FRIAS Workshop on Variation, February 2011

The role of genealogical units in explaining linguistic distributions:

a case study on referential density

Balthasar Bickel, U Leipzig

Genealogical units in linguistic research

- Dialect/language/family as the basic units of data representation:
  - dialect/language/family X has value ("type") A on variable V1, degree .9 on variable V2, etc.
  - vector of values $V_1 \ldots V_n$ characterizes dialect/language/family X best
  - etc.
- Typically, statements like these require massive and highly problematic data reduction (Bickel 2007, Waechli 2009):
  - constructional variation is reduced (e.g. "basic" word orders)
  - speech samples are aggregated (e.g. "mean" orders)

So why do it?

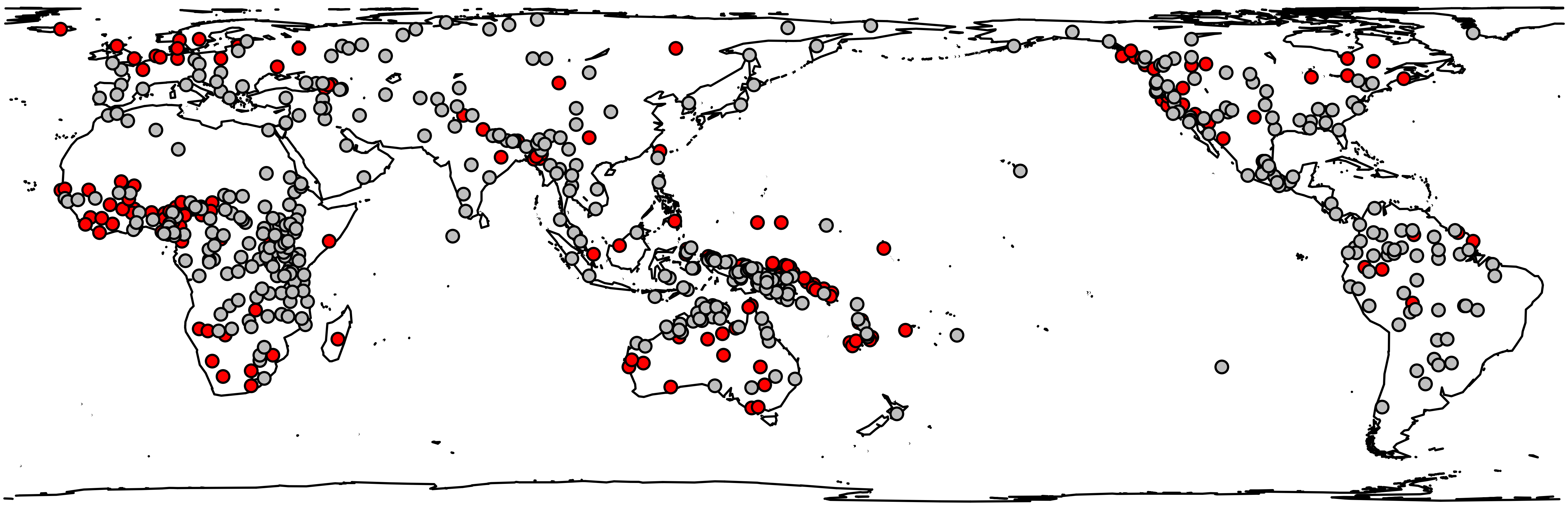
Collecting data at the level of genealogical units

1. **Descriptive convenience:** we need labels to identify the speech samples or constructions we analyze.

- Look at this in a case study on NP use in discourse and see what we get

Case study: referential density (RD)

- Point of departure: a universal preference for ‘pro-drop’



Case study: referential density (RD)

- But to very different degrees: Pear story experiments

Belhare (Kiranti, Sino-Tibetan)

pʌila . . . aː . . . ambibu phig-he kinahuŋgo
first PTCL mango [ABS] [3s.A-]pick.from.above-PT[3O] SEQ

otutuiʔ = na jhola-e ukt-he
quite.big = ART[S] bag-LOC [3s.A-]take.down-PT[3O]

inetnahuŋgo dhaki-e leŋs-e
then closely.weaved.basket-LOC [3s.A-]put-PT[3O]

il-lam il-lam sas-sa-ba leŋs-e ʌni . . .
DIST:DEM-MED DIST:DEM-MED pull-CONV-LOC [3s.A-]put-PT[3O] and.then

riksa, eː saikil-lamma, saikil-lamma ta-he
rikshaw PTCL bicycle-MED bicycle-MED [3s.S-]come-PT

kinahuŋgo . . . ⟨B99.4.1–5⟩
SEQ

‘First, . . . uh . . . [someone] picked mangos and took [them] down in a big bag. Then [s/he] put [them] into a basket. [Someone] moved over [an animal] by pulling from over there, and then [someone] came on a rikshaw, uh . . . on a bike, on a bike and then . . .’

Maithili (Indo-Aryan, Indo-European; Nepal)

ek-ṭā ām-ke gāch rah-ai. ā . . . a . . . a . . .
one-CL mango-GEN tree[NOM] be-3NH.NOM[PR] PTCL

ām me ek e-goṭā chaurā ām tor-ait
mango in one one-CL boy[NOM] mango[NOM] pluck-IP

rah-ai
AUX-3NH.NOM[-3NH.NONNOM.PR]

ā . . . u ām toir-ke ṭokari me rakh-ne
PTCL 3NH.NOM mango[NOM] pluck-CONV basket in keep-INF

jāi che-l-ai. omaharse e-goṭā chaurā
AUX AUX-PT-3NH.NOM[-3NH.NONNOM] and.then one-CL boy[NOM]

e-l-ai,
come-PT-3NH.NOM

laḍkā sāikal par caḍh-ne, ā . . . u ek-ṭā am-ke
boy.H[NOM] bike on ride-INF PTCL 3NH.NOM one-CL mango-GEN

ṭokari corā-ke cail ge-l-ai . . . ⟨M3.6.1–6⟩
basket[NOM] steal-CONV move.IP AUX-PT-3NH.NOM

‘There is a mango tree and . . . uh . . . uh . . . in the mangos, one, a boy is picking mangos. And when picking mangos, he put them into a basket. Then a boy came, a young man riding on a bike, and he stole one basket of mangos, and took off . . .’

