Are there universal principles determinig phonological word size?

Balthasar Bickel[1], Kristine Hildebrandt[2], and René Schiering[1]

[1]University of Leipzig

[2]University of Manchester

The word in theory

- A universal: all languages have exactly one phonological domain between the foot and the phrase, and this is the p-word (Nespor & Vogel 1986, Dixon & Aikhenvald 2002, etc), which serves as a domain for sound patterns (and in some theories has a minimal length of two moras).
- But what kind of universal is this, absolute or statistical?

2

The word as an absolute universal

- Absolute universals are necessarily true because they follow from the axioms and primitives of one's theory/ metalanguage:

    Both Nespor & Vogel's (1986) and Dixon & Aikhenvald's (2002) metalanguages include the word as a primitive, *a priori* term (on a par with terms like 'contrastive feature' or 'segment'). Call this the 'A PRIORI WORD' theory.

- Empirical challenges cannot come from typological surveys but can only ever arise when the theory makes contra- dictory predictions for the analysis of a single language.

The challenge from Limbu (Kiranti; Sino-Tibetan)

- If we assume the A PRIORI WORD theory, we end up with a contradictory analysis of Limbu because the Limbu word both includes and excludes prefixes at the same time:
  - pf-[stem-sf-cl], domain of Liquid Alternation and ʔ-Insertion

    kɛ-[**L**eː-**L**e=**L**o] > kɛ[**l**eː**r**e**r**o] 'your penis!'
    2sPOSS-penis-GEN=PTCL

  - [pf-stem-sf-cl], domain of Coronal Assimilation and Stress

    [mɛ-**n**-mɛ**t**-paŋ] > [mɛ**m**mɛ**p**paŋ] 'We did not tell him'
    nsA-NEG-tell-1>3.PST

- Any rescue?

Schiering et al, Ms. submitted; Hildebrandt 2007 in *J. Himalayan Linguistics* 8, 1 -34

Trying to rescue the word as an absolute universal

- Claim that one Limbu word is the real one; the other is not really a prosodic domain but is an epiphenomenon of lexical properties of affixes or due to something else

  No evidence for this. Both patterns are fully general across the lexicon, and if their description is to be adequate, it must include a proper domain delimitation.

- Posit strata: prefixes apply at a different stratum than suffixes.

  In Limbu, genuine clitics (phrasal affixes, lacking stem subcategorization) are included in both domains, so we would have to posit two postlexical domains, one including prefixes, one excluding prefixes. This shifts the problem from the word to the clitic-group domain, but it does not solve it.

- Claim recursive structure: [ω [ω]]

  But that wrongly predicts that the two word domains have the same phonological properties.

- Relativize prosodic structure to sound patterns, e.g. tone vs. quantity (Hyman et al.'s 1987 proposal for multiple word domains in Luganda)

  But that wrongly predicts that the two word domains relate to different types of phonological patterns.

Alternative: the word as a statistical universal

- This presupposes a typological variable, whose possible values are the language-specific word domains, e.g.
  - The Limbu Coronal Assimilation Word
  - The Limbu Liquid Alternation Word
  - The Kyirong Tibetan Tone Word
  - etc.
- This was the point of departure of the Leipzig Word Project:
  - collect information about individual words
  - then, explore universal trends within this, including the old claim about domains between foot and phrase, **but now as a probabilistic hypothesis:**

    Languages ***tend to*** *have exactly one domain*.

Building a database of phonological words

- Working definition: pw-pattern = any sound pattern that
  - is delimited by some morphological structure,
  - includes up to one stem (i.e. ignore compounds, for now)
  - is general across the lexicon (for now)
- NB: this excludes smaller domains like the foot (as feet don't reference morphology) and the phrase (as phrases license more than one stem).

7

PW-patterns in a bottom-up, AUTOTYP database

138 **LID** 674 Limbu **UNIT** P Domain

Ppattern_ID 448 Exactly 1 _

Exactly 1 Main Stress

Coding:

stress | stress | suprasegmental

NB: Information about the nature of the p-pattern (kind of pattern, resolution patterns, if any) are now stored in ppatterns_def, as proprties of each ppattern.

PROCESS_STRATA

If source is loanword:

LID1 LID2 Source

ppattern_ID	word_type_def::Pword_Definition	word_type_def::word_type	ppattern1	ppatte
429	The vowel of a final syllable is lengthened in open syllables and those	Final Vowel Lengthening	quantity	quantity
430	no velar onset consonant in this domain	*C Onset if Velar	weakening	process
431	A series of 2 homorganic consonants adjacent across some kind of	*C.C if homorganic	dissimilation	process
432	A series of 2 homorganic consonants adjacent across some kind of	*C.C if homorganic	deletion	process
433	A series of 2 consonants adjacent across some kind of phonological	*C.C	assimilation	process
434	List of onsets not permitted in this domain	*C onset if /ğ, c, f, j, l, m, n, r, z/	insertion	process
435	List of codas not permitted in this domain	*C coda if /b, c, g/	weakening	process
436	The coda velar plosive unasp unvoiced is banned/dispreferred in this	*C if k	weakening	process
437	ambisyllabic germinate only here	Geminate only here	quantity	quantity
438	consonant clusters prohibited in this domain	*CC	constraint	constrai
439	consonant clusters prohibited in this domain	*CC	insertion	process
440	A series of 3 adjacent consonants (regardless of syllable/foot	CCC series (ambisyllabic) only here	constraint	constrai
441	List of codas not permitted in this domain	*C coda if /p, ç, t/	weakening	process
442	List of codas not permitted in this domain	*C coda if /p, ç, t/	quantity	quantity
443	An obstruent assimilates for voicing/phonaton/laryngeal features of	Obs Voicing Assim	weakening	process
444	There is an unspecified type of vowel harmony in this domain	Vowel Harmony	assimilation	process
445	2 syllables with series of vowel-only, or where first syllable is open and	*V.V	insertion	process
446	List of codas not permitted in this domain	*C coda if /b, c, g/	constraint	constrai

