Contact vs. inheritance vs selection in northern Eurasia:
What corpus work can contribute

Johanna Nichols, UC Berkeley and HSE and U. Helsinki
johanna@berkeley.edu

Goals

Problem: Using corpus data to bigger interesting ends
Proposed solution: Do it!

Use comparative corpus frequency data to:

- determine probabilities of various changes
- rates of change for a number of typological variables
- how those respond to sociolinguistic and typological context
- evidence of previous language contacts and shifts

And propose further applications of Lab corpus and survey data and expertise that can address further questions of great interest

Big questions, big-data solutions

Introduction: Large-scale areality in northern Eurasia

Multivariate (composite, complex) variables and what they reveal

Gradual, incremental differences and similarities

(Contrast: gross, saltatory differences with classic variables)

Kinds of geolinguistic distribution

North Pacific Rim: Peak in NE Asia to W North America

High northern: Cline with peak in far west (Europe) or far east (E. N. America) of the northern hemisphere higher latitudes

Global: Cline along trajectory from Africa to Australia or South America

Questions:

What mechanism created these continuities?

Over what time frame?

Outline

Some typological variables that reveal large-scale contact patterns
and some additional promising ones

Corpus work needed
why, how, where

Risks, problems; technological needs

Inflectional person

Person markers behave like inflectional morphemes:

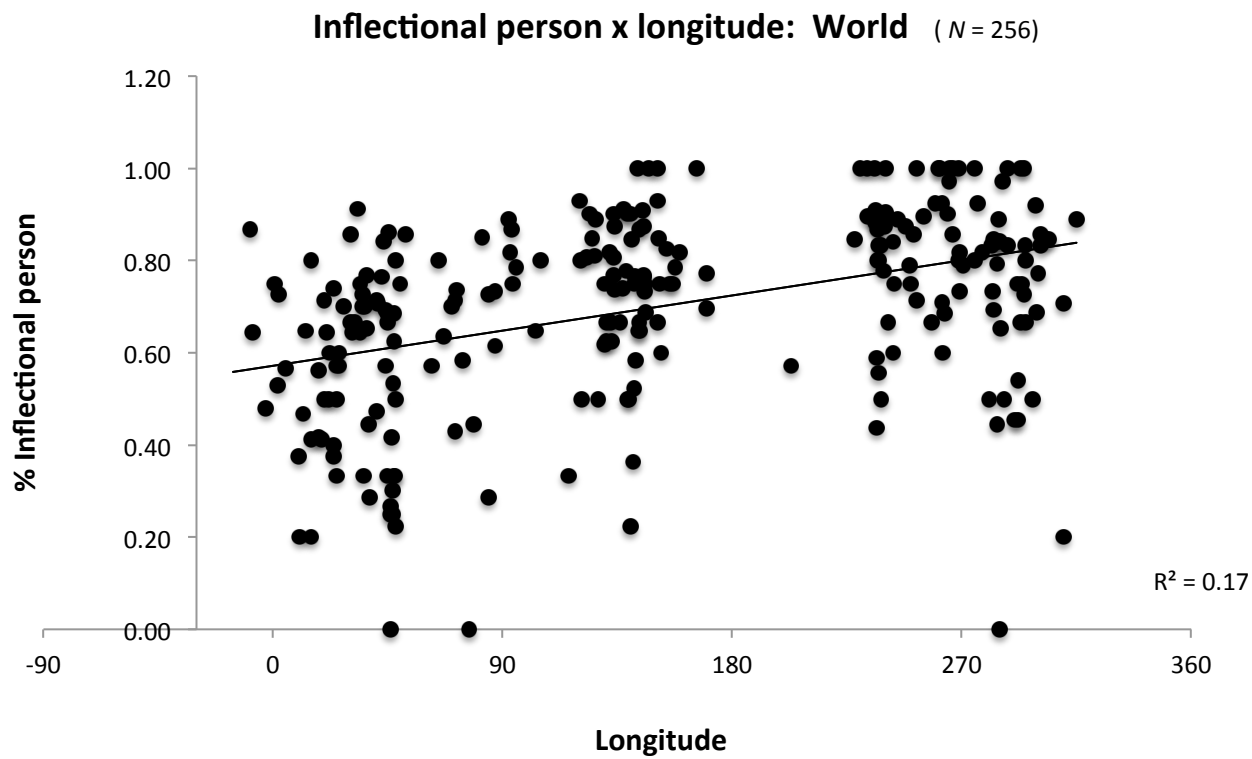
- Indexation on verb, possessed noun, etc.
- Set of person markers forms an inflectional paradigm
  - small, closed; phonosymbolically marked by rhyme, alliteration, etc.
- Formally distinct from number
- Multiple exponence

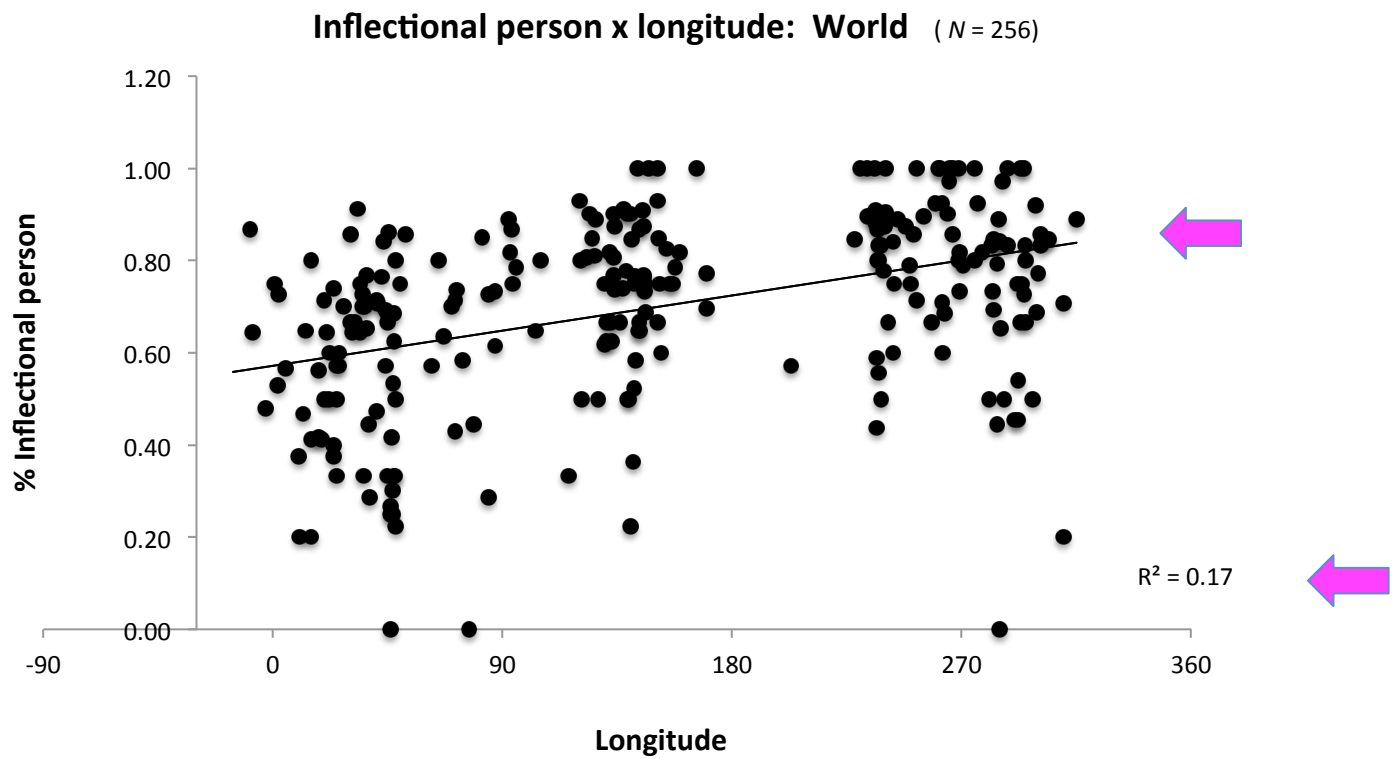
Person markers behave like lexical items:

- Same inflectional categories as nouns (e.g. case, number)
- Same inflectional classes as nouns (same endings, same ablaut)
- Portmanteau with gender, etc. (inherent noun categories)
- Open set

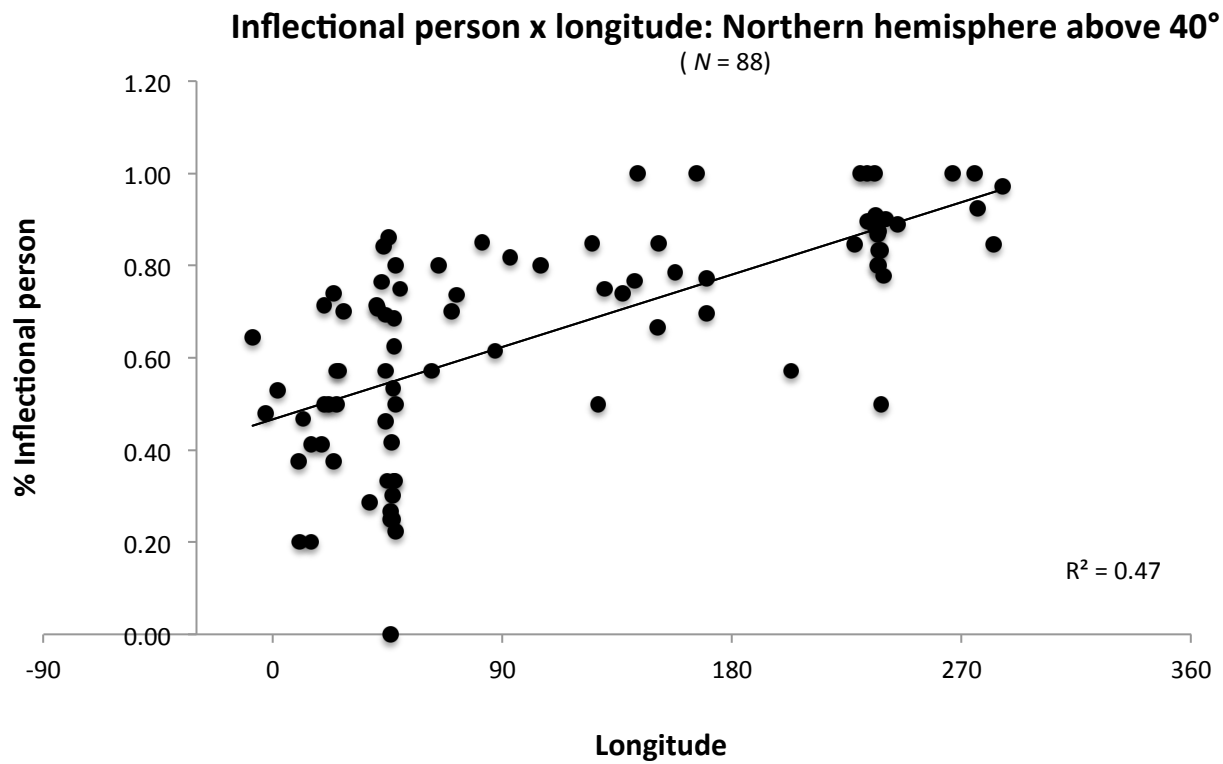
40-item questionnaire; typology is % of the items that behave inflectionally