Case study: referential density (RD)

Belhare (Kiranti, Sino-Tibetan)

pʌila . . .	aː . . .	ambibu		phig-he	kinahuŋgo
first	PTCL	mango	[ABS]	[3s.A-]pick.from.above-PT[3O]	SEQ

otutuiʔ = na	jhola-e	ukt-he
quite.big = ART[s]	bag-LOC	[3s.A-]take.down-PT[3O]

inetnahuŋgo	dhaki-e	leŋs-e
then	closely.weaved.basket-LOC	[3s.A-]put-PT[3O]

il-lam	il-lam	sas-sa-ba	leŋs-e	ʌni . . .
DIST:DEM-MED	DIST:DEM-MED	pull-CONV-LOC	[3s.A-]put-PT[3O]	and.then

riksa,	eː	saikil-lamma,	saikil-lamma	ta-he
rikshaw	PTCL	bicycle-MED	bicycle-MED	[3s.S-]come-PT

kinahuŋgo . . .	⟨B99.4.1–5⟩
SEQ	

‘First, . . . uh . . . [someone] picked mangos and took [them] down in a big bag. Then [s/he] put [them] into a basket. [Someone] moved over [an animal] by pulling from over there, and then [someone] came on a rikshaw, uh . . . on a bike, on a bike and then . . .’

Case study: referential density (RD)

- Maithili (Indo-Aryan, Indo-European; Nepal)

ek-ṭā ām-ke gāch rah-ai. ā . . . a . . . a . . .
one-CL mango-GEN tree[NOM] be-3NH.NOM[PR] PTCL

ām me ek e-goṭā chaurā ām tor-ait
mango in one one-CL boy[NOM] mango[NOM] pluck-IP

rah-ai
AUX-3NH.NOM[-3NH.NONNOM.PR]

ā . . . u ām toir-ke ṭokari me rakh-ne
PTCL 3NH.NOM mango[NOM] pluck-CONV basket in keep-INF

jāi che-l-ai. omaharse e-goṭā chaurā
AUX AUX-PT-3NH.NOM[-3NH.NONNOM] and.then one-CL boy[NOM]

e-l-ai,
come-PT-3NH.NOM

laḍkā sāikal par caḍh-ne, ā . . . u ek-ṭā am-ke
boy.H[NOM] bike on ride-INF PTCL 3NH.NOM one-CL mango-GEN

ṭokari corā-ke cail ge-l-ai . . . ⟨M3.6.1–6⟩
basket[NOM] steal-CONV move.IP AUX-PT-3NH.NOM

‘There is a mango tree and . . . uh . . . uh . . . in the mangos, one, a boy is picking mangos. And when picking mangos, he put them into a basket. Then a boy came, a young man riding on a bike, and he stole one basket of mangos, and took off . . .’

Case study: referential density (RD)

- But to very different degrees: Pear story experiments

Belhare (Kiranti, Sino-Tibetan)

pʌila . . . aː . . . ambibu phig-he kinahuŋgo
first PTCL mango [ABS] [3s.A-]pick.from.above-PT[3O] SEQ

otutuiʔ = na jhola-e ukt-he
quite.big = ART[S] bag-LOC [3s.A-]take.down-PT[3O]

inetnahuŋgo dhaki-e leŋs-e
then closely.weaved.basket-LOC [3s.A-]put-PT[3O]

il-lam il-lam sas-sa-ba leŋs-e ʌni . . .
DIST:DEM-MED DIST:DEM-MED pull-CONV-LOC [3s.A-]put-PT[3O] and.then

riksa, eː saikil-lamma, saikil-lamma ta-he
rikshaw PTCL bicycle-MED bicycle-MED [3s.S-]come-PT

kinahuŋgo . . . ⟨B99.4.1–5⟩
SEQ

‘First, . . . uh . . . [someone] picked mangos and took [them] down in a big bag. Then [s/he] put [them] into a basket. [Someone] moved over [an animal] by pulling from over there, and then [someone] came on a rikshaw, uh . . . on a bike, on a bike and then . . .’

Maithili (Indo-Aryan, Indo-European; Nepal)

ek-ṭā ām-ke gāch rah-ai. ā . . . a . . . a . . .
one-CL mango-GEN tree[NOM] be-3NH.NOM[PR] PTCL

ām me ek e-goṭā chaurā ām tor-ait
mango in one one-CL boy[NOM] mango[NOM] pluck-IP

rah-ai
AUX-3NH.NOM[-3NH.NONNOM.PR]

ā . . . u ām toir-ke ṭokari me rakh-ne
PTCL 3NH.NOM mango[NOM] pluck-CONV basket in keep-INF

jāi che-l-ai. omaharse e-goṭā chaurā
AUX AUX-PT-3NH.NOM[-3NH.NONNOM] and.then one-CL boy[NOM]

e-l-ai,
come-PT-3NH.NOM

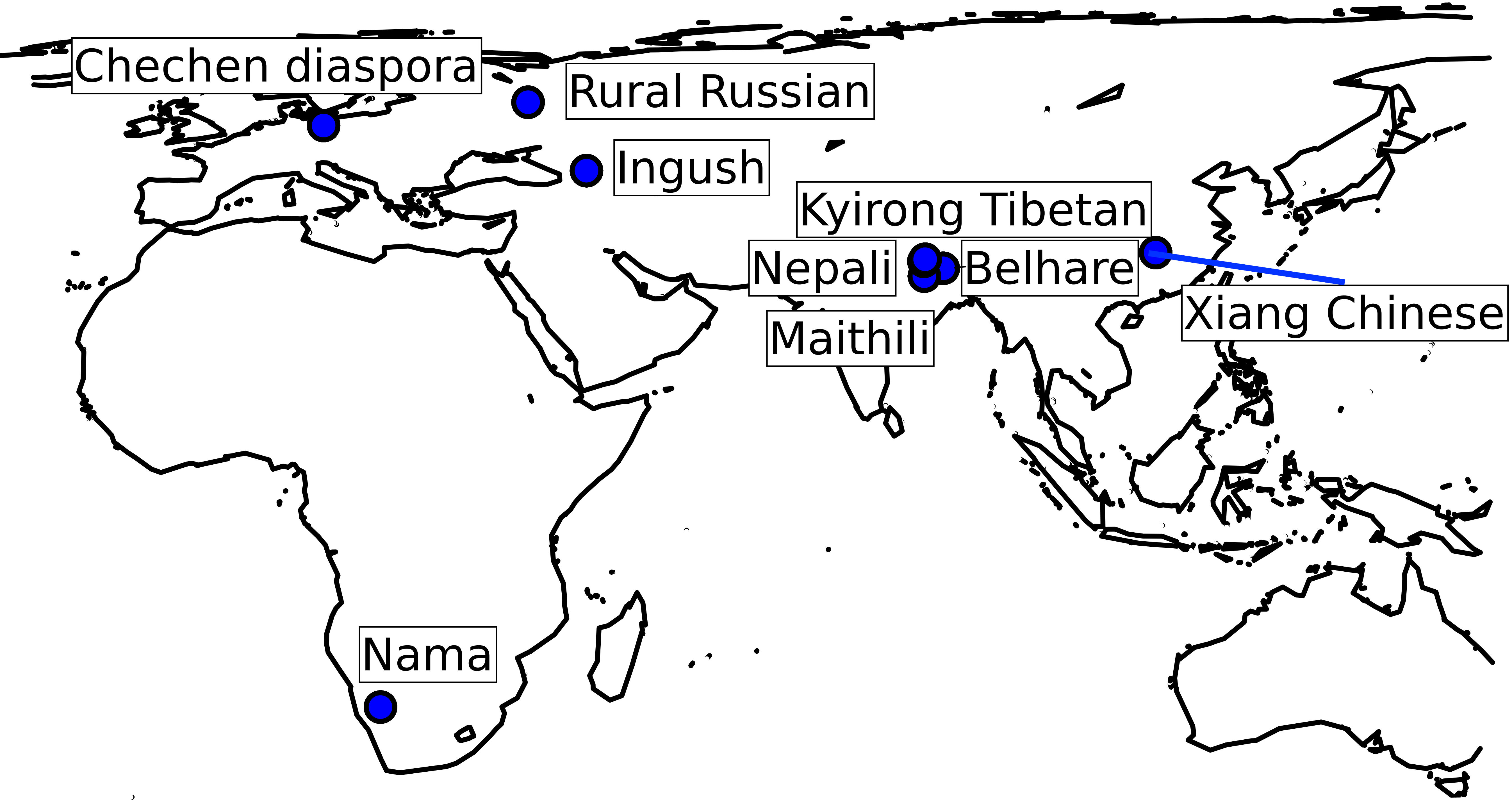
laḍkā sāikal par caḍh-ne, ā . . . u ek-ṭā am-ke
boy.H[NOM] bike on ride-INF PTCL 3NH.NOM one-CL mango-GEN

