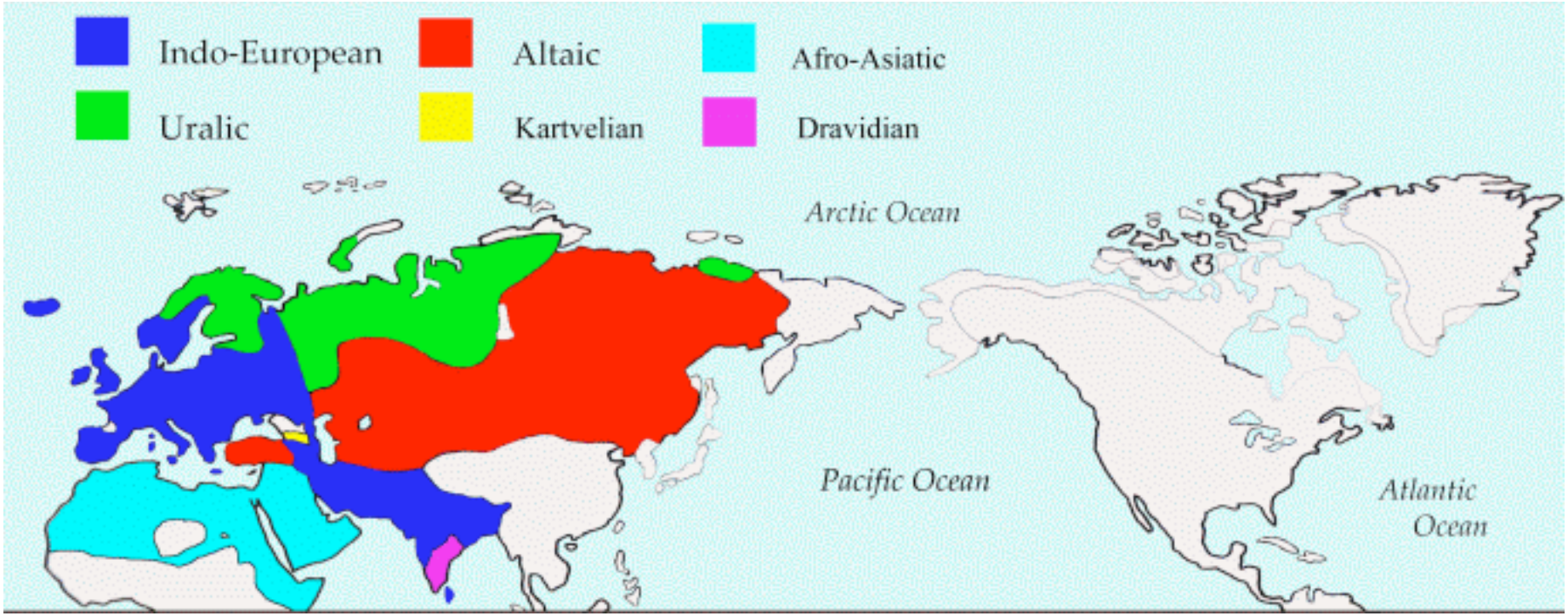
Dating and Relationships



(or...)

Computational historical linguistics and long-range reconstruction

Eric Smith, Santa Fe Institute

Part of the Evolution of Human Languages project at SFI
<http://ehl.santafe.edu/>
in collaboration with the Tower of Babel project
<http://starling.rinet.ru/main.html>
Thanks to Murray Gell-Mann, George Starostin, Ilia Peiros,
in memory of Sergei Starostin

Work done jointly with Tanmoy Bhattacharya, Jon Wilkins, Dan Hruschka, William Croft, Ian Maddieson, Logan Sutton, and Mark Pagel

Outline

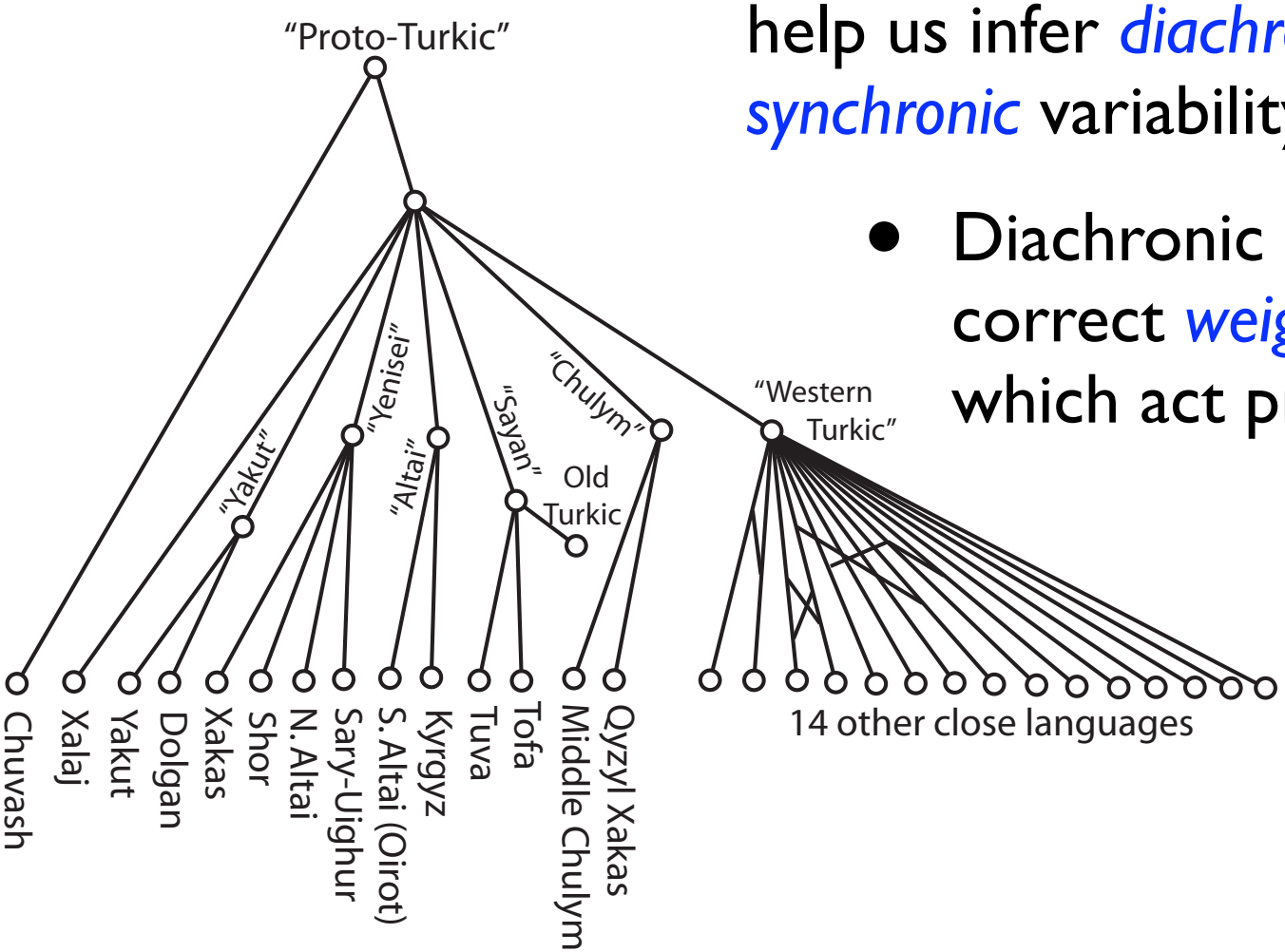
- Goals of historical linguistics
- The classical comparative method
- Attempts at deep reconstruction
- New observations change the landscape
- Our attempts at quantitative reconstruction