Nichols 2017

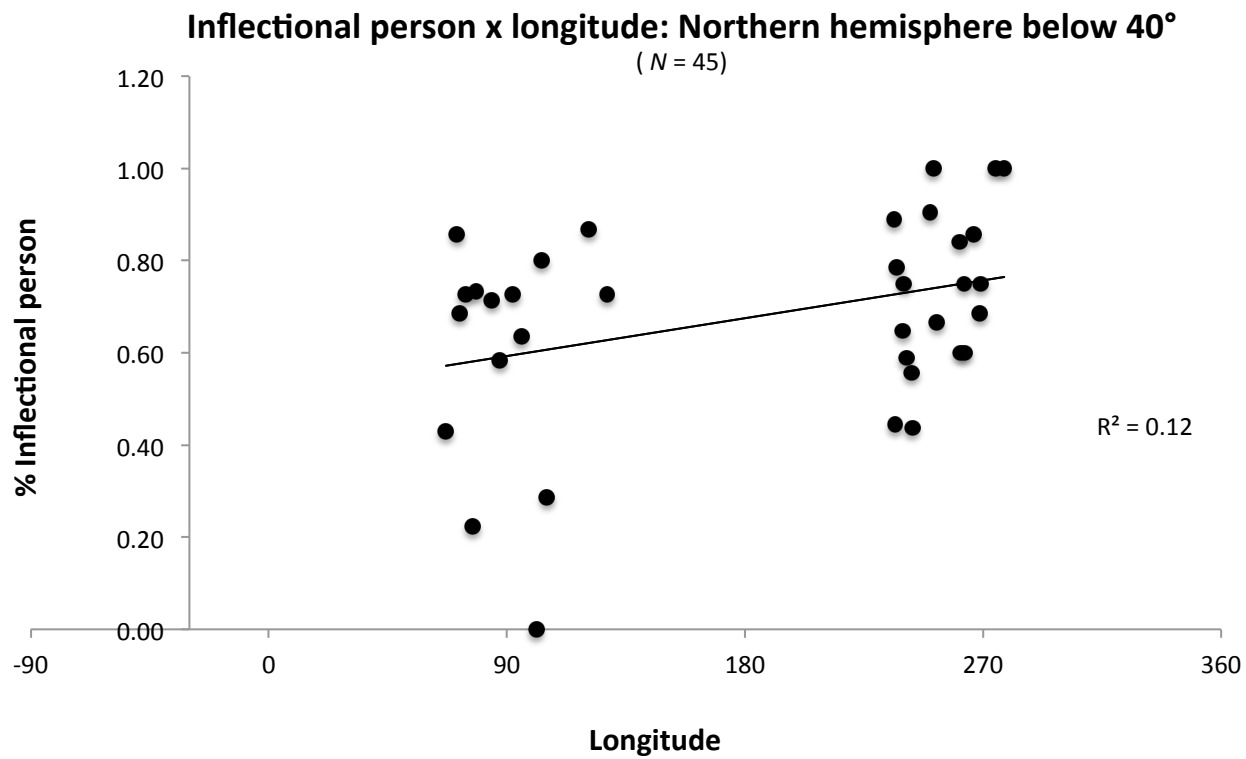

Inflectional person x longitude: World (N = 256)
% Inflectional person
1.20
1.00
0.80
0.60
0.40
0.20
0.00
-90
0
90
180
270
360
Longitude
R² = 0.17



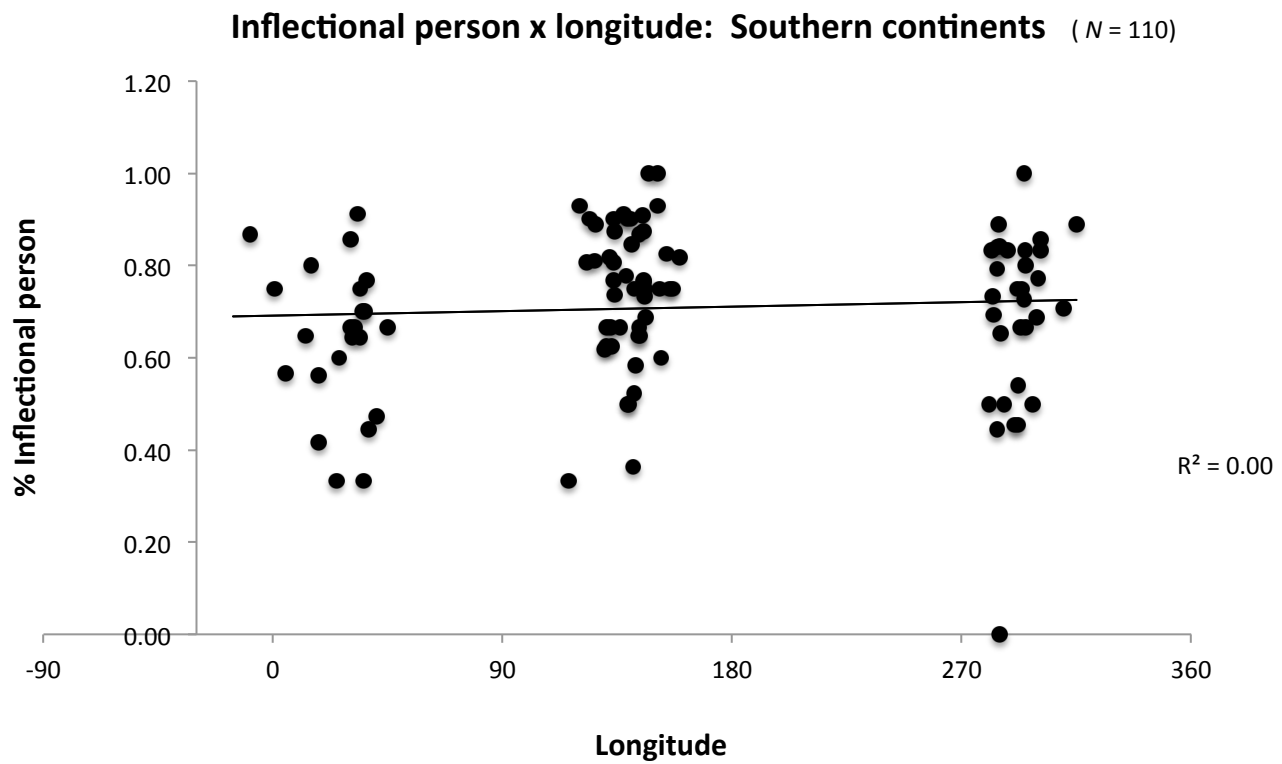
What to look for in these graphs:
Trendline slope. Is there a cline?
R2 number: Strength of correlation. 0-0.5 none 0.10 weak 0.15 moderate 0.20 strong
(Linear trendline. Correlation coefficient is 0.412.)



Corr. ceff. 0.684

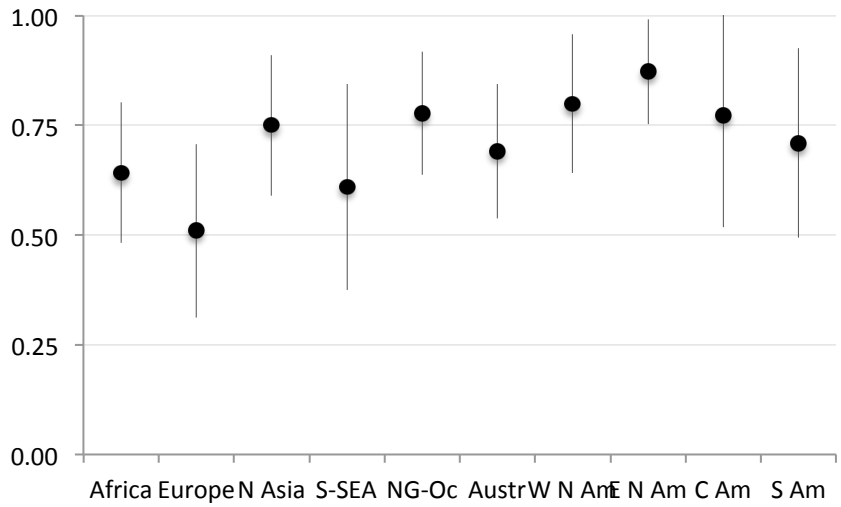


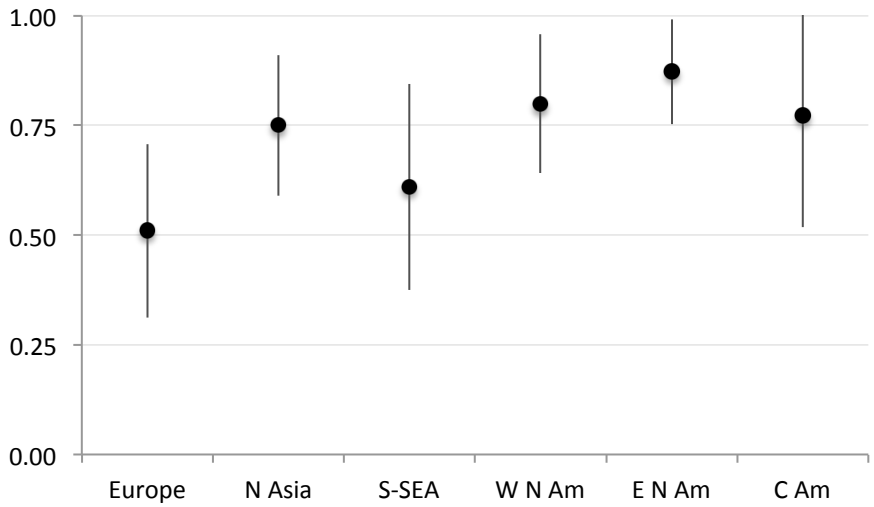
Corr. ceff. 0.377



Corr. ceff. 0.087

Inflectional person: Continental means (±1 sd).
Left, all continents (*N* = 255); right, northern hemisphere only.



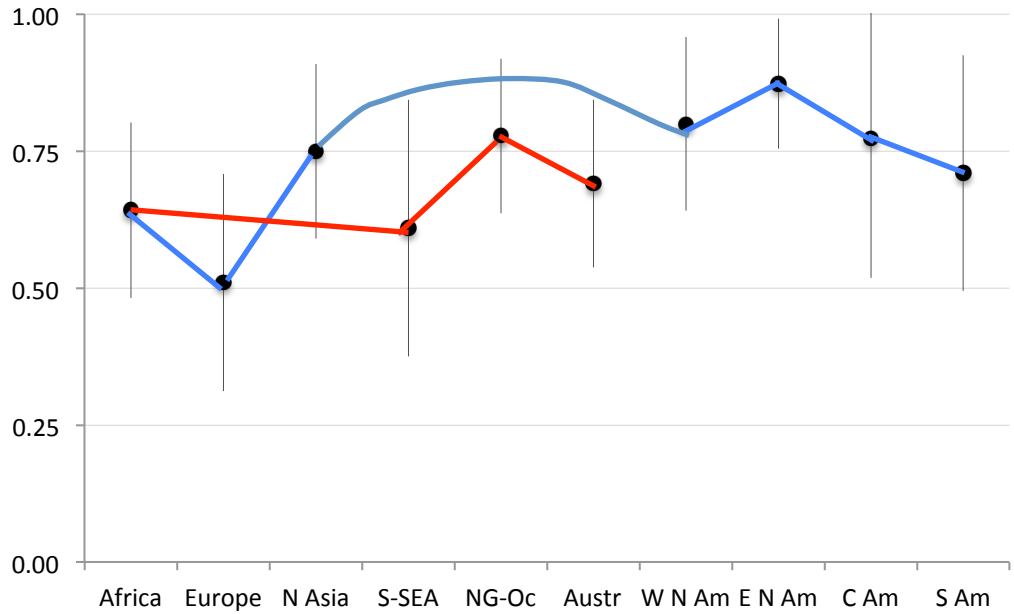


Distribution: High northern.

Continental breakdown: AUTOTYP (Nichols et al. 2013)

Inflectional person: Continental means (±1 sd).

Blue: Northern trajectory from Africa to the Americas.
Red: Southern trajectory from Africa to Australia.



Continental breakdown: AUTOTYP (Nichols et al. 2013)

The causative alternation

Pairs such as:

Non-causal	Causal
'fear'	'scare, frighten'
'be/get angry'	'anger, make angry'
'sit'	'seat'
'break'	'break (something)'

Constant semantics and argument structure; **derivations vary**.

Ingush	wa-xou down-sit	'sit down'	wa-xoa-**d.u** down-sit-CAUS	'seat, have sit'
Spanish	sentar-**se** seat-REFL		sentar seat	

The causative alternation

Pairs such as:

Non-causal	Causal
'fear'	'scare, frighten'
'be/get angry'	'anger, make angry'
'sit'	'seat'
'break'	'break (something)'

Constant semantics and argument structure; derivations vary. Base in box.