ṭokari corā-ke cail ge-l-ai . . . ⟨M3.6.1–6⟩
basket[NOM] steal-CONV move.IP AUX-PT-3NH.NOM

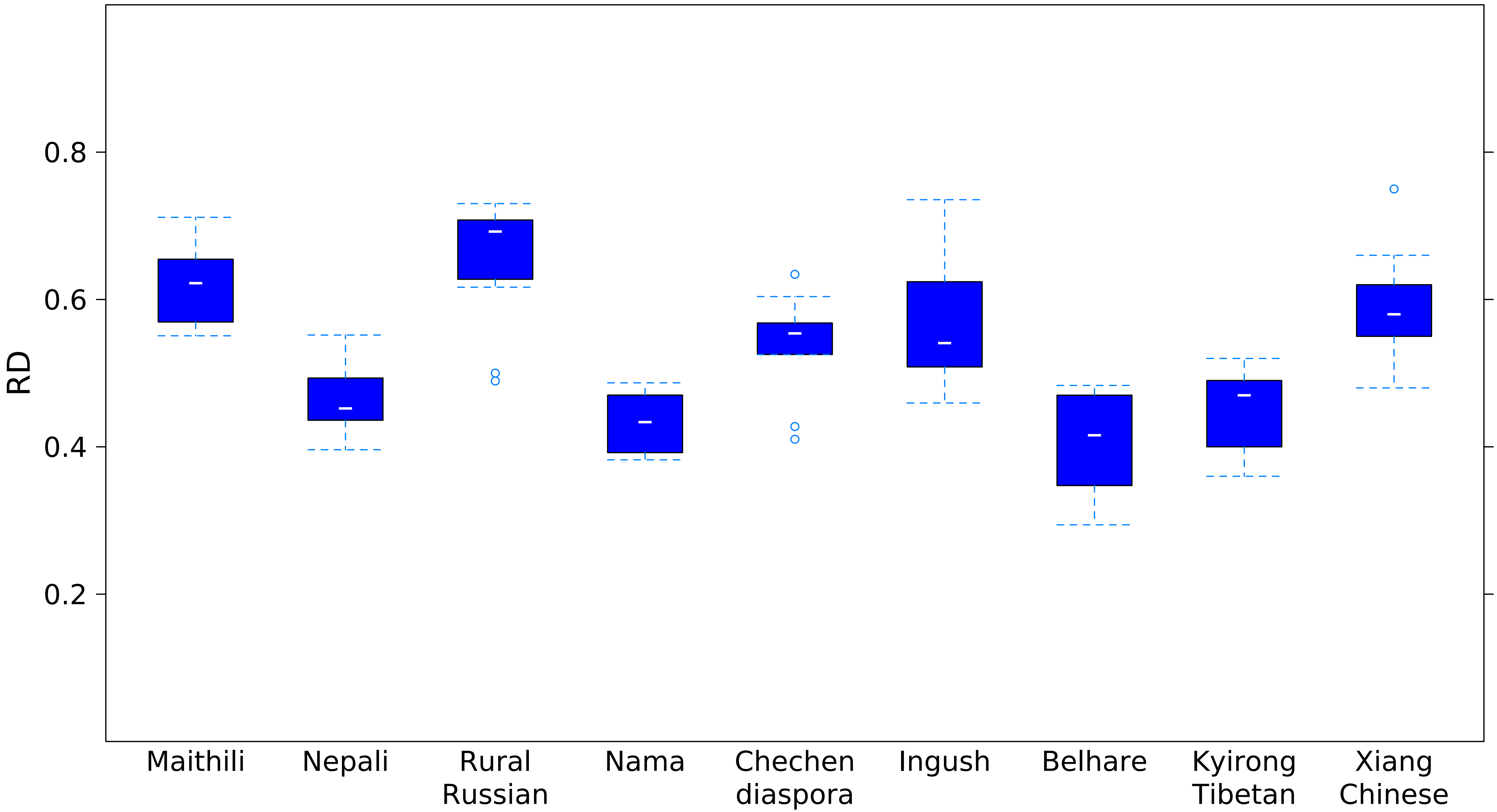
‘There is a mango tree and . . . uh . . . uh . . . in the mangos, one, a boy is picking mangos. And when picking mangos, he put them into a basket. Then a boy came, a young man riding on a bike, and he stole one basket of mangos, and took off . . .’

$$\mathrm{RD} = \frac{N\ \text{(overt argument NPs)}}{N\ \text{(available argument positions)}}$$