Goals of historical linguistics

- To understand how languages change, and how they have changed historically
- To identify relations among languages due to common ancestry or cultural contact
- To reconstruct the languages of past speakers
- To contribute to an understanding of human populations and migrations
- To understand what is possible in language as a window on cognitive constraints

The interaction of history with process

- History-dependent phenomena combine *lawful dynamics* with *historical accident*
  - Accidents make branching processes -- help us infer *diachronic* relations from *synchronic* variability
    - Diachronic relations assign the correct *weights* to processes which act probabilistically



The classical comparative method of historical linguistics: *to interpret innovations*

- A hypothesis of relationship among a set of languages.
- Cognate identification to correctly group elements (word roots or other morphemes) that have a common origin as evidenced by sound structure and meaning.
- Regular sound correspondences that describe relations among phonological segments in different languages across the lexicon; these may depend on phonetic context.
- Reconstruction of the ancestral sounds, morphemes and lexical items of a protolanguage that serves as a model of the ancestor of the proposed group.
- Establishing the innovations that created the descendant languages from the protolanguage.
- Classification of the languages within the family, using shared innovations to identify a structure of subfamilies.
- Construction of an etymological dictionary that traces semantic shifts and borrowings.

Language structure: many kinds of innovation can motivate hypotheses of relatedness

- Phonology:
  - the sound system used by a language ("phones" or un-analyzed segment)
  - the sound sets (phonemes) recognized as carrying distinctions in meaning
- Lexicon:
  - the map from root meanings to words (strings of phones or phonemes)
  - the overlap structure of meanings (so, what's in a word)
- Morphology:
  - "word shape": modifications to indicate case, tense, aspect, etc.
- Syntax:
  - word categories and composition rules to determine phrase structure, etc.
- Typology:
  - major structure categories of languages: word order, implicational groups

Helpful glossary at <http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/>

Rare innovations versus clusters of common innovations

- Rare innovations: single features with ~0 probability to occur by chance
  - Imply common descent or borrowing, even w/o mathematics
  - Only seen once: hard to assign probabilities from frequencies
  - Common in morpho-syntactic features
  - Useless for dating; do not support induction
- Common variations:
  - Examples: sound shift and meaning shift in core lexicon
  - Individually uninformative, but can assign probabilities from data
  - Require math to handle, but do support induction, and can be informative about dates if change processes are regular

Morphology, syntax, typology: regular functions; differing representations

Word order

(Russian borrowing)

Oglum shkolada oe:renip turar

my-son at-school is studying-3

(S O V)

Case

nom	a book
nomnar	books
nomnaram	my books
nomnaramnung	of my books

Vowel harmony

nom	a book
inek	a cow
nomnar	books
inekter	cows

Tense

ajtyr	to ask
ajtyrar men	I will ask
ajtyryp tur men	I am asking
ajtyrbas tur men	I am not asking

Turkic family: examples of "inflectional" and "agglutinating" language TYPES

Word lists as the key to lexical (= phonological / semantic) reconstruction

http://starling.rinet.ru/cgi-bin/response.cgi?root=config&morpho=0&basename=\data\alt\turcet&first=1

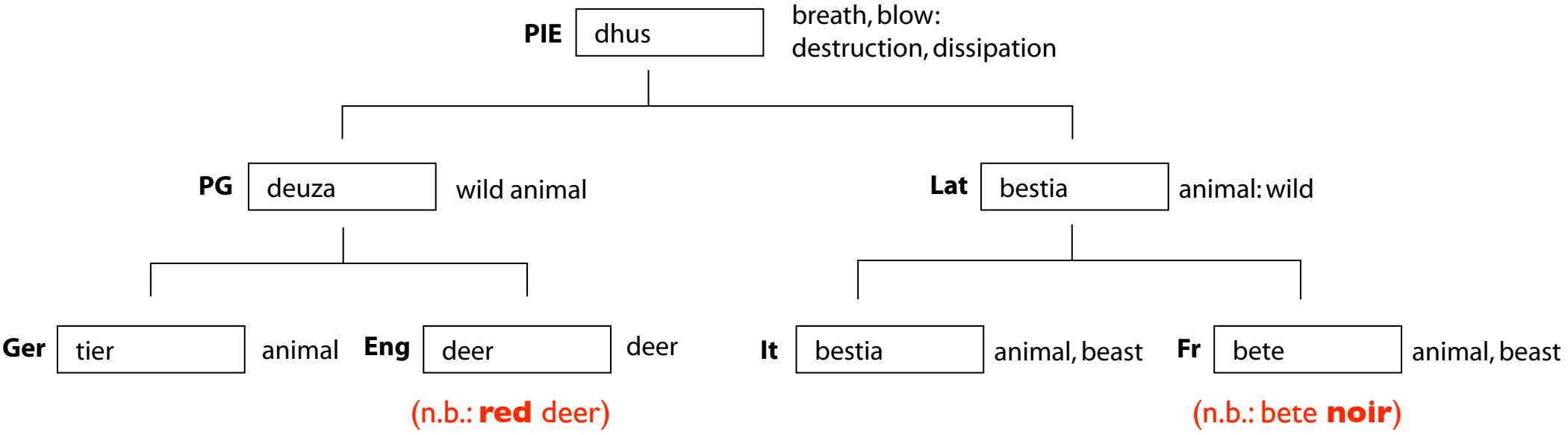
Turkic etymology :

New query
Total of 2017 records 101 page
Pages: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Forward: 1 20 50 100