Max number categories in sequence

MORPH DOMAIN Align ID 2 Domain ID 173 Size: 4

left edge

stem ± prefix ± suffix ± postposed particle

Less strict domain definition ("DomMrg"): 56

Position 1	n/a	**Type 1**	stem	**Restriction 1**	unrestricted
Strata_1					
Position 2	prae	**Type 2**	formative	**Restriction 2**	
Strata_2					
Position 3	post	**Type 3**	formative	**Restriction 3**	
Strata_3					
Position 4	post	**Type 4**	formative	**Restriction 4**	unrestricted
Strata_4					
Position 5		**Type 5**		**Restriction 5**	
Strata_5					
Position 6		**Type 6**		**Restriction 6**	
Strata_6					
Position 7		**Type 7**		**Restriction 7**	
Strata_7					

M…	Morpheme…	Definition
1	formative	Marker of an inflectional category that cannot occur as an independent part of spe
3	any	Can be used both ways, e.g. SEA versatile verbs/coverbs
0	n/a	In domain_def, use if no other morph domain parts relevant in domain definition
7	unknown	type is unknown at this time
13	stem part	An element of a stem (as defined under ID 12) that is not delimited by general
15	none	There is no overt manifestation of the exemplar for whcih the record is made

Mad

Stres

WghtStr1

…	RestrValue	RestrDef
2	semi-restricted	Can occur wtih some, but not all POS/head elements (so, NOT a 'phrasal affix', but rather
3	unrestricted	Can occurs with anything (e.g. Turkish mi)
4	n/a	not applicable
5	unknown	degree of restriction unknown at this time
6	Part circum/simulfix	even more restricted than restricted; used with simulfixation & circumfixation
7	restricted: Head	Something like affix only to a head element (in a narrowed sense of RestrID 1
8	restricted: Phrase	Can occur with whatever POS element of that phrase it is adjacent to like Manange NP

Binary Recode of this, as used in scripts ("1" = structure preserving): 1

oherence (relative to possible size) 1

Phonology Process Filter (word_type_def) 5

No phrasal stuff (value= '2') (domain_def) subphrasal

NOTES Affixes (and enclitics I suppose) are ususally not stressed. Verbs and deverbatives are stressed on the root, nouns and other parts of speech are stressed on the first syllable. [RS]

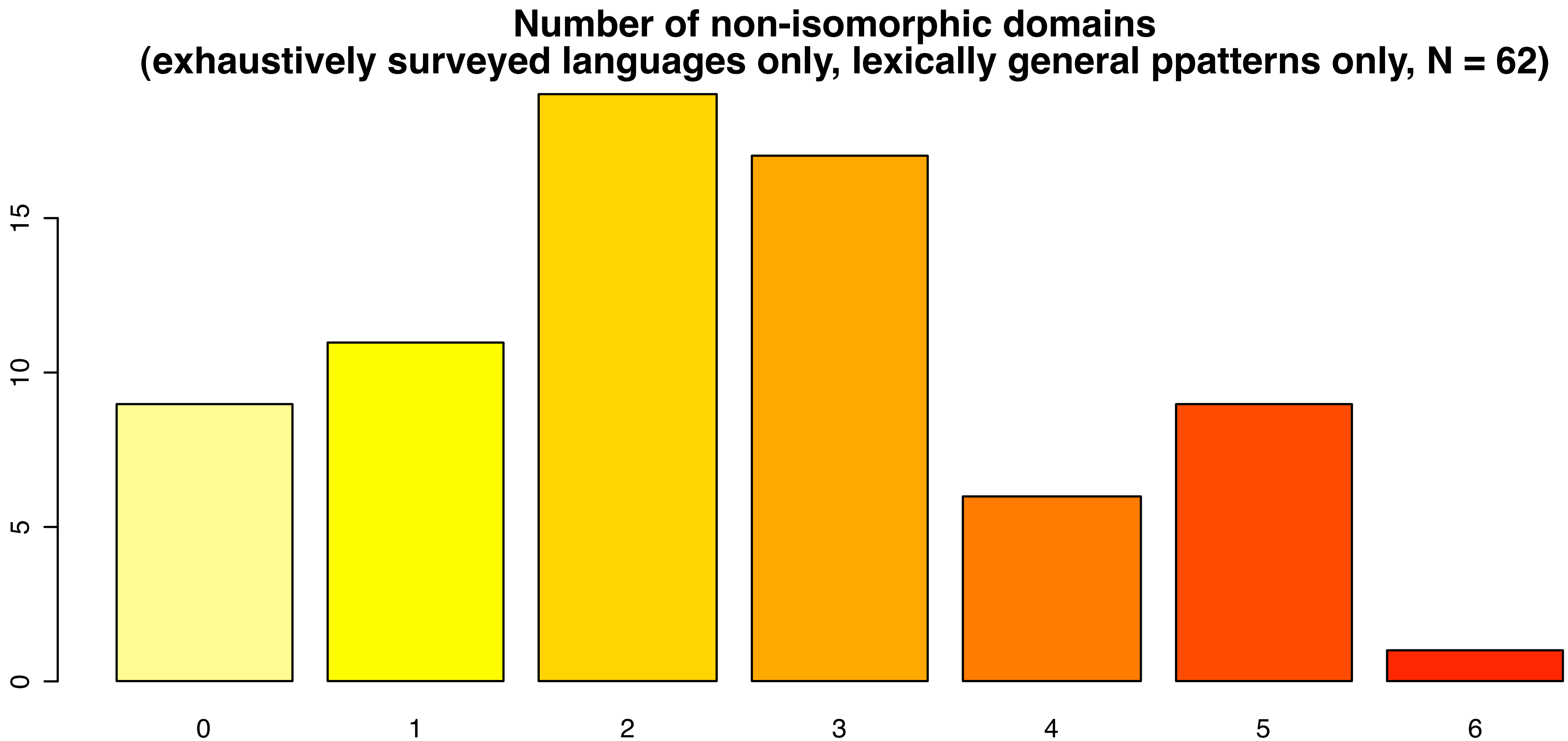
Data coverage

- 72 languages
- In 9 of these, we have not found any evidence for pw-patterns because no known sound pattern is strictly sub-phrasal and fully general across the lexicon.
- The other 63 languages have
  - between 1 and 19 pw-patterns, most between 1 and 5
  - between 2 and 7 morpheme types, most between 2 and 4

9

Hypothesis I

- A statistical universal: *languages tend to have exactly one domain between foot and phrase*
- The reality:



10

A new question

- If there are no categorical clusters on which pw-patterns converge, are there probabilitistic clusters depending on the type of phonological pattern involved?
- To find out, we need

1. a means of comparing word domains across languages
2. a taxonomy of phonological pattern types

11

Coherence: a measurement for comparing word domains

- How many morpheme types are included in the domain? (stem alone? stem plus prefix? plus prefix and suffix? etc.)
- Obviously, this depends on what is available in a language. Therefore, for each pw-pattern p in each language L, compute:

$$c\,(p, L) = \frac{N\,(\text{morpheme types referenced by } p)}{N\,(\text{morpheme types in } L)}$$

12

Measuring coherence: examples

- Limbu Coronal Assimilation:

a. /mɛ-**n-m**ɛ**t-p**ɛŋ/ [mɛ**mm**ɛ**pp**aŋ] ‘I did not tell him’

nsA-NEG-tell-1s>3.PST

b. /hɛ**n = p**hɛlle/ [hɛ**mb**hɛlle] ‘What?’

what-QUOT

$$\frac{4\ \text{(prefix-stem-suffix=clitic)}}{4\ \text{(prefix-stem-suffix=clitic)}}$$

→ *c* (Limbu Coronal-to-Labial Assimilation) = 1

13

Measuring coherence: examples

- Limbu Liquid Alternation

a. /nɛ**l**ɛt/ [nɛ**r**ɛt] ‘heart’

b. /pha-**l**e siŋ/ [pha-**r**e siŋ] (bamboo-GEN wood) ‘the wood of bamboo’

c. /peːg-i=**l**oː/ [peːg-i=**r**oː] (go-p=ASS) ‘Come on, let’s go!’

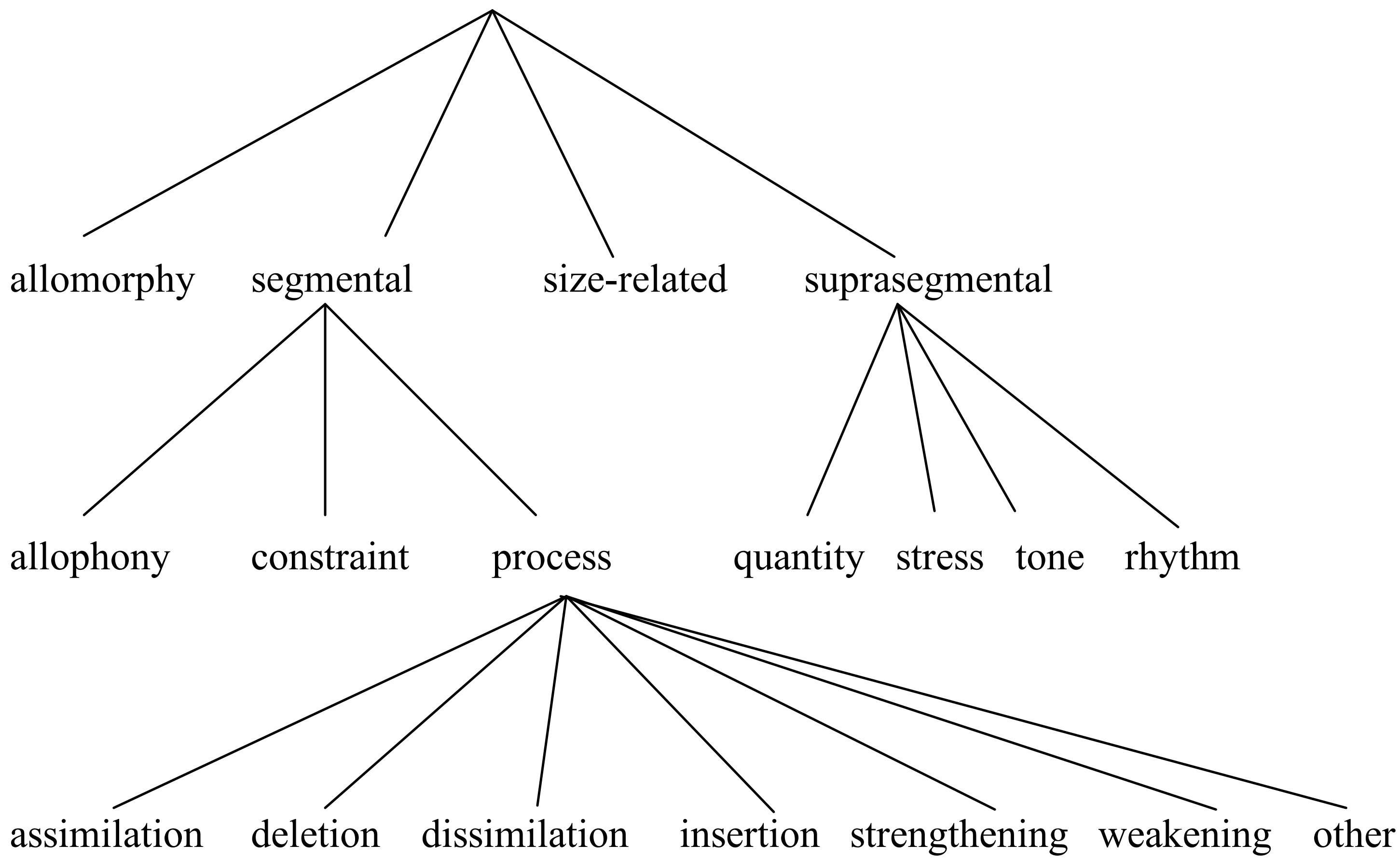
d. /kɛ-**l**ɔʔ/ [kɛ-**l**ɔʔ] (2-say) ‘you say’

$$\frac{3 \text{ (stem-suffix=enclitic)}}{4 \text{ (prefix-stem-suffix=enclitic)}}$$