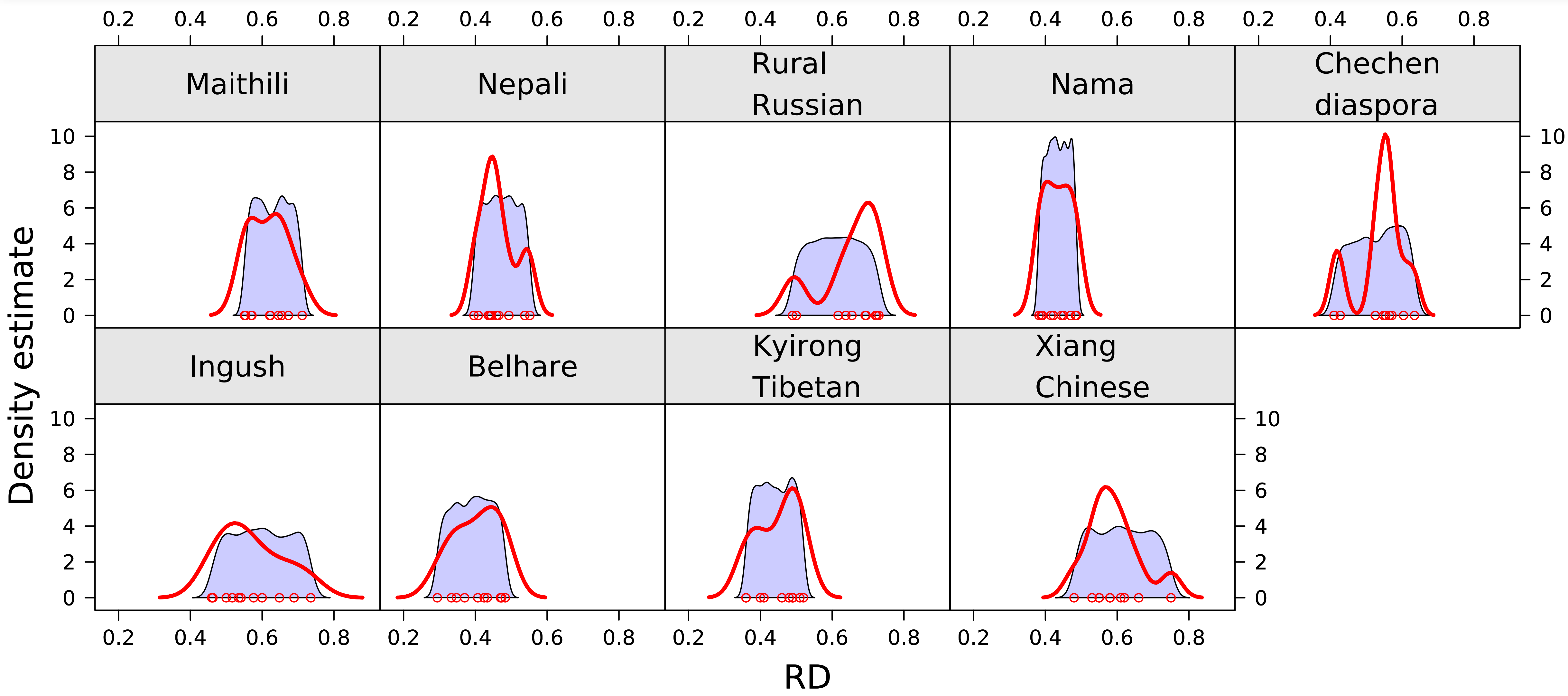
Case study: referential density (RD)



Case study: referential density



Case study: referential density

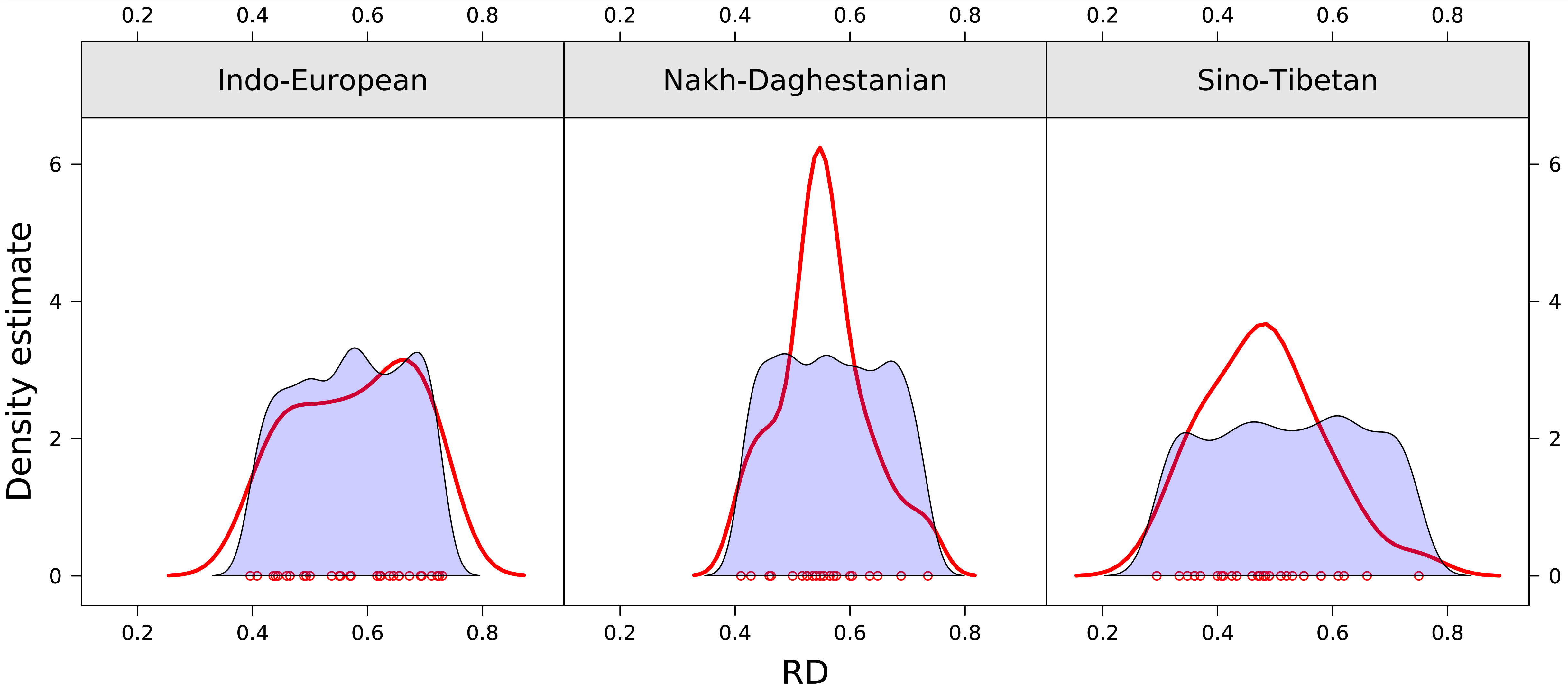


- in most languages flat distributions, no clear "normative mean":
- Variance test against H_0: $\mathcal{U}(\min(RD_L), \max(RD_L))$, i.e. with a H_0 independent of the overall sample location: all $p > .1$,

  adopting Coeurjolly et al.'s (2009) robust test based on the statistic

  $\hat{\theta} = \frac{\hat{\sigma}^2 - \sigma_0^2}{\sqrt{Var(\hat{\sigma}^2)}}, \quad \sigma_0^2 = \frac{1}{12}(\max(RD_L) - \min(RD_L))^2$

Case study: referential density



- Variance test again against H_0: $\mathcal{U}(\min(RD_{family}), \max(RD_{family}))$:
  - Indo-European: $\hat{\sigma}^2 = .01, \hat{\theta} = .98, p = .835$
  - Nakh-Daghestanian: $\hat{\sigma}^2 = .006, \hat{\theta} = -1.35, p = .088$
  - Sino-Tibetan: $\hat{\sigma}^2 = .01, \hat{\theta} = -2.36, p = .009$

Collecting data at the level of genealogical units (*cont'd*)

- So: in many cases, no evidence for a trend towards a mean (at least not with the small sample sizes I have here, $N_L = 10$)
- no evidence so far for typical or characteristic values per language or family, no “rhetorical norms” per unit!