Proto-Turkic	Altaic etymology	Meaning	Russian meaning	Old Turkic	Karakhanid	Turkish	Tatar	Middle Turkic	Uzbek	Uighur	Sary-Yughur	Azerbaidzhan	Tur
*Äb ⊞	Altaic etymology ⊞	hunt, chase	охота	*ab* (Orkh.), *av* (OUygh.)	*av* (MK)	*av*	*aw*	*aw* (Pav. C.)	ɔv	*aw*, dial. *ɑ*		*ɑv*	äv
*ab- ⊞	Altaic etymology ⊞	to crowd, come together	собираться, встречаться	av- (OUygh.)	av- (MK, KB)								
*abuč ⊞	Altaic etymology ⊞	handful	пригоршня		avut (MK), avut-ča, avuč-ča (KB), avuč (Tefs.)	avuč	*uč*	avuč (MA, Sangl., Бор. Бад.)	ɔwuč	*oč*	*oš*	ovuč	ovu jan-
*Abuč-ka ⊞	Altaic etymology ⊞	1 husband, old man 2 foster-mother 3 elder sister 4 uncle	1 муж, старик 2 кормилица 3 старшая сестра 4 дядя	avïčya, abučya 1, abučqa 2 (OUygh.)	avïčya 1 (MK, KB)	abuš 3 dial.	abušqa, awucqa 1 dial. (Sib.)	abušqa, avušqa 4 (Abush., Sangl.)					

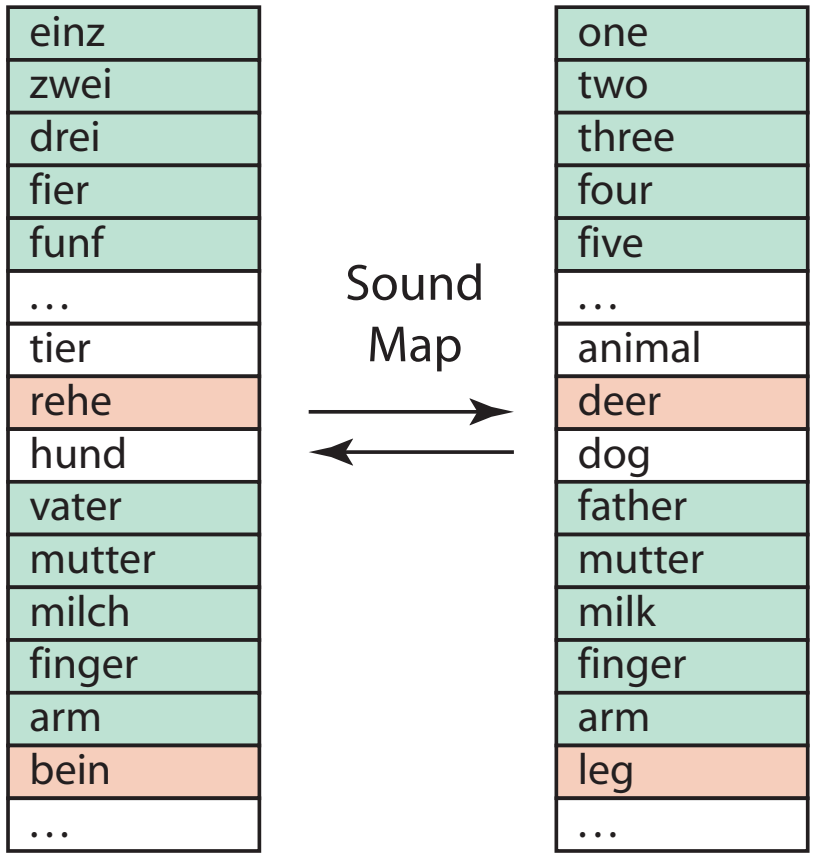
Fundamental object is the history of a given word's sound and meaning

- n.b., word forms are attested; meanings are indirectly inferred, and often ambiguous
- easy to trace a form; but inadequate to infer history from forms alone



Representing sound and meaning "innovation" in the comparative method

- Suppose that some stable meaning categories can be identified
- Identify primary words for each meaning
- Try to exclude "borrowed" terms; suppose that what is left has been transmitted through vertical descent
- Identify systematic sound relations and try to infer historical sound changes
- Associate semantic innovations with in-language substitutions within meaning categories



Preserved meanings suggest sound maps

PIE
Hoi-(wo, ko, no)
duoh
trei
k^wetuor
finfi

(Numbers give an example in which we can treat primary meanings as language-universal and historically relatively stable)

PG
ainaz
twai
δrjiz
fidwor
finfi

(Innovation)

(Sound similarity)

Use preservation of meaning to infer regular relations of sounds: here, e.g., thr <> tr

Lat
unus
duo
tres
quattuor
quinque

(Sound similarity)