Ingush	wa-xou down-sit	'sit down'	wa-xoa-**d.u** down-sit-CAUS	'seat, have sit'
Spanish	sentar-**se** seat-REFL		sentar seat	

The causative alternation

(Relevant derivational morphology bold and blue)

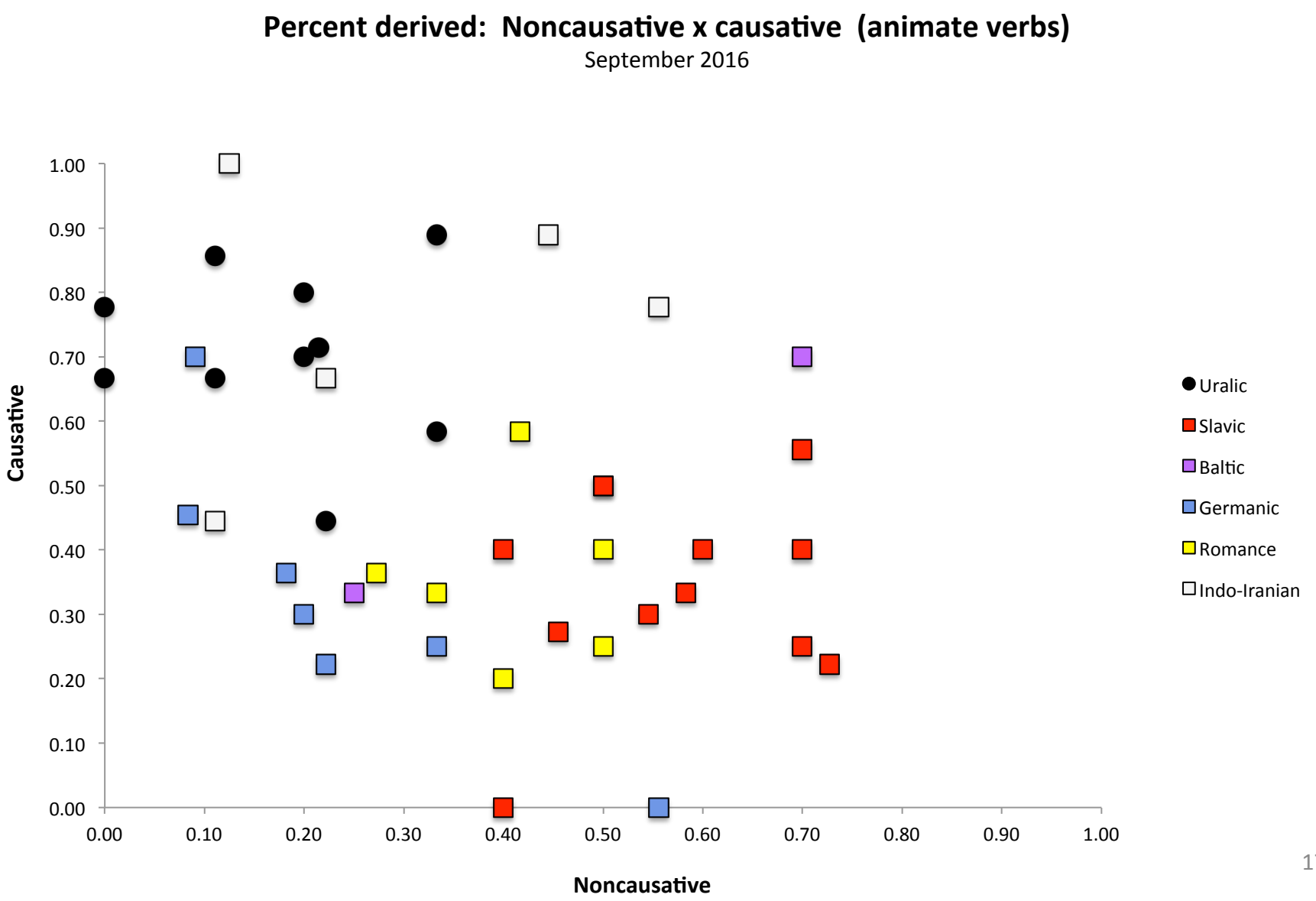
	Non-causal	Causal	
	'fear'	'frighten'	
Macedonian	**se** plaši	plaši	Decausative
Russian	bojat'-**sja**	pugat'	Decaus.; suppletion
Polish	bać **się**	przestra**sz-y-ć**	Double; suppletion
Estonian	hirmu-ma	hirmu-**ta**-ma	Causativization
English	fear; afraid	scare	Suppletion
	'learn'	'teach'	
Macedonian	nauči	nauči	Ambitransitive

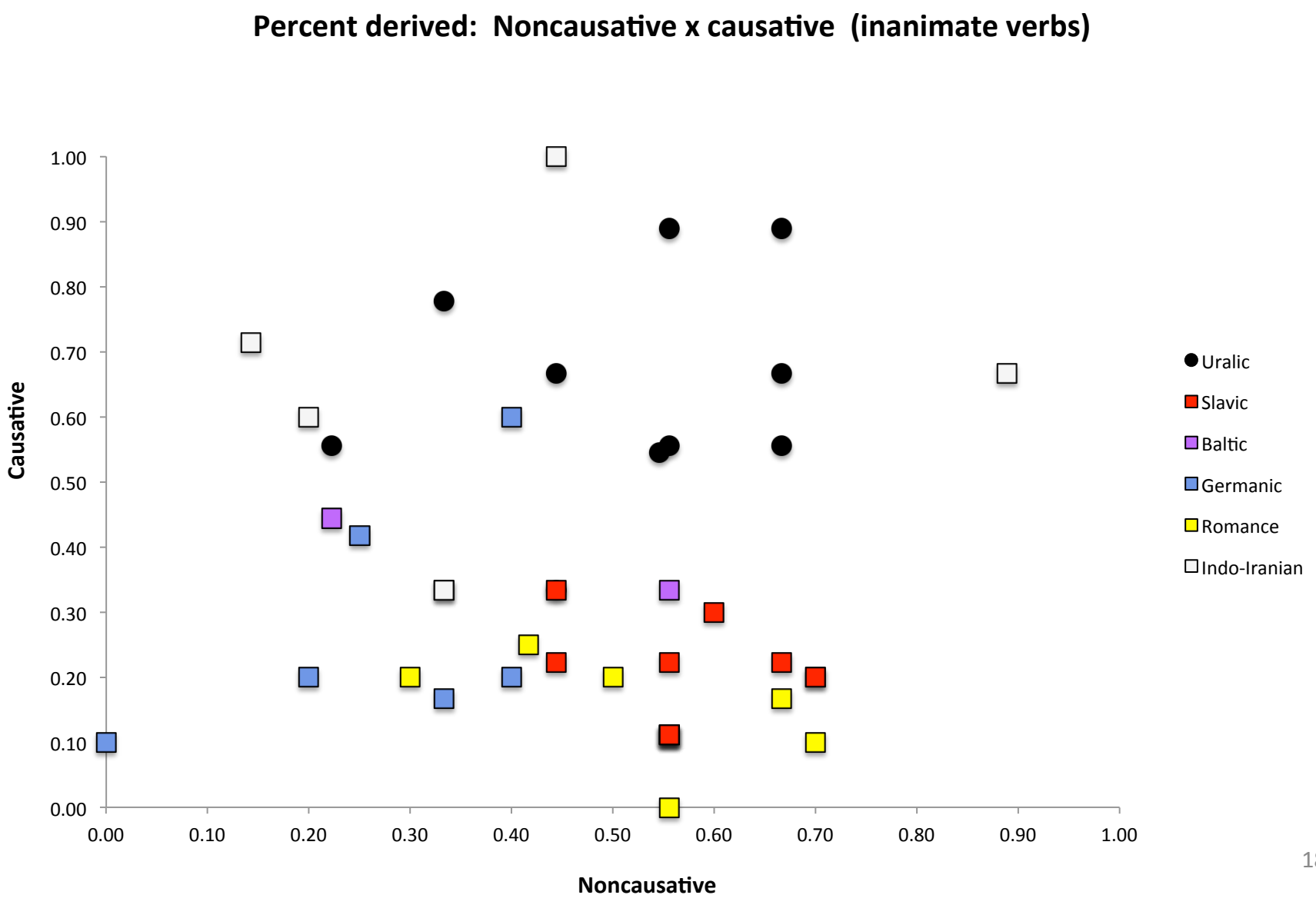
Avar and the Andic languages are very consistently causativizing languages. Some Avar pairs:

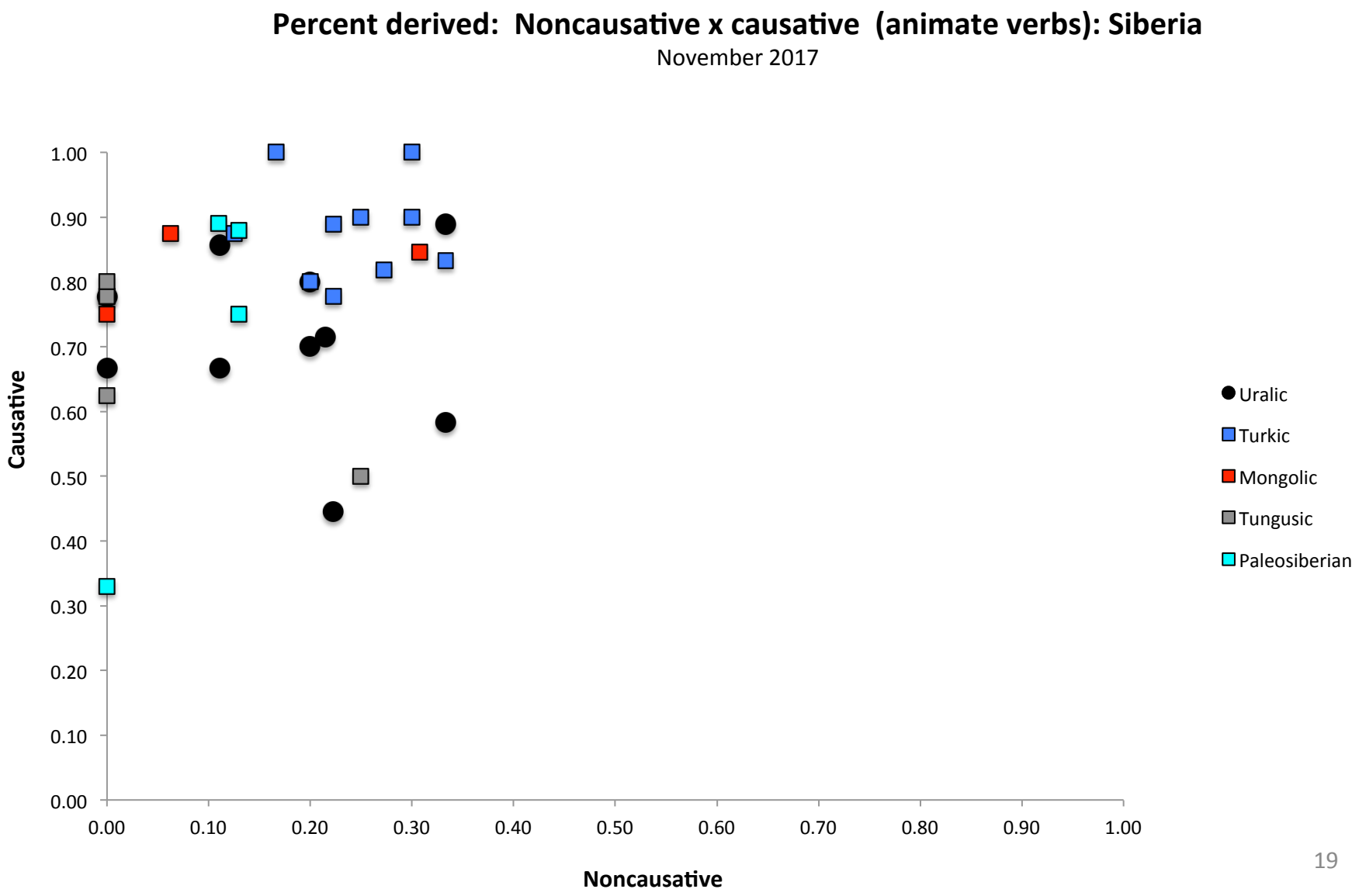
laugh, make laugh	belh-ize	belh-iz-**ab**-ize
sit, seat	k'us-ize	k'us-iz-**ab**-ize
eat, feed	k'wan-aze	k'wan-az-**ab**-ize
see, show	bix-ize	bix-iz-**ab**-ize
get angry, anger	ccin+daxx-ine	ccin+daxx-in-**ab**-ize
fear, frighten	h^{c}inq'-ize	h^{c}inq'-iz-**ab**-ize
hide	baxch-ize	baxch-ize; baxch-iz-**ab**-ize

etc.

