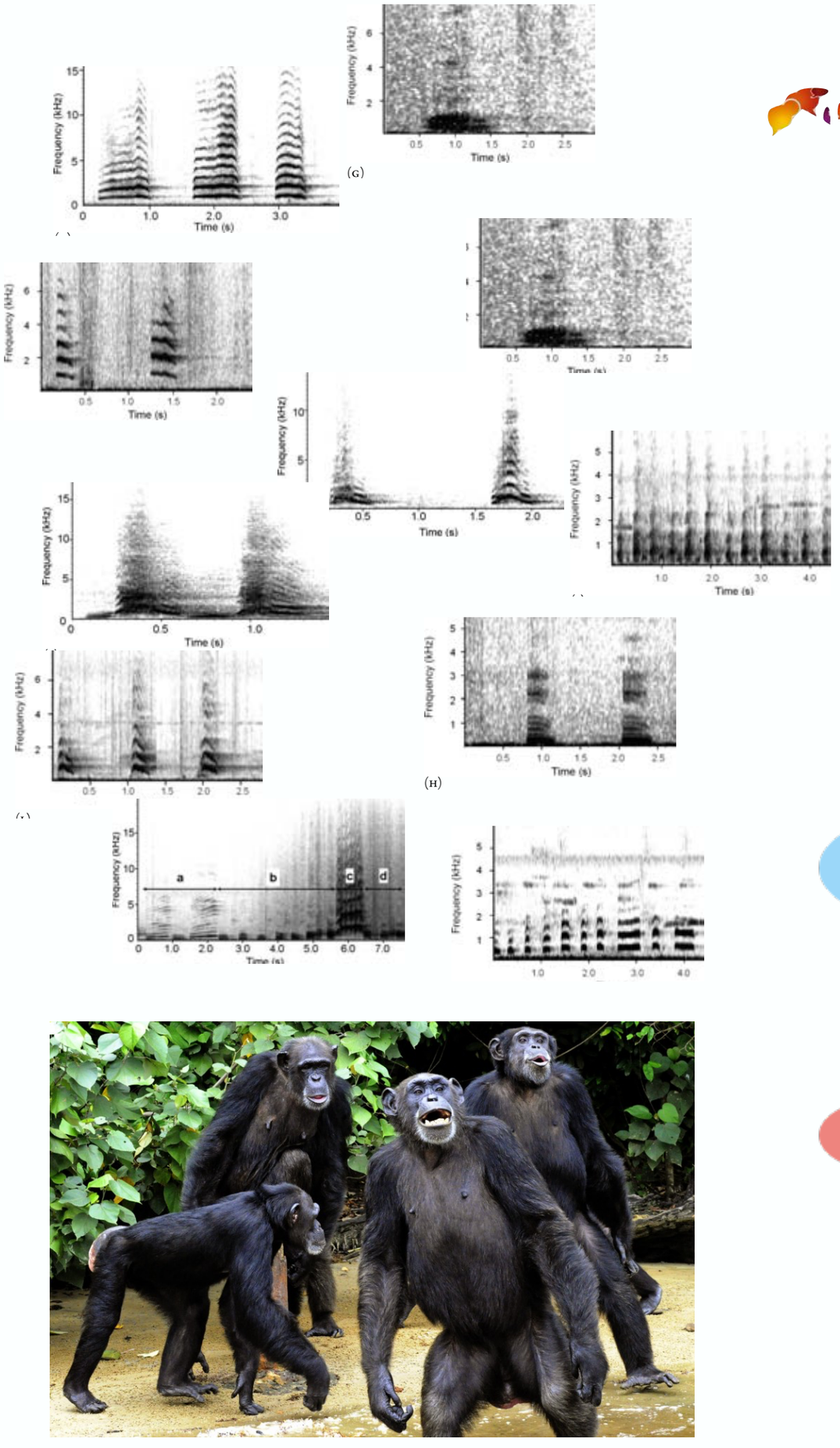
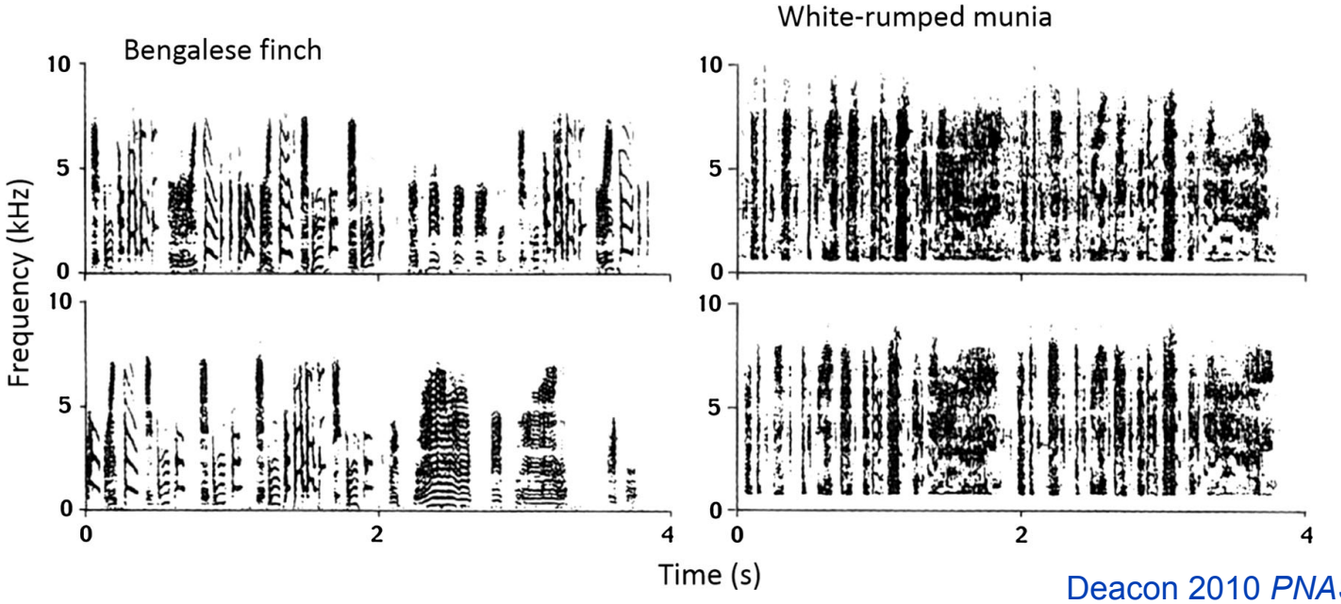
University
of Zurich[UZH]
Department of Comparative Linguistics

Balthasar Bickel

Cross-linguistic corpora reveal constraints on language dynamics

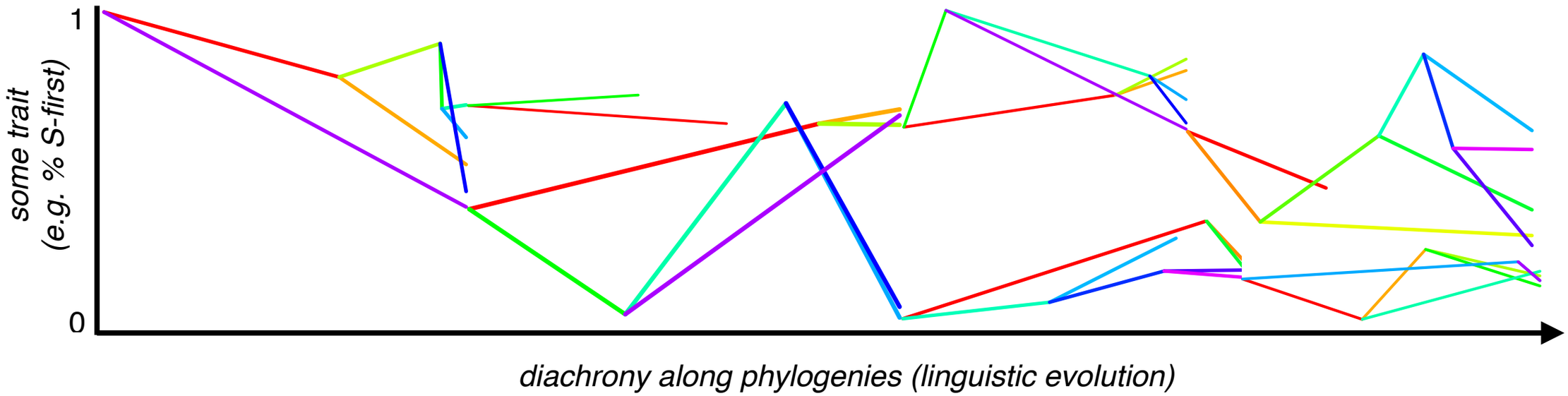

Frequency (kHz)
Time (s)
(G)
(H)
a
b
c
d

Diversification under relaxed selection?

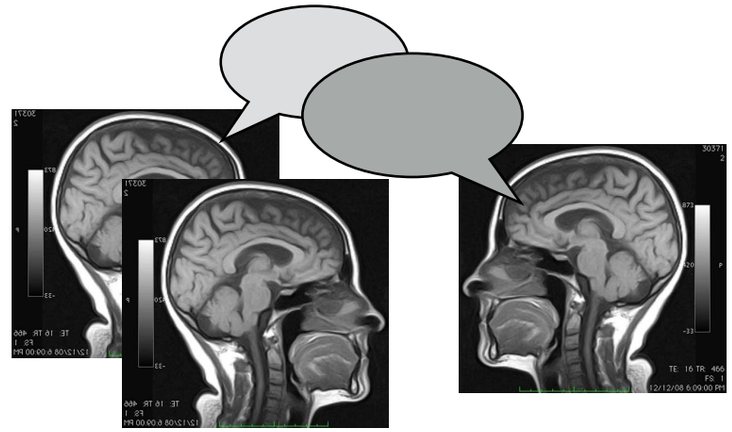


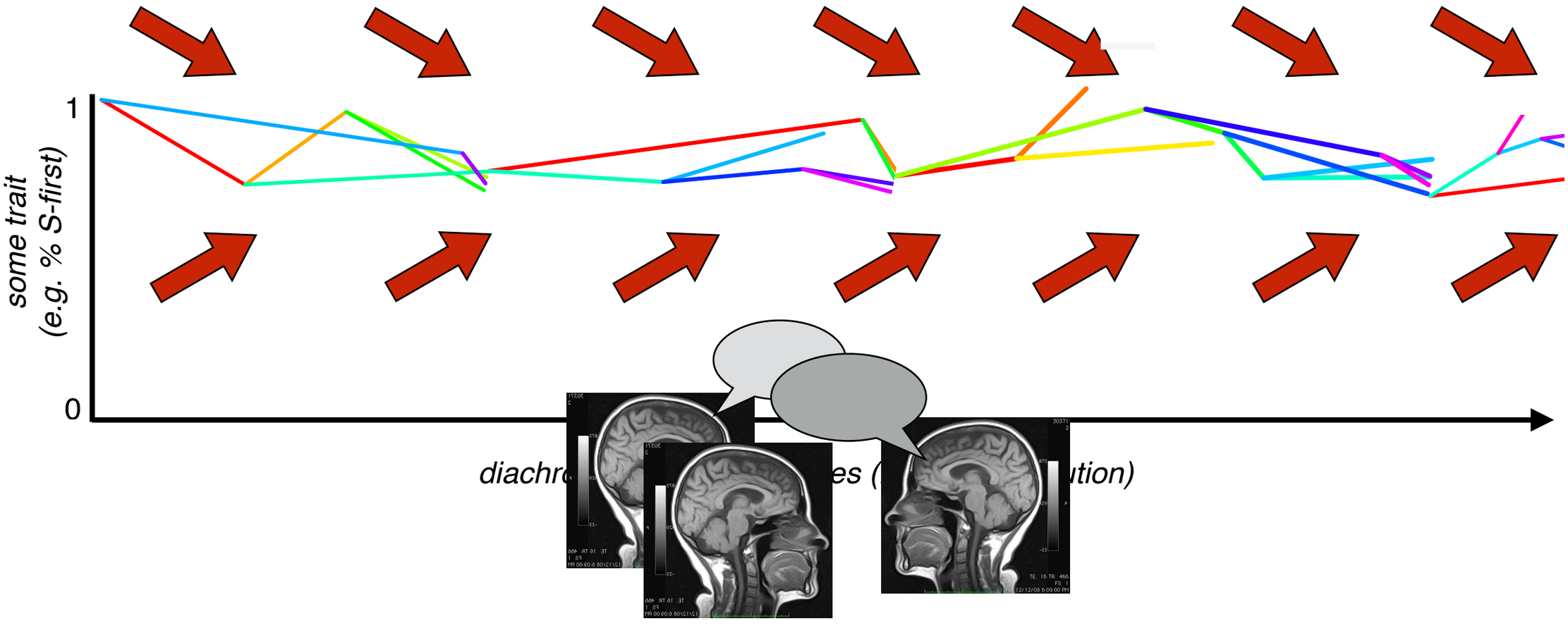
Deacon 2010 *PNAS*; Okanoya 2015 *J Ornithol*

Diversification under relaxed selection?