→ *c* (Limbu [l] ~ [r] domain) = .75

14

A taxonomy of pw-pattern types



Combining coherence and type

::Language	word_type.def::word_type	::ppattern1	full_id1	coh_rltv	::plevel	::unit	::IE…	::Reliability	::statu
Kusunda	*C Coda	constraint	constraint_Kusund526	.333333333	subphrasal	P Domain		Questionnaire	pword exhau
Kusunda	*V-Initial syllable	allomorphy	allomorphy_Kusund527	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	*V: unless Here	quantity	quantity_Kusund528	.666666667	subphrasal	P Domain		Questionnaire	pword exhau
Kusunda	*V.V	strengthening	strengthening_Kusund52	.666666667	subphrasal	P Domain		Questionnaire	pword exhau
Kusunda	*V.V	weakening	weakening_Kusund530	.666666667	subphrasal	P Domain		Questionnaire	pword exhau
Kusunda	*C Coda	constraint	constraint_Kusund531	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	Vowel Pharyngealization	assimilation	assimilation_Kusund532	.333333333	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	Vowel Nasalization	assimilation	assimilation_Kusund533	.333333333	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	*C if p^h	constraint	constraint_Kusund534	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	*C if ɢ	constraint	constraint_Kusund535	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	Nasal Segment Palatalization	assimilation	assimilation_Kusund537	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	Uvular /q/ Voicing Assimilation	assimilation	assimilation_Kusund538	.666666667	subphrasal	P Domain		Grammar Explicit	pword exhau
Kusunda	Voiced Uvular Plosive Manner Assimilation	weakening	weakening_Kusund539	.333333333	subphrasal	P Domain		Grammar Explicit	pword exhau
Lahu	Stress Reduction	stress	stress_Lahu127	.5	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Lahu	Tone Change	tone	tone_Lahu128	.5	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	Min VC	allomorphy	allomorphy_Limbu123	0	n/a	P Domain	1	Grammar Implicit	pword exhau
Limbu	C POA Assim	assimilation	assimilation_Limbu124	1	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	*C if Velar Nasal	constraint	constraint_Limbu126	.75	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	Exactly 1 Main Stress	stress	stress_Limbu138	1	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	Exactly 1 Main Stress	stress	stress_Limbu326	.5	subphrasal	P Domain	1	Field Notes	pword exhau
Limbu	*C if r	constraint	constraint_Limbu377	1	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	/l/ > [r] alternation	allophony	allophony_Limbu1026	.75	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	/l/ > [r] alternation	allophony	allophony_Limbu1027	.5	phrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	*V-Initial syllable	insertion	insertion_Limbu1031	.75	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	*V-Initial syllable	insertion	insertion_Limbu1032	.25	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	C POA Assim	assimilation	assimilation_Limbu1033	.5	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	Glottal Stop /ʔ/ assimilation	assimilation	assimilation_Limbu1034	.5	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Limbu	C POA Assim	assimilation	assimilation_Limbu1037	.5	subphrasal	P Domain	1	Grammar Explicit	pword exhau
Lithuanian	Superheavy VVC only Here	constraint	constraint_Lithua636	.25	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	Superheavy V:C only Here	constraint	constraint_Lithua637	.25	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	Superheavy V:C only Here	constraint	constraint_Lithua638	.5	phrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	*V.V	insertion	insertion_Lithua645	.25	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	*V.V	deletion	deletion_Lithua646	.5	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	Onset clusters dispreferred & restricted	constraint	constraint_Lithua657	.25	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	Exactly 1 Main Stress	stress	stress_Lithua658	1	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	Exactly 1 Main Stress	stress	stress_Lithua659	1	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	*C^j (Palatalized C)	constraint	constraint_Lithua660	.25	subphrasal	P Domain	1	Questionnaire	pword exhau
Lithuanian	C Palatalization	assimilation	assimilation_Lithua661	.5	subphrasal	P Domain	1	Questionnaire	pword exhau

16

Exploring structure in the coherence data

1. Calculate a distance matrix

	constraint Nepali 81	constraint Arabic 82	weakening Lithuanian 673	deletion Lithuanian 674	stress Sko 675	size-related Semelai 881	constraint Mon 936
constraint Nepali 81	0						
constraint Arabic 82	0.36	0					
weakening Lithuanian 673	0	0.36	0				
deletion Lithuanian 674	0	0.36	0	0			
stress Sko 675	0.5	0.86	0.5	0.5	0		
size-related Semelai 881	0.21	0.57	0.21	0.21	0.29	0	
constraint Mon 936	0.3	0.06	0.3	0.3	0.8	0.51	0