▶ the units may not be suitable units of data aggregation (so far)

- But even if we find significant trends towards a mean in a unit, the aggregation may be
  - right/beautiful/fascinating, but ...

▶ **So, using units for convencience either lacks justification or interest or both**

Collecting data at the level of genealogical units

2. **Statistical control:** we need to control for influences of individual quirks and “historical accidents” when testing universals (cf. Dryer 1989, 2000, 2009):

- If two speech samples or constructions are from family *F*, they might share features because of this, not because of universals,

  e.g. both have OV&Po because proto-*F* happened to have had *OV&Po, not because OV prefers Po

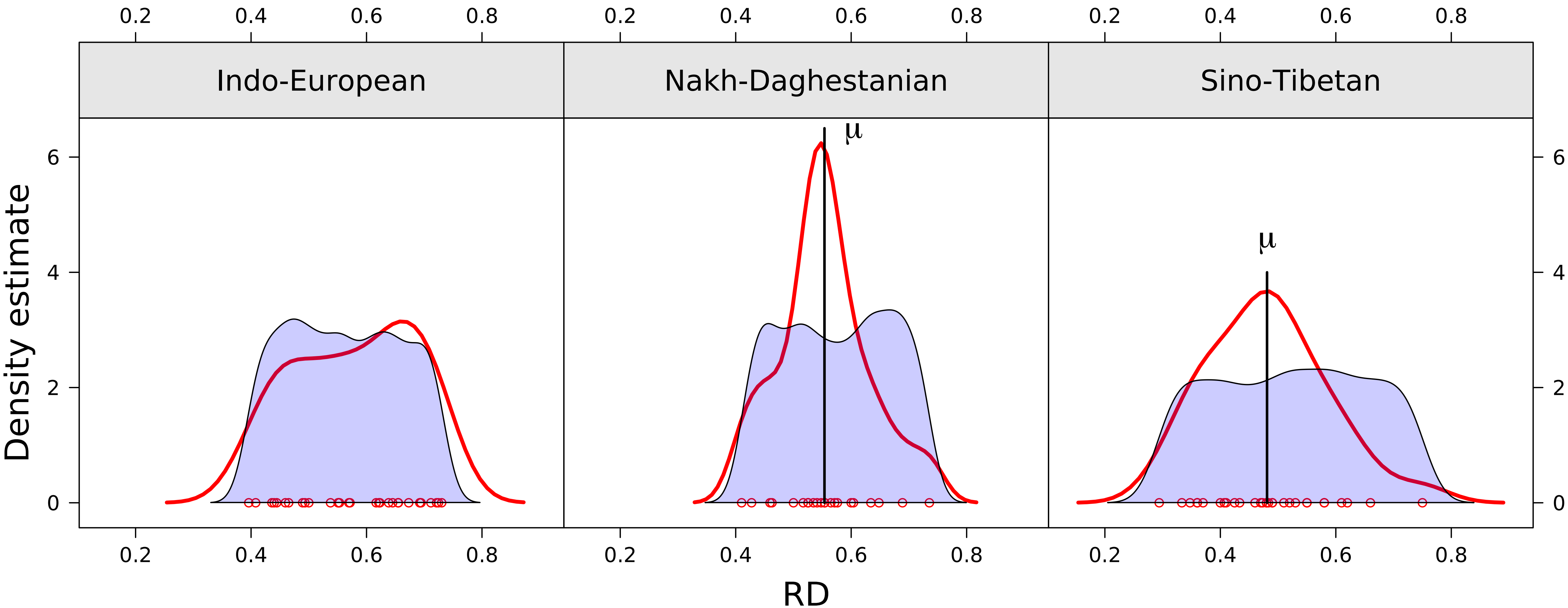
- If two speech samples or constructions are from language *L*, they might share features because of this, not because of universals

  e.g. both have similar RD values because *L* happens to have such a RD value **as a norm**

Genealogical units as statistical controls

- This equates means/norms/biases/trends/preferences within units with **hi-fi replication**, i.e.
  - “blind inheritance” within families
  - “normativity” within languages/dialects

(which are really the same processes)



Genealogical units as statistical controls

- But when things are replicated, this is
  - not always just because of lazy inertia and conservatism
  - but because they are good for the brain or because we like them (where we live) (Maslova 2000, Bickel 2008, 2011)
- i.e. the trend towards a mean in Sino-Tibetan (and perhaps Nakh-Daghestanian) can have many reasons, such as
  - universal stability (intrinsic, principled normativity)
  - universal preferences
  - areal diffusion
- These allow true explanations, but stating that two samples or constructions share values because they belong to “Chinese” or “Sino-Tibetan” does not explain anything.

Genealogical units as statistical controls

- In fact, this all follows from the definition of genealogical units through **individual-identifying features** (Nichols 1996):
- Saussurian form/meaning pairs whose similarity patterns
  - are unexpected from random sound developments:
    - finding {*sum* ‘3’ ∧ *li* ‘4‘ ∧ *ŋa*|*ŋɑ* ‘5’} several times across several speech samples is far below $p < .01$ and
    - suggests that we really only have **one single individual** unit: the proto-dialect/language/family plus non-random developments from it
  - and **cannot be explained** by universals or contact/ diffusion

▶ **Genealogical units are unexplained quirks by definition**

Quirks in general

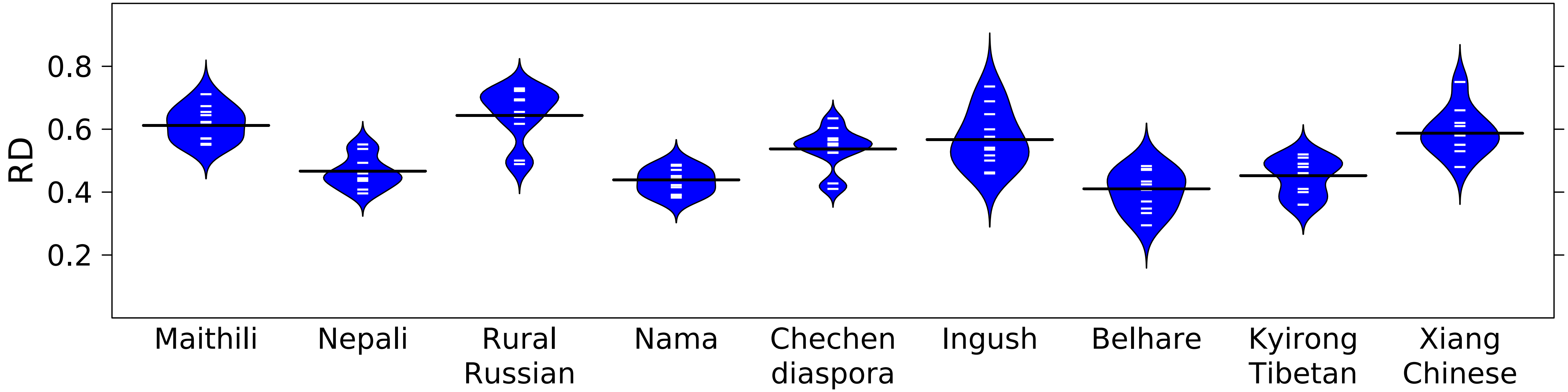
- Unexplained quirks can account for much of the variance (like speaker quirks in experiments, cf. Baayen 2008:259):
- Modeling language or family as a random factor, i.e. comparing