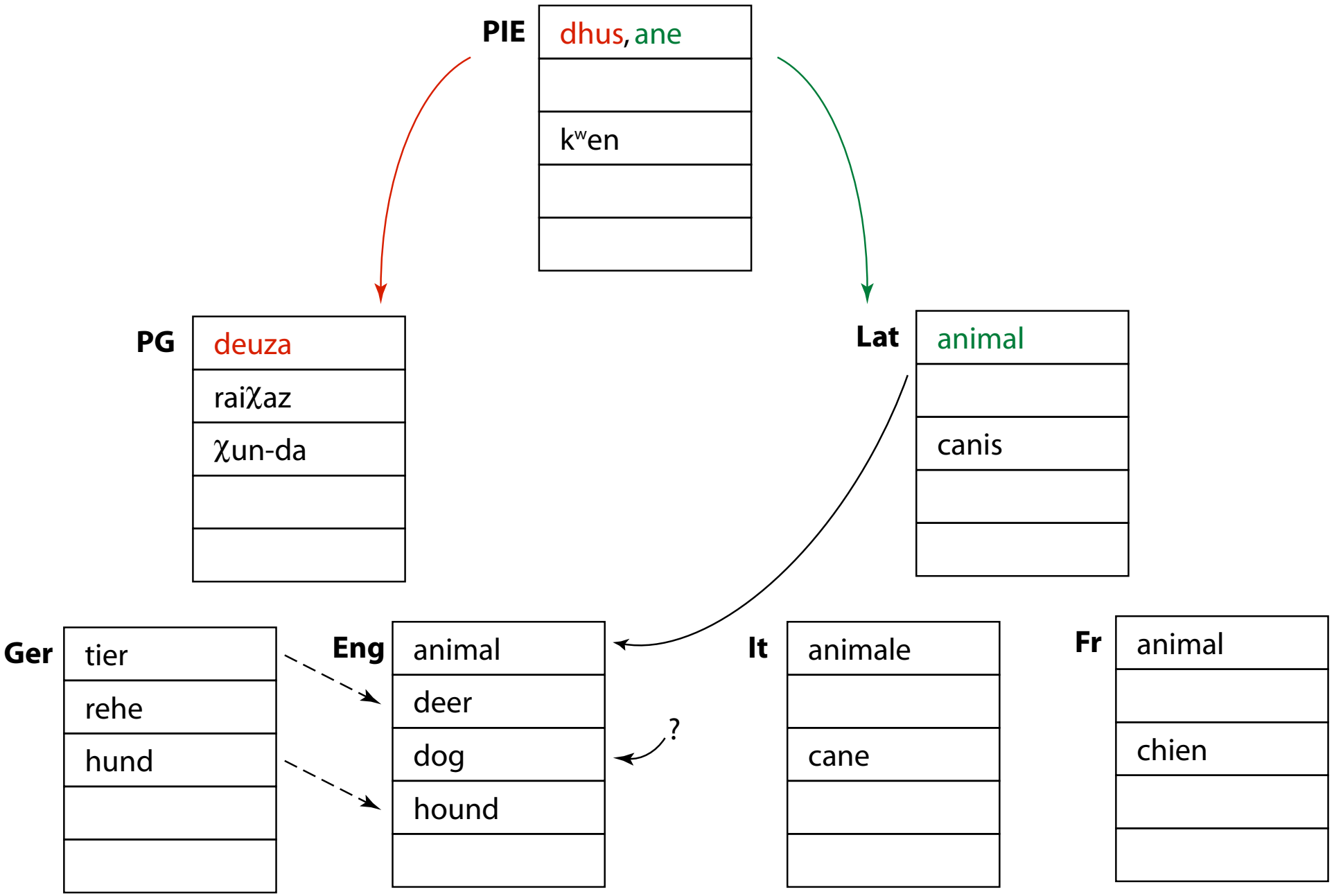
Ger
einz
zwei
drei
fier
funf

Eng
one
two
three
four
five

It
uno
due
tre
quattro
cinque

Fr
un
duex
trios
quatre
cinque

Sound maps help identify meaning shifts



But sound changes can depend on context

Split of Old English /k/

Stage I	katt	keaff	kinn
Stage II	katt	tʃeaff	tʃinn
Stage III	katt	tʃaff	tʃinn

Split of Latin/s/

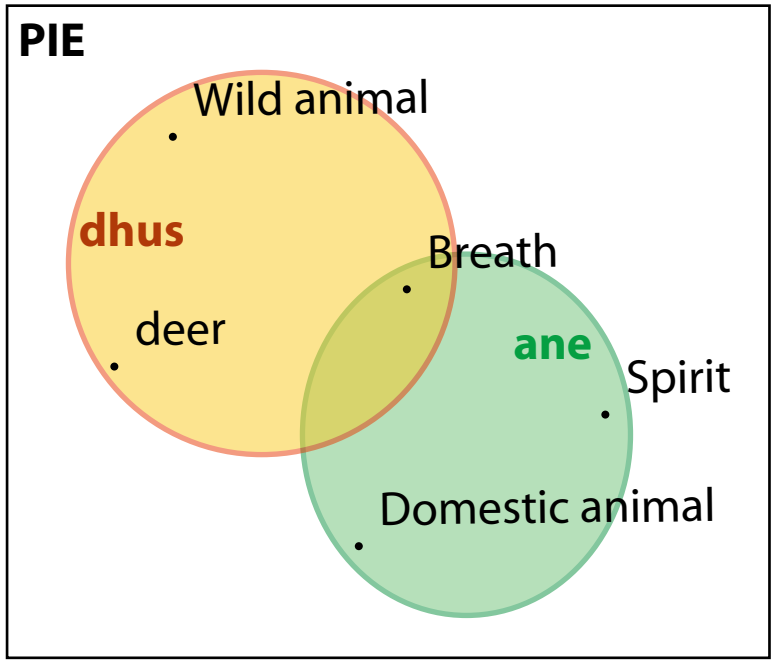
Stage I	ka:ra	flo:s	flo:ses
Stage II	ka:ra	flo:s	flo:zes
Stage III	ka:ra	flo:s	flo:res

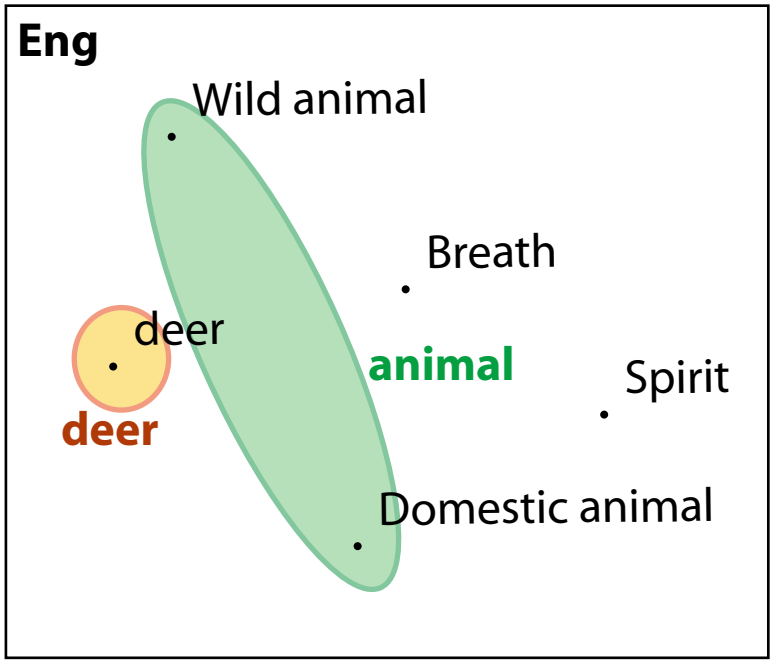
From R.L. Trask, "Language change"

- Even if true, not uniformly available, and unrealistic to rely on at larger time depths
- Eventually context discovery requires more information than the language pattern yields