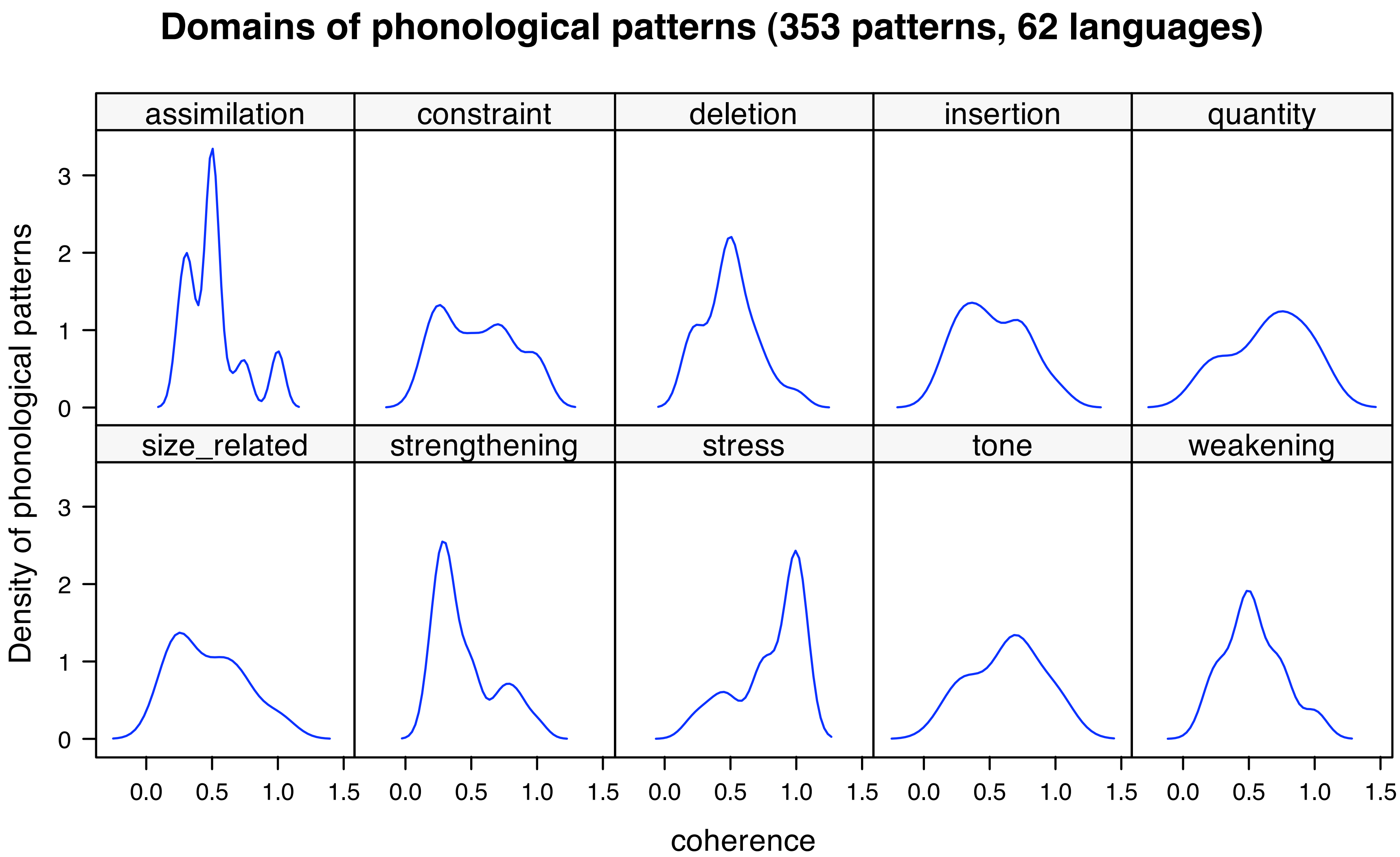
17

2. Multidimensional Scaling

18

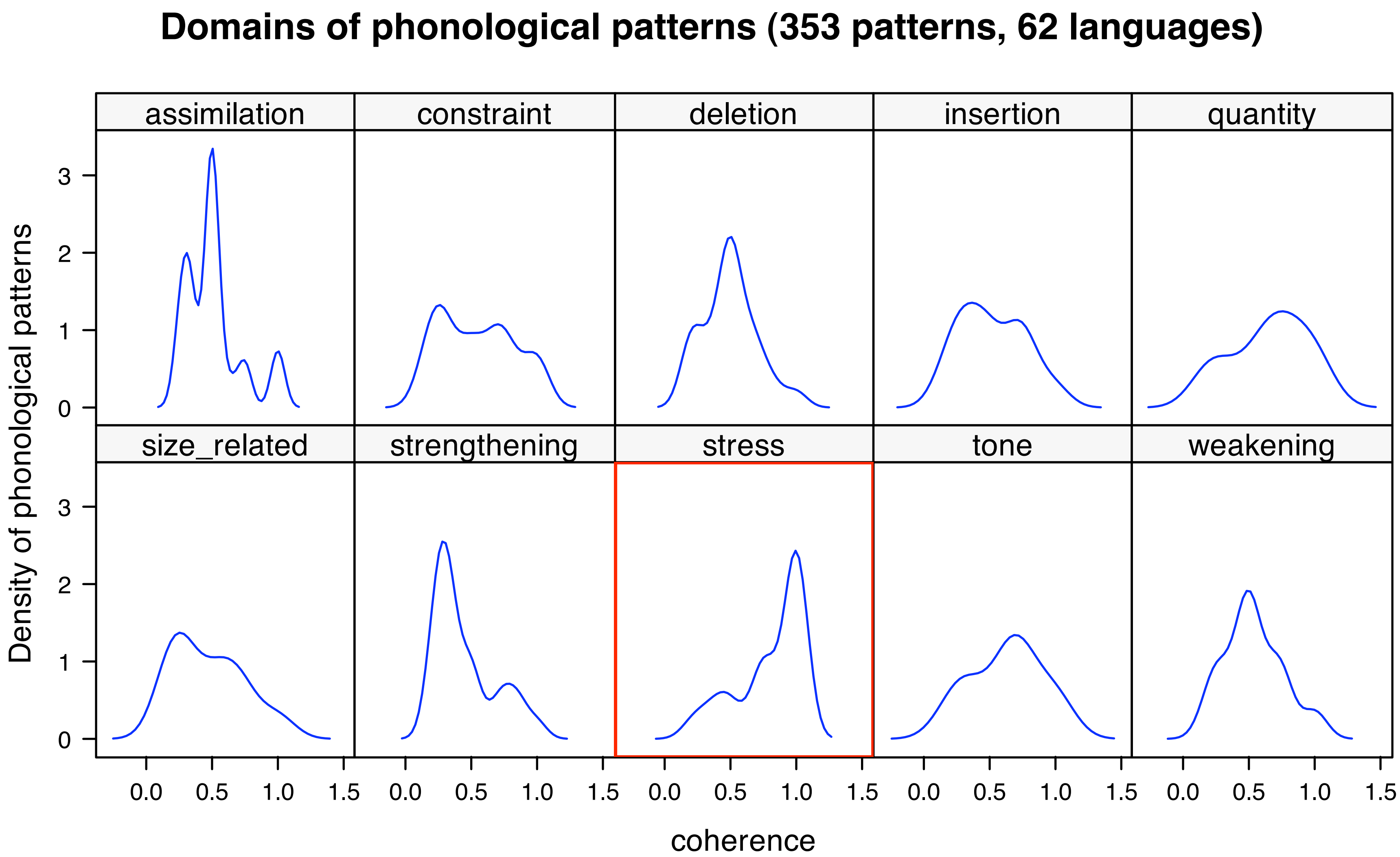
Results

Taking coherence as the measurement, we discover a probablistic cluster of stress-defined pw-patterns:



Results

Taking coherence as the measurement, we discover a probablistic cluster of stress-defined pw-patterns:



Hypothesis II: a statistical universal

Stress-related domains tend to be universally larger than other domains.

