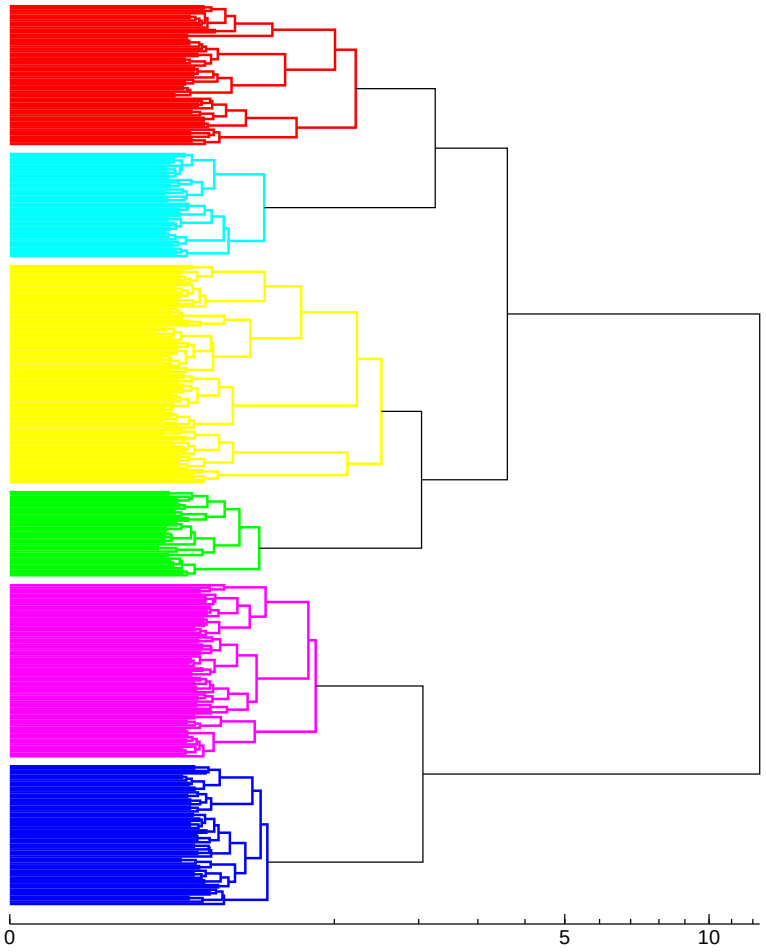


RuG

Symbols

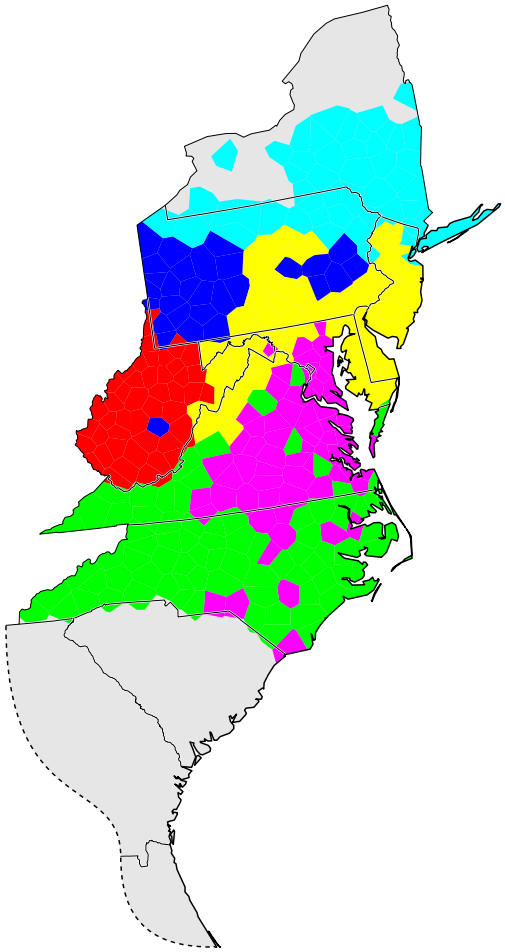
LSA Dialects

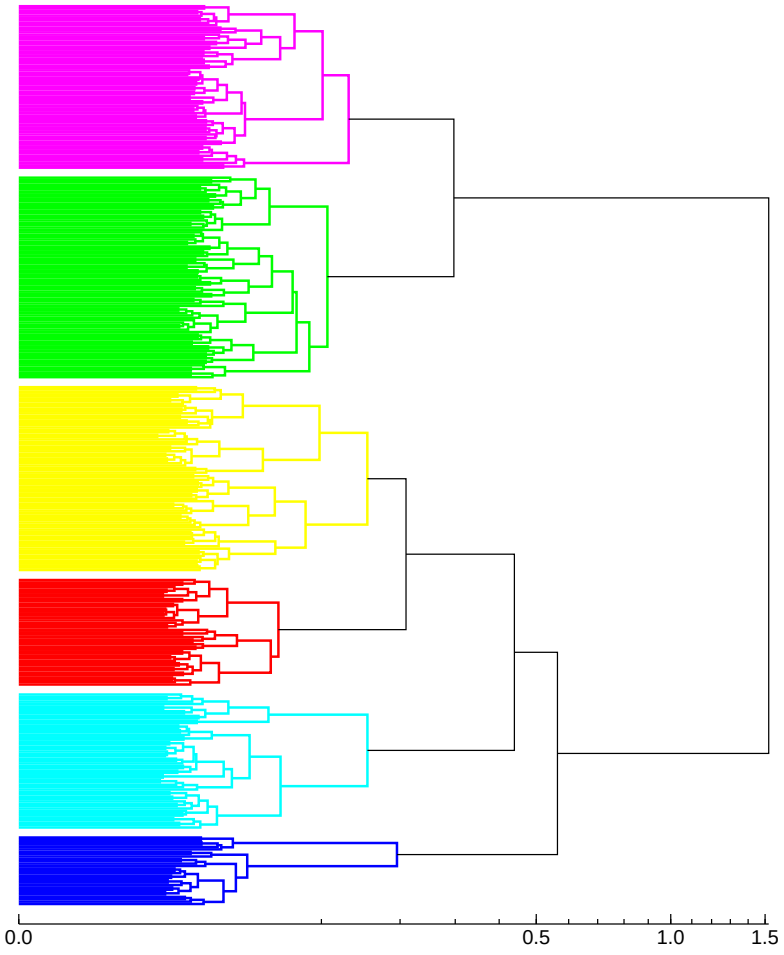




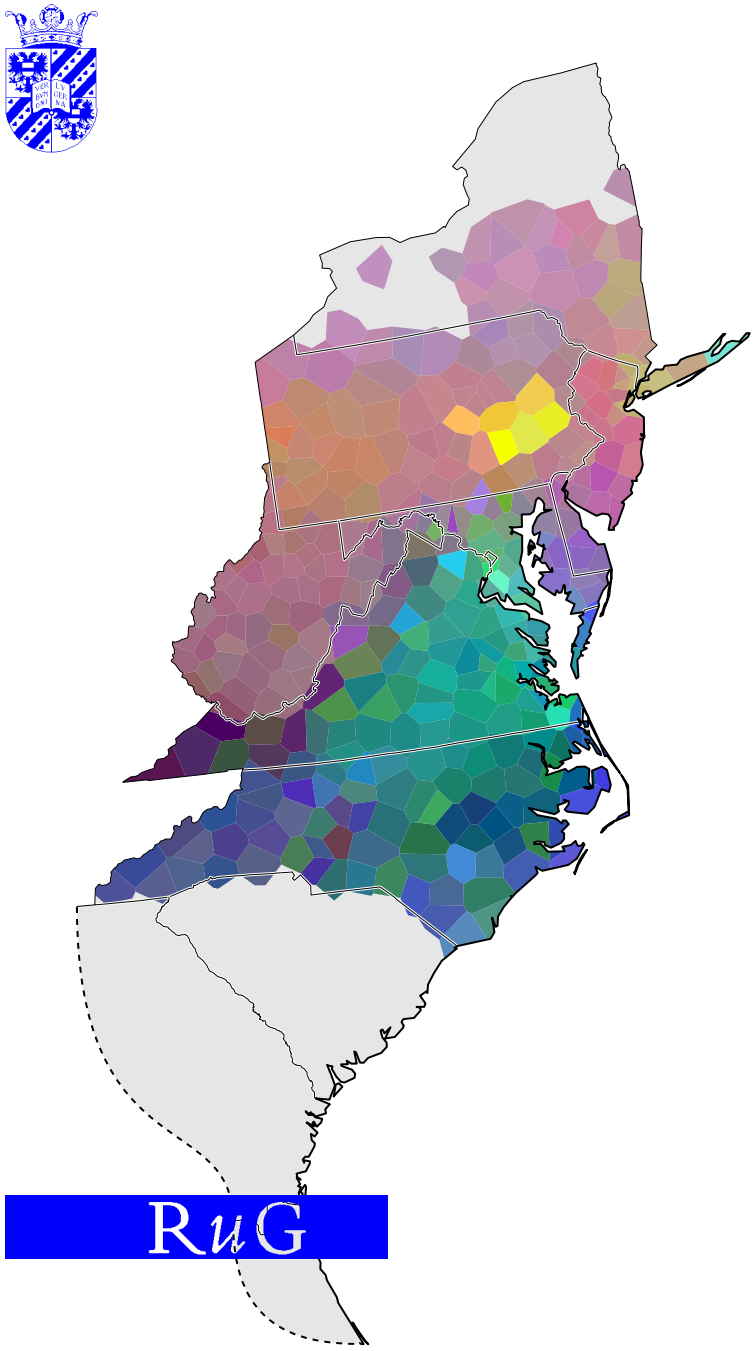
RuG

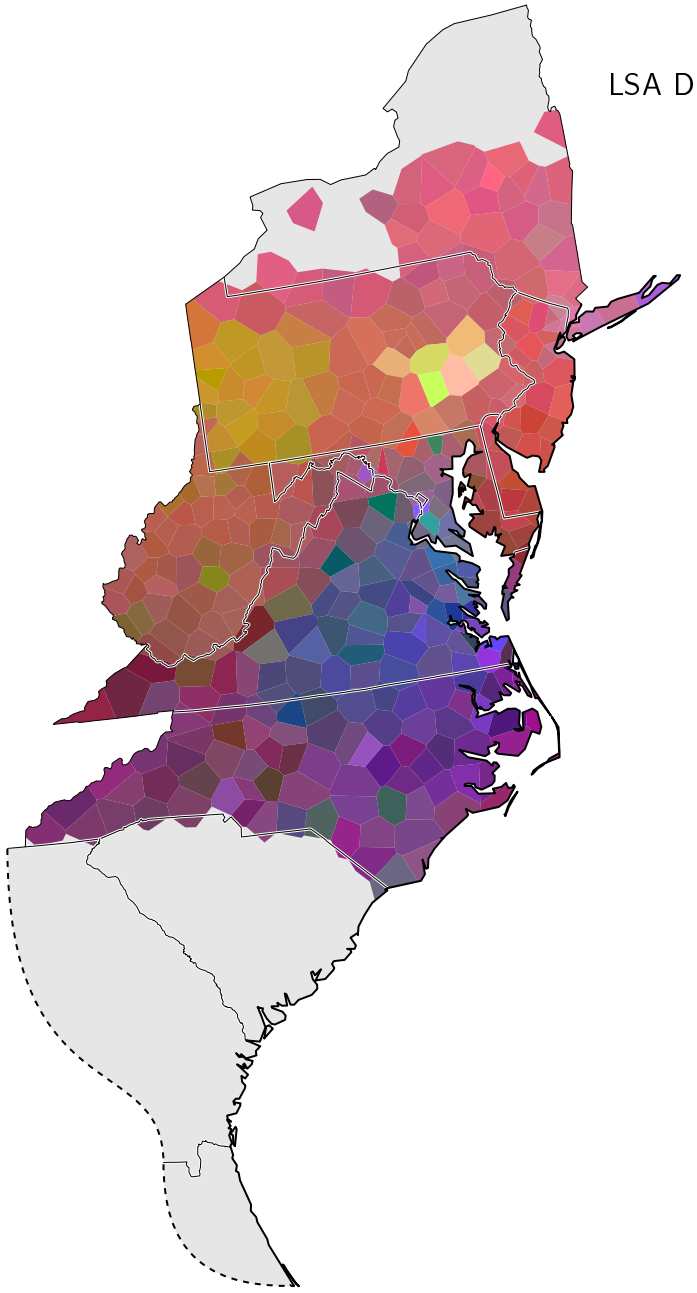
Features





RuG





LSA Dialects

RuG

Conclusions

- Levenshtein distances provides a useful way to measure pronunciation distance among dialect atlas data, alllowing cartography emphasizing dialect areas as well as dialect continua
- Puzzle: the more senstive feature-based techniques do not appear to lead to much improved analyses. Either the analyses are little better or the situation is very noisy (or both)
- We need more rigorous assessment of which techniques are probative