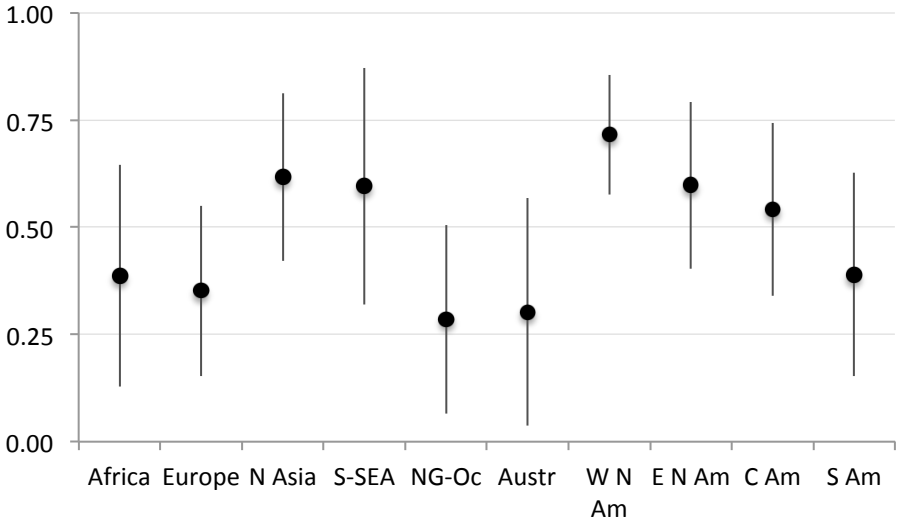
Creissels 2014

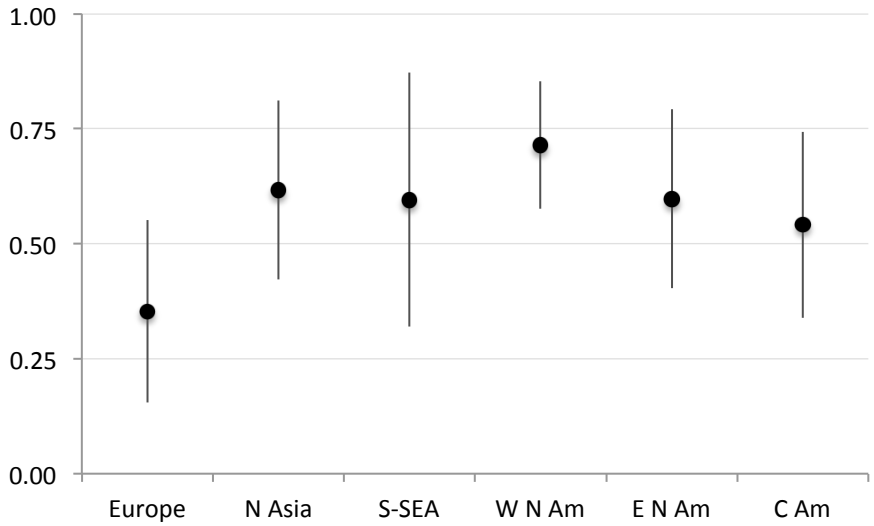

Percent derived: Noncausative x causative (animate verbs)
September 2016
Causative
Noncausative
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Uralic
Slavic
Baltic
Germanic
Romance
Indo-Iranian


Percent derived: Noncausative x causative (inanimate verbs)
Causative
Noncausative
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Uralic
Slavic
Baltic
Germanic
Romance
Indo-Iranian


Percent derived: Noncausative x causative (animate verbs): Siberia
November 2017
Causative
Noncausative
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Uralic
Turkic
Mongolic
Tungusic
Paleosiberian

Causative alternation: Continental mean percent causativized.
Left: all continents; right: N. hemisphere. (Bottom: 2nd order polynomial trendline.)
Distribution: N Pacific Rim.

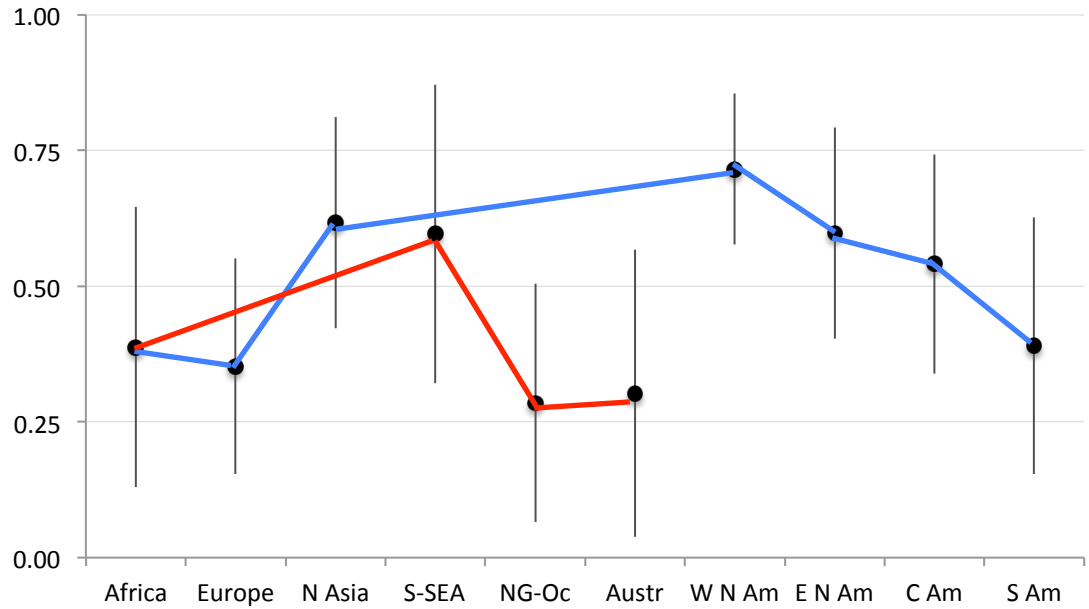




Distribution: Fortescue 1998

Causative alternation: Continental mean percent causativized.

Blue: Northern trajectory from Africa to the Americas.
Red: Southern trajectory from Africa to Australia.



