Cluster analysis of phonological word domains

Balthasar Bickel*, Kristine A. Hildebrandt#

& René Schiering*

*University of Leipzig, #University of Manchester

Introduction

Theoretical Predictions (Selkirk 1984, Nespor & Vogel 1986, etc.)

U
|
I
|
P
|
ω
|
φ
|
σ
|
μ

Clustering:
Phonological Domains cluster on the set of domains enshrined in the Prosodic Hierarchy (i.e. one and only one ω domain)

Strict Succession (Proper Headedness):
Each level *L* is followed by (at least) one level *L-1* until the terminal level *L=0* (i.e. (at least) one ω in any prosodic tree)

Proper Bracketing:
No language will exhibit non-stacking domains (i.e. no overlapping ω domains)

A problem

The facts on the ground: Limbu (Kiranti, Sino-Tibetan)

P	**Phrase**: voicing assimilation, e.g. /p/ → [b]
|	*pe:kmaʔ* ***bo:ŋ*** ‘it’s time to go’
Prefix Stem Suffix Clitic	**ω_4**: e.g. one stress per word
|	(*ku-ˈtaŋ=mɛ*) ‘it’s horn on the contrary’
Prefix Stem Suffix Clitic	**ω_3**: e.g. [ʔ]-insertion
|	(**ʔ***a-*)(**ʔ***iːr-ɛ*) ‘we wandered’
Prefix Stem Suffix Clitic	**ω_2**: e.g. /m/ → [ŋ]
|	*ha**ŋ**-ŋʔna* ‘being sent’
Prefix Stem	**ω_1**: e.g. restructured stress (prefix-stem)
|	(*ˈku-laːp*) ‘it’s wing’
ϕ	**Foot**: trochaic rhythm (secondary stress)
|	*ʔaˈʔoŋ ˌŋe:* ‘my brother in law!’
σ	**Syllable:** C(G)V(C)

A problem

The facts on the ground: Limbu

U
|
I
|
P
|
ω
|
ϕ
|
σ
|
μ

***Clustering**:
Phonological domains in Limbu cluster on more domains than provided by the PH
(i.e. four ω domains)

***Strict Succession (Proper Headedness):**
A level ω may be followed by another level ω in Limbu
(i.e. ω is multiplied in every prosodic tree)

***Proper Bracketing:**
ω_1 and ω_2 constitute non-stacking domains
(i.e. overlapping ω domains)

Possible solutions

- One Limbu ω is the real one; the others are not really prosodic domains but lexical properties of affixes or due to something else

  In fact ω_1 is coerced by a constraint against end stress and ω_2 is limited to some lexically specified affixes — but ω_3 (glottal insertion excluding prefixes) and ω_4 (stress including prefixes) remain!

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

  In Limbu, clitics are included in both ω_3 and ω_4 domains, so both would be postlexical strata. But there is no evidence that affixes are postlexical in Limbu.

- Recursive structure: [ω [ω]]

  But that wrongly predicts that ω_3 and ω_4 have the same phonological properties!

- Relativize prosodic structure to sound patterns, e.g. tone vs. quantity (Hyman et al. 1987)

Possible solutions

Tone and quantity in Luganda (cf. Hyman et al. 1987)

a. QD $((\text{tú-ly-áá})_{\omega}\ (\text{kô})_{\omega})_{C}$ ‘we eat a little’

TD $((\text{tú-ly-áá})_{\omega}\ (\text{kô})_{\omega})_{C}$

b. QD $((\text{te-tú-ly-à})_{\omega})_{C}\ ((\text{mu-púùnga})_{\omega})_{C}$ ‘we don’t eat rice’

TD $((\text{te-tú-ly-à})_{\omega})_{C}\ ((\text{mu-púùnga})_{\omega})_{C}$

c. QD $((\text{tú-ly-á})_{\omega})_{C}\ ((\text{mú-púùnga})_{\omega})_{C}$ ‘we eat rice’

TD $((\text{tú-ly-á})_{\omega}\ (\text{mú-púùnga})_{\omega})_{C}$

d. QD $((\text{te-tú-ly-àà})_{\omega}\ (\text{kô})_{\omega})_{C}$ ‘we don’t eat rice’

TD $((\text{te-tú-ly-àà})_{\omega})_{C}\ ((\text{kô})_{\omega})_{C}$

- Prosodic structure is independently construed on different phonological tiers (tone vs. quantity)
- **But there is no evidence that Limbu domains differ as to tier!**

Goals

- Turn the PH from a UG declaration into a hypothesis of what structures languages actually evidence, i.e. turn the ‘word’ from a universal a priori into a typological variable

- Explore what factors govern word structures in a cross-linguistic sample:
  - explore sound pattern type by standard methods of Dissimilarity Analysis (Multidimensional Scaling, Clustering, Neighbornet)
  - test the effects of sound pattern type controlling for areal and genealogical factors

- discuss the consequences of the typological findings for theory architecture.

Database and coding

AUTOTYP database on 72 languages

- word-defining phonological patterns, e.g. stress, tone, segmental rules, phonotactic constraints, etc.

  Range = (1,26), *Mean* = 9.5, *Mode* = 12

- morpheme types, e.g. postposed, restricted formatives (‘suffixes’), preposed, unrestricted formatives (‘proclitics’, ‘particles’), stems, etc.

  Range = (2,7), *Mean* = 4.25, *Mode* = 4

- domain types, i.e. what strings of morpheme types are referenced by a phonological pattern

  Range: (1,10), *Mean* = 3.87, *Mode* = 4

Database and coding

- Measuring coherence: 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:

$$c = \frac{N(\text{morpheme types in domain})}{N(\text{available morpheme types})}$$

Range = (.14, 1), *Mean* = .54, *Mode* = .5

Database and coding

Examples:

- Limbu stress domain: *c* = 1

  /mɛ-'thaŋ-e=aŋ/ 'they come up and …'

  4 (prefix-stem-suffix=particle)
  4 (prefix-stem-suffix=particle)

- Limbu coronal to labial assimilation: *c* = 1

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

  /hɛ**n**=**p**hɛlle/ [hɛ**mb**hɛlle] 'What?'