  $$RD \sim \alpha + \alpha | L$$

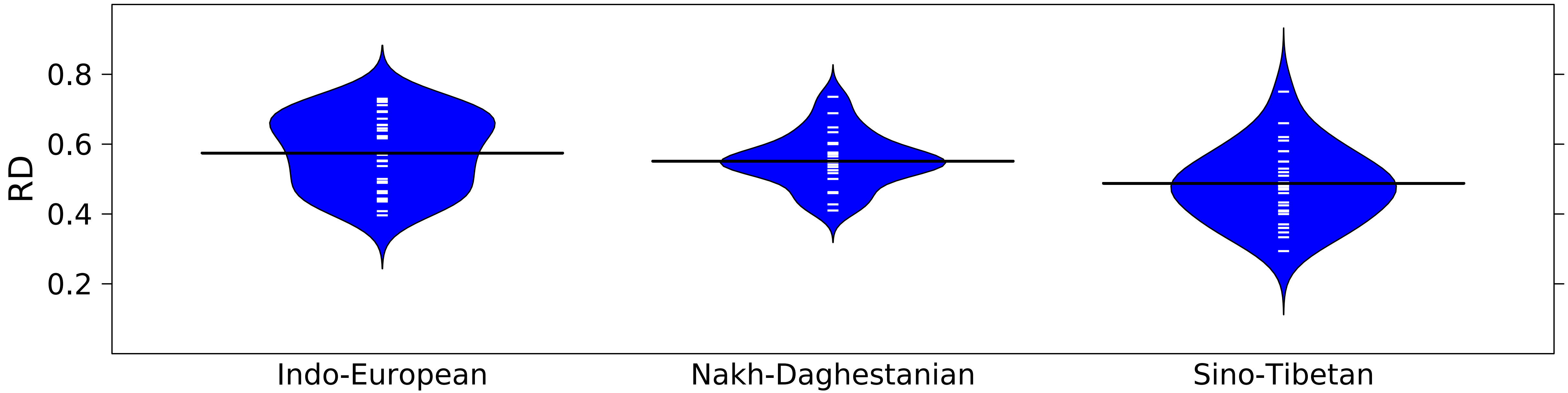
  $$RD \sim \alpha$$

Quirks in general

- Random factor language: *LR* = 57.82, $p_{(\chi^2)} < .001$, $R^2 = .63$



- Random factor family: *LR* = 13.46, $p_{(\chi^2)} < .001$, $R^2 = .24$



▸ **But they obviously don't explain anything...**

Beyond per-unit aggregation

- Speech samples and constructions as basic datapoints, on which we can directly model possible effects:

RD	gender	length	agreement syntax	social.network	language	stock
0.55	f	57	case-based	loose	Chechen diaspora	Indo-European
0.58	f	58	case-based	close	Ingush	Nakh-Daghestanian
0.62	f	88	case-based	close	Maithili	Indo-European
0.49	f	39	case-based	close	Nepali	Indo-European
0.36	f	47	other	close	Kyirong Tibetan	Sino-Tibetan
0.57	f	47	case-based	close	Maithili	Indo-European
0.69	f	92	case-based	close	Ingush	Nakh-Daghestanian
0.56	m	119	case-based	loose	Chechen diaspora	Indo-European
0.61	f	57	case-based	loose	Chechen diaspora	Indo-European
0.55	f	69	other	loose	Xiang Chinese	Sino-Tibetan

1. Sociology of communication: close-knit vs. loose

- Common observation in the Ethnography of Speaking: people who know each other (‘close-knit society’) tend to presuppose more information than strangers.
- This habituates them into presupposing knowledge even when talking about the unknown, as in the Pear Story experiment.
- Predictions :
  - close-knit ➜ low RD
  - loose ➜ high RD
- Coding on individual level, based on the relationship to the listener in the Pear Story experiments

Possibly relevant factors

2. **Some structural property of grammar:** case-based agreement requires NP information, and this primes activation of NP structures in production (Bickel 2003)

Case-based agreement in Maithili (IE)

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?’

Non-case-based agreement in Belhare (ST)

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?’

Possibly relevant factors

A construction primes a structurally associated construction (e.g. V-agr → NP; cf. Lu et al 2001 for parallels).

↓

This construction primes its subsequent re-use (e.g. NP → NP; Bock 1986 etc.)

↓

Long-term persistence and habituation effects (cf. Bock & Griffin 2000)

Possibly relevant factors

Other suspects:

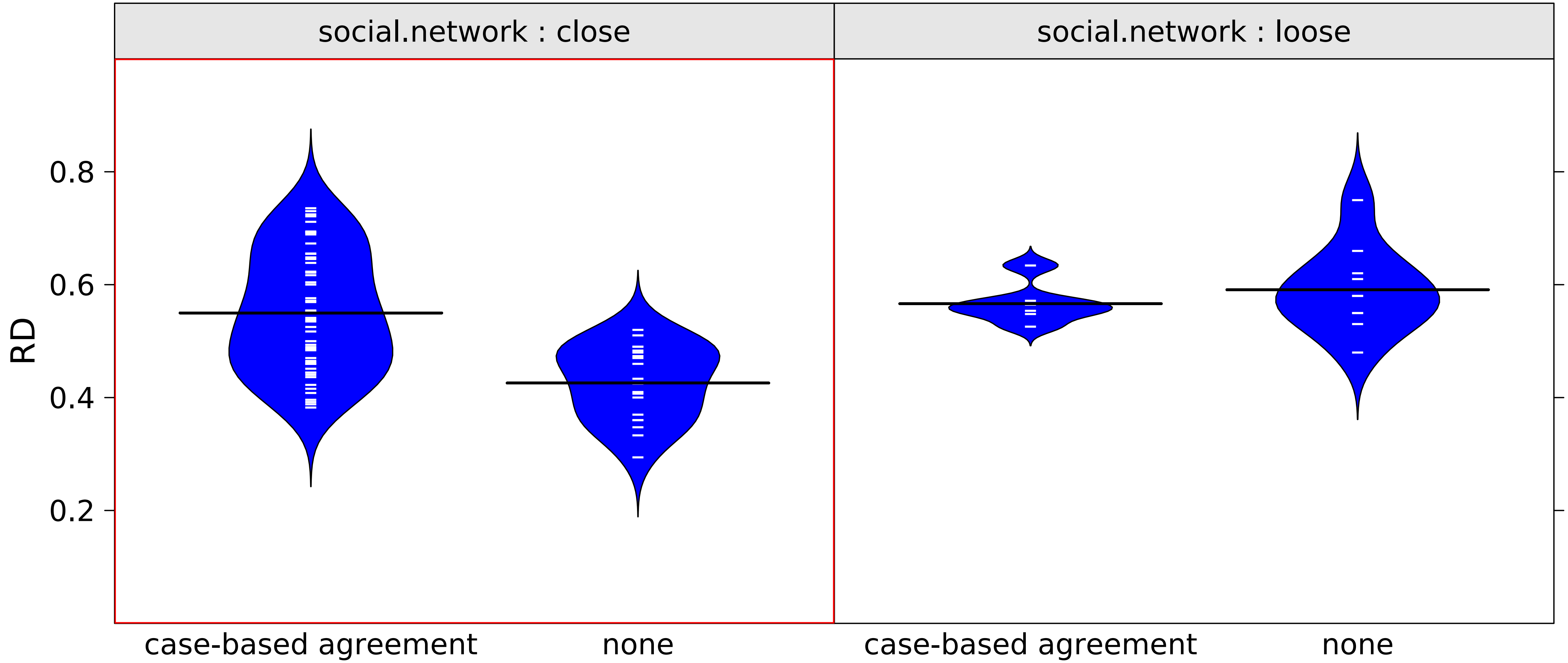
- Text length: talkative vs. non-talkative narrators
- Gender: marginal but unexplained effect noted in Bickel 2003 and again in Seifart et al. 2010

Modeling

- $\mu(\mathrm{RD}) = \alpha + \beta_1\mathrm{SOC}+\beta_2\mathrm{SYN}+\beta_3\mathrm{LENGTH}+\beta_4\mathrm{GENDER}$ ...
- No evidence for any interaction
- expect for SOC x SYN, $F = 7.30$, $p = .008$
- where high RD values of each factor blur the effects of the other factor

Factorial analysis

- Syntax effect only in the absence of social network effect



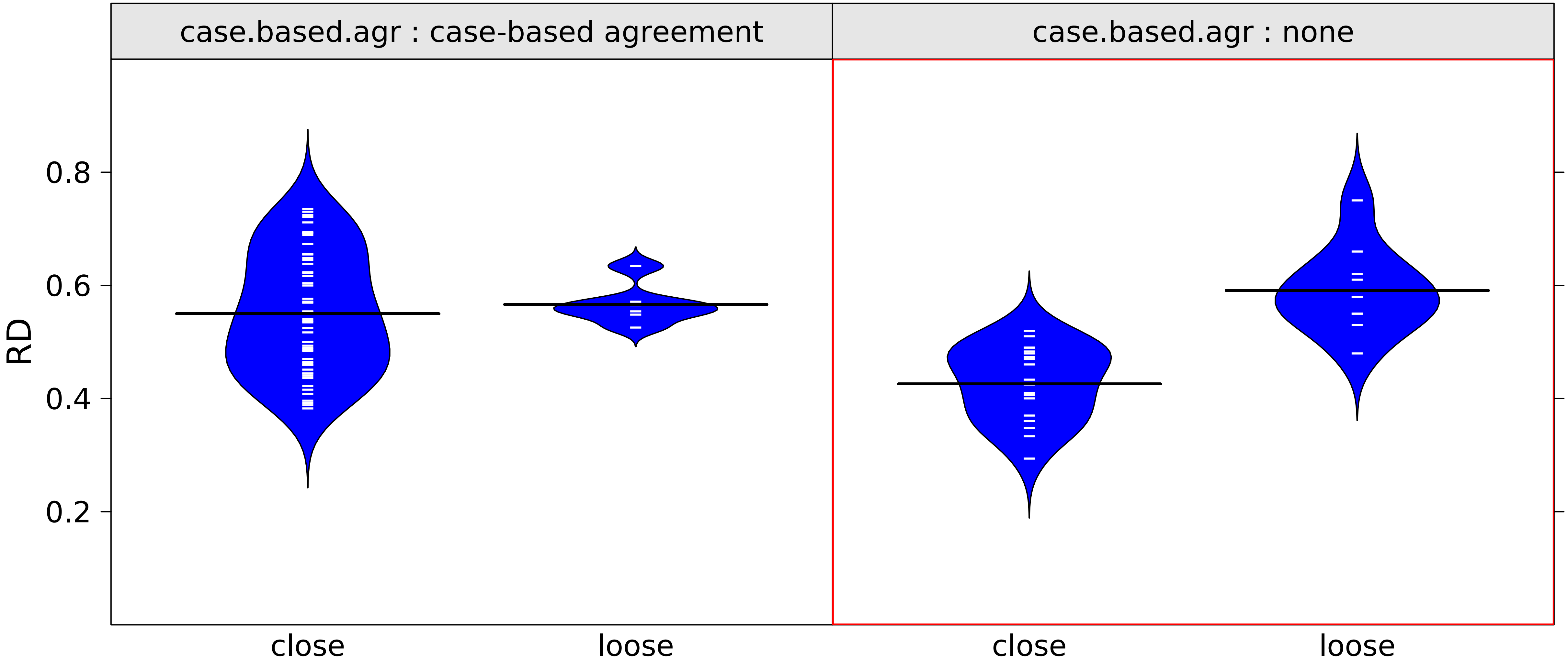
All further interactions, gender and length effects $p > .1$

Syntax: $F(1,73) = 24.44$, $p < .001$; $R^2 = .25$

no significant effects at all

Factorial analysis

- Social network effect only in the absence of syntax effect:



no significant effects at all

All further interactions, gender and length effects $p > .1$

Social network: $F(1,73) = 38.63$, $p < .001$; $R^2 = .58$

Interim summary

- RD can be modeled by interacting effects of
  1. syntactic practice: habitual activation of NPs
  2. social network: habitual expectations about hearer knowledge
- This model explains less variance ($R^2 = .28$) than a model based on language ($R^2 = .63$), but the language model assumes **per-unit norms/trends** without any evidence
- except perhaps in Sino-Tibetan (and Nakh-Daghestanian)
    - more research needed on possibly historical norms in the Sino-Tibetan family
    - better control for areal diffusion of RD patterns in the Sino-Tibetan area

Discussion

- So, should we completely ignore genealogical units beyond their practical (i.e. library catalogue) use?
- No!
  - Genealogical units are defined as data sets in which all similarities and all dissimilarities must have arisen by maintaining or changing norms.
  - As such, they allow estimating diachronic biases in this.
  - If biases are systematic (universally or areally, conditionally or unconditionally), this demands explanation.
  - Therefore, the history of typological distributions can be examined by estimating biases within genealogical units (= the Family Bias Method: Bickel 2008, 2011)

Three ways in which linguistic distributions can be shaped

A. **Via biases**, i.e. through effects on language change or resistance against change:

- what is preferred by some individuals becomes the norm
- and results in a bias for an entire language and possibly any further groups that split off from it
- ▶ biases within genealogical units
- if biases are systematic, there might be principled effects
- *Examples:* any kind of trend in constructional choices, e.g. universal preference for A-before-P word order; areal preference for relative pronouns in Europe, etc.