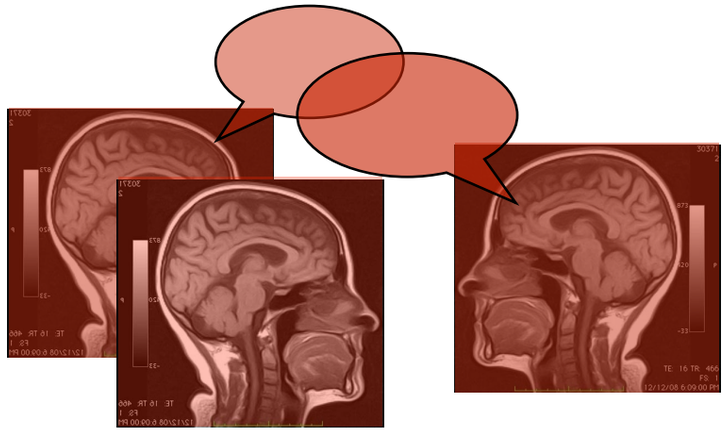


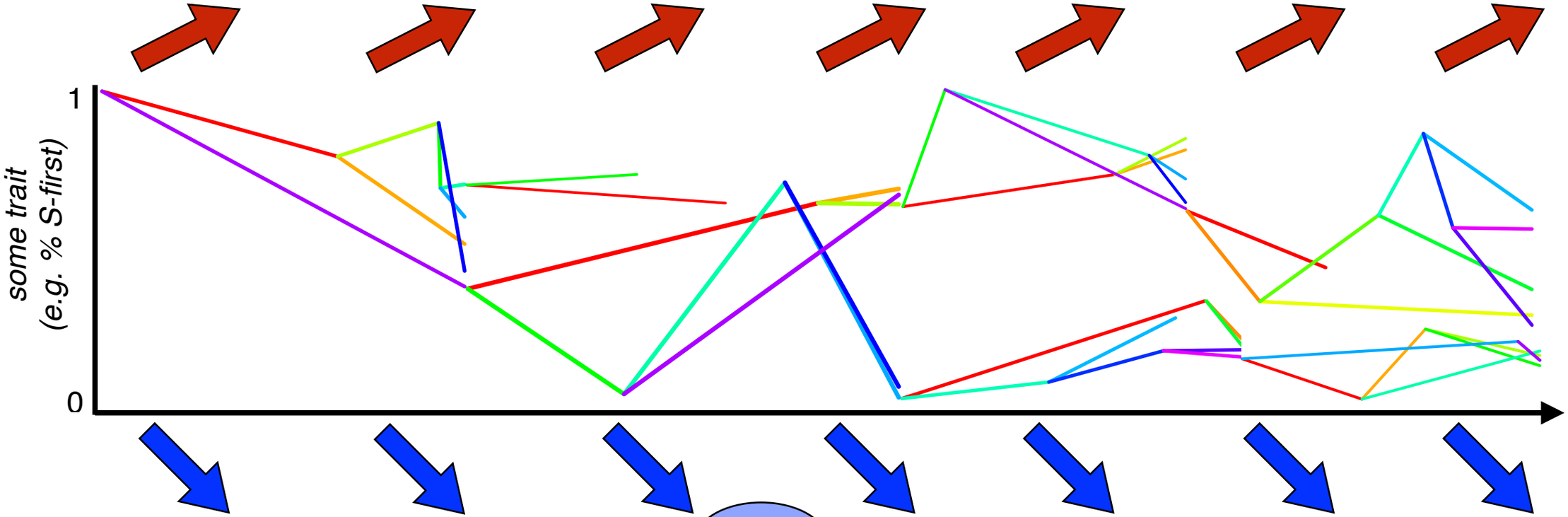
Or constrained evolution?

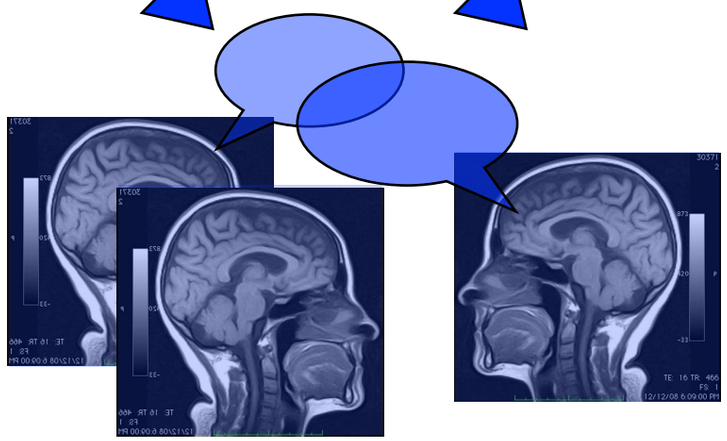




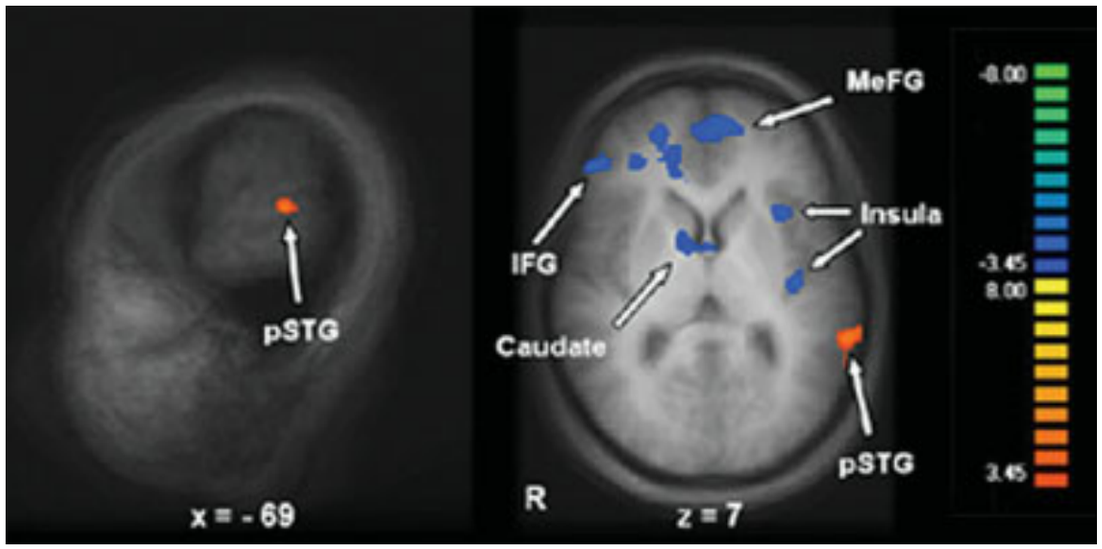
But brain and behavior can also adapt!







Example 1:
Exposure to lexical tone shapes pitch processing (Wong et al. 2007ff)



Example 2:
Exposure to case-based agreement syntax shapes referential density (Bickel 2003ff)

Belhare (Sino-Tibetan)

a. *(han)* *khar-e-**ga*** *i?*
2s**NOM** go-PST-**2sS** Q

‘Did you go?’

b. *(han-**na**) un* *lur-he-**ga*** *i?*
2s-**ERG** 3sNOM [3sA-]tell-PST-**2sA** Q

‘Did you tell him/her?’

c. *ciya* *(han-**naha**) n-niũa* *tis-e-**ga*** *i?*
tea.NOM 2s-**GEN** 2sPOSS-mind please-PST-**2sA** Q

‘Did you like the tea?’

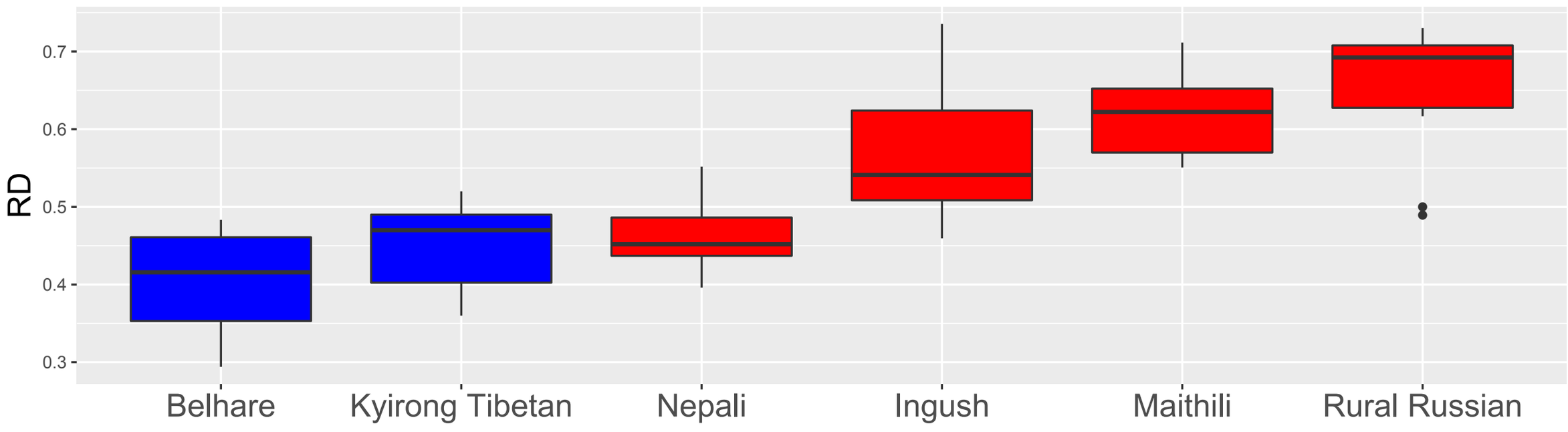
Maithili (Indo-European)

a. *(**tũ**)* *bimār ch-**æ**?*
2nh**NOM** sick be-2nh**NOM**

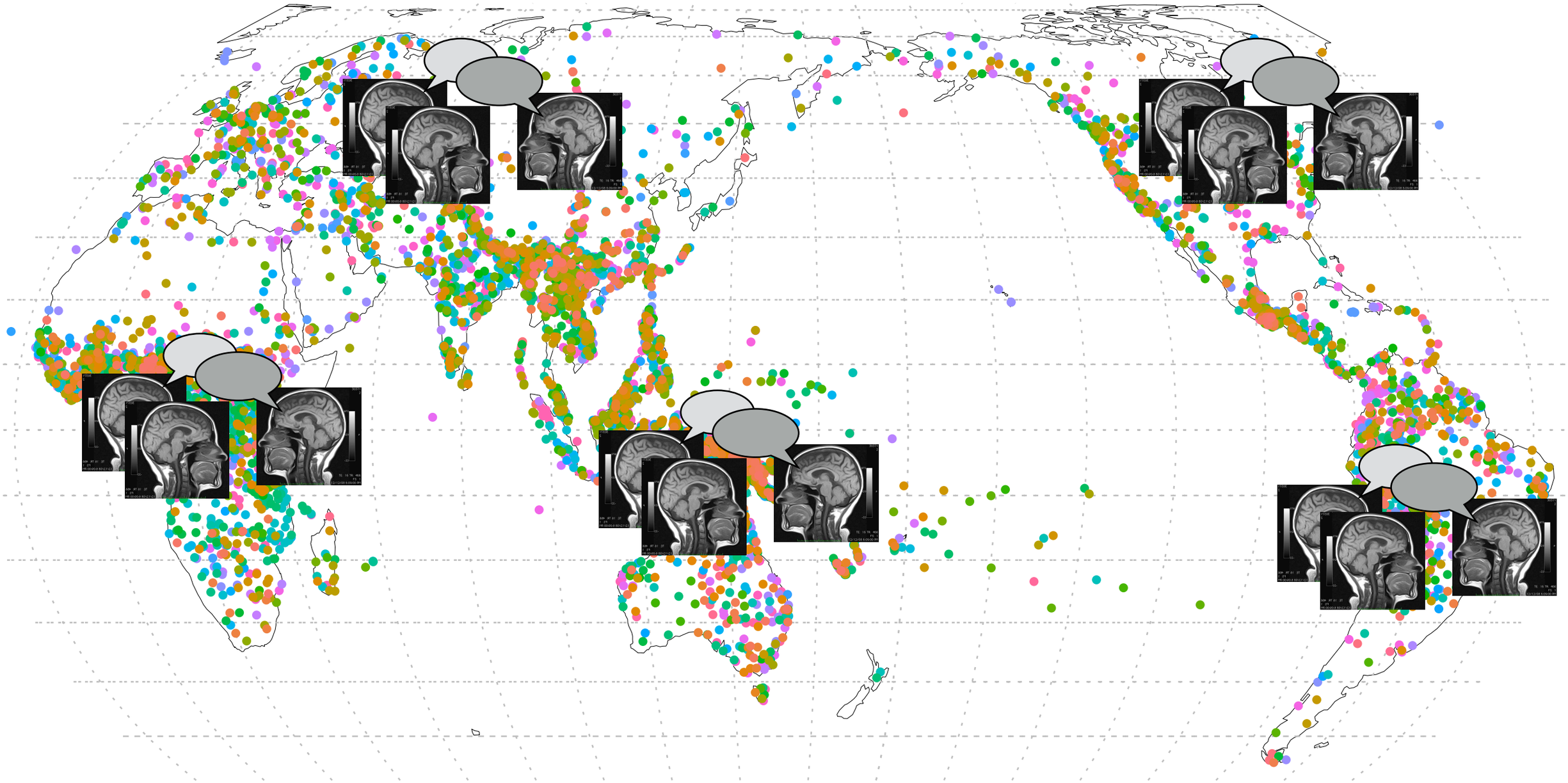
‘Are you sick?’

b. *(**torā**)* *khuśi* *ch-**au**?*
2nh**DAT** happy 2nh-**NONNOM**

‘Are you happy?’