  4 (prefix-stem-suffix=particle)
  4 (prefix-stem-suffix=particle)

Database and coding

- Lahu (Lolo-Burmese, ST) stress domain: *c* = .5

  (ɔ̀-ˈu) NMLZ-lay.egg

  (ˈvì)-(ˈtā) buy-PFPM

  $\frac{\text{2 (prefix-stem)}}{\text{4 (prefix-stem-suffix=particle)}}$

- Lahu tone domain: *c* = .5

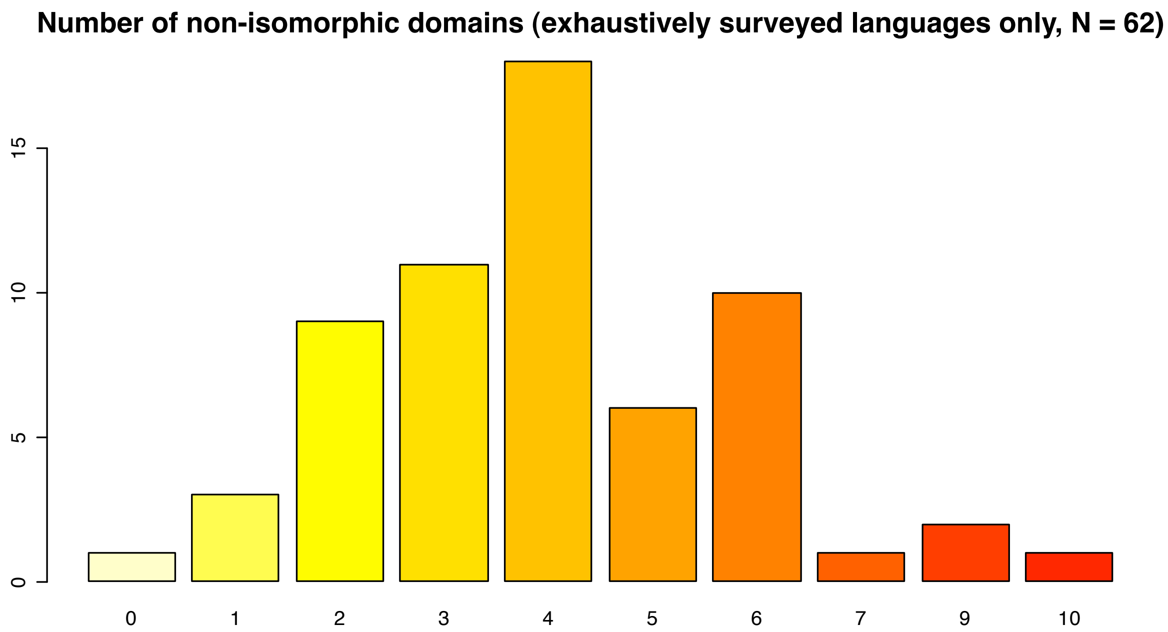
  /ši-ɛ̀/ [ší-ɛ̀] yellow-ADVLZ

  /á-qhâ/ [á-qhâ] NFP-ragweed

  $\frac{\text{2 (stem-suffix)}}{\text{4 (prefix-stem-suffix=particle)}}$

Domain clustering?

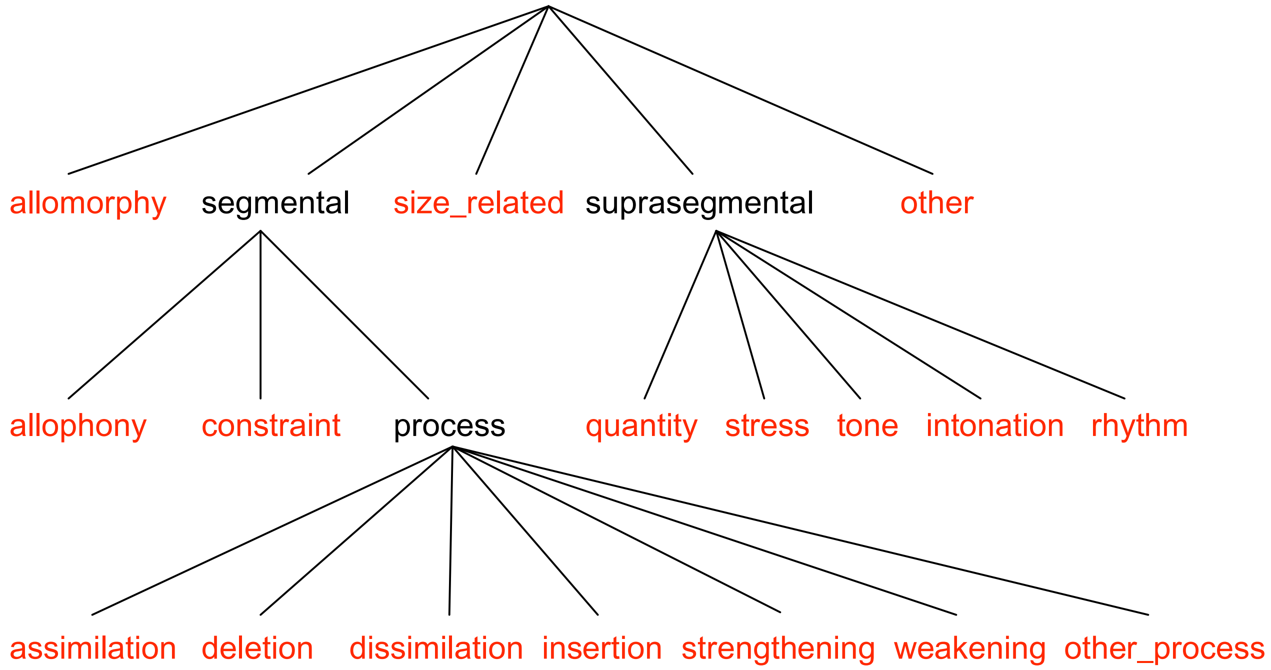
- Most languages violate the Clustering Hypothesis, i.e. have more than one non-isomorphic domain:



- Question: instead of categorical clusters, are there probabilistic clusters depending on sound pattern type? I.e. all tone-defined domains converge on one domain, all assimilation-defined domains on another domain?