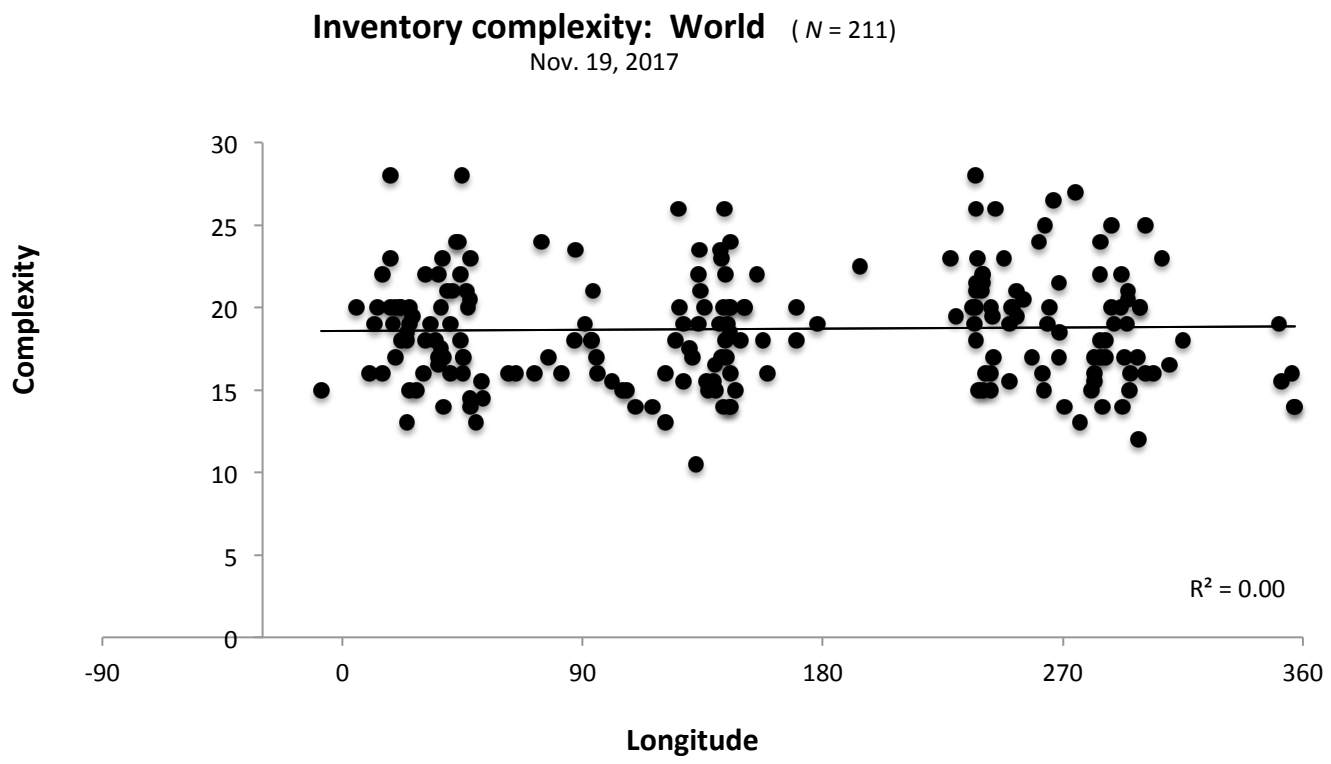
Distribution: Fortescue 1998

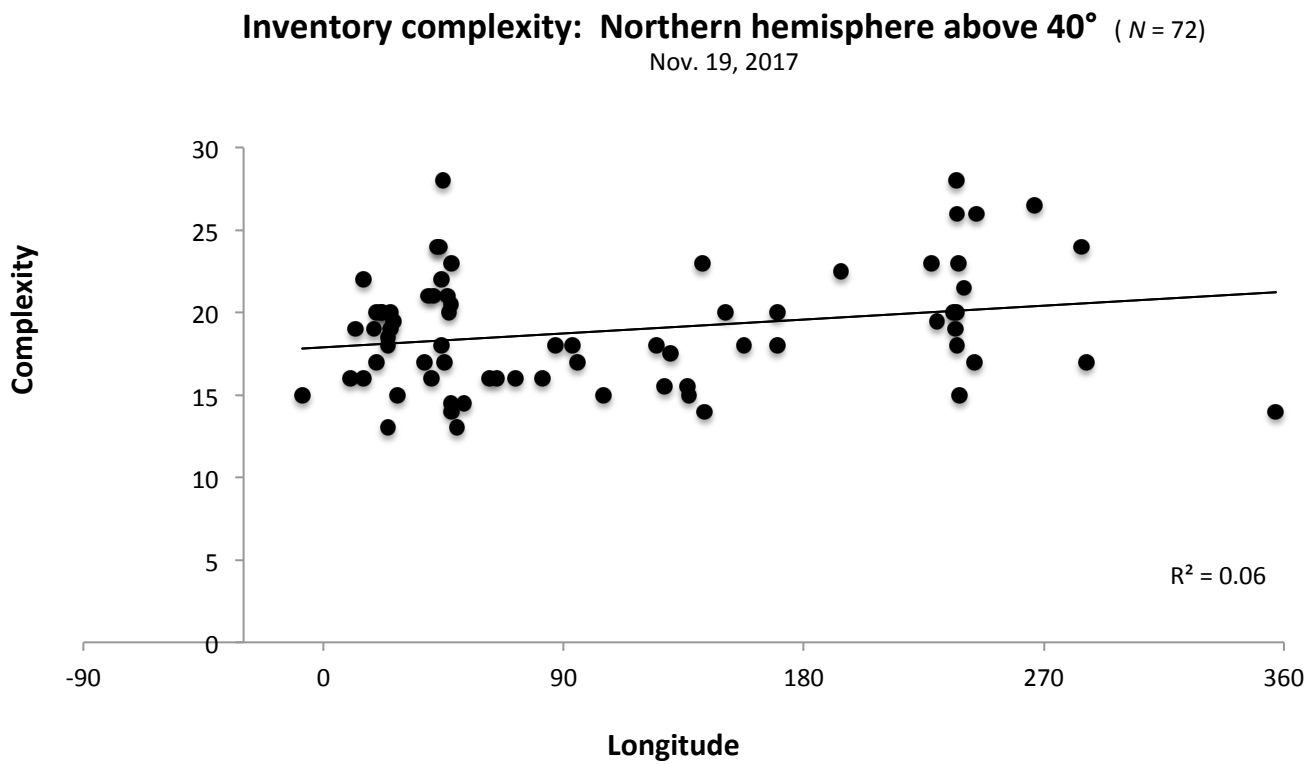
Inventory complexity

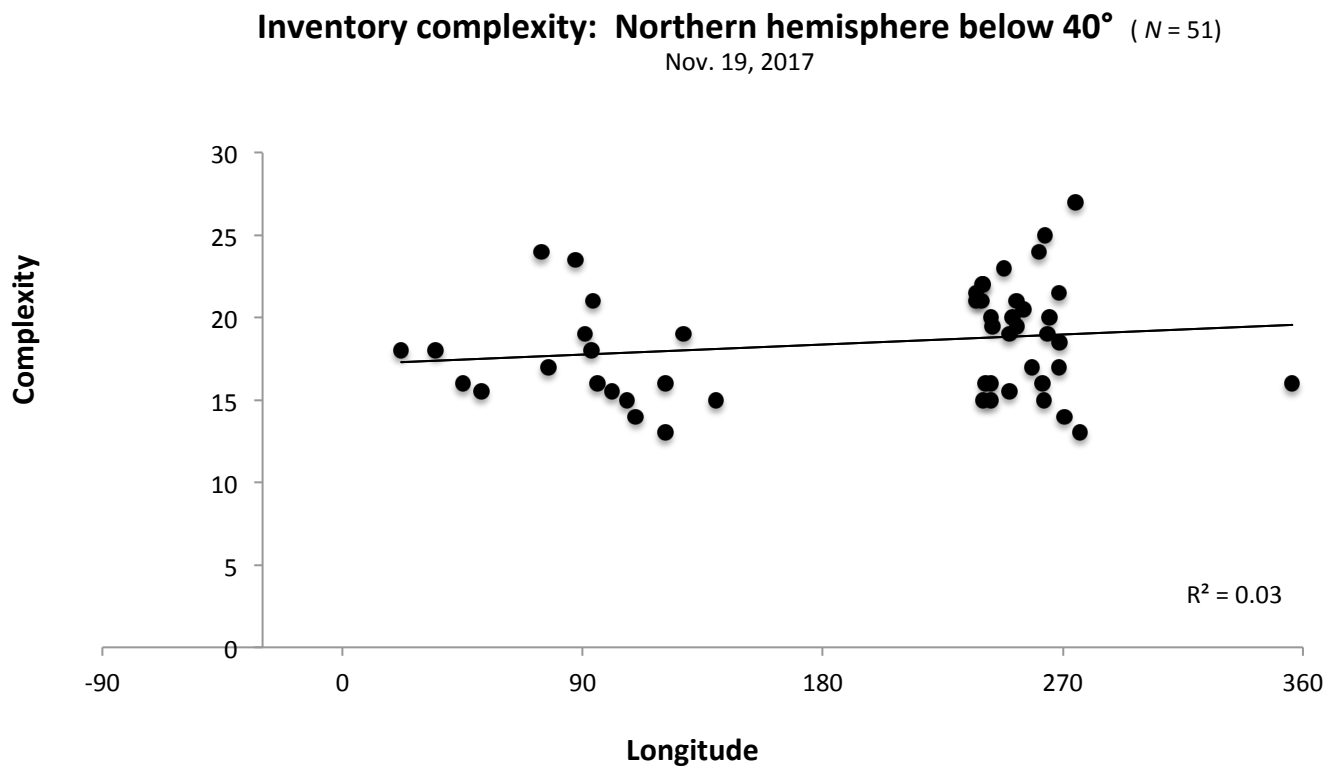
The number of elements in a system, for such systems as:

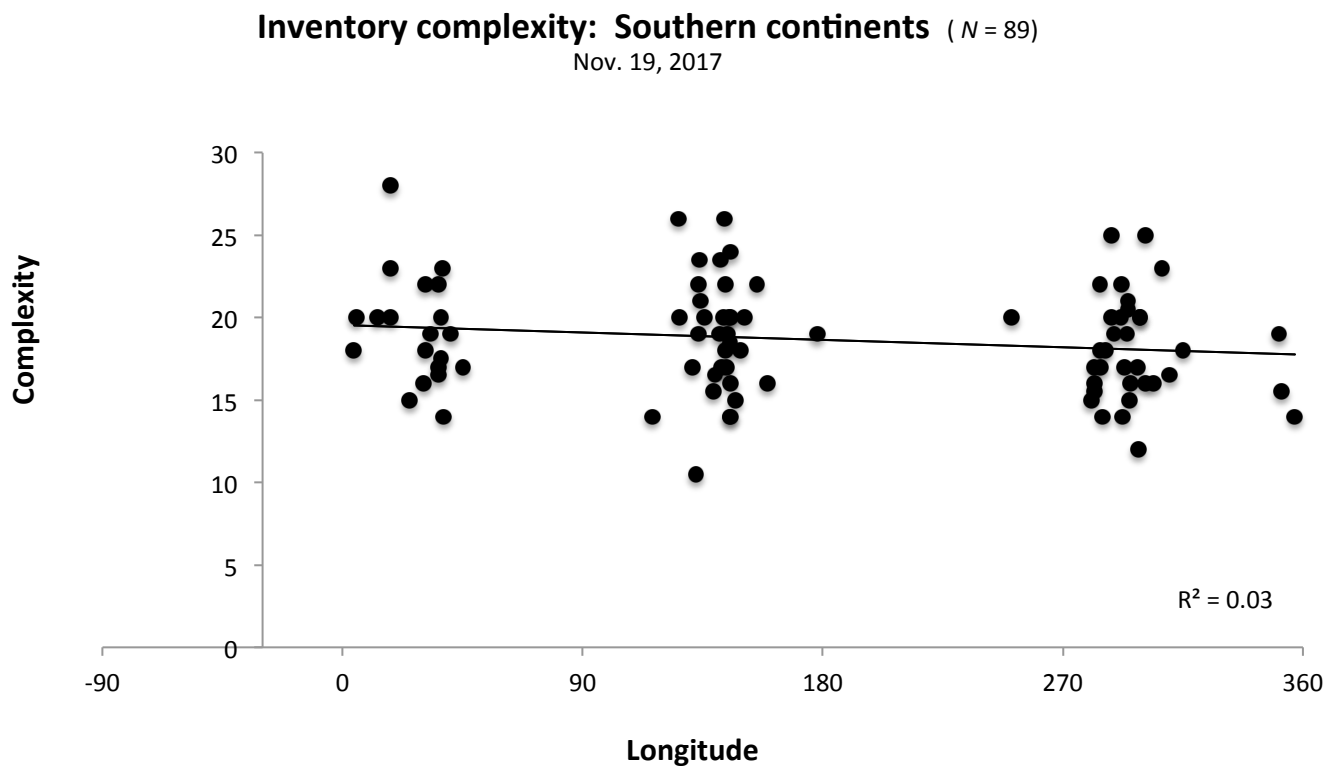
Phonemes. Number of vowel qualities, consonant manners
Phonation types, tones, etc.
Syllable structure: Maximum number of consonants per syllable
Verb inflectional synthesis (number of inflectional categories)
Noun inflectional synthesis (" " " ")
Classes. Number of genders, of numeral classifiers, of possessive classes
Number of basic or default alignments
Number of basic or default word orders

Typology:
Number of elements per trait out of a set questionnaire
Each trait can be binned into high/medium/low

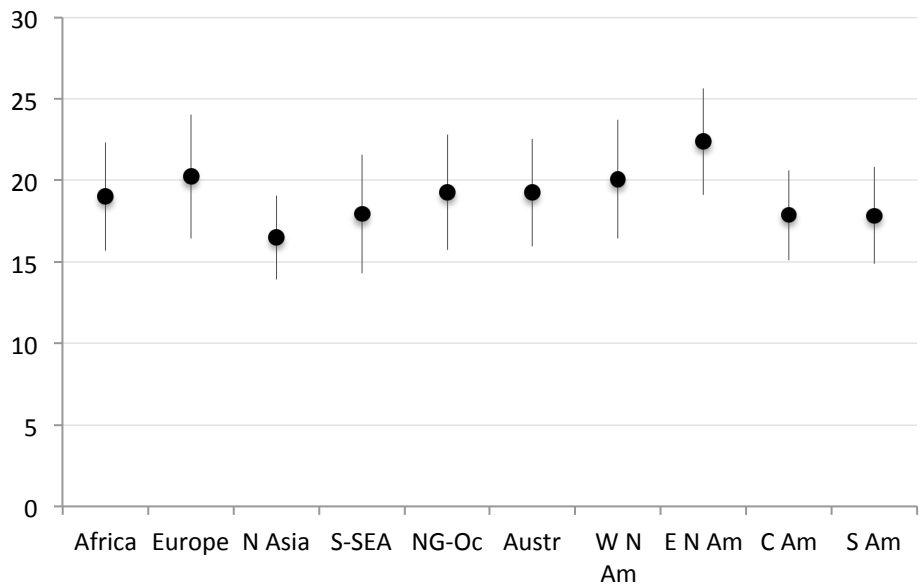

Inventory complexity: World (N = 211)
Nov. 19, 2017
Complexity
30
25
20
15
10
5
0
-90
0
90
180
270
360
Longitude
R² = 0.00

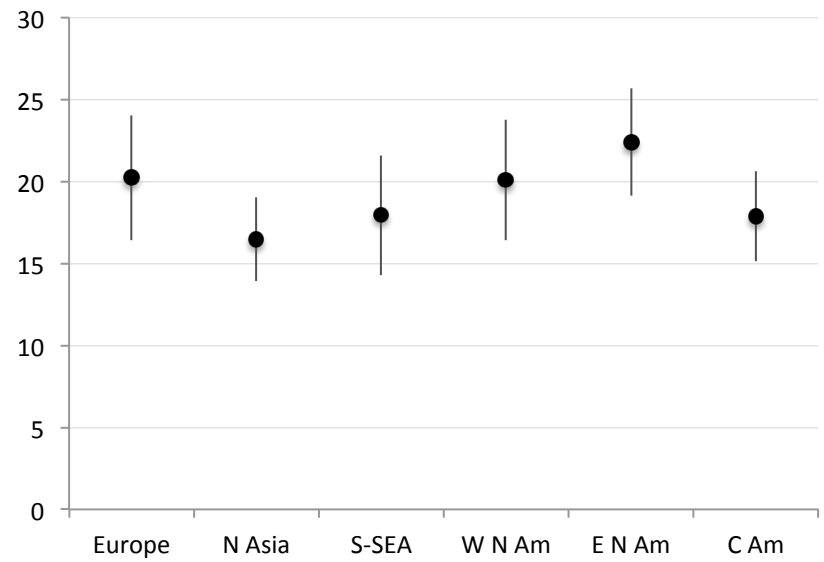

Inventory complexity: Northern hemisphere above 40° (N = 72)
Nov. 19, 2017
Complexity
30
25
20
15
10
5
0
-90
0
90
180
270
360
Longitude
R² = 0.06


Inventory complexity: Northern hemisphere below 40° (N = 51)
Nov. 19, 2017
Complexity
30
25
20
15
10
5
0
-90
0
90
180
270
360
Longitude
R² = 0.03


Inventory complexity: Southern continents (N = 89)
Nov. 19, 2017
Complexity
30
25
20
15
10
5
0
-90
0
90
180
270
360
Longitude
R² = 0.03