So we need to test stability of brain and communication in non-WEIRD samples



And we can use corpora as natural language production experiments!

Three case studies from recent work:

Constraints on the global evolution of

1. **affixation** (with Frank Seifart et al)

2. **affix order** (with Sabine Stoll and John Mansfield)

3. **word order** (with Damián Blasi and Jing Yingqi)

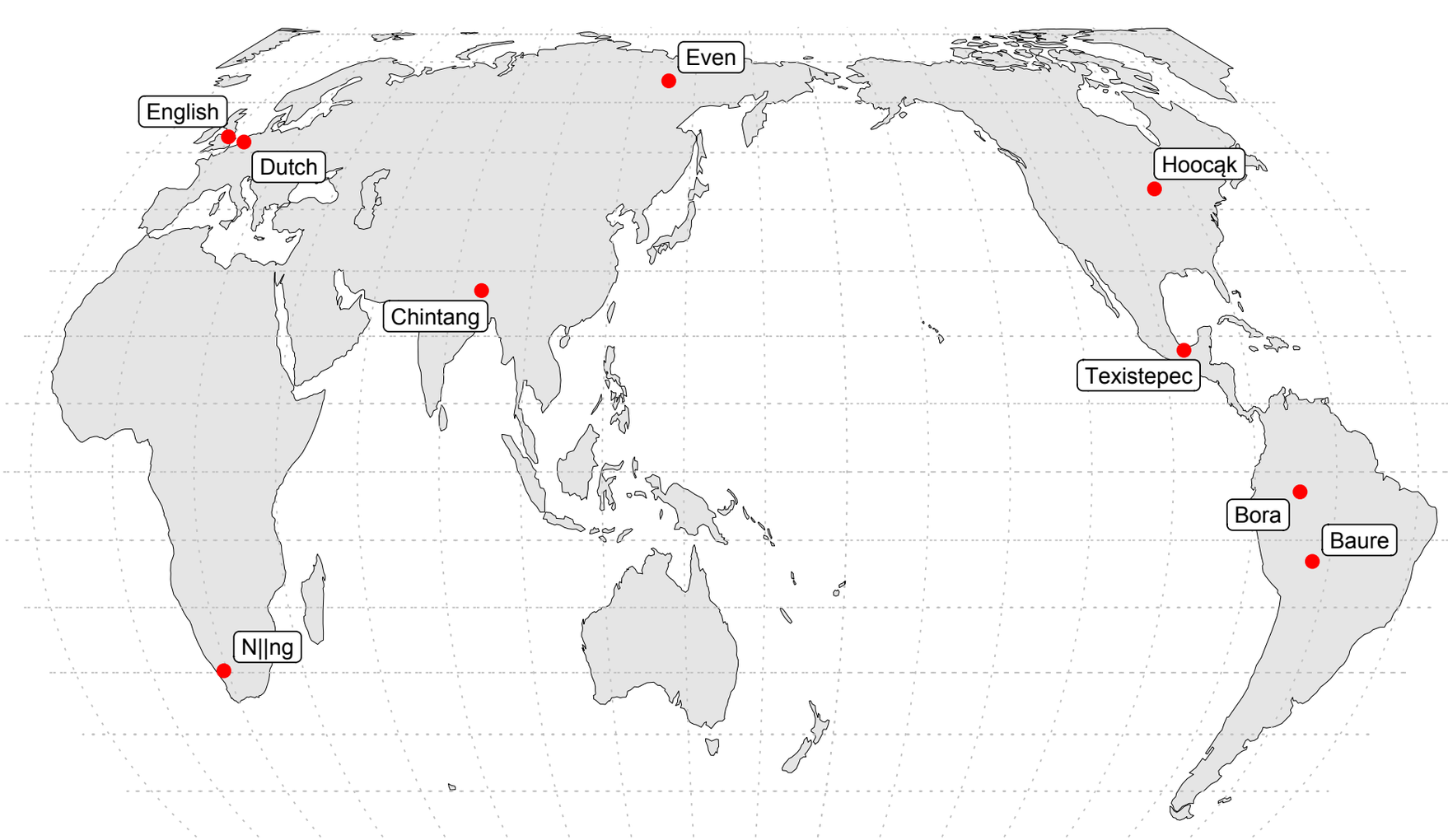
Study 1: Constraints on affix evolution

PNAS

Nouns slow down speech across structurally and culturally diverse languages

Frank Seifart[a,b,c,1], Jan Strunk[b], Swintha Danielsen[d], Iren Hartmann[d], Brigitte Pakendorf[c], Søren Wichmann[e,f], Alena Witzlack-Makarevich[g], Nivja H. de Jong[e,h], and Balthasar Bickel[i]

[a]Amsterdam Center for Language and Communication, University of Amsterdam, 1012 VT Amsterdam, The Netherlands; [b]Institut für Linguistik, University of Cologne, 50923 Cologne, Germany; [c]Laboratoire Dynamique du Langage, UMR5596, CNRS & Université de Lyon, 69007 Lyon, France; [d]Institut für Linguistik, University of Leipzig, D-04107 Leipzig, Germany; [e]Leiden University Centre for Linguistics, Leiden University, 2311 BX Leiden, The Netherlands; [f]Laboratory of Quantitative Linguistics, Kazan Federal University, 420000 Kazan, Russia; [g]Abteilung für Allgemeine Sprachwissenschaft, Institute for Scandinavian Studies, Frisian Studies, and General Linguistics, Kiel University, 24098 Kiel, Germany; [h]Leiden University Graduate School of Teaching, Leiden University, 2333 BN Leiden, The Netherlands; and [i]Department of Comparative Linguistics, University of Zurich, 8032 Zurich, Switzerland



Corpora

Language	Family	Speakers	Texts	Words	Reference
Baure	Arawakan	12	37	17,652	Danielsen et al. (2009)[1]
Bora	Boran	46	37	29,802	Seifart (2009)[2]
Chintang	Sino-Tibetan	74	40	37,737	Bickel et al. (2011)[3]
Dutch	Indo-European	42	17	39,519	CGN-consortium (2003)[4]
English	Indo-European	80	47	56,135	Calhoun et al. (2009)[5]
Even	Tungusic	32	67	37,430	Pakendorf et al. (2010)[6]
Hoocąk	Siouan	28	62	23,191	Hartmann (2013)[7]
N‖ng	!Ui-Taa	7	33	26,061	Güldemann et al. (2011)[8]
Texistepec	Mixe-Zoquean	1	6	21,321	Wichmann (1996)[9]
Sum		**322**	**346**	**288,848**	

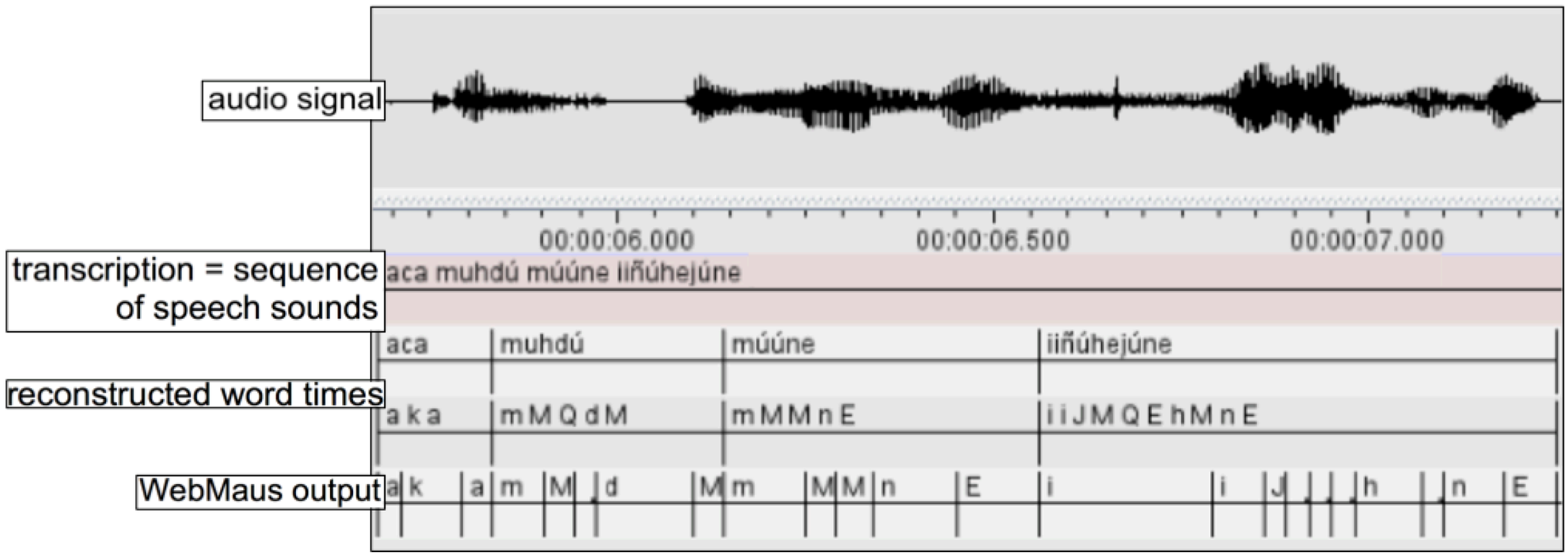
Seifart at al. 2018 *PNAS*

Semi-automated analysis

	aa-bé = váa	*tsá-ijyu*	*íjtsámeí*	*í-llí-mútsi-kye*
	CON-M.SG=QUOT.PAST	one-day	think	3-child-M.DU-ACC
	no-ni-cli-cli	**adv**-clf	**v**	ni-**n**-ni-ni
→	**PRO**	**OTHER**	**V**	**N**