And the nature of meaning shift is not mathematically understood

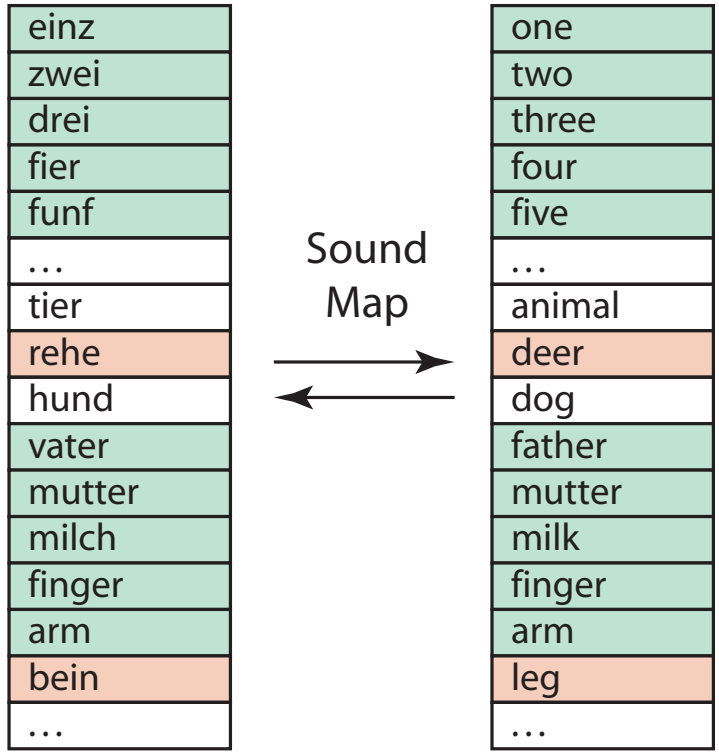
- Phonological and semantic constraints interact with polysemy and synonymy to structure sound and meaning change
- Semantic categories, split, join, and move in some "space" which we do not know





Attempts at long-range reconstruction: lexicostatistics and glottochronology

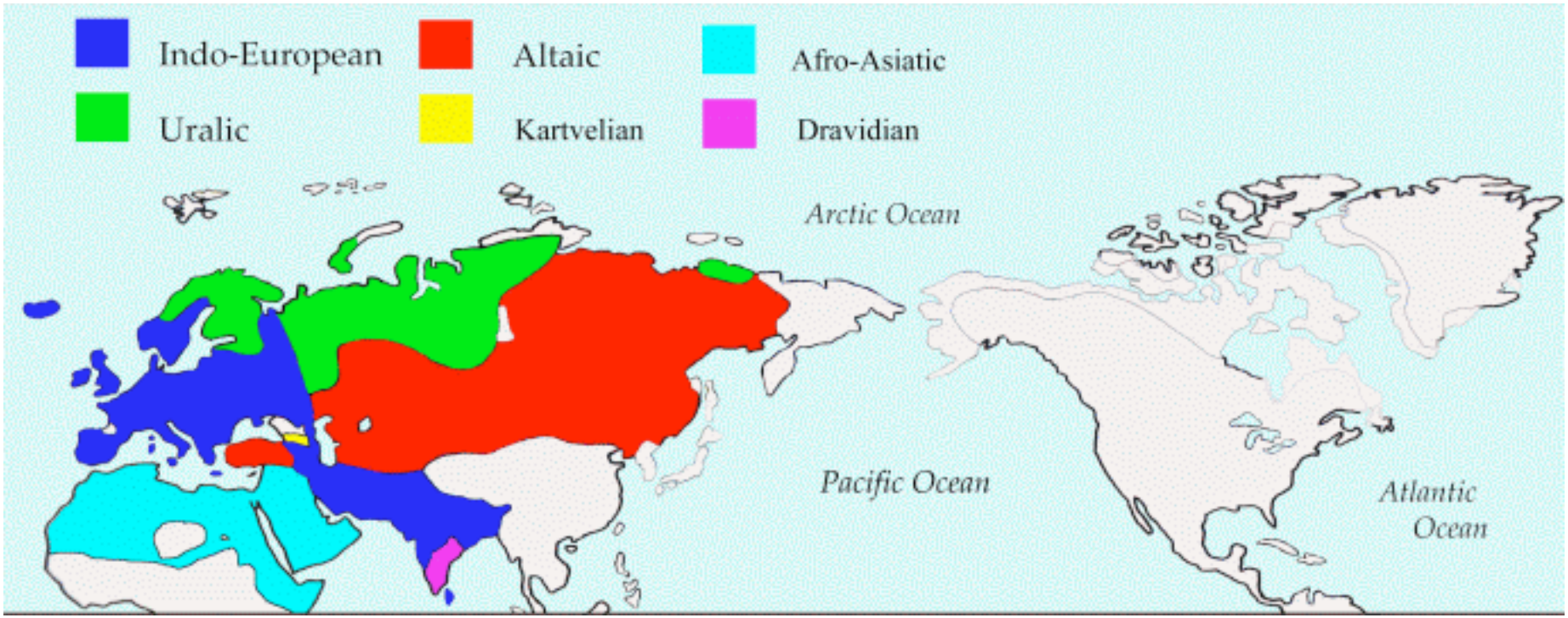
- Assign any sound map without penalty, but require regularity
- Exclude borrowed items in either language from consideration
- Identify fraction preserved cognates; convert to separation time (penalty)
- Attempt to fit separation times to an ultrametric structure (tree)



$$P_{\mathrm{Preserve}}(\mathrm{word}) = e^{-t/\tau}$$

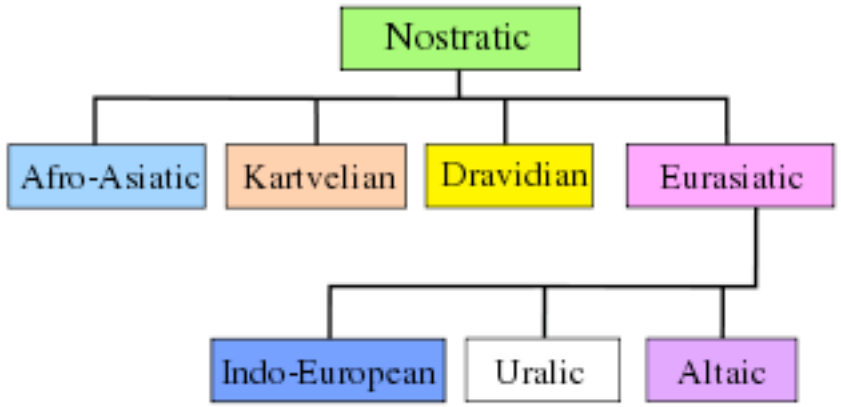
$$\Delta t_{\mathrm{sep}} = -\tau \log(\mathrm{frac.\ \ preserved})$$