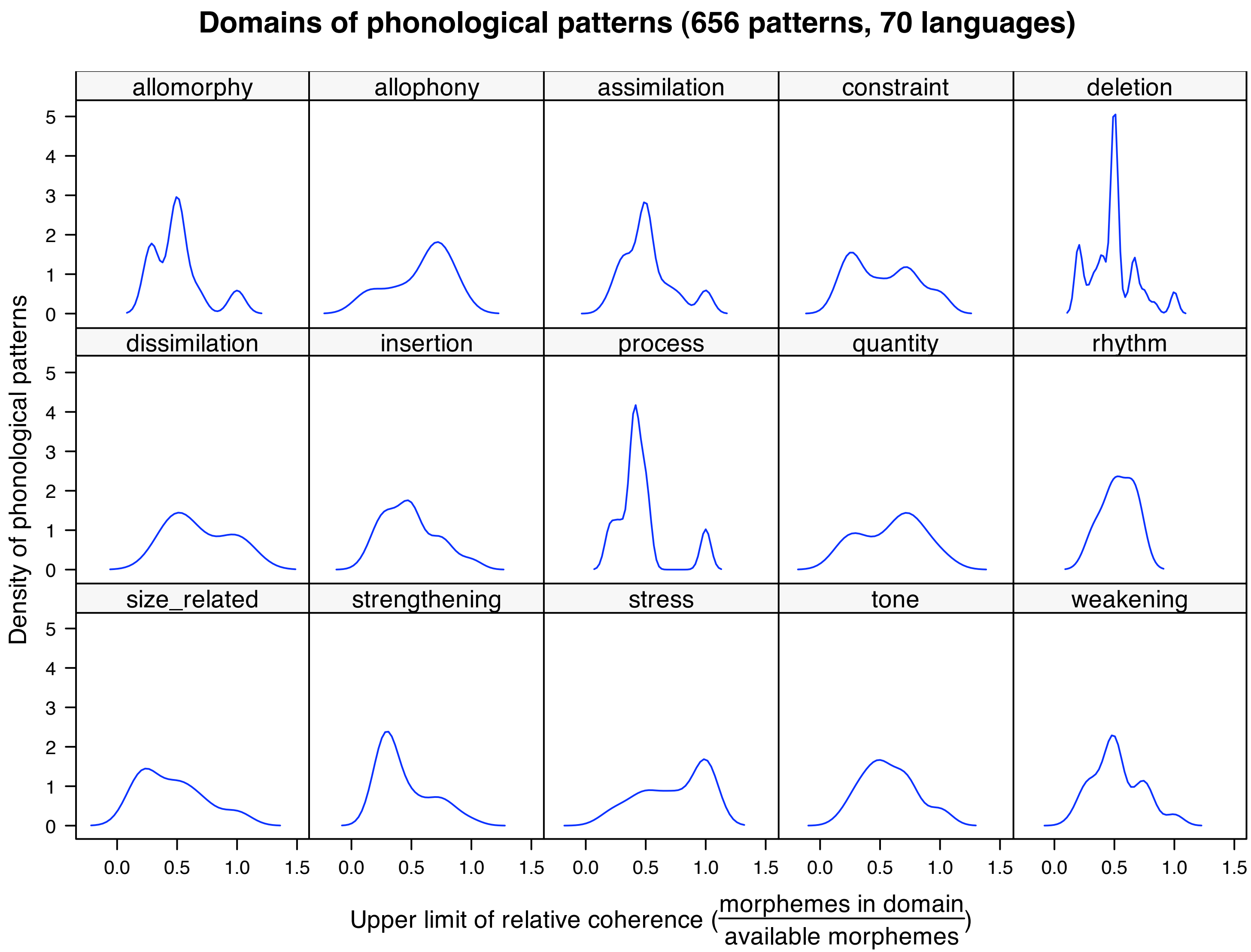
Dissimilarity Analysis: methods

1. Code individual phonological patterns into a taxonomy of sound patterns types (“ppatterns”) on various levels of resolution, e.g. collapsing all segmental types into one.



NB: A low-level taxonomy of 17 types reveals all structure that higher-level taxonomies (e.g. with only 9 types) reveal, and we present results from this only.

Dissimilarity Analysis: methods



Dissimilarity Analysis: methods

3. Which ppatterns target domains of similar coherence?

4. Ignore those ppatterns which happen not to co-occur in any language of the sample, e.g. special alternations like Limbu *l~r* resulting from reanalysis (coded in our taxonomy as ‘allophony’ instead of say ‘weakening’ or ‘assimilation’)

5. Table of ppattern coherence per language:

Dissimilarity Analysis: methods

	Arabic (Egyptian)	Armenian	Belhare	Burmese	Burushaski	Cambodian	Cantonese
assimilation	0.28	0.5	0.33, 0.33, 0.66, 0.5	0.5	0.5	0.33, 0.5	NA
constraint	0.14, 0.28, 0.28, 0.14, 0.71	0.75	NA	NA	0.75, 0.75	0.16, 0.5, 0.5	NA
deletion	NA	NA	1	NA	0.5, 0.75	NA	NA
insertion	NA	0.5, 0.5	0.66	NA	0.5, 0.5	NA	NA
quantity	NA	NA	NA	NA	NA	NA	NA
rhythm	NA	NA	NA	NA	NA	NA	NA
size_related	NA	NA	NA	NA	NA	0.16, 0.5	NA
strengthening	0.28	NA	NA	NA	0.25	NA	NA
stress	1	1	0.33, 0.66	NA	0.75	0.5, 0.5	NA
tone	NA	NA	NA	0.5	NA	NA	0.33
weakening	0.28, 0.28	0.25, 0.75	0.66	NA	0.25, 0.5, 0.5, 0.5, 0.5, 0.5	NA	NA

Where there are several ppattern types in a single language, take the mean, e.g.

Belhare assimilation rules target domains (0.33, 0.33, .66, .5), μ = .46

Dissimilarity Analysis: results

Compute dissimilarities wrt coherence:

- dist = 0: ‘targets a domain with the same coherence degree’ (predicted by Prosodic Hierarchy Theory)
- dist > 0: ‘targets domains with different coherence degrees’

	assimilation	constraint	deletion	insertion	quantity	rhythm	size_related	strengthening	stress	tone
constraint	1.81									
deletion	1.78	2.03								
insertion	1.58	1.65	2.31							
quantity	2.16	1.85	3.24	2.42						
rhythm	0.00	0.60	1.19	2.20	1.39					
size_related	1.52	2.90	2.29	2.41	3.39	1.61				
strengthening	1.77	1.99	2.42	1.26	1.32	1.78	2.52			
stress	3.46	3.66	3.66	3.81	3.20	2.20	4.14	4.48		
tone	1.28	1.49	0.40	1.25	2.04	0.00	2.55	1.54	2.79	
weakening	1.14	1.91	1.62	1.53	2.06	1.39	2.56	1.64	3.61	1.29