Inventory complexity: Continental means

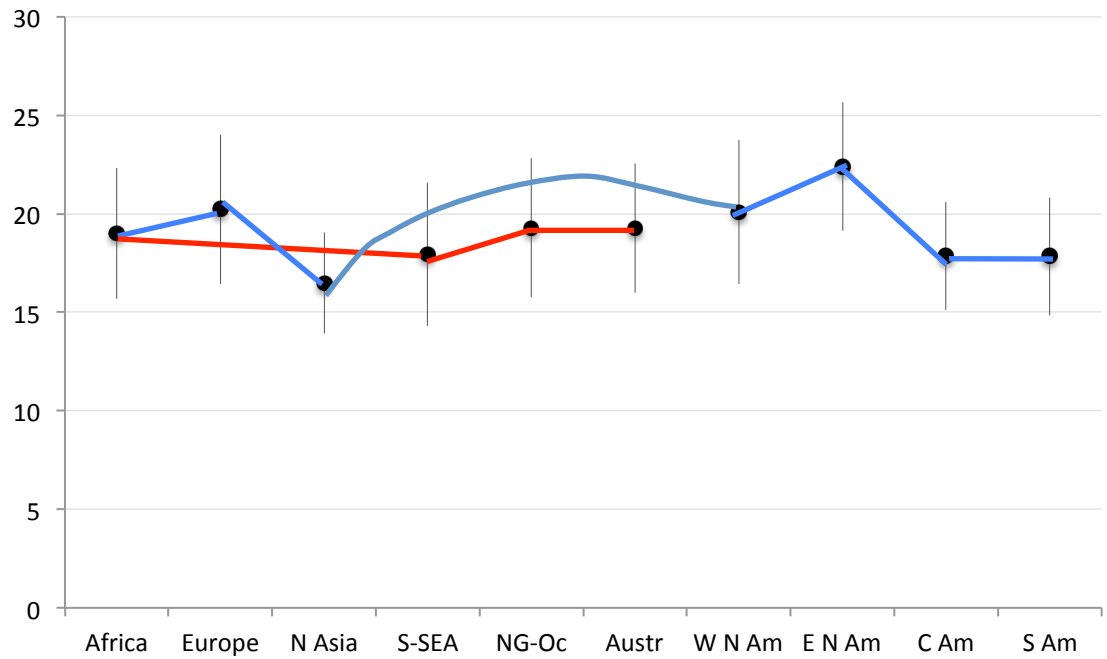




Distribution: **None, or easterly peak of N. Pacific Rim.**
Complexity is sensitive to sociolinguistic factors, which are very local.

Inventory complexity: Continental means

Blue: Northern trajectory from Africa to the Americas.
Red: Southern trajectory from Africa to Australia.



Distribution: **None, or easterly peak of N. Pacific Rim.**
Complexity is sensitive to sociolinguistic factors, which are very local.

Inventory complexity: Findings, questions

Affected by sociolinguistics:

isolation (Trudgill 2011, Nichols & Bentz in press); symbiosis (Nichols 2017)

– this is a much stronger factor than geography

High overall in the Americas (Nichols 2009)

Still debated: Affected by population size?

Base part of speech

Base of derivational paradigm				IE root
Verb:	Russ.	bojat'sja	boj-a-	verb
		smejat'sja	*směj-a- / smьj-	verb
Noun:	Pol.	przestraszyć	*strax-i-	verb
	Russ.	smešit'	*směx-i-	verb

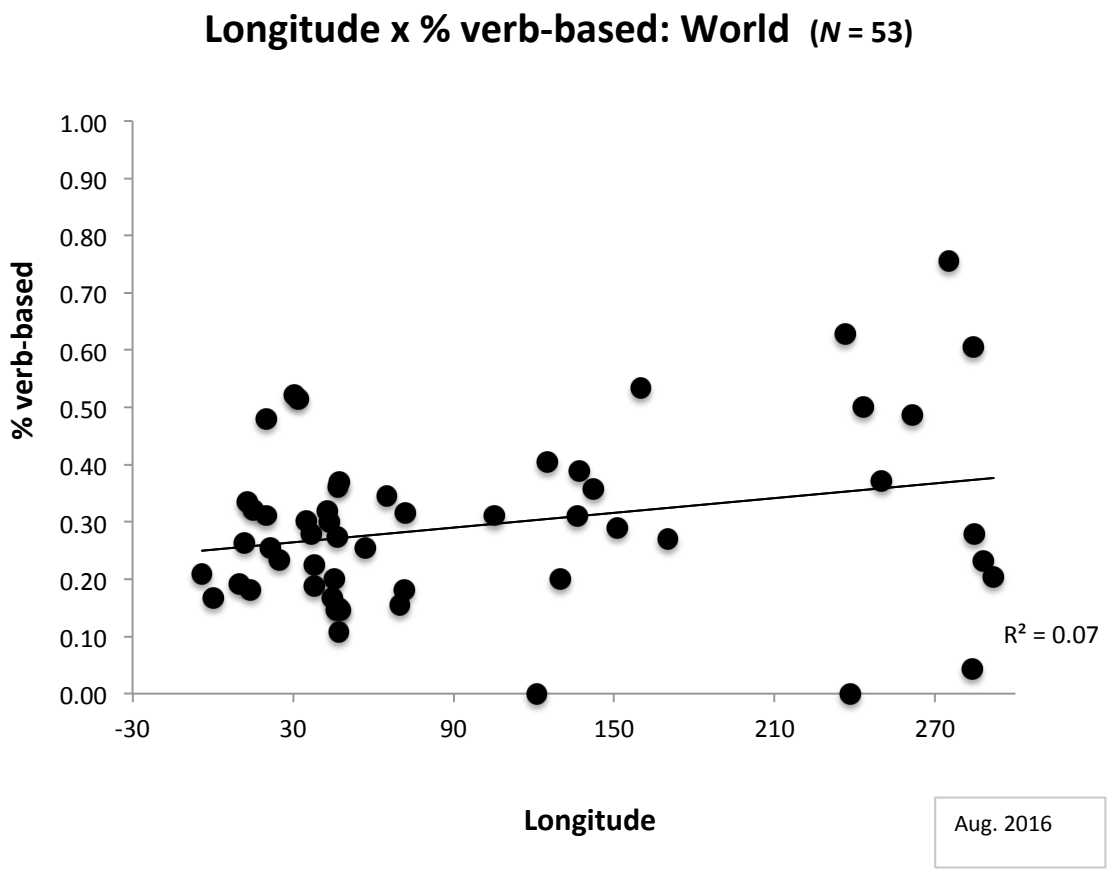
(Derivational paradigm: Based on synchronic morphology and morphophonology.)

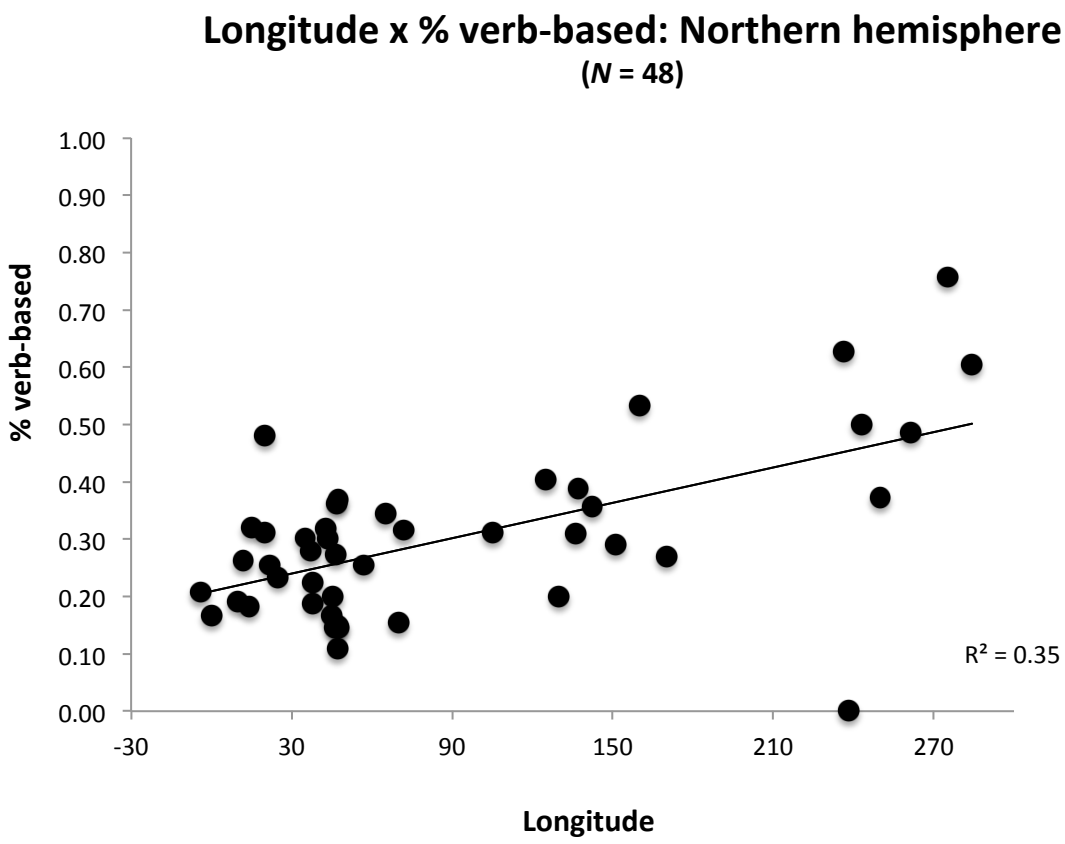
One or another base dominates in many languages.

Noun-based: Tajik, Nakh-Daghestanian languages
Verb-based: PIE?, many North American languages
Flexible: Tagalog, Kharia; English

50-item wordlist. For each item, find derivational paradigm and base.

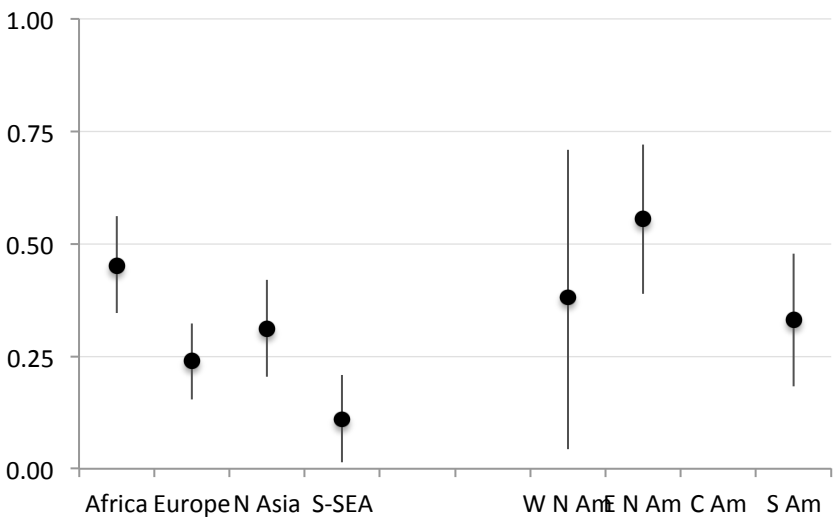
Nichols 2016, Foley in prep.

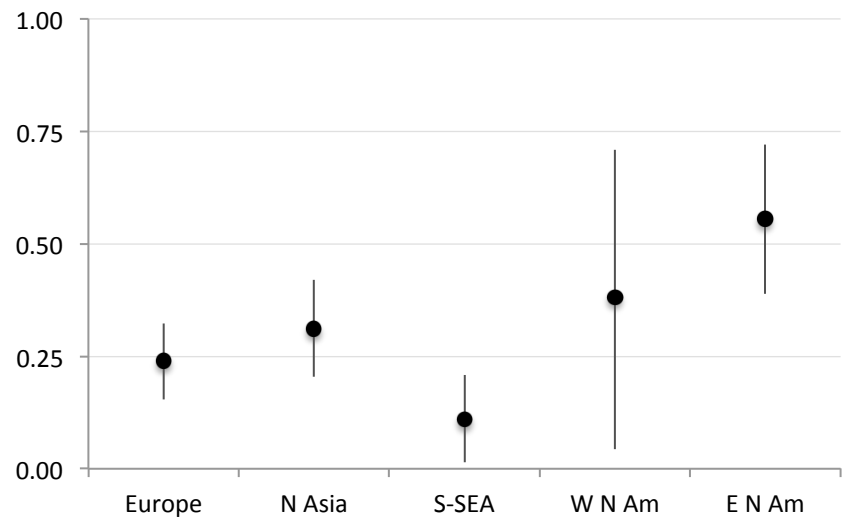

Longitude x % verb-based: World (N = 53)
% verb-based
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-30
30
90
150
210
270
R² = 0.07
Longitude
Aug. 2016



N. Eurasia alone:
$R^2 = 0.16$;
trendline identical.

% Verb-based: Continental means (±1 sd).
Left, all continents; right, northern hemisphere only.
(No data yet from Australia-New Guinea or Mexico-Central America.)