Three ways in which linguistic distributions can be shaped

B. **Via habits:** no per-unit bias but individual linguistic patterns are selected by speakers' habits because of common effects:

- systematic habits yield systematic responses
- *Examples:*
  - habitual activation of NP information and habitual expectations systematically affect RD values, but no language-wide norm
  - habitual use of absolute vs. relative coordinate systems systematically affect nonlinguistic spatial cognition (Pederson et al. 1998, Levinson 2003)
    - no large-scale test of this, but tentative evidence from Pederson 1995:

Three ways in which linguistic distributions can be shaped

- Pederson 1995:
  - Two speech samples within the same unit (a variety of Tamil), differing only wrt spatial language
  - strong and sign. correlations with spatial cognition
  - but no community-wide (Tamil-wide) norm

Three ways in which linguistic distributions can be shaped

C. **Online:** no per-unit bias, nor habits, but linguistic patterns directly reflect some relevant principle of processing

- perhaps trends in MLUs and other chunking effects

Could the RD effects be online rather than mediated by habits?

Another look at RD effects

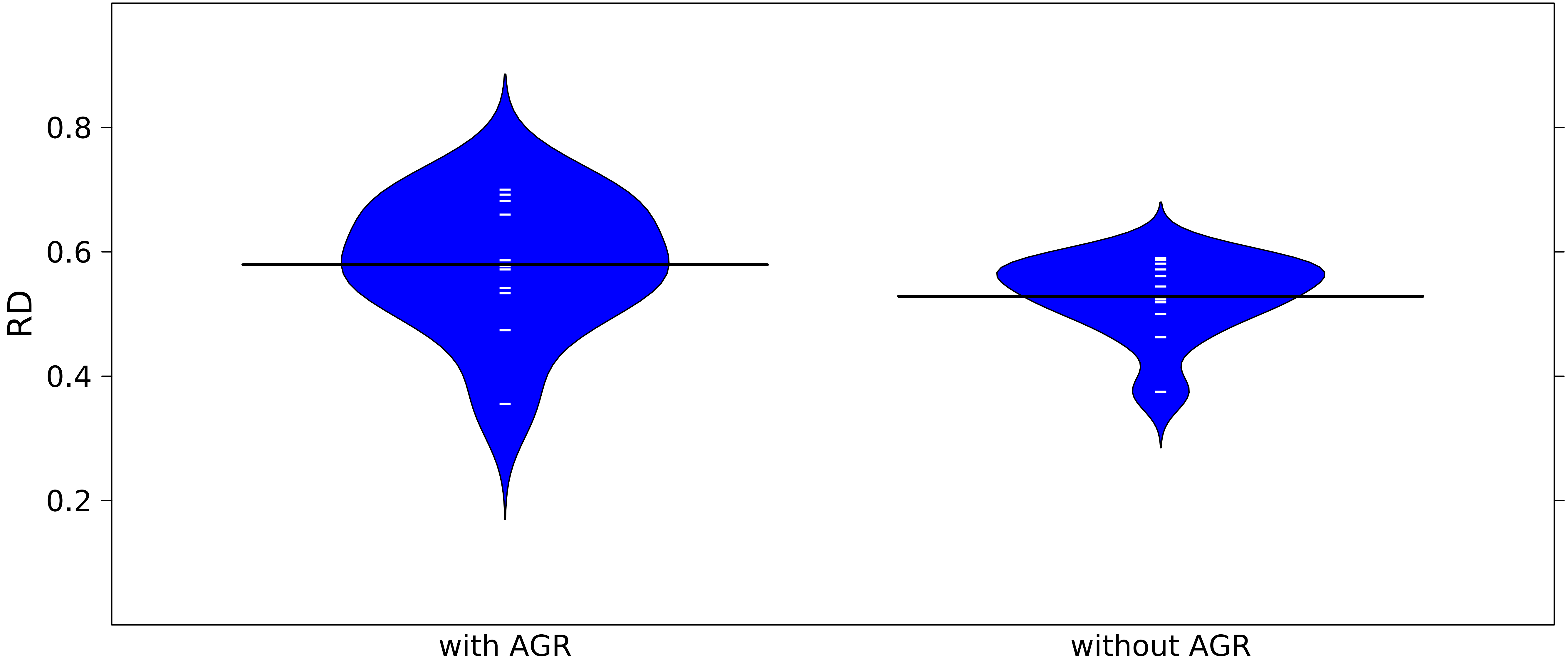
Test case in Chechen: some verbs show overt agreement, others don’t:

a.	*suuna*	*iz*	*v-eez-a.*
	1sDAT	3sNOM(V)	V-love-PRS
	‘I love him.’		
b.	*suuna*	*iz*	*go.*
	1sDAT	3sNOM(V)	see.PRS
	‘I see him.’		

Another look at RD effects

But no evidence for agreeing verbs triggering more overt NPs than non-agreeing verbs:

paired *t*-test, $t = -1.54$, $df = 10$, $p = .155$



Another look at RD effects

- So far now evidence for online effects (although clearly more data are needed to establish this.)
- Best-fitting model assumes habituation effects of both syntax and social network

Conclusions: negative

- Genealogical units are not explanatory factors and should not be modelled as such, e.g.
  - not as random factors in linear models
  - not as control strata in sampling (Dryer 1989)
- They may or may not be suitable units for data aggregation (depending on how the data are distributed within them)

  (e.g. probably not suitable in the case of RD)
- And they may hide insight into other factors (by blurring all possible effects)

Conclusions: positive

- Genealogical units define datasets in which we can estimate the presence of biases (norms) that may reflect systematic effects of some external factor (universally or areally) → key evidence for any such effect (Maslova 2000, Bickel 2008, 2011)
- But linguistic distributions can also be affected
  - via habits: RD affected by syntactic and social habits
  - online: possibly MLUs
- Linguistics needs to move beyond collecting or aggregating statements per genealogical unit

Acknowledgments

- Mary Erbaugh: Xiang pear stories
- Brigitte Huber: Kyirong pear stories
- Lan Li: Xiang pear stories
- Yan Luo: Xiang pear stories
- Zarina Molochieva: Chechen and Ingush pear stories
- Johanna Nichols: Chechen and Ingush pear stories
- Sabine Stoll: Russian pear stories
- Alena Witzlack-Makarevich: Nama pear stories