Dissimilarity Analysis: results

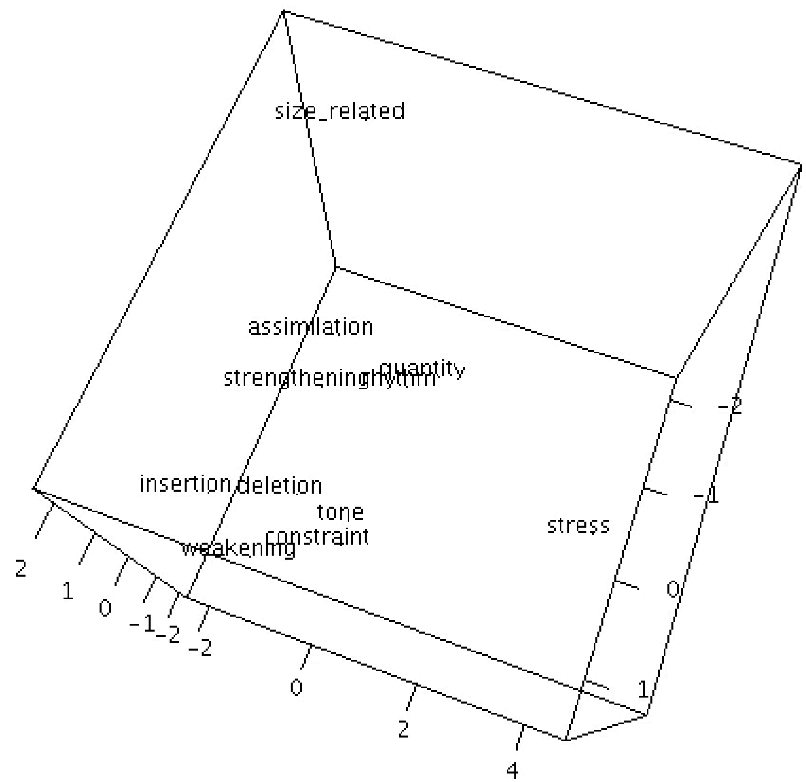
Compute dissimilarities wrt coherence:

- dist = 0: 'targets a domain with the same coherence degree' (predicted by Prosodic Hierarchy Theory)
- dist > 0: 'targets domains with different coherence degrees'

	assimilation	constraint	deletion	insertion	quantity	rhythm	size_related	strengthening	**stress**	tone
constraint	1.81									
deletion	1.78	2.03								
insertion	1.58	1.65	2.31							
quantity	2.16	1.85	3.24	2.42						
rhythm	0.00	0.60	1.19	2.20	1.39					
size_related	1.52	2.90	2.29	2.41	3.39	1.61				
strengthening	1.77	1.99	2.42	1.26	1.32	1.78	2.52			
stress	**3.46**	**3.66**	**3.66**	**3.81**	**3.20**	**2.20**	**4.14**	**4.48**		
tone	1.28	1.49	0.40	1.25	2.04	0.00	2.55	1.54	**2.79**	
weakening	1.14	1.91	1.62	1.53	2.06	1.39	2.56	1.64	**3.61**	1.29

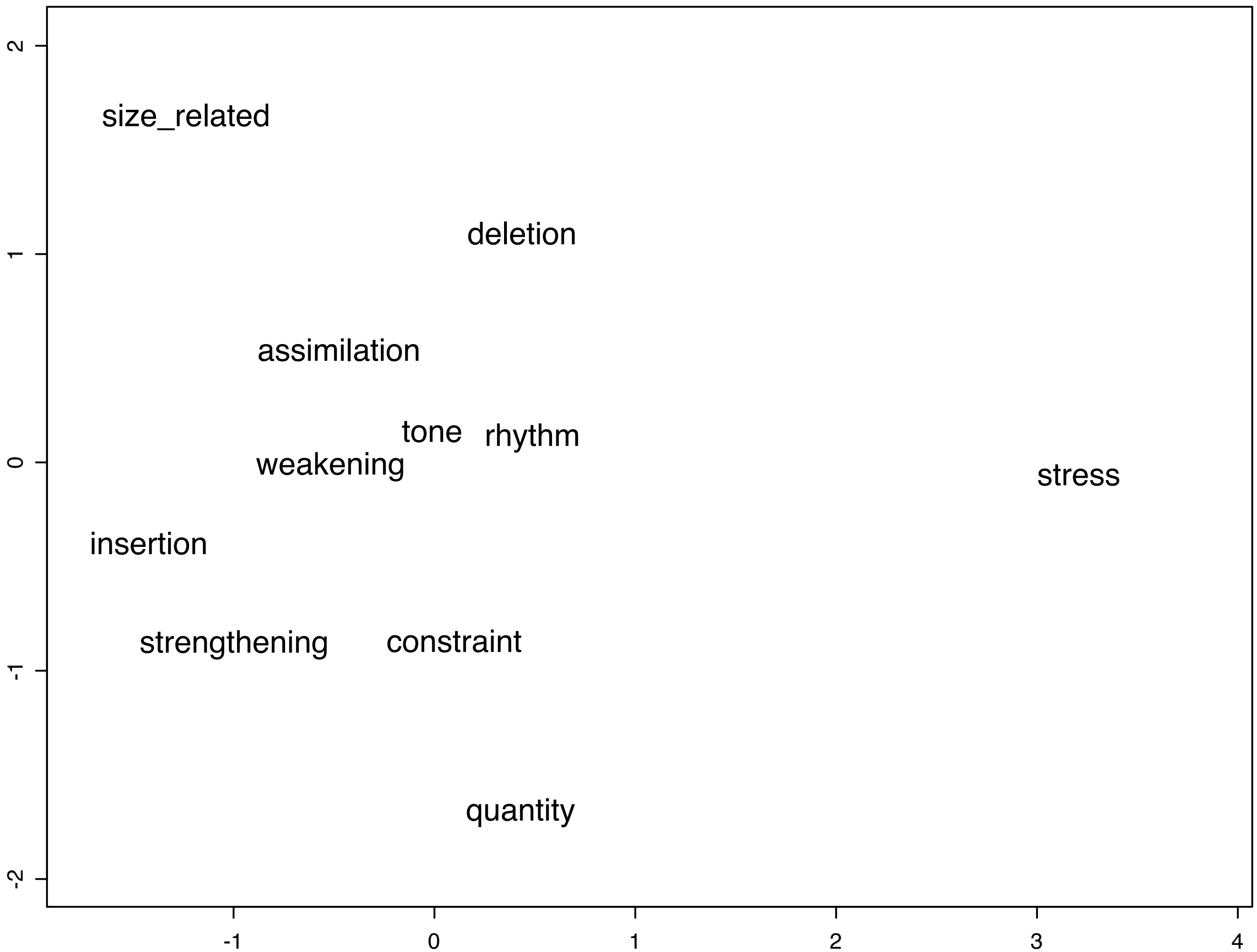
Dissimilarity Analysis: results

Better visualization of dissimilarities by Multidimensional Scaling:



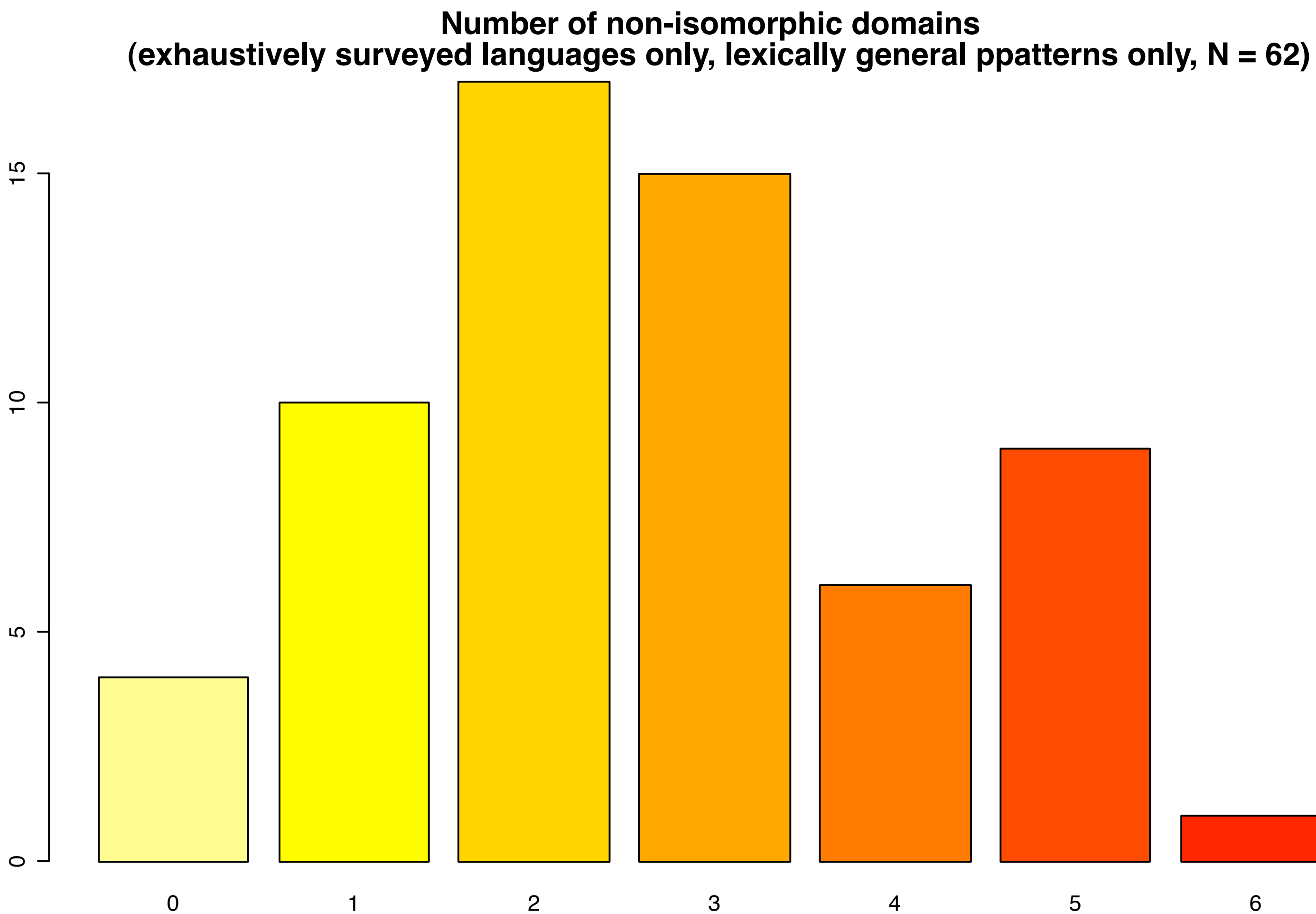
Dissimilarity Analysis: results

Projected onto 2D:



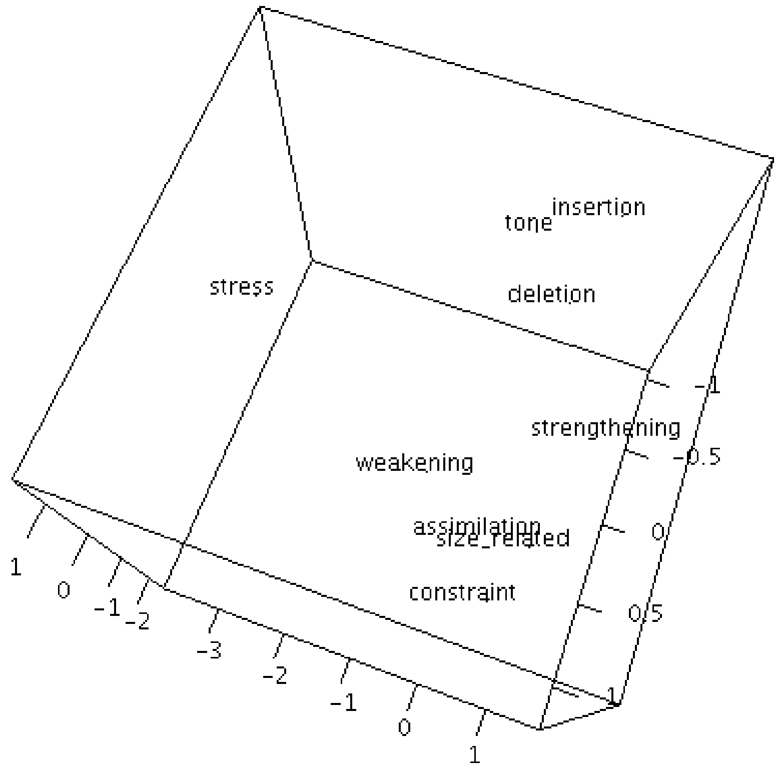
Dissimilarity Analysis: follow-up

What if we considered only lexically general patterns?