Example: the Nostratic Hypothesis



http://starling.rinet.ru/maps/maps.php?lan=en

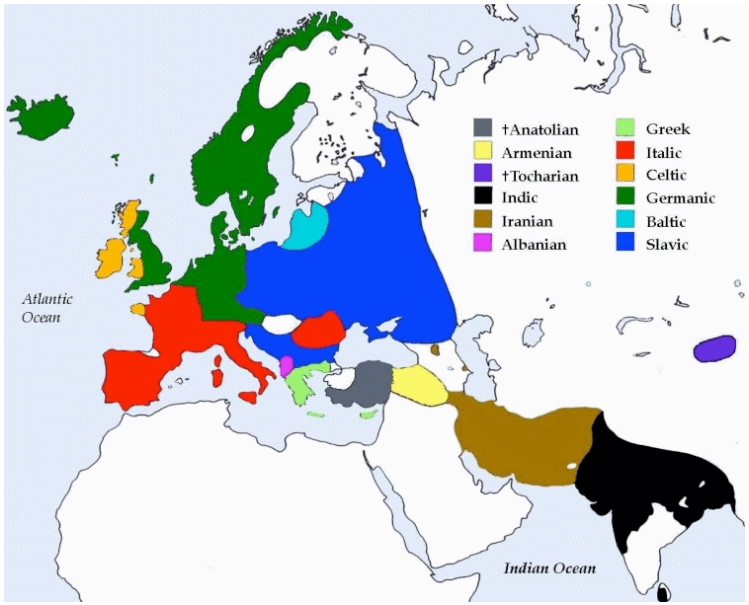
- Coined by Holger Pedersen (1903)
- Modern form of the hypothesis by Vladislav Ilitch-Svitych and Aharon Dolgopolsky (1960s -- present)
- Estimated 12,000-15,000 BCE