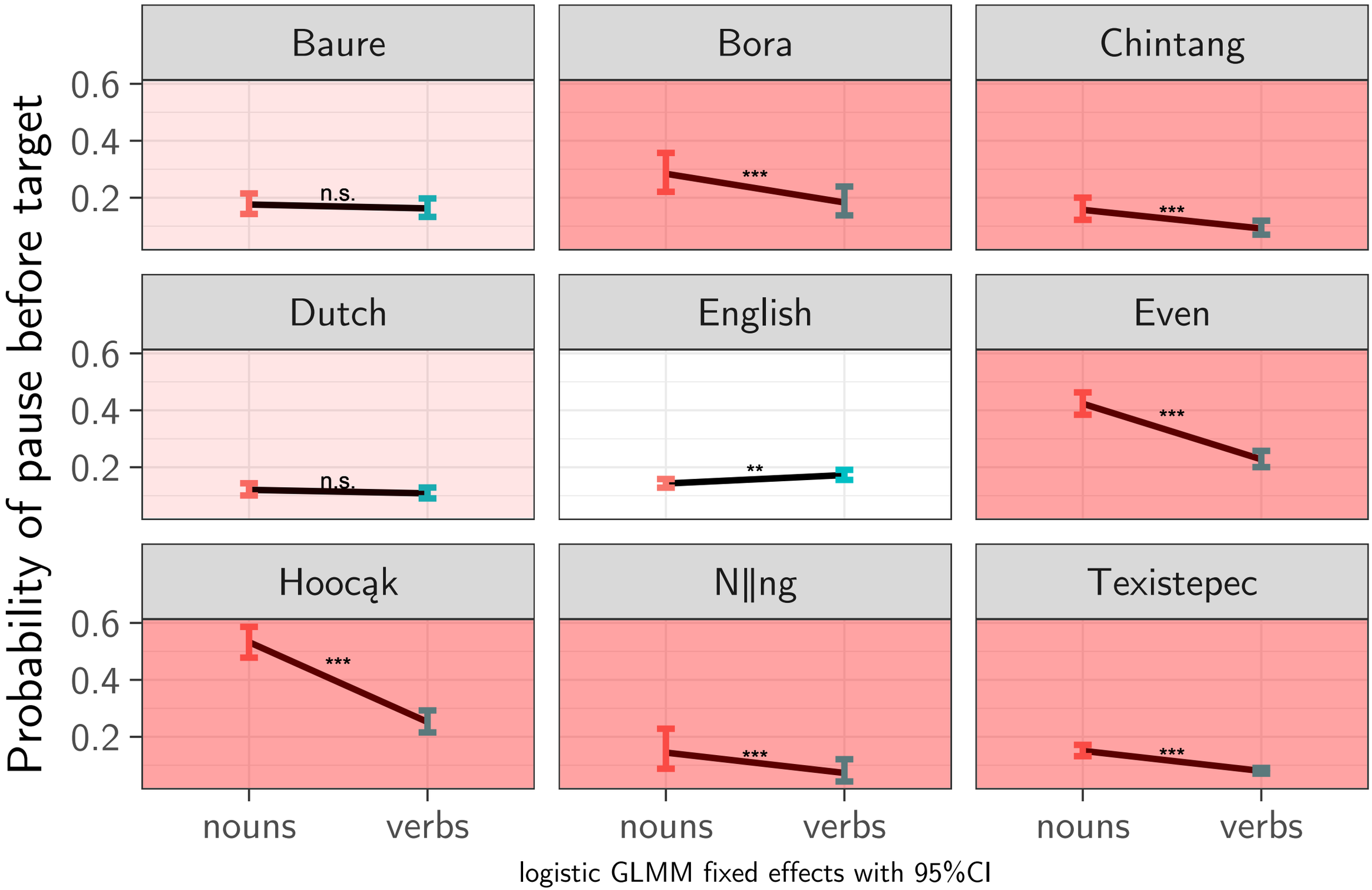
	iámejcá-nu*-í-ñe,*	*wallee*	*wajpii*	*íjcya-ne*
	festival-VBZ:DO-FUT-3	woman	man	be-3
	n-nd-vi-ni	**n**	**n**	**v**-vi
→	**N(V)**	**N**	**N**	**V**

'And one day he thought of making a festival for his two children, who were a girl and a boy' [piivyeebe_ayju 005]



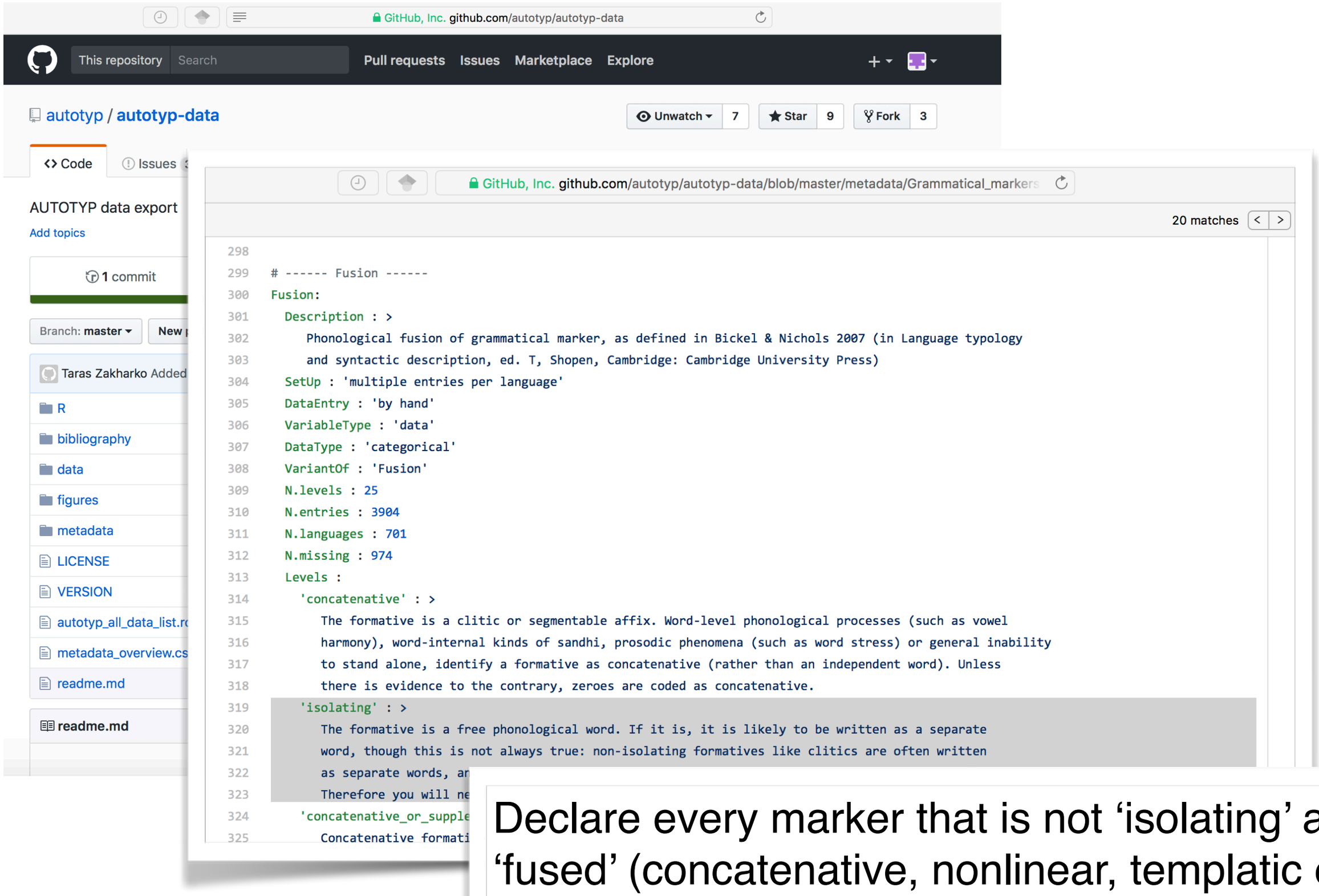
Seifart at al. 2018 *PNAS*

An asymmetry in lexical planning



Seifart at al. 2018 *PNAS*

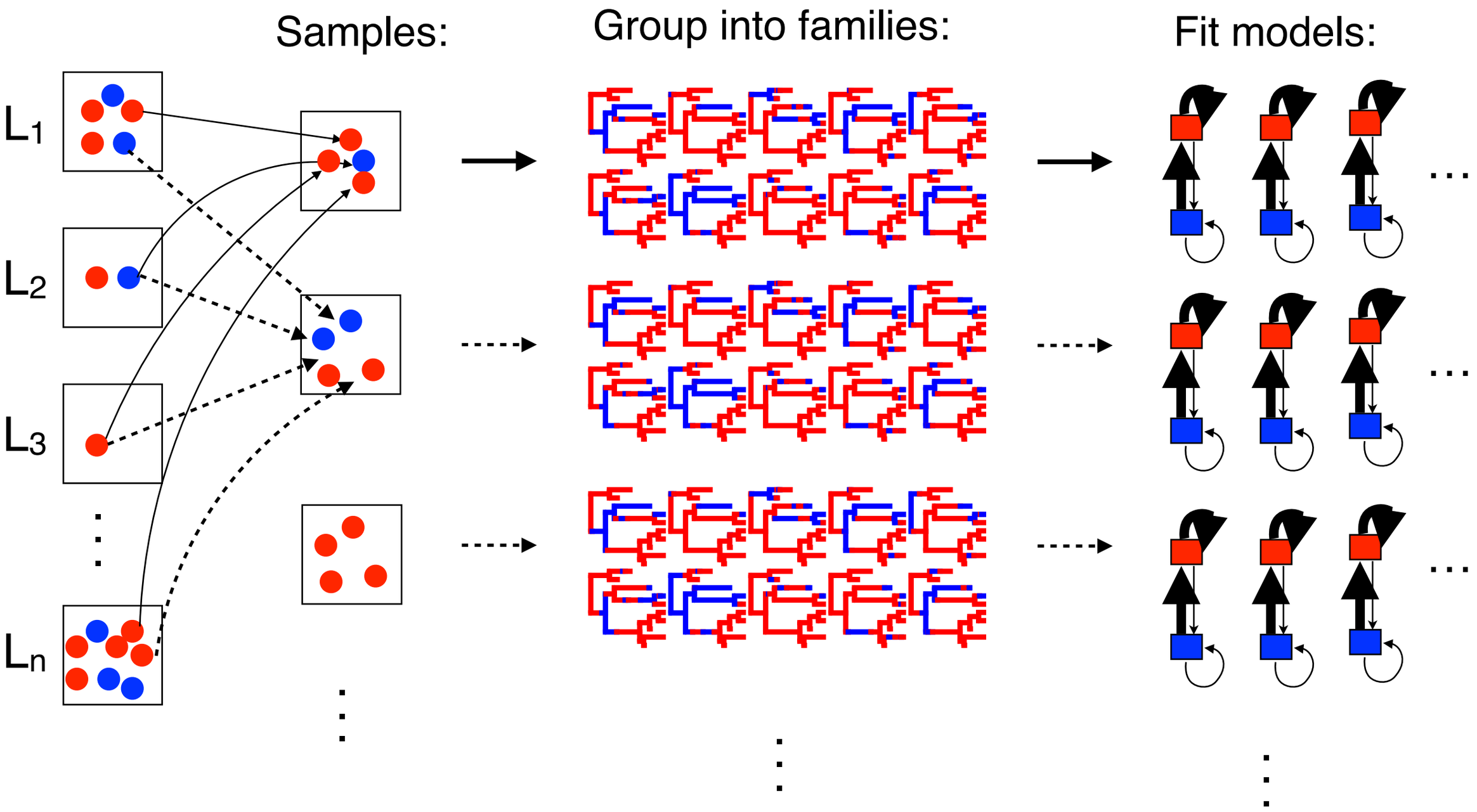
A correlated asymmetry in diachrony?



Declare every marker that is not ‘isolating’ as ‘fused’ (concatenative, nonlinear, templatic etc)

Seifart & Bickel 2017 *ALT*

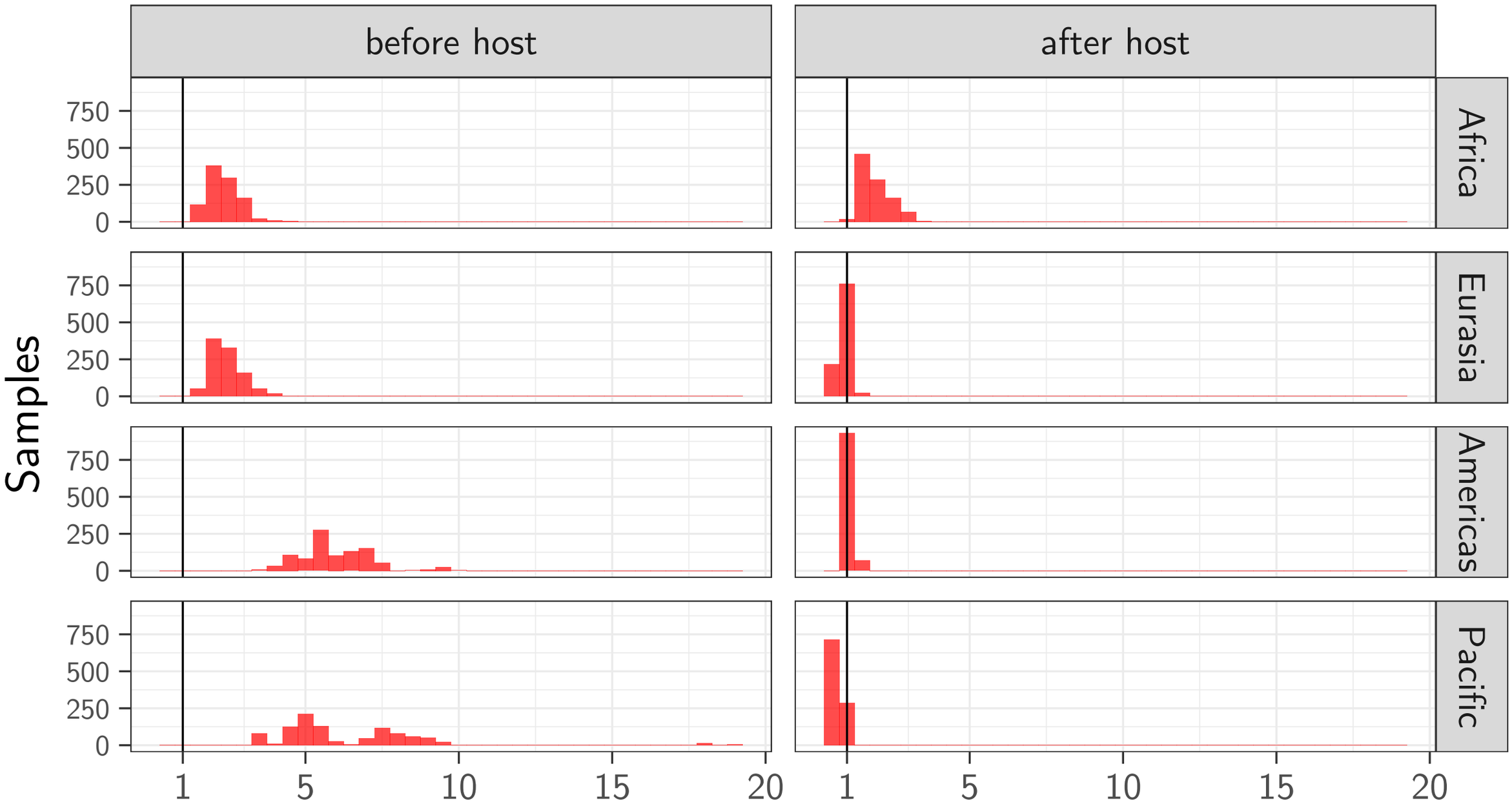
Treat language-internal variation as uncertainty: sampling markers