- *Hypothesized to be very common:*

  Limbu (Sino-Tibetan) Stress: [prefix-ˈstem-suffix=clitic]

  [mɛ-ˈthaŋ-e=aŋ]
  3ns-come.up-PST=and

- *Hypothesized to be much less common:*

  Mon (Austroasiatic) Stress: [ˈcl]=[pf<infix>ˈstem]=[ˈcl]

  [k<ə>ˈlɒʔ]
  <CAUS>cross

  [ˈkɒ]=[ˈklɒʔ]
  CAUS=cross

21

Testing Hypothesis II

- Apart from the difference between stress-defined vs other pw-patterns, two other factors are likely to affect the shape of phonological word domains:
  - **areality**: for example, South-East Asia is known for its ‘prosodic diffusibility’ (Matisoff 2001)
  - **families**: phonologies tend to be conservative within genealogical units (Blevins 2004)
- Therefore, test the effects of each factor and of each interaction in a multiple regression model:

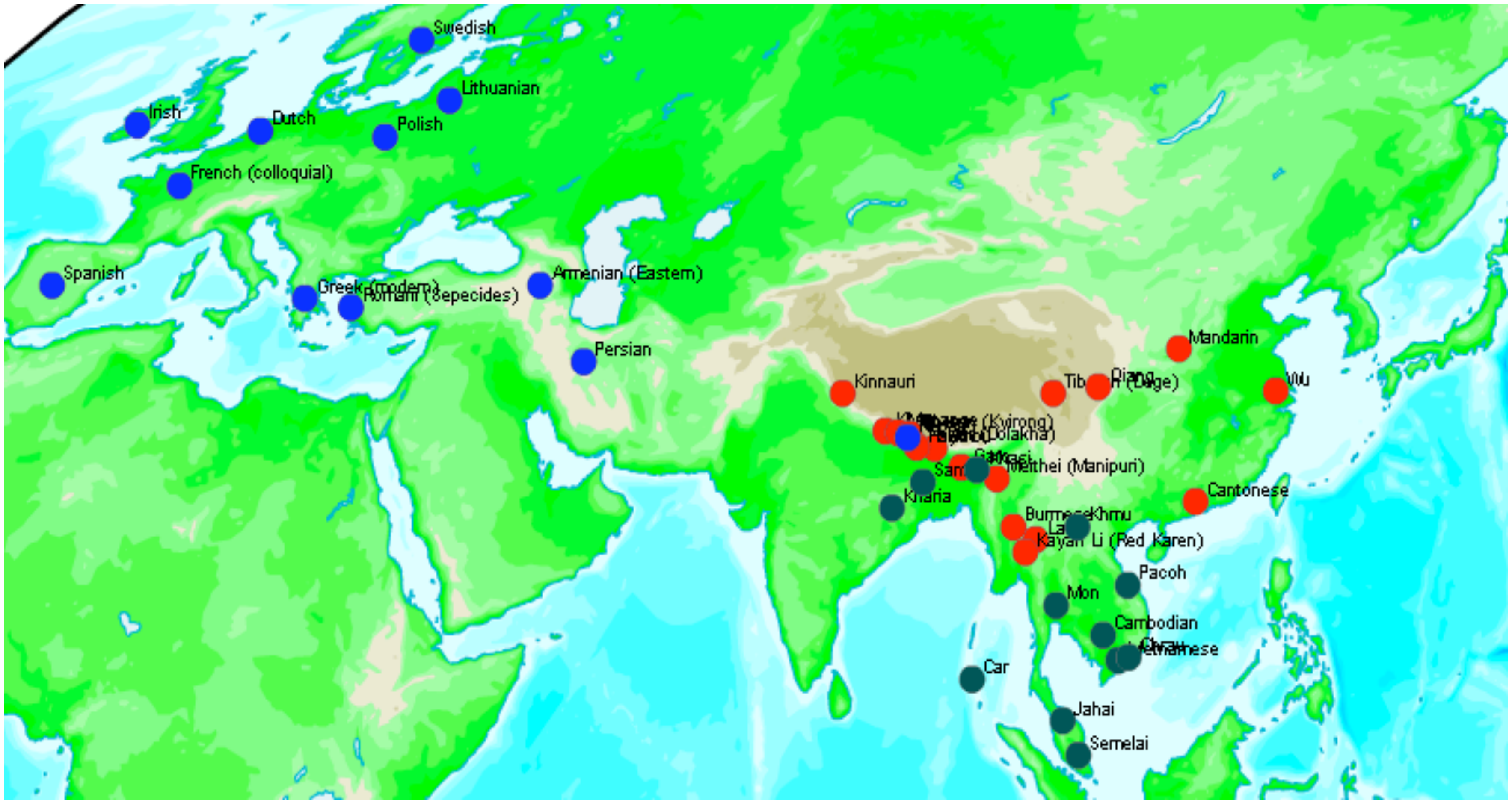
$$\mu(c) \sim \alpha + \beta[\text{PW-PATTERN}] \times \gamma[\text{FAMILY}] \times \delta[\text{AREA}]$$

- Test this against a sample that is stratified for family and area, as follows:

22

Factor FAMILY

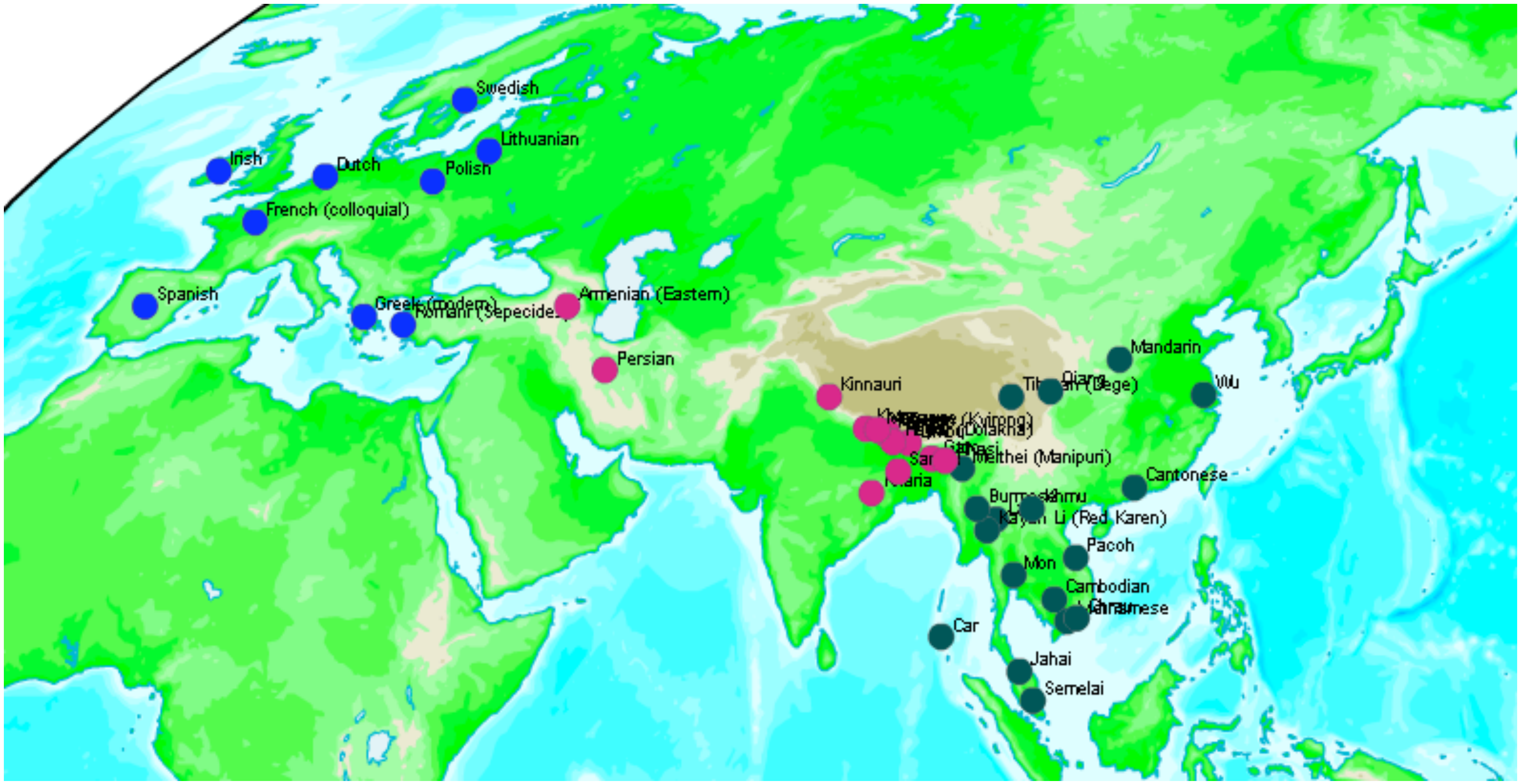
For this, take one representative per sub-branch of major branches in three families (or two if phonologies known to be diverse and data are sufficient): Austroasiatic (11), Indo-European (12), Sino-Tibetan (17)



23

Factor AREA

For this, take standard AUTOTYP linguistic area definitions, reassigning stray (e.g. Armenian) and border languages (e.g. Romani), though this had no impact on any result.



24

Results

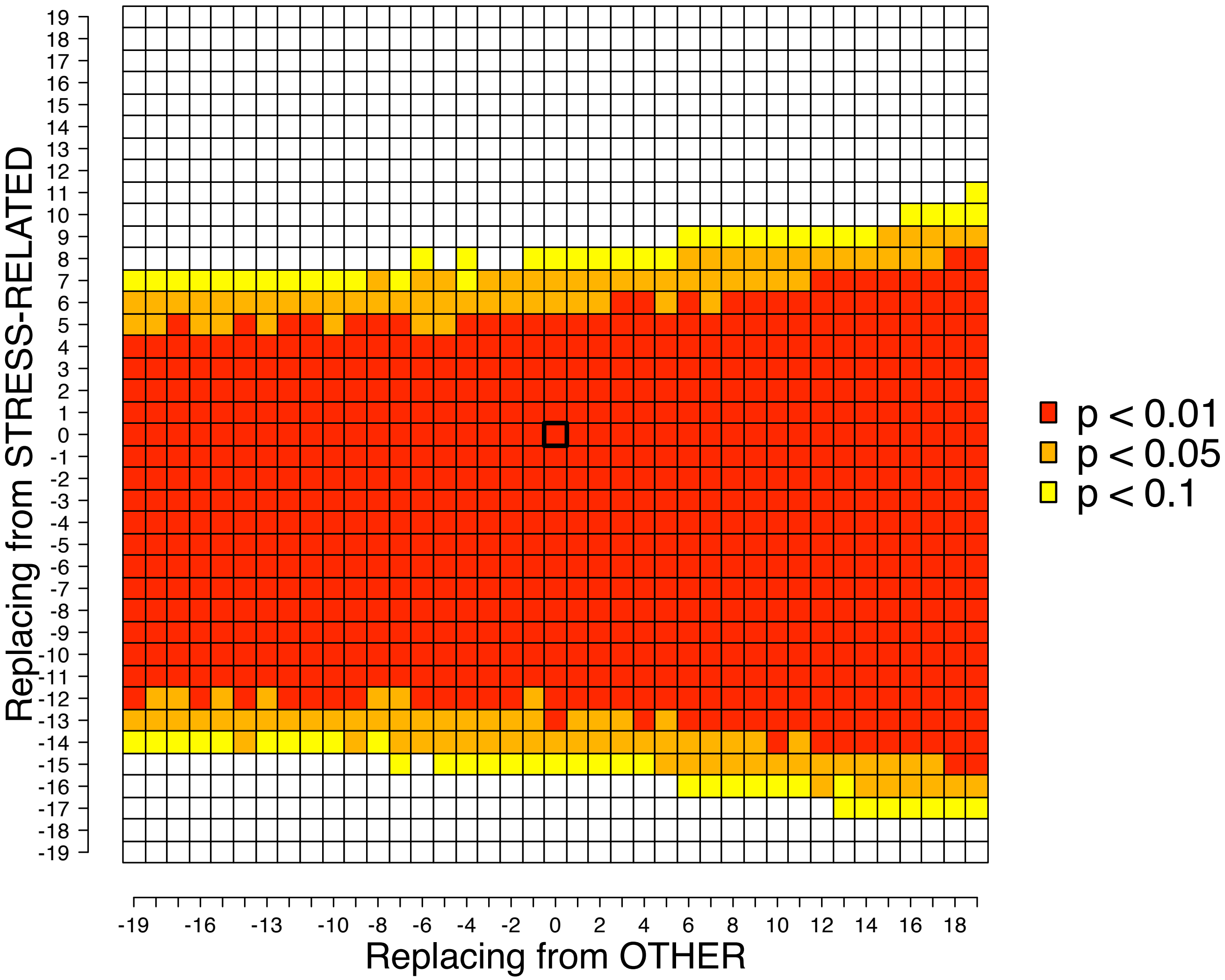
Based on 238 pw-patterns in 40 languages, using Randomization tests (Janssen et al. 2006), we find:

- no evidence for any interactions between any factors;
- no evidence for AREA effect ($F(2)=.92$, $p=.51$); also when removing the areal borderline languages of our sample, i.e. Romani, Armenian, and Persian ($F(2)=.92$, $p=.39$);
- a significant main effect of FAMILY ($F(2)=11.03$, $p<.0001$)
- a significant main effect of PW-PATTERN ($F(1)=20.99$, $p=.0001$)

25

Reliability Analysis

Since there are many less stress-related pw-patterns (19) than others (222), we also performed a Reliability Analysis (Janssen et al 2006), replacing critical values of *c* by their grand mean:

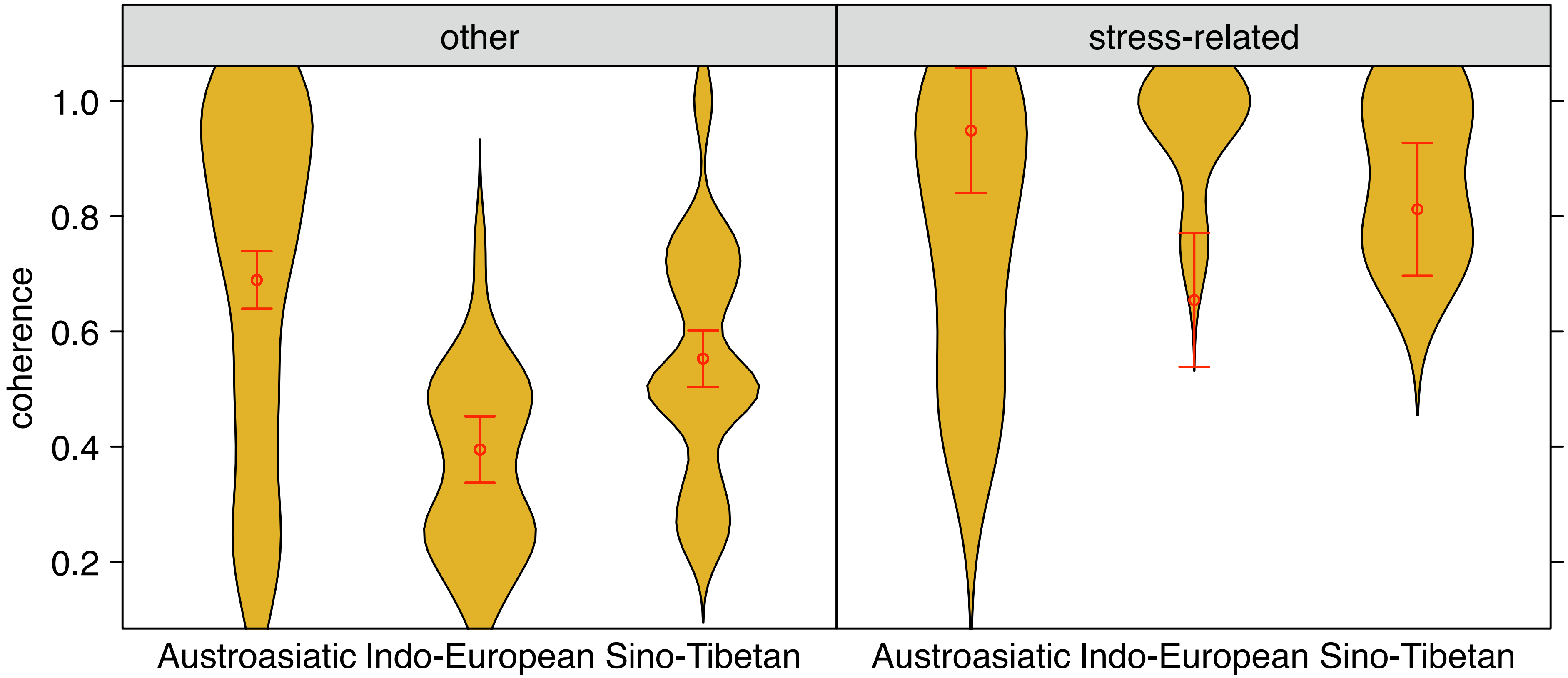


26

Summary

The best-fitting model is

$\mu(c) = .69 + .26$ [STRESS vs OTHER] - .30[IE vs AA] - 1.4 [ST vs AA]



27

Conclusions

- Stress-defined domains tend to be significantly larger than other domains.
- No other pw-pattern has a systematic impact on domain size (coherence); tone, for example, does not target different sizes than any segmental pattern.
- This finding is compatible with traditional conceptions of prosodic structure in which only stress and intonation are necessarily included in hierarchical structures (e.g. Pike 1945)

Conclusions (*cont'd*)

- Family relations also have significant effect on coherence, but this effect is independent of the effect form stress.
- The family effect is likely to reflect a general inertia in phonological change.
- Interestingly, despite the known 'prosodic diffusibility' especially of Southeast Asia, we find no evidence for areal spreads of coherence!

29

Acknowledgments

- Thanks to our student assistants for help in data collection and database programming: Thomas Goldammer, Franziska Crell, Sven Siegmund, Taras Zakharko, Jenny Seeg, Sebastian Hellmann, Josh Wilbur.
- Thanks to the DFG for funding this research (DFG Grant Nos. BI 799/2 and 799/2-3).
- http://www.uni-leipzig.de/~autotyp/projects/wd_dom/wd_dom.html
- All statistical analysis and all plots were done in R 2.4.1 (R Development Core Team 2006).
- Maps were created running Hansjörg Bibiko’s iAtlas tool on our FileMaker Pro database.