http://en.wikipedia.org/wiki/Nostratic_languages

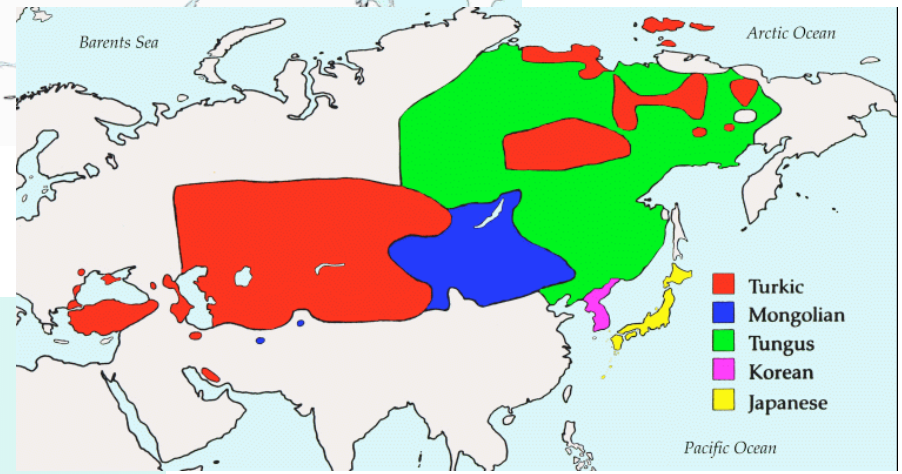
Sub-families within Nostratic

Indo-Hittite



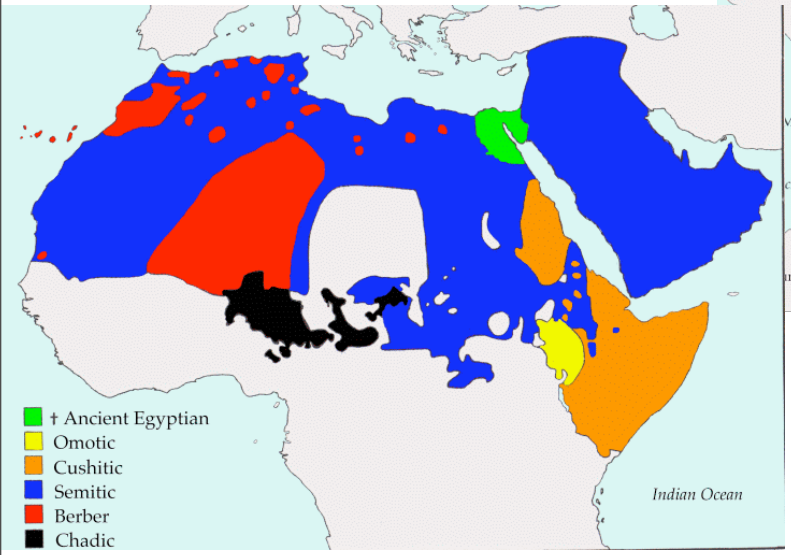
Uralic

Altaic

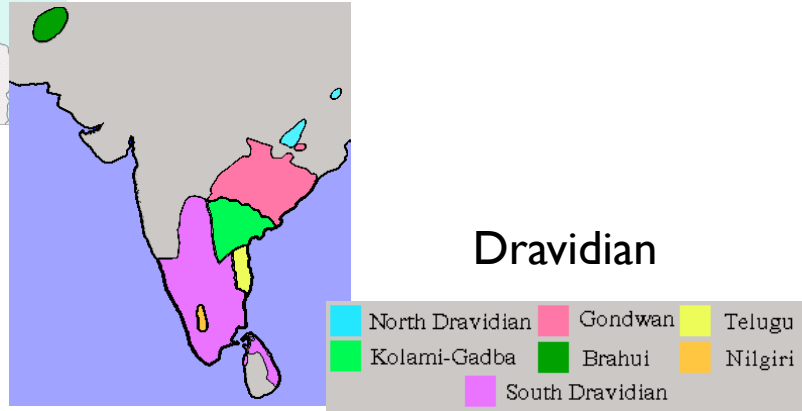


Kartvelian

Afro-Asiatic



Dravidian



New quantitative observations: finer process models for language change

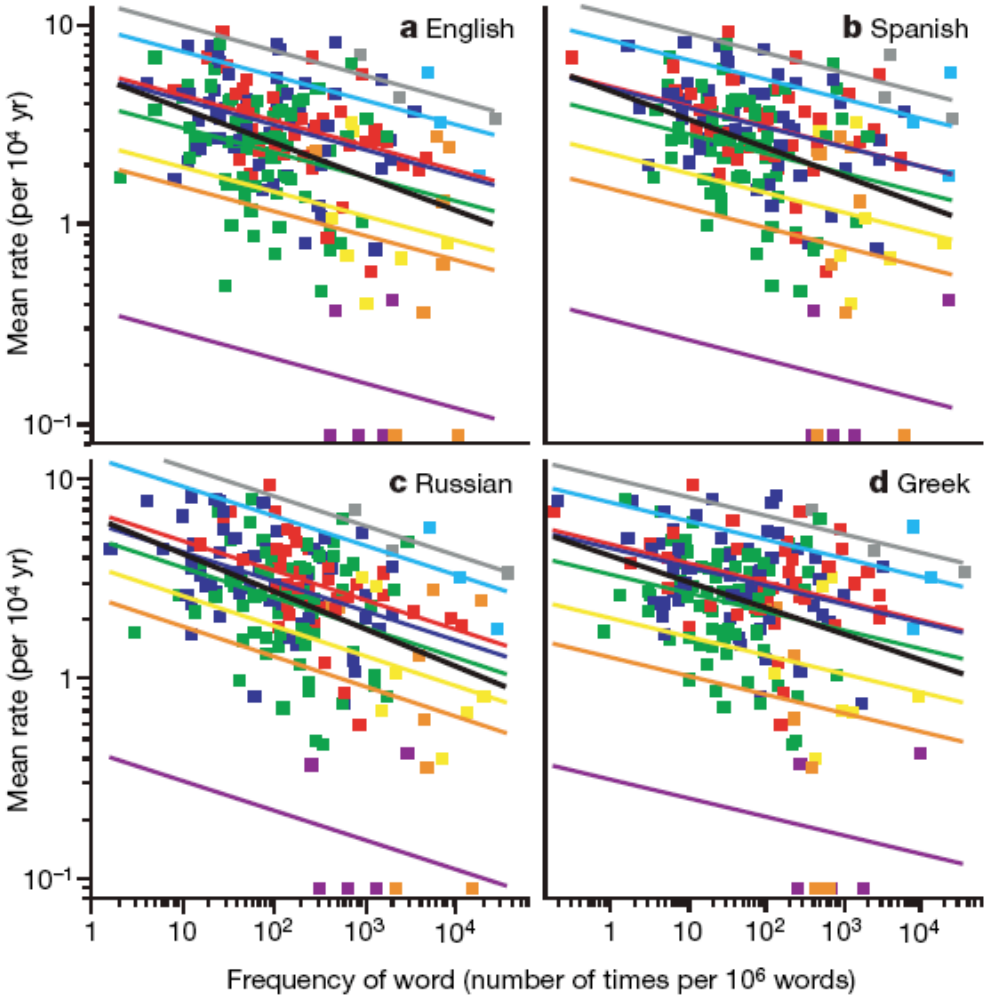
- Rates of change in semantics, phonology, morphology
- “Point processes” associated with branching

Typology I: frequency of use and rates of change

Frequency of word-use predicts rates of lexical evolution throughout Indo-European history

Mark Pagel[1,2], Quentin D. Atkinson[1] & Andrew Meade[1]