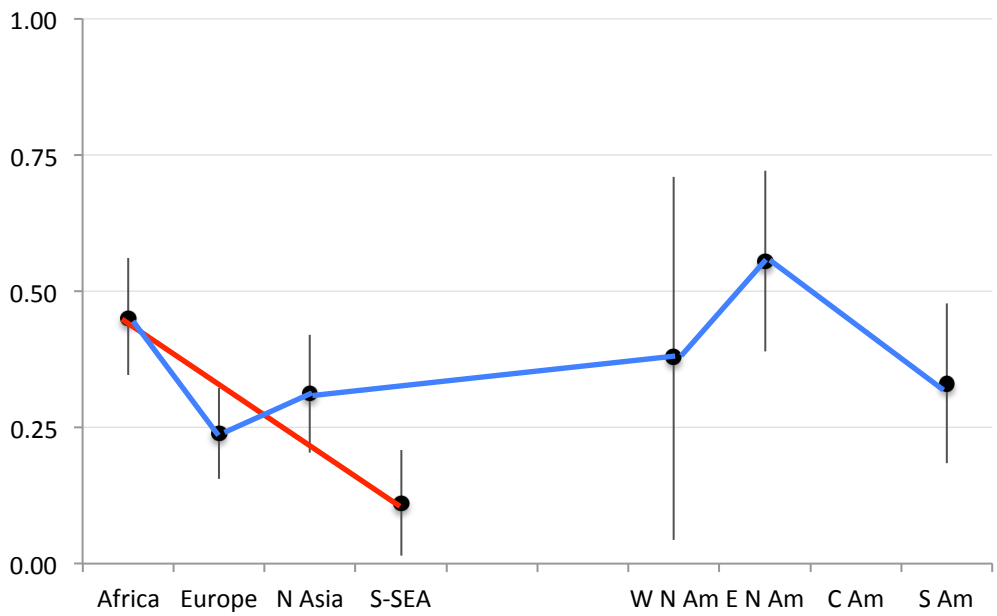




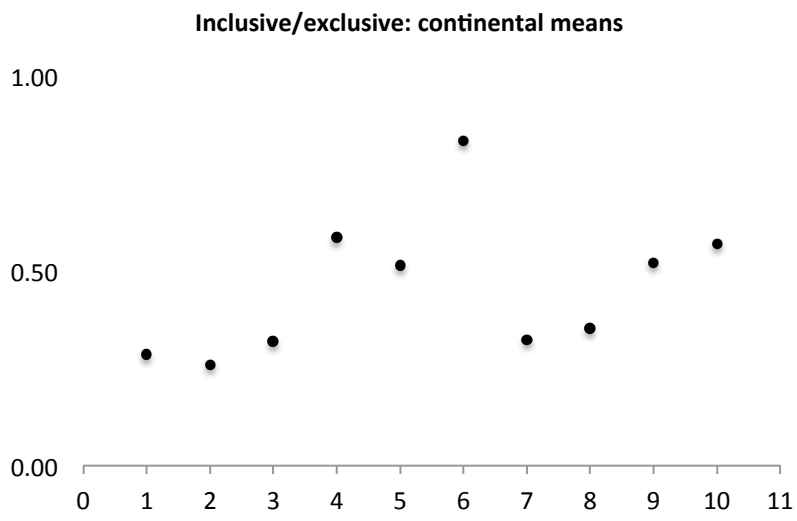
Distribution: High northern.

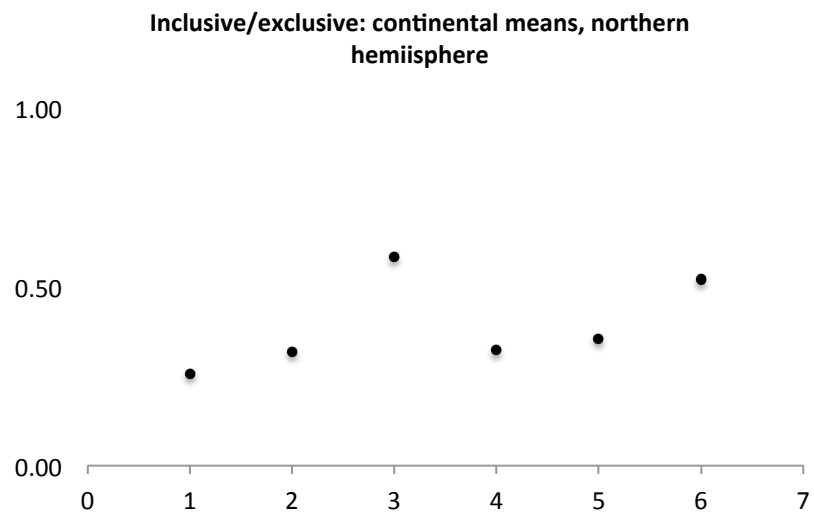
% Verb-based: Continental means (±1 sd).
(No data yet from Australia-New Guinea or Mexico-Central America.)

Blue: Northern trajectory from Africa to the Americas.
Red: Southern trajectory from Africa to Australia.



For comparison: Inclusive/exclusive pronouns. Continental % "Yes".
Left, all continents (*N* = 471); right, northern hemisphere only.

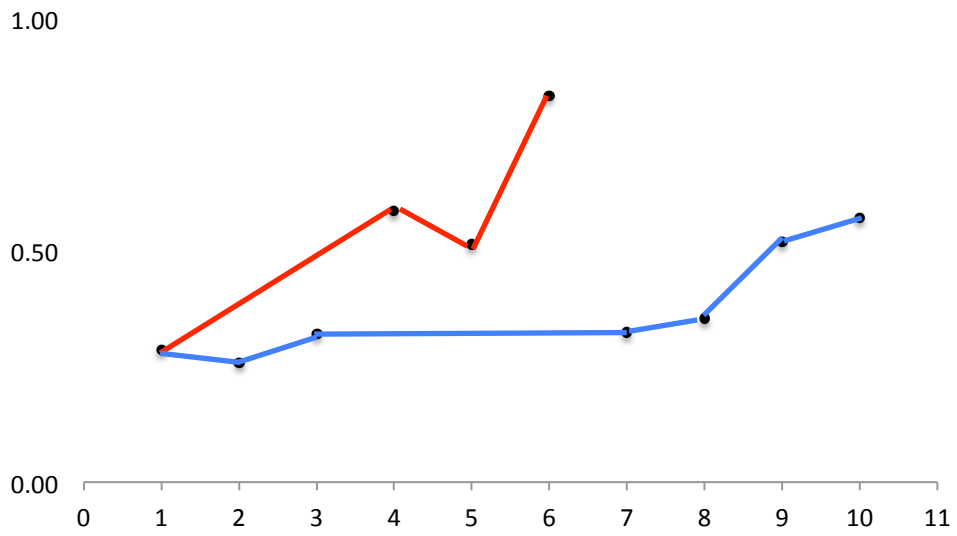




End points of the two trajectories out of Africa (Australia, South America) are highest values for those trajectories. **Distribution: Global.**

For comparison: Inclusive/exclusive pronouns. Continental % "Yes".

Blue: Northern trajectory from Africa to the Americas.
Red: Southern trajectory from Africa to Australia.



End points of the two trajectories out of Africa (Australia, South America) are highest values for those trajectories. **Distribution: Global.**

Basic event structure

Verb derivational paradigms can be based on:

	State or activity	Inchoative	Causative
Russian	sid-e-t'	**ses-t'** {**sed**-)	(po-, u-)sad-i-t'
English	**sit**	sit down	seat
Spanish	estar sentado	sentarse	**sentar**

Basic event structure

Possible general terms: Continuous vs. bounded.
(state, activity) (inchoative, ingressive, telic, punctual)
Posture verbs (survey: 90 languages, chiefly Eurasia and North America).

Proportion of survey verbs with that base. *N* = 90.

Base:	Cont.	Bound.	Caus.
Western Eurasia	0.34	0.44	0.18
Eastern Eurasia	0.16	0.83	0.02
North America	0.46	0.37	0.00

Bounded: Peak in Eastern Eurasia
Continuous: Trough in Eastern Eurasia
Causative: Unique to Europe

Referential density

Ratio of overt NP's to possible overt NP's (core arguments only)

e.g.	I went out this morning and (I) bought groceries.	possible
	I went out this morning ___ to buy groceries.	impossible

Ratio low in some Himalayan languages (ST), high in Nepali (IE) (Bickel 2003)
High in several Nakh-Daghestanian languages (Forker et al. n.d.)
High in European languages

Genealogically stable (even when contact effects change other traits) (Bickel 2003)

Bickel 2003

Referential density (Bickel 2003:728)