Seifart & Bickel 2017 *ALT*

A correlated asymmetry in diachrony

Re-sampling from nearly 4000 grammatical markers in AUTOTYP, fitting evolutionary models on each sample and analyze directional biases in this models as GLMMs:



Seifart & Bickel 2017 *ALT*

Study 1 Summary

1. Non-WEIRD corpora reveal lower pause probability before verbs than before nouns

2. This increases the odds for prefixes before verbs compared to nouns in language change

Study 2: Constraints on affix order evolution (with John Mansfield & Sabine Stoll)

A natural experiment in Chintang (Sino-Tibetan, Nepal): free prefix order!

a. *u-kha-ma-cop-yokt-e*
3snA-1nsP-NEG-see-NEG-PST

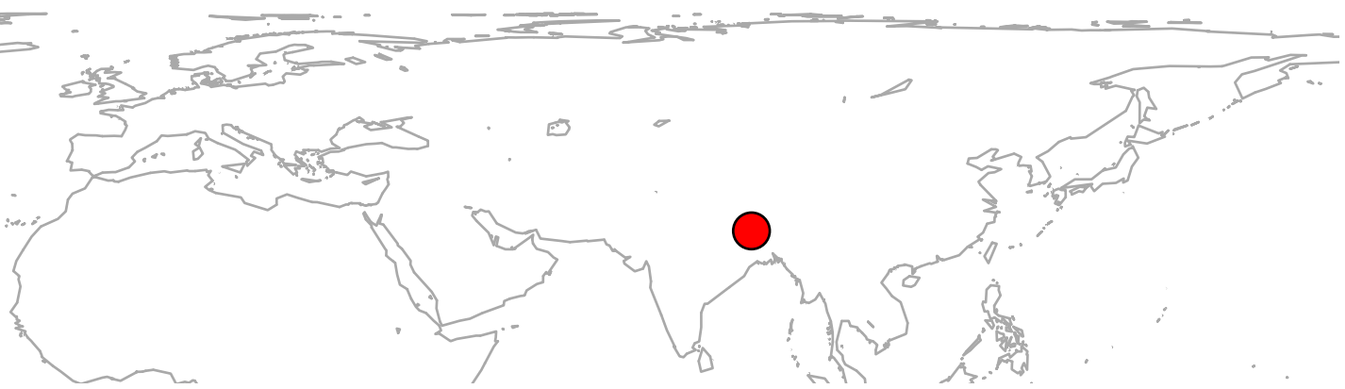
b. *u-ma-kha-cop-yokt-e*
3snA-NEG-1nsP-see-NEG-PST

c. *kha-u-ma-cop-yokt-e*
1nsP-3snA-NEG-see-NEG-PST

d. *ma-u-kha-ma-cop-yokt-e*
NEG-3snA-1nsPsee-NEG-PST

e. …

All: 'They didn't see us.'



Chintang prefixes

Category	Prefix	Meaning	(Village)
NEG	***mai- ~ ma-***	NEG	
SUBJ	***a-***	2.S/A	
	u-	3ns.S/A; 3.A (if P = 1s)	
OBJ	***kha-***	1ns.P	Sambugaũ
	ma-	1ns.excl.P	Mulgaũ
	mai-	1ns.incl.P	
A>P	***na-***	3>2	

Mansfield, Stoll & Bickel 2019 *ALT*

Not Secure — clrp.uzh.ch

clrp

overview
language
people and culture
language acquisition
- research projects
- ma theses
linguistic analysis
- phd theses
- ma theses
publications
archive and access
team
contact

The Chintang language research program
छिन्ताङ भाषा अनुसन्धान कार्यक्रम

The Chintang Language Research Program aims at a rich documentation and in-depth analysis of Chintang, a language of the Kiranti subgroup of Sino-Tibetan spoken in Eastern Nepal. CLRP is the successor of an earlier project that was funded by the Volkswagen Foundation (Documentation of Endangered Languages Program) between 2004 and 2009 and included the development of a corpus of Chintang and one other Kiranti language, Puma (see the Chintang and Puma Documentation Project).

CLRP was started in 2009 and includes two components:

- a linguistic component devoted to analyzing grammar, lexicon and language use
- a language acquisition component devoted to analyzing how children learn the language

CLRP is carried by a team of researchers headed by Sabine Stoll and Balthasar Bickel at the University of Zurich. The program cooperates with the Central Department of Linguistics and the Centre for Nepal and Asia Studies at Tribhuvan University, Kirtipur and is part of LiNSuN (the Linguistic Survey of Ne

576 instances of prefix bigrams

The corpus contains recordings of c

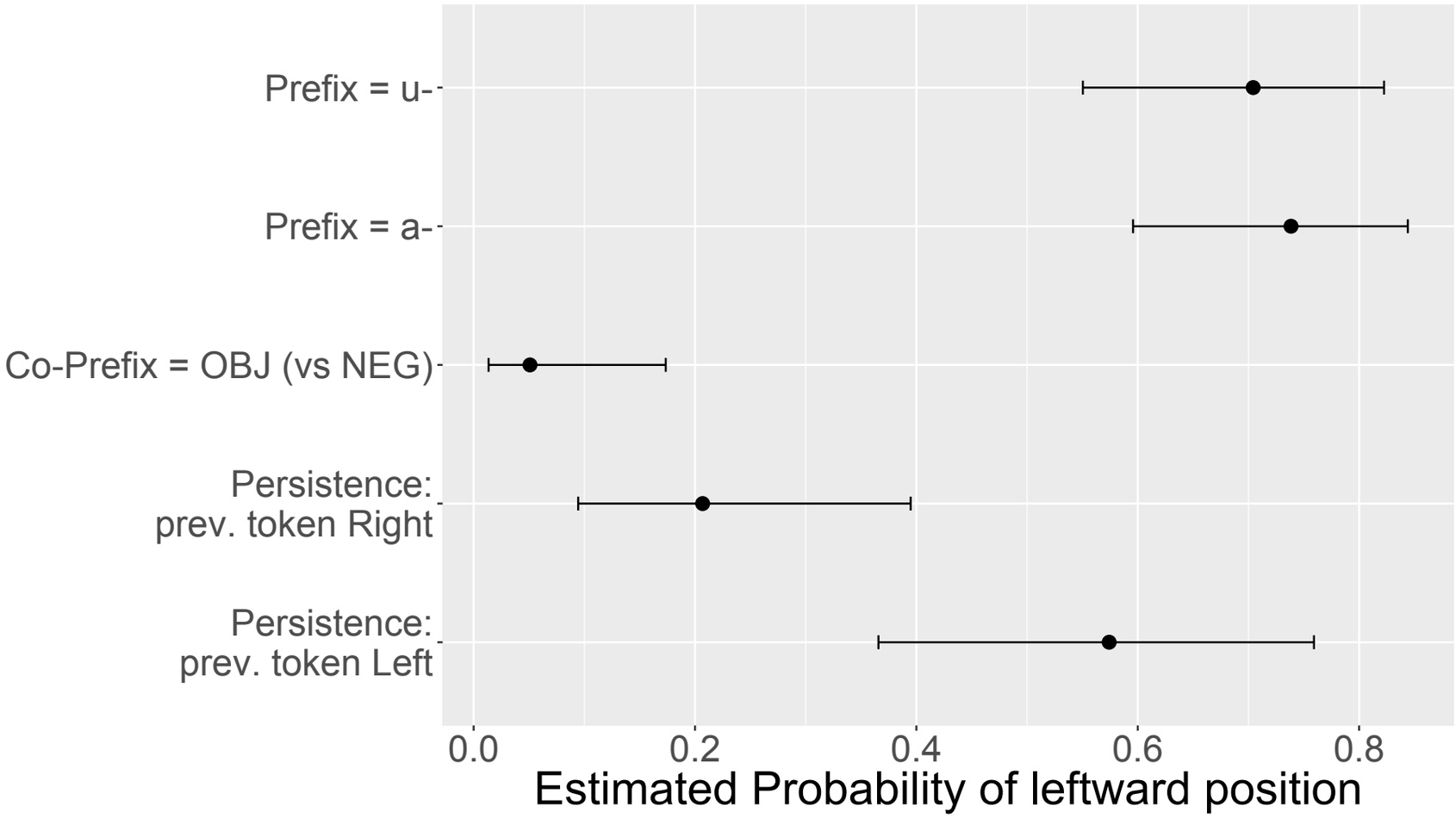
Genre	Transcribed duration	Transcribed no. of words	Translated duration	Translated no. of words	Glossed duration	Glossed no. of words
conversation	232:44:54	1,064,109	232:39:18	1,045,254	207:23:27	903,645
description	3:09:10	20,934	3:06:24	20,617	1:52:56	14,433
narrative	6:15:13	46,044	6:12:12	45,826	5:54:52	42,922
experimental	4:11:31	43,780	3:57:37	33,273	2:48:08	24,110

Impressum

Chintang prefixes

Given a bigram of an agreement prefix and its co-prefix: What is the probability of the prefix being placed on the left

- if the prefix is *u-* ‘3’ vs *a-* ‘2’ (**paradigmatic alignment: all together**)
- if its co-prefix is OBJ vs NEG (**featural coherence: coherent slots**)
- if the same order occured before (persistance, priming)?



OBJ>SUBJ>NEG

Mansfield, Stoll & Bickel 2019 *ALT*

A correlated bias in diachrony?

	+ COHERENT	– COHERENT
+ ALIGNED	STEM-(A_1|A_2)-(P_1|P_2)	STEM-(A_1|A_2|P_1|P_2) STEM-(A_1>P_2|A_2>P_1)
– ALIGNED	STEM-A_1-(P_1|P_2)-A_2 STEM-($A_{1\alpha}$|A_2)-(P_1|P_2)-$A_{1\beta}$	STEM-(A_1|P_1)-(A_2|P_2) STEM-(A_1>$P_{2\alpha}$)-($P_{2\beta}$)