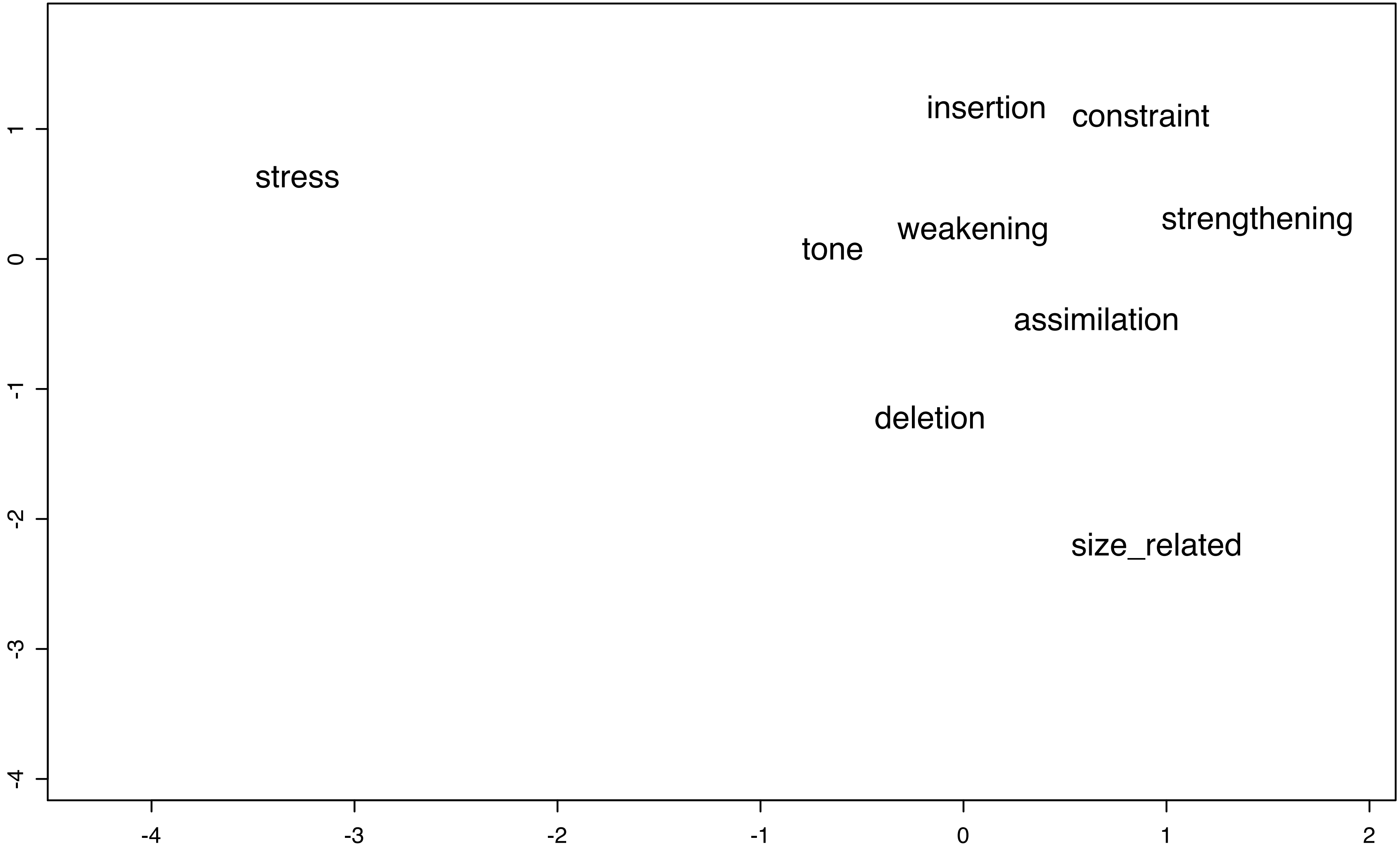
Dissimilarity Analysis: follow-up

Multidimensional scaling down to 3D:



Dissimilarity Analysis: follow-up

Multidimensional scaling down to 2D:



Interim conclusion

1. Stress-defined domains cluster on similarly-sized domains.

2. No other ppatterns seem to form clusters of similarly-sized domains.

3. Closer inspection suggests that stress-defined domains tend to be larger than others.

4. Test this as a hypothesized empirical universal, *controlling for genealogical and areal affiliation*

Factorial Analysis - methods

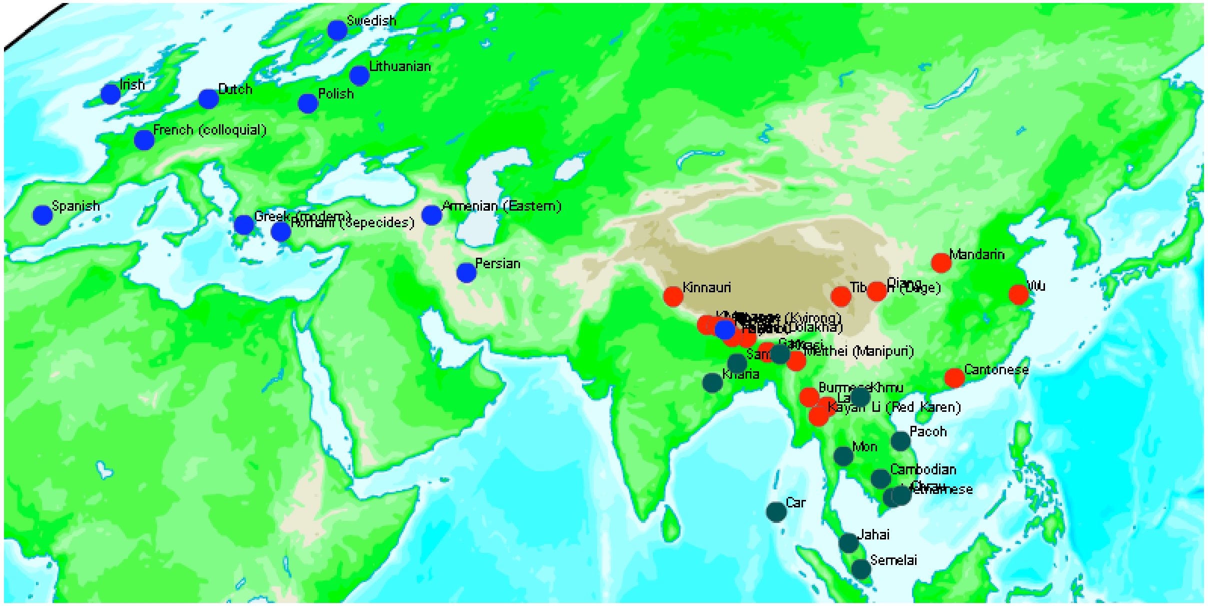
Factor 1: STRESS

stress-defined (*N*=38) vs. other (*N*=367) ppatterns

Factorial Analysis - methods

Factor 2: genealogical STOCK (inherited domain types)