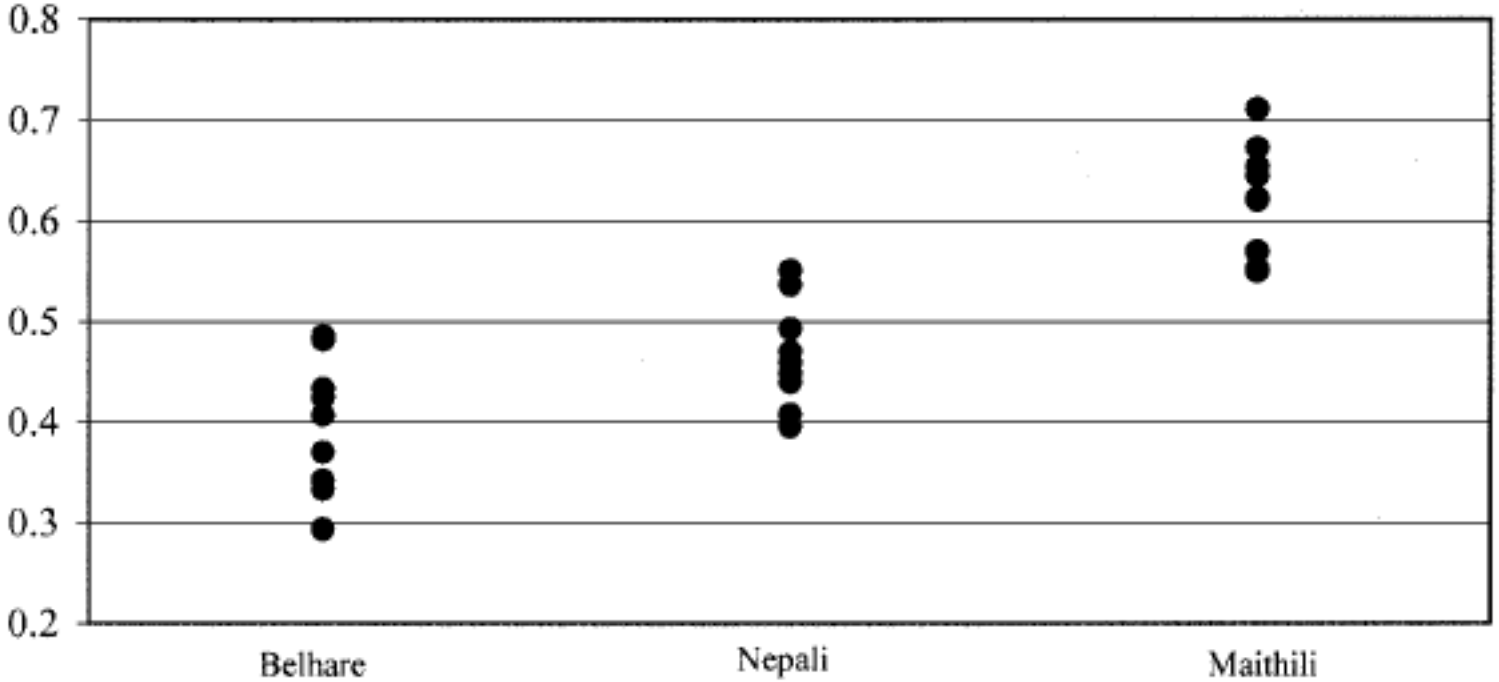


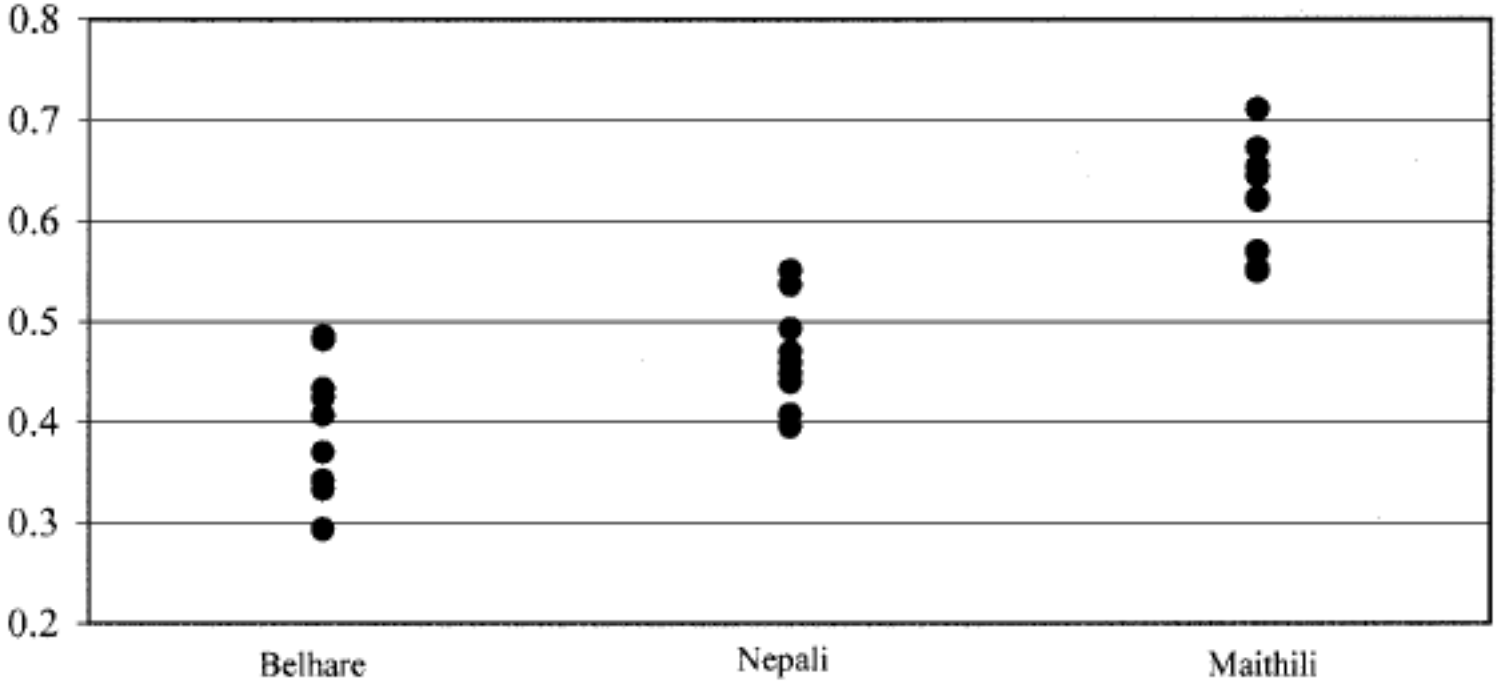
FIGURE 1. Referential density across speakers of three languages.

		CONTRAST	STD. ERROR	t	p (2-tailed)
Belhare	Nepali	.06	0.03	2.39	0.024
Belhare	Maithili	.22	0.03	8.25	< 0.001
Nepali	Maithili	.15	0.03	5.87	< 0.001

TABLE 3. Contrast analyses.

Referential density (Bickel 2003:728)

plus Ingush –>
(Nichols in press)



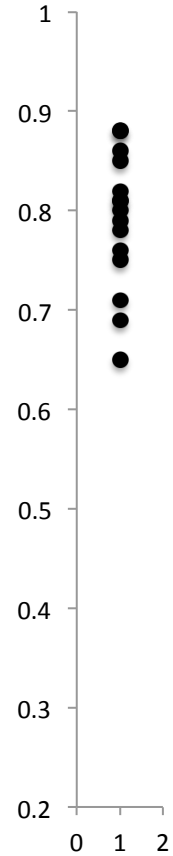


FIGURE 1. Referential density across speakers of three languages.

		CONTRAST	STD. ERROR	t	p (2-tailed)
Belhare	Nepali	.06	0.03	2.39	0.024
Belhare	Maithili	.22	0.03	8.25	< 0.001
Nepali	Maithili	.15	0.03	5.87	< 0.001

TABLE 3. Contrast analyses.

Proof of concept above. How to formulate and test hypotheses?

More detailed pictures of distributions
- More languages, more dialects
- Larger wordlists (impracticable)

Rates of change per variable:
- within families
- in various typological contexts
  - e.g. CausAlt in N-based vs. V-based languages;
  - Event structure in base-causal vs. base-noncausal languages
- within areas
- in different sociolinguistic contexts
  - (e.g. spreading language, receding language, stable multilingualism)

Directions and types of change per variable, and their probabilities:
- same contexts

Grammar survey yields gradual clines, but granularity too large to capture differences between dialects, communities, generations, etc.

Corpus work needed. Plan:

Start with key contact situations where history, sociolinguistics, and language are well understood. Minimal contrasts:

Balkan Peninsula and Avar sphere (NE Caucasus): Symbiotic vs. non-symbiotic contact

SE Circum-Baltic and Ustja River Basin (Russian dialects with and without known substrata)

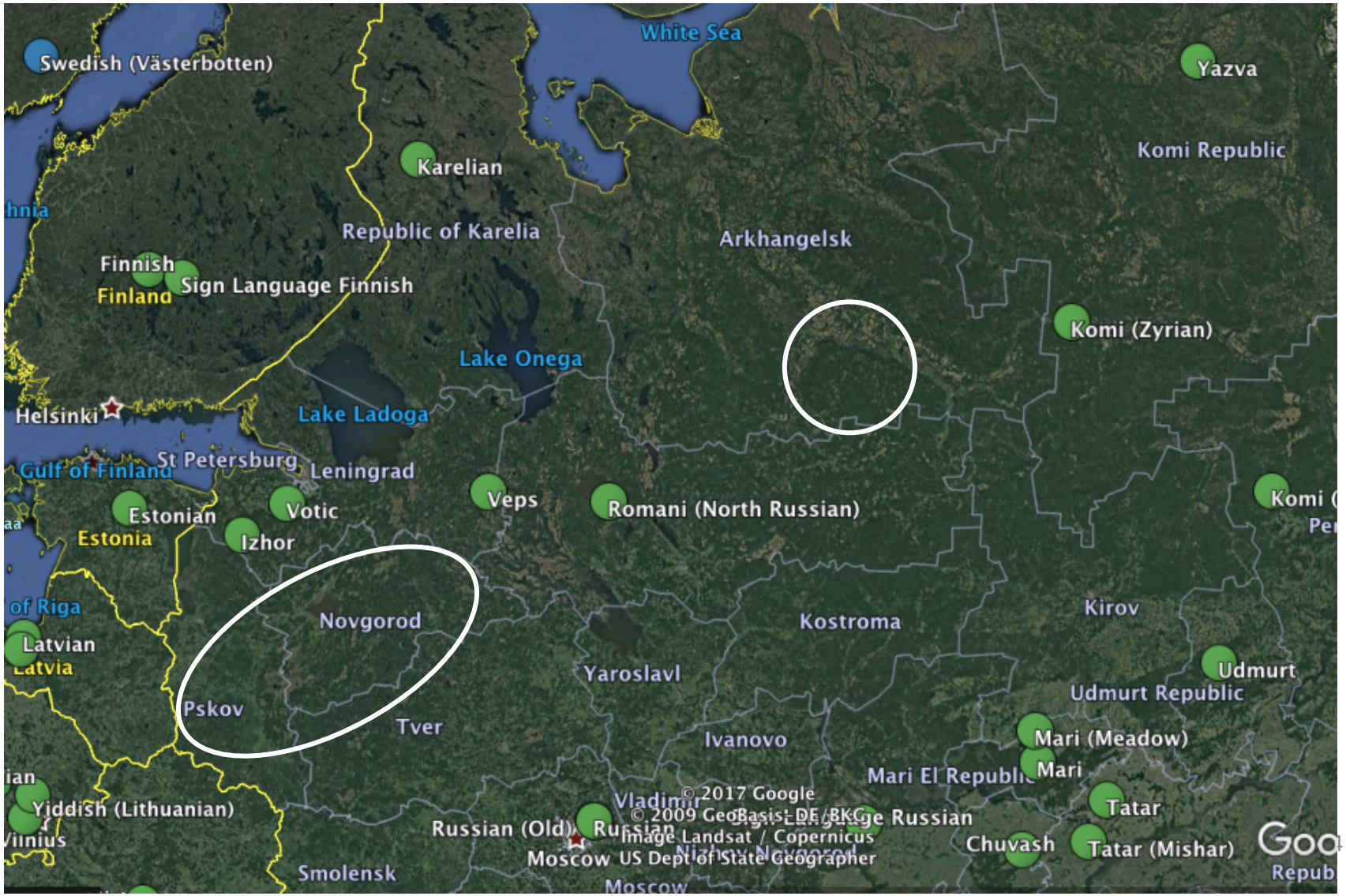
Then: accretion zones of Eurasia (much contact, less substratum)

Start with the best-worked-out variables: RD, CausAlt, POS.

Also event structure, since Russian aspect is knowable for every verb form.

Russian dialects: substratum and non-substratum

Substratum: Novgorod-Pskov Non-substratum: Ustja River basin



Accretion zones of northern Eurasia

Accretion zone (residual zone): pile-up zone where language diversity builds up. These ring the Eurasian steppe, a spread zone.

Risks and problems

Most of these variables are extremely time-consuming to survey
Corpus work even more so

– Automated coding needed.

Key corpus features vary with speaker, genre, typological context, sociolinguistic context, …

But for the key languages and dialects we lack parallel texts
Solution: restrict to (e.g.) sequenced narrative clauses

Corpora exist and more being created

– probably many of them large enough

But corpora lack the annotations needed for these projects.

Annotated lexicons of the type needed do not exist

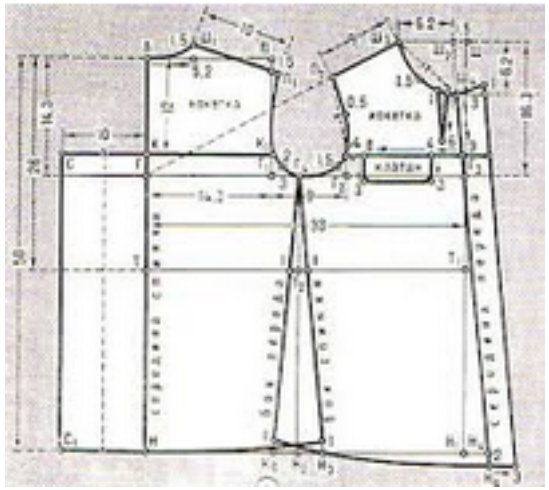
Massive task to do it by hand
Information on word formation is often minimal

Technical capacities needed if massive comparison is to be done

From text corpus, identify:	Needed for:
• clause boundaries	all
• main vs. non-main clauses	Event
• argument dependencies	All
• argument roles	All
• valence	All
• zeroes	RD, CausAlt, Person
• coordination and subordination rules per language	RD
• person categories	Person
• morpheme types (clitic, affix, independent, etc.)	RD, Person
• predicates, their inflectional categories	Event
• discourse functions	Event

From annotated lexicon, identify:	
• derivational paradigms	CausAlt, POS
• base per derivational paradigm	CausAlt, POS
• predicates and aktionsart types	CausAlt, Event

We have our work cut out for us.



Acknowledgments

NSF 9222294
Kone Foundation
Russian Academic Excellence Project 5-100