Mansfield, Stoll & Bickel 2019 *ALT*

Paradigmatic Alignment

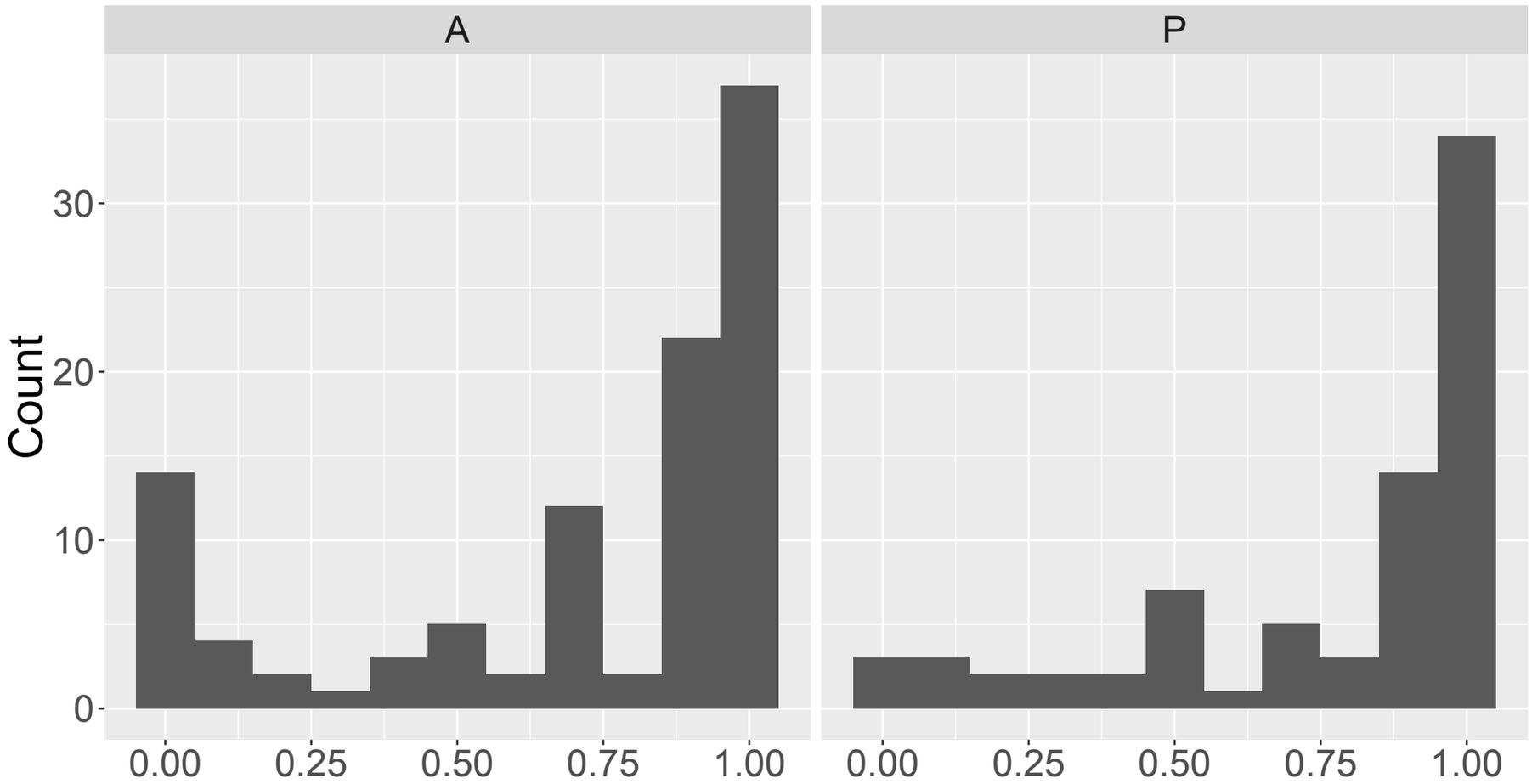
Example:

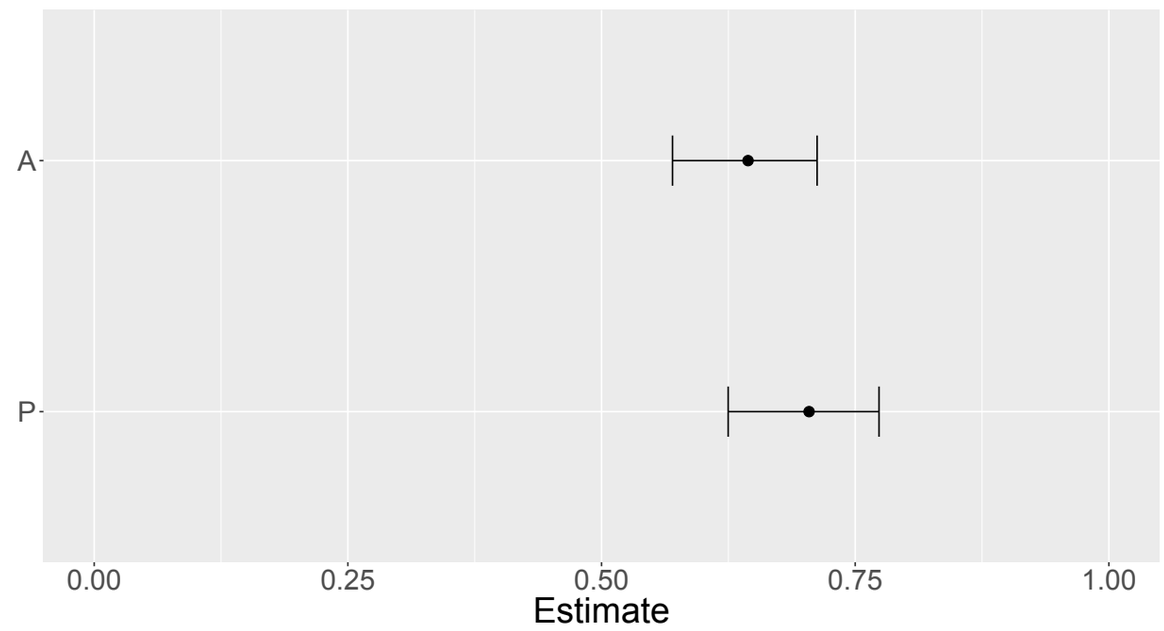
Reyesano A affix allocations (Guillaume 2009)

	Σ-2	Σ-1	Σ	Σ+1
1s	*m-*			
1p	*k-*			
2s	*mi-*			
2p	*mik-*			
3				*-ta*

Partition	Example allocations			N allocations	Pr	H	Cum Pr	Alignment Index
	Σ-2	Σ-1	Σ+1					
{5}	1s,1p,2s,2p,3 - -	- 1s,1p,2s,2p,3 -	- - 1s,1p,2s,2p,3	3	0.01	0	0.01	0.99
{4,1}	**1s,1p,2s,2p** 1s,1p,2s,3 *etc…* - - 3 2p *etc…*	– 2p 1s,1p,2s,2p 1s,1p,2s,3 *etc…* 1s,1p,2s,2p 1s,1p,2s,3	**3** - 3 2p - -	30	0.12	0.72	0.13	0.87
{3,2}	1s,1p,2s 1s,1p,2p 1s,2s,2p *etc…* - - -	2p,3 2s,3 1p,3 1s,1p,2s 1s,1p,2p 1s,2s,2p *etc…*	- - - 2p,3 2s,3 1p,3	60	0.25	0.97	0.38	0.62
{3,1,1}	1s,1p,2s 1s,1p,2p 1s,2s,2p *etc…* 1s,1p,2s 1s,1p,2p 1s,2s,2p *etc…*	2p 2s 1p 3 3 3	3 3 3 2p 2s 1p	60	0.25	1.37	0.63	0.37
{2,2,1}	1s,1p 1s,2s 1s,2p 1s,1p 1s,2s *etc…*	2s,2p 1p,2p 1p,2s 2s,3 1p,3	3 3 3 2p 2p	90	0.37	1.52	1	0
SUM				243	1.00			