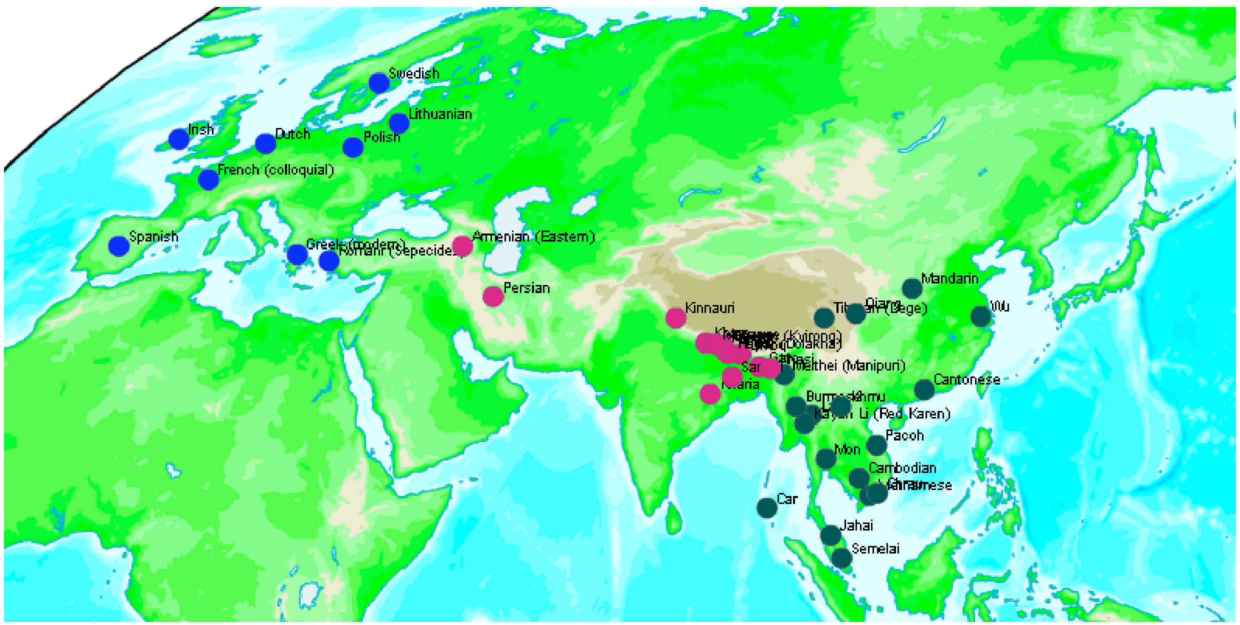
For this, take one representative per sub-branch of major branches in three families (or two if phonologies known to be diverse and data are sufficient): Austroasiatic (11), Indo-European (12), Sino-Tibetan (17)



Factorial Analysis - methods

Factor 3: AREA affiliation (spread domain types)

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



Factorial Analysis: results

Design:

2 (STRESS) x 3 (STOCK) x 3 (AREA)

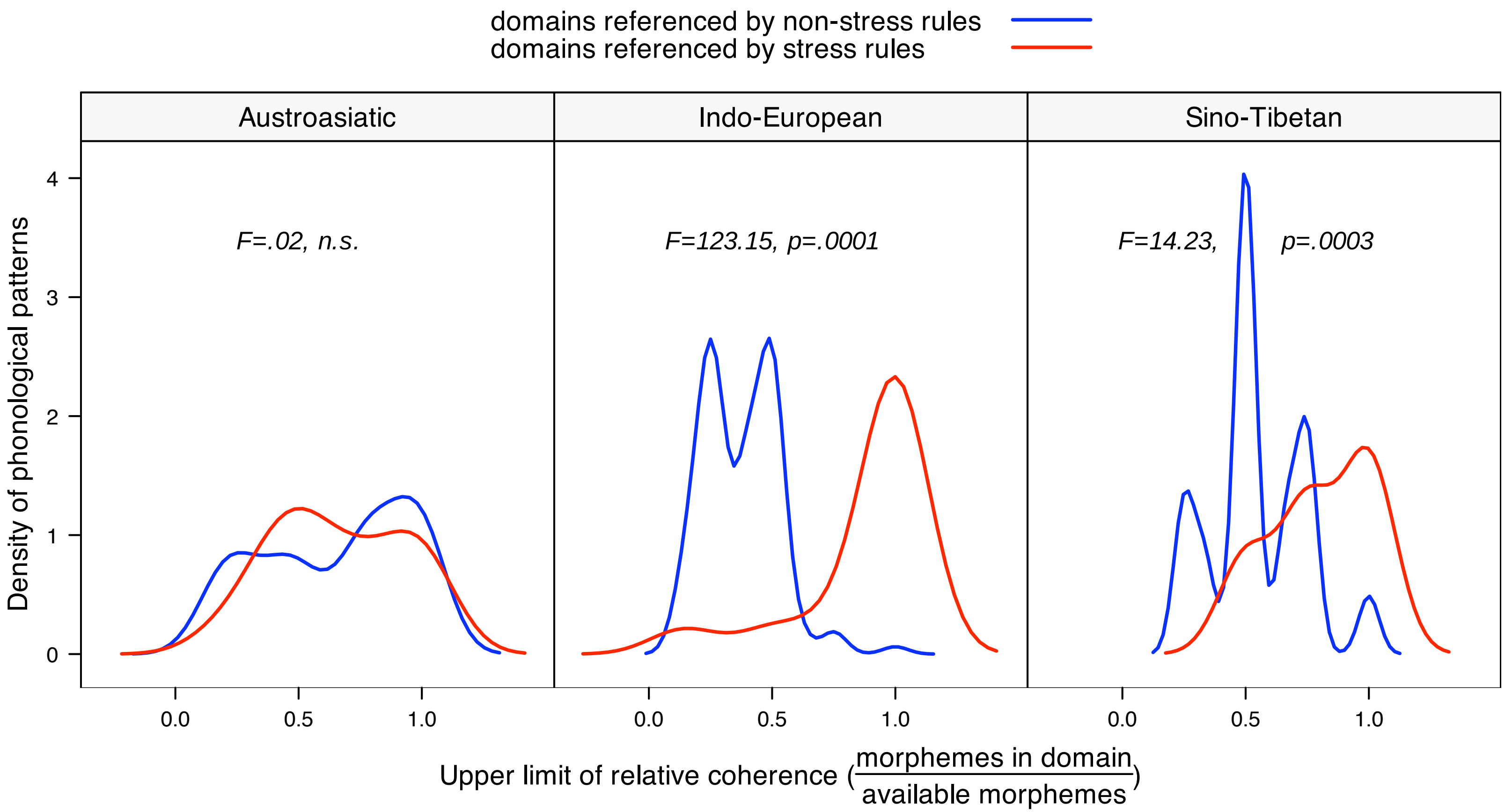
Procedure:

Randomization-based ANOVAs (Janssen, Bickel & Zúñiga 2006)

Results (405 ppatterns in 40 languages)

1. No significant three-way interaction.
2. Borderline evidence for two-way interaction between STRESS and STOCK ($F(2)=3.27$, $p=.09$), and for no other interaction.
3. Significant main effects of STRESS -- but not of AREA -- within Indo-European and Sino-Tibetan, but not Austroasiatic:

Factorial Analysis: results



Reliability analysis (Janssen et al. 2006):

- in IE, reliable at p<.01 up to replacing the 9 (out of 14) largest stress-defined and up to any number of the smallest other-defined domains by the grand mean
- in ST, reliable at p<.01 up to replacing the 4 (out of 9) the largest stress-defined and up to any number of the smallest other-defined domains by the grand mean. The 4 critical datapoints were triple-checked.

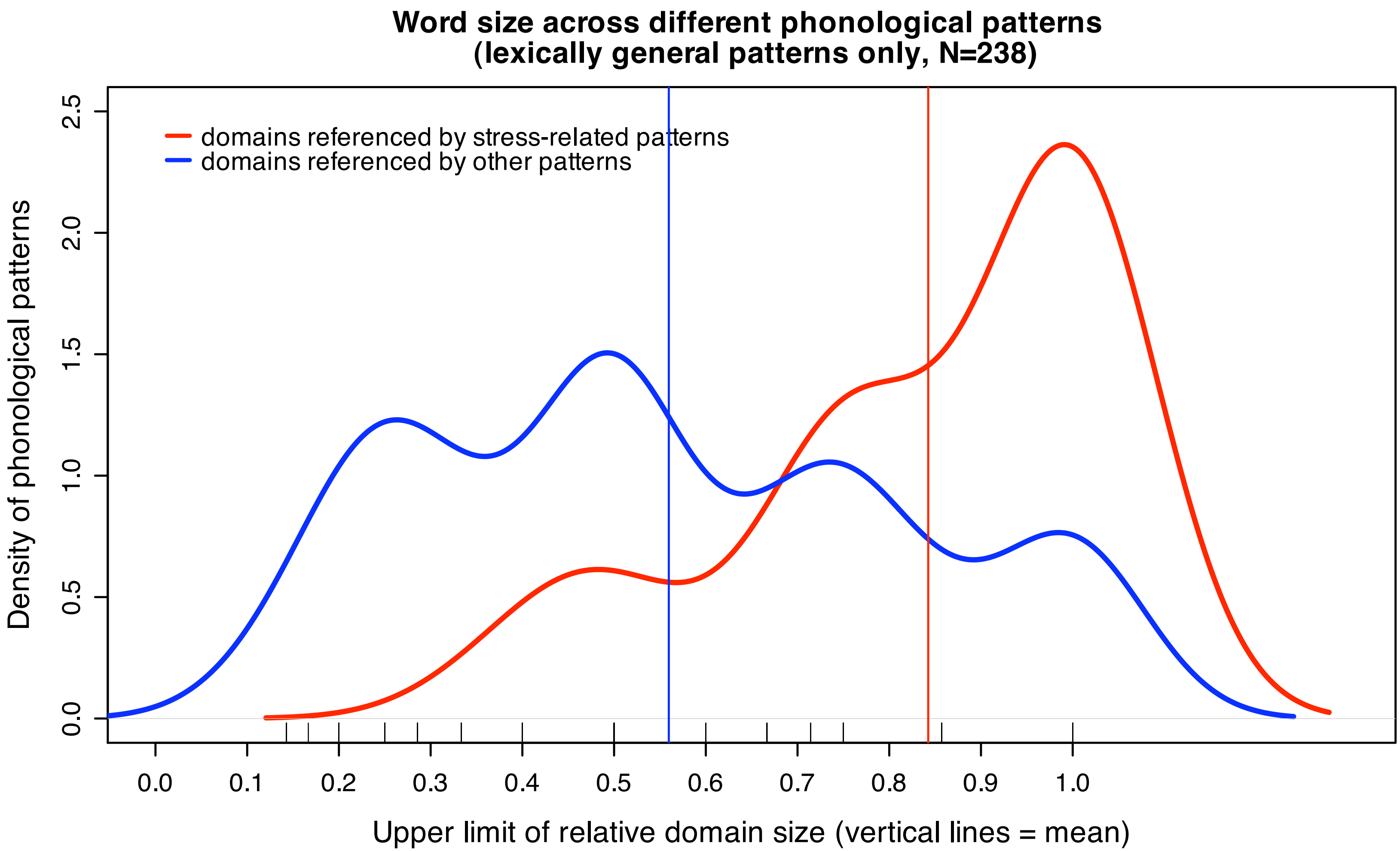
Factorial Analysis: results

Concentrating on ppatterns that are lexically general (lacking strata-specifications), the STRESS*STOCK interaction effect is lost ($F(2)=2.04$, $p=.24$); and no other interaction effect either.

Main effects (238 ppatterns in 40 languages):

1. No evidence for AREA effect ($F(2)=.77$, $p=.57$)
2. Main effect of STOCK ($F(2)=10.55$, $p<.0001$)
3. Main effect of STRESS ($F(1)=20.99$, $p=.0001$)

Factorial Analysis: results



Reliability analysis (Janssen et al. 2006):

- reliable at p<.01 up to replacing the 5 (out of 19) largest stress-defined and up to any number of the smallest other-defined domains by the grand mean

Conclusions

If we limit the evidence, as is generally done, to ppatterns that are not tied to specific lexical information, we find robust statistical support for the following universal:

Stress-defined domains tend to be significantly larger than other domains.

But no other ppattern 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 pre-generative conceptions of prosodic structure in which only stress and intonation are necessarily included in hierarchical structures (e.g. Pike 1945)

Acknowledgments

- **Thanks to our student assistants for help in data collection and computational issues: Thomas Goldammer, Franziska Crell, Sven Siegmund, Taras Zakharko, Jenny Seeg, Sebastian Hellmann, Anja Gampe, Josh Wilbur.**

- **Thanks to the DFG for funding this research (DFG Grant Nos. BI 799/2 and 799/2-3).**

- **All statistical analysis and all plots were done in R 2.4.1 (with added packages *trotter*, *rgl, MASS,* and *lattice*).**

- **Maps were created running Hansjörg Bibiko's iAtlas tool on our FileMaker Pro database.**