Paradigmatic Alignment





Beta regression with random family intercepts, $p < .001$

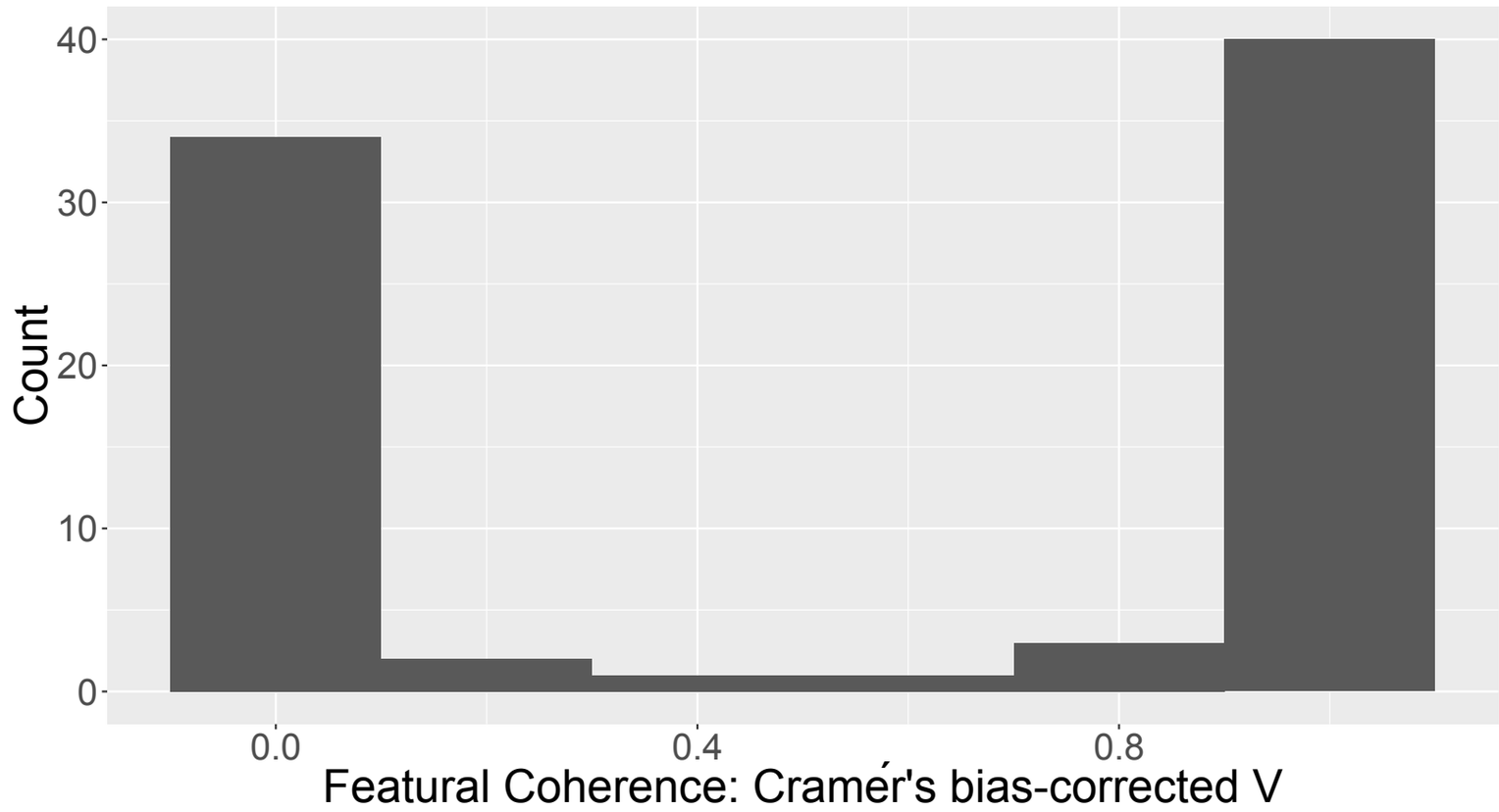
Mansfield, Stoll & Bickel 2019 *ALT*

Featural Coherence

	Σ-1	Σ+1	Σ+2
A	1	0	5
P	0	4	0

	Σ-2	Σ-1	Σ+1
A	1	4	1
P	0	3	1

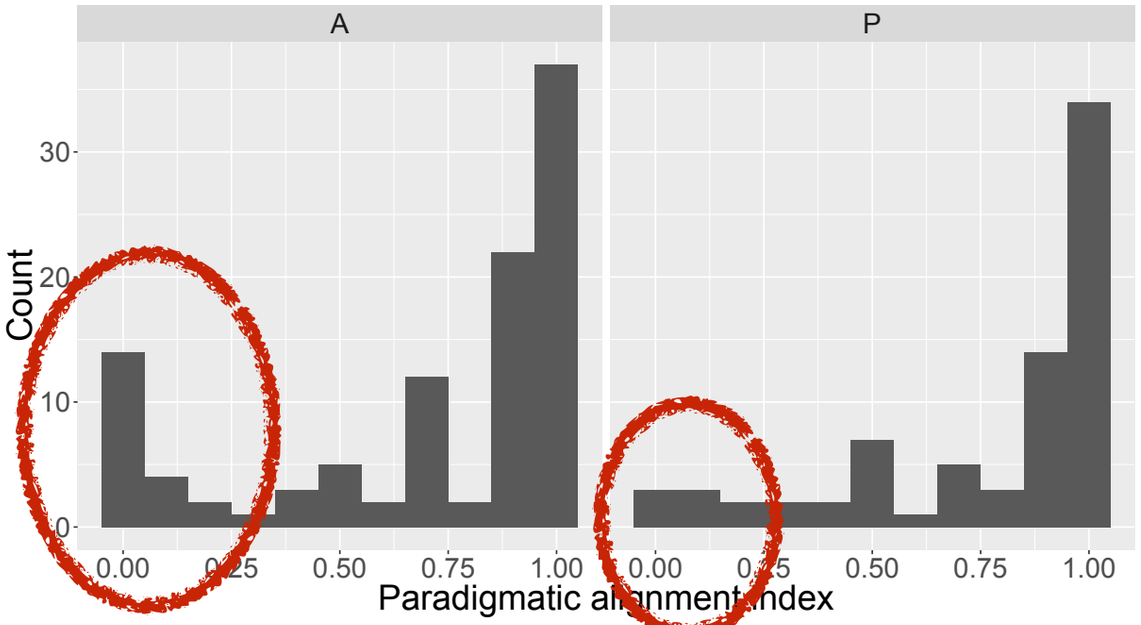
	Σ-2	Σ-1	Σ+1
A	1	2	2
P	1	2	1

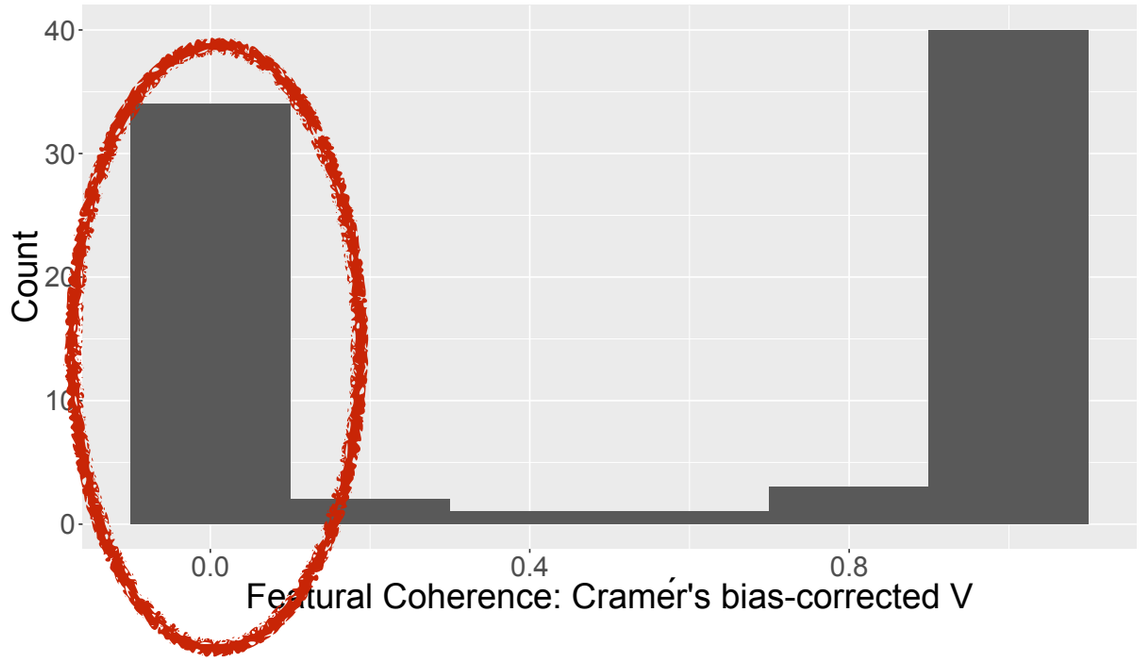


Beta regression with random family intercept,
$\hat{V}$ = .78, 95% CI = [.62, .88], $p < .01$

Mansfield, Stoll & Bickel 2019 *ALT*

Exceptions





Probably two main sources:

- person- rather than role-defined positions
- distributed exponence

Belhare (Kiranti, Bickel 1996)

lui-t-u-m-chi-m-ga

tell-NPST-3P-nsA-nsP-nsA-2

‘You will tell them’

Cree (Algonquian, Dahlstrom 1986)

ki-pēhtaw-iti-n

2-hear-1>2>s1/2

‘I hear you’

Mansfield, Stoll & Bickel 2019 *ALT*

Study 2 Summary

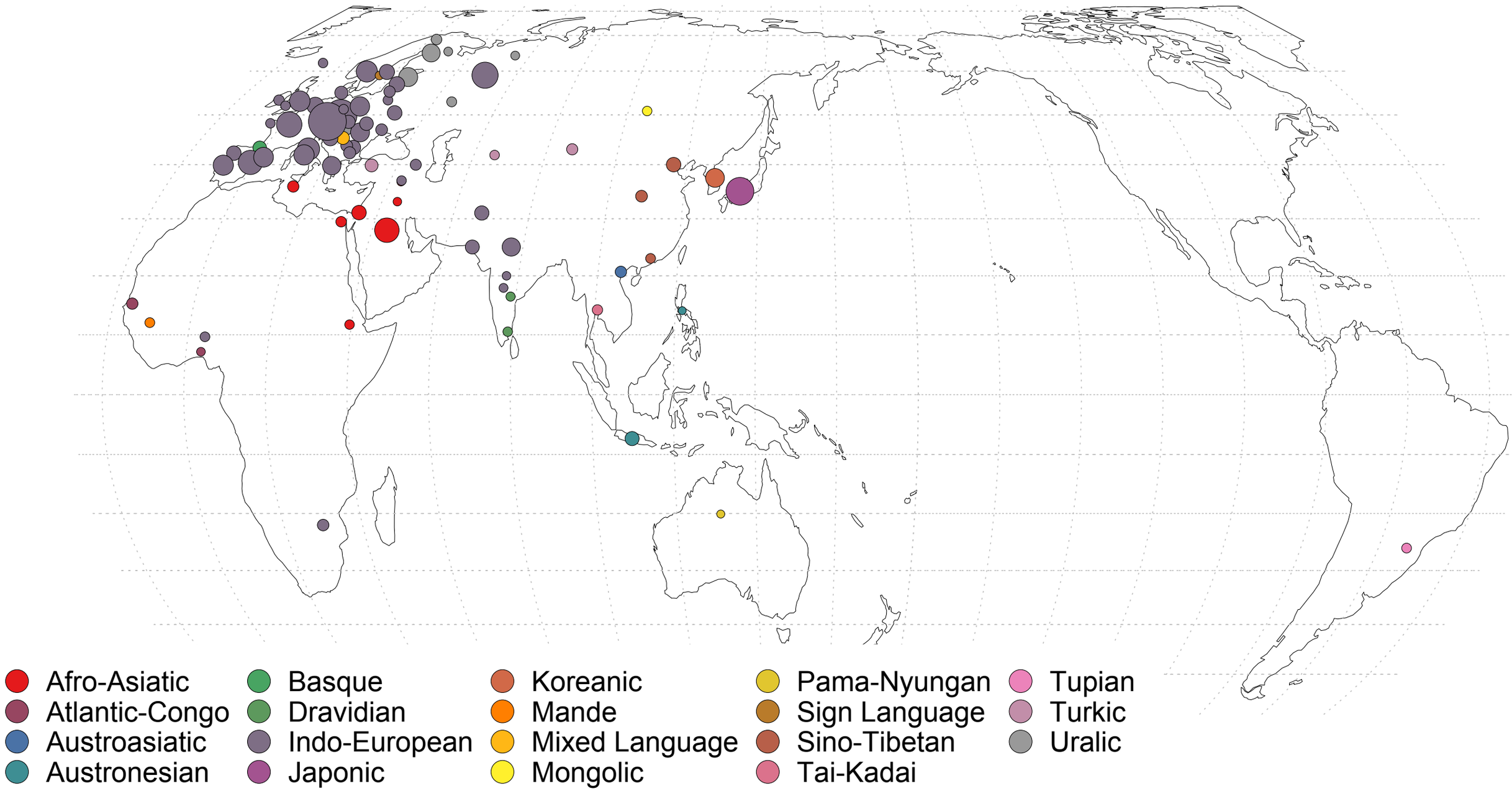
1. Bias towards clustering when a grammar allows variation, possibly because this facilitates learning and prediction

2. The same bias drives clustering of A and P markers when languages evolve over time, with two principled exceptions

Mansfield, Stoll & Bickel 2019 *ALT*

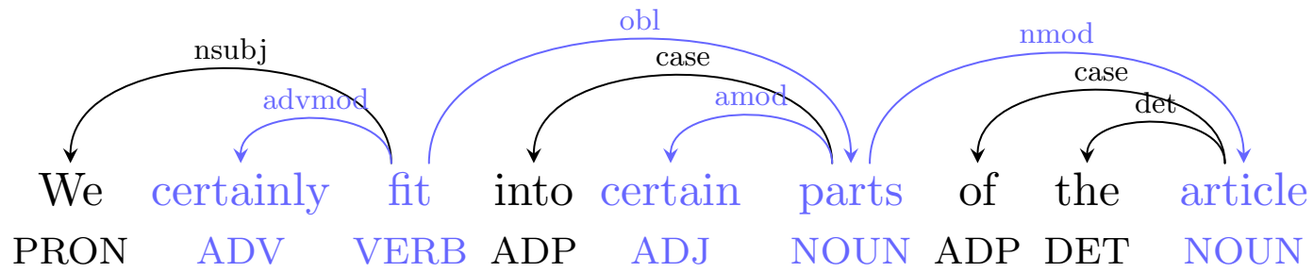
Study 3: Constraints on word order evolution (with Damián Blasi and Jing Yingqi)

UD 2.4 (Nivre et al. 2019):

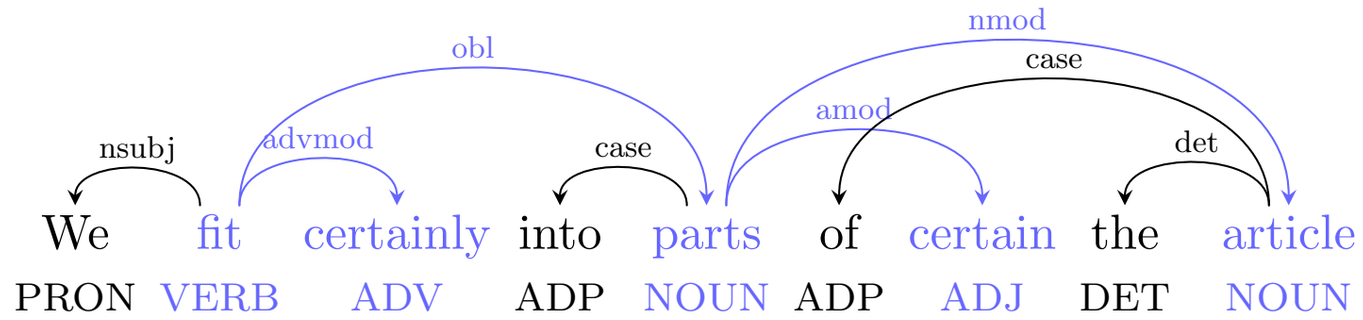


Baselines

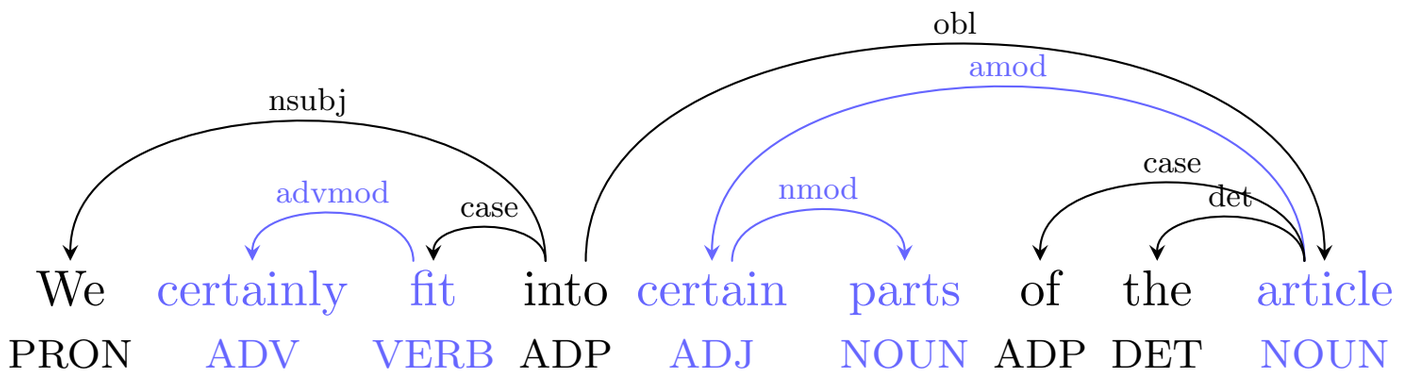
Observed:



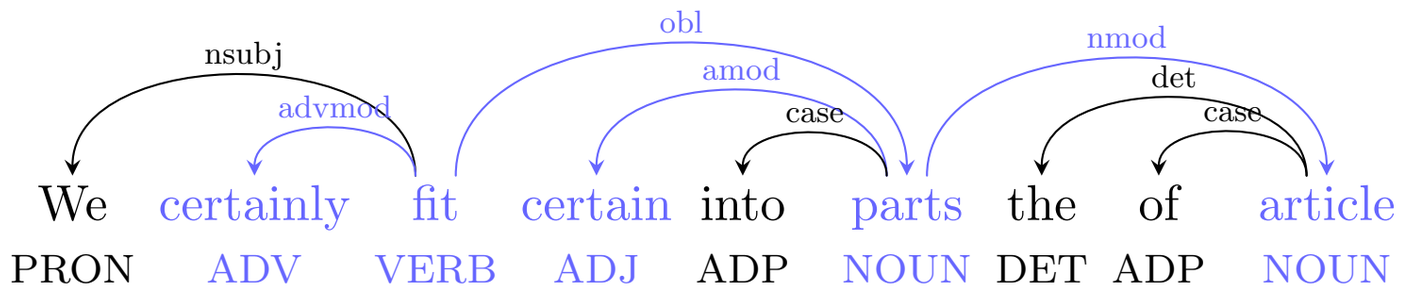
Ferrer-i-Cancho 2004:



Liu 2008:

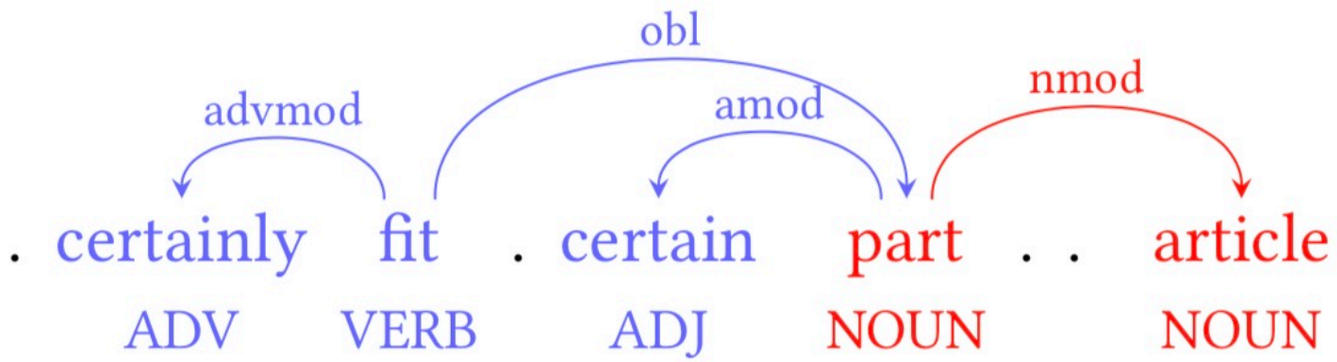


Futtrell et al. 2015:

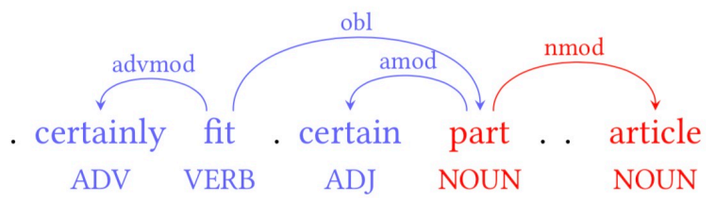


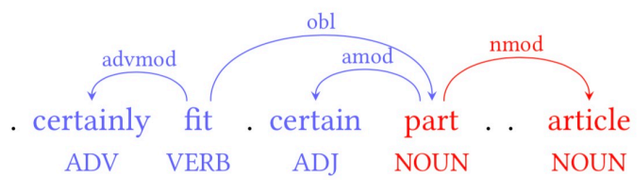
Jing, Blasi & Bickel 2017 *ALT*

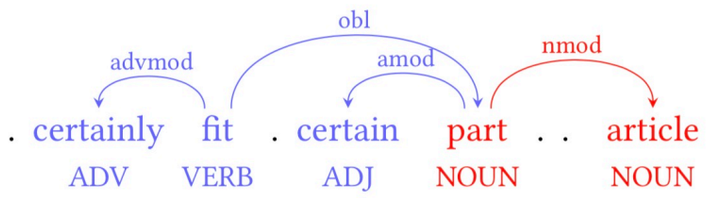
A psycholinguistically informed baseline: produce what you learnt without further production constraints (like DLM)!

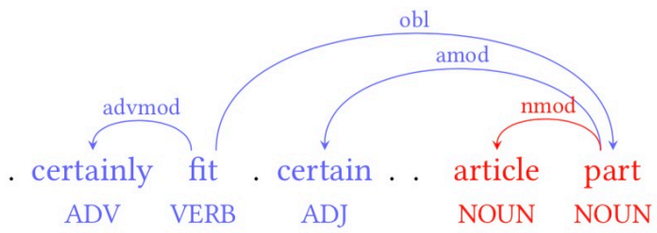


VO: 97.7%
VS: 14.76%
NGen: 90.34%
NAdj: 1.41%
⋮





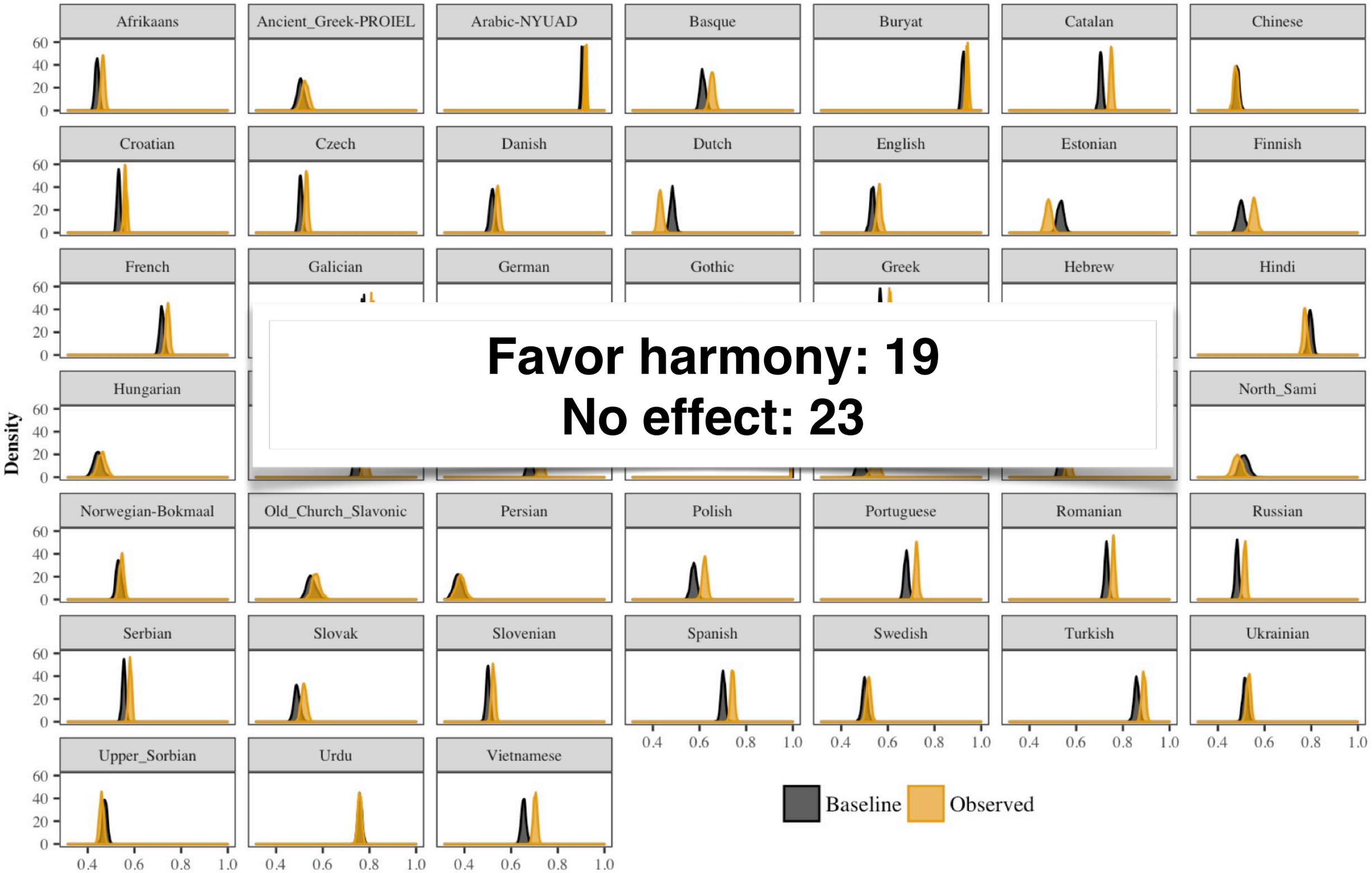




...

Jing, Blasi & Bickel 2017 *ALT*

Harmony in dependency bigrams?



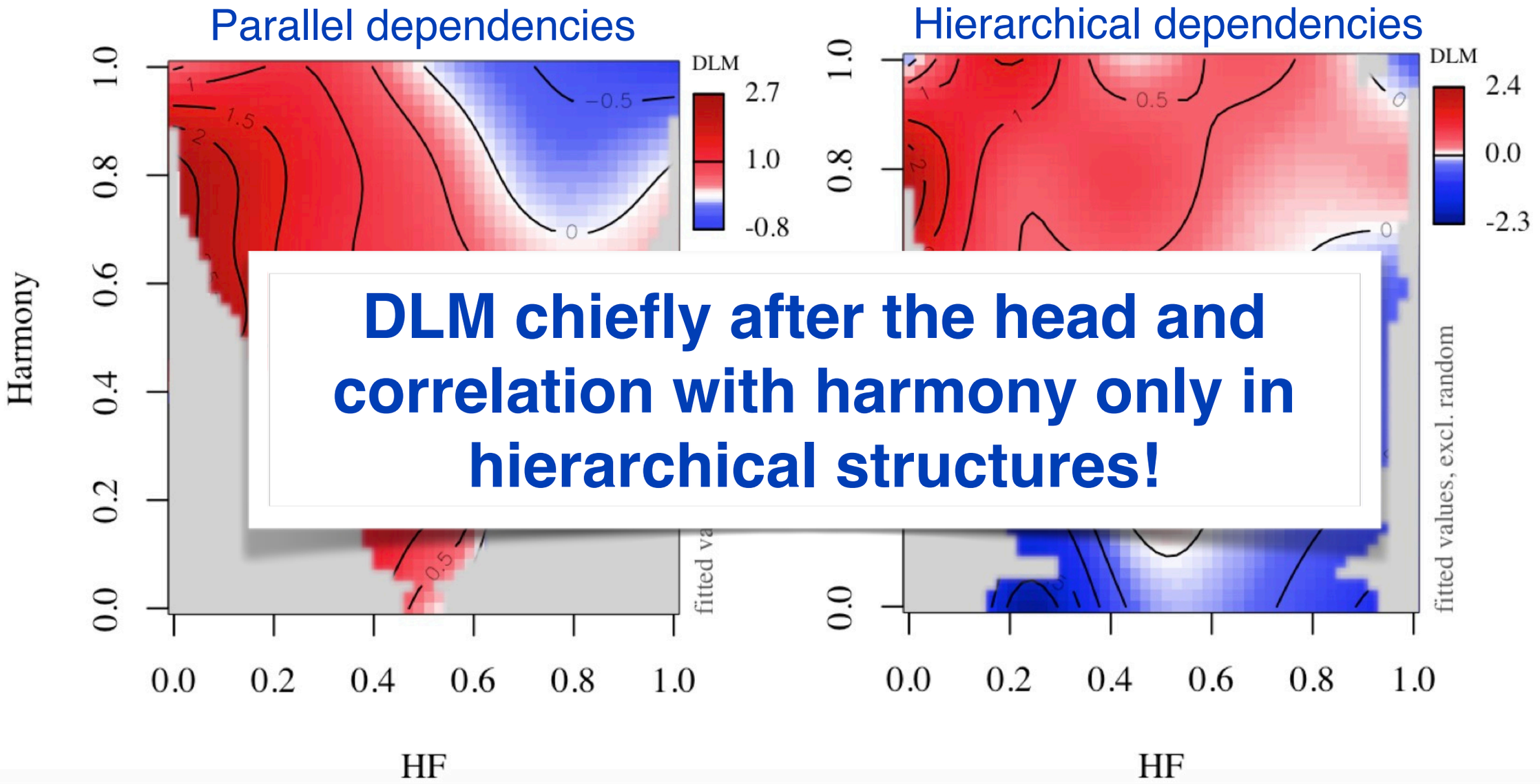
Jing, Blasi & Bickel 2017 *ALT*

Dependency Length Minimization (DLM)?

DLM = Pr($DL_{obs} \leq DL_{baseline}$)

DLM ~ Pr(Harmony) × Pr(HF) + random lang., per sentence

binomial GAM; only languages with Pr(HF) = [.2, .8]



DLM chiefly after the head and correlation with harmony only in hierarchical structures!

Jing, Blasi & Bickel in prep.

Study 3 Summary

1. Our baseline asks about whether we need to postulate anything above and beyond a simple mechanism of reproducing structures in proportion to the frequencies they are learned with

2. On this basis, we need fewer mechanisms for harmony and DLM than current theories predict.

Jing, Blasi & Bickel in prep.

Conclusions

1. Claims about onstraints on language require testing in non-WEIRD samples, and these samples allow new discoveries (e.g. pause and affix order probabilities).

2. Corpora are fantastic natural production experiments, but they deserve psycholinguistically informed baselines, not just any randomization.