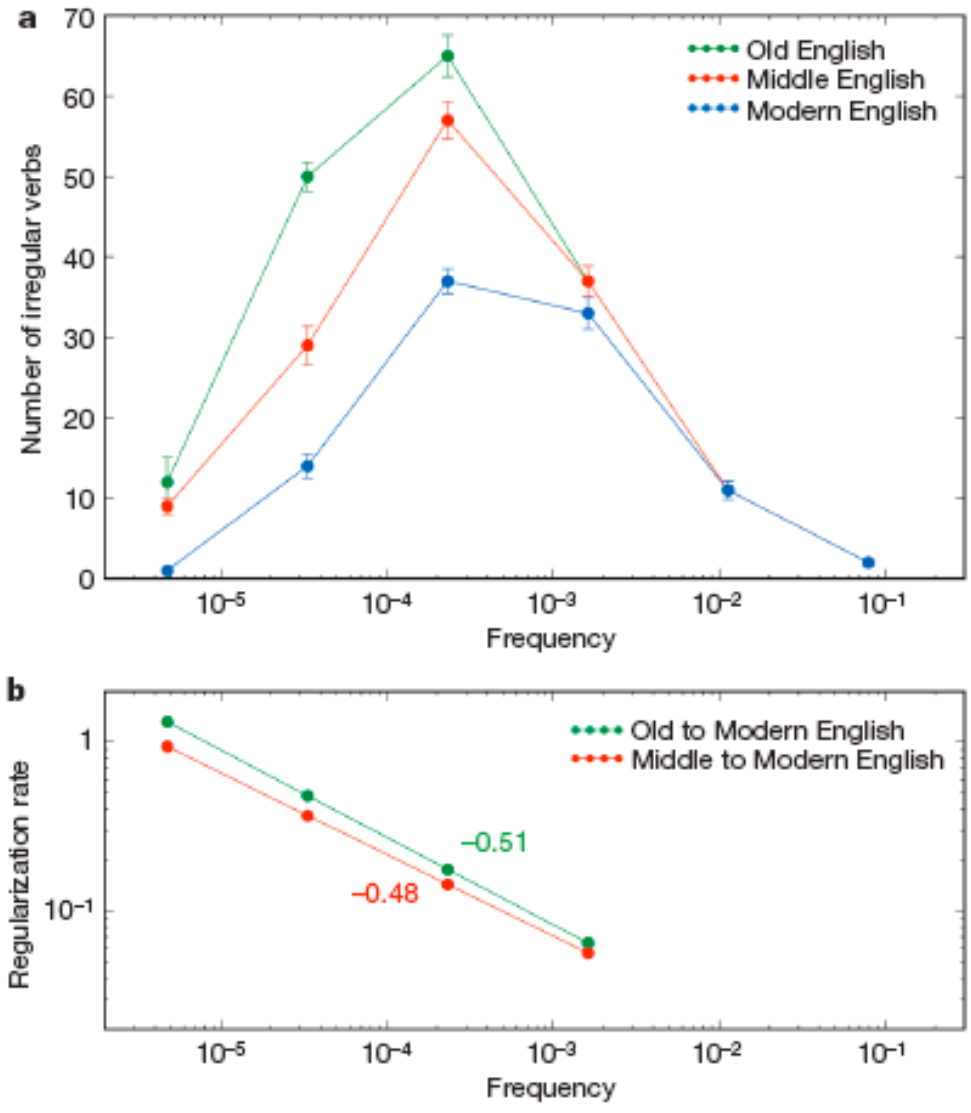
NATURE|Vol 449|11 October 2007



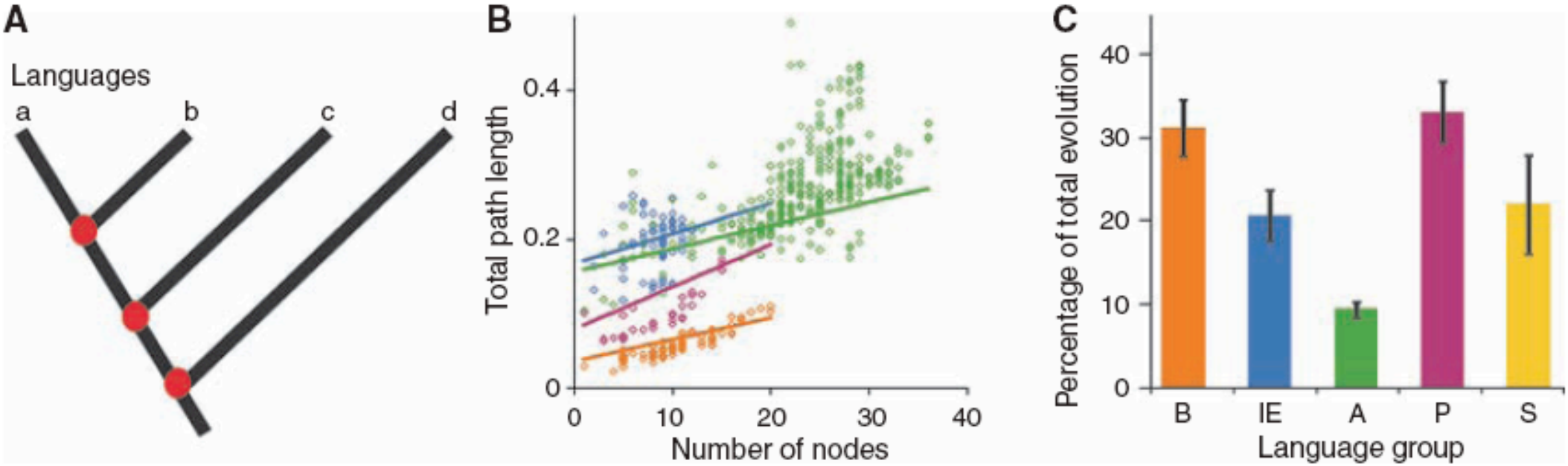
Quantifying the evolutionary dynamics of language

Erez Lieberman[1,2,3]*, Jean-Baptiste Michel[1,4]*, Joe Jackson[1], Tina Tang[1] & Martin A. Nowak[1]

NATURE|Vol 449|11 October 2007



Linguistic punctuated equilibrium: an interaction of phylogeny with underlying process?



Languages Evolve in Punctuational Bursts

Quentin D. Atkinson,[1]* Andrew Meade,[1] Chris Venditti,[1] Simon J. Greenhill,[2] Mark Pagel[1,3]†

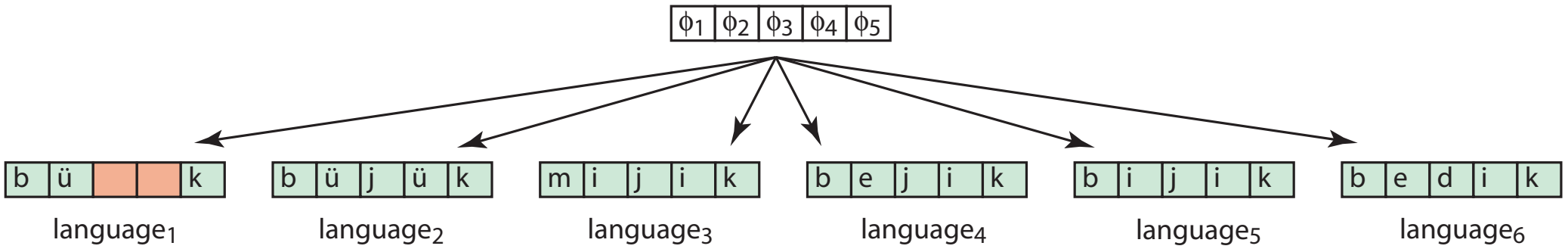
1 FEBRUARY 2008 VOL 319 **SCIENCE** www.sciencemag.org

Our work to quantify the comparative method

- Alignment
- Regular sound correspondence
- Minimum-bias methods for unexplained variation
- Context-dependence and information criteria
- Attempts to map the semantic space

Maximum-likelihood estimation of history and process

- Align words in daughter languages
- Propose phoneme assignments to aligned positions in the ancestor (with probabilities)
- Estimate regular correspondence of ancestor to daughter phonemes (w/ or w/o probabilities)
- Estimate random violations (with probabilities)



Alignments inferred without prior knowledge

Etymology 4 "belly"

q	q	k	q	q	q	q	q	G	G	x	q	q	q	x	x	x	x	q	q	q	q	q	q	q	q	q	q
a	a	a	a	a	ɔ	e	a	a	a	a	a	a	ā	ɨ	a	ɨ	ɨ	a	a	a	a	a	a	a	a	a	a
r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r
ɨ	ɨ	ɨ	ɨ	ɨ	i	i	ɨ	ɨ	ɨ	ɨ		ɨ	ɨ	ъ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ
n	n	n	n	n	n	n	n	n	n	n	n	n	n	m	n	n	n	n	n	n	n	n	n	n	n	n	n
											ɨ																

Etymology 5 "big, high"

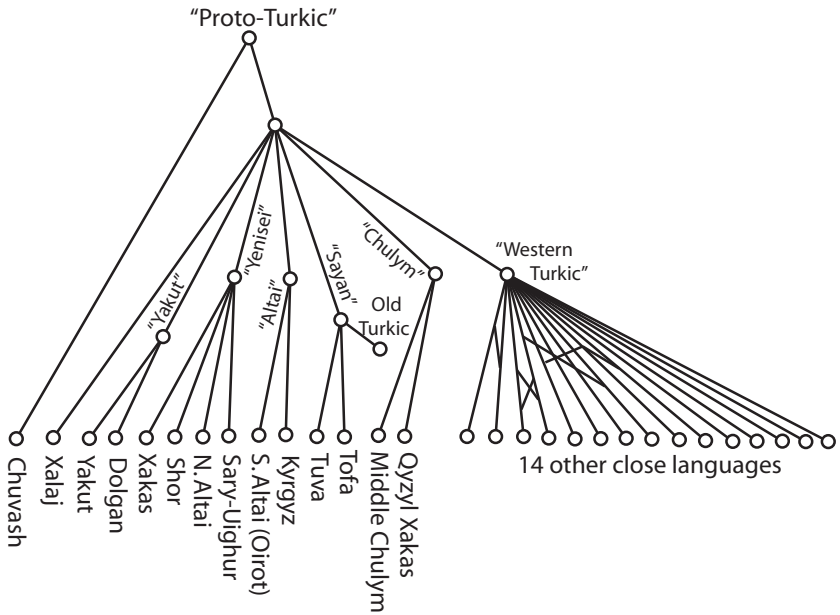
b	b	b	b	b	b	b	b	b	b	p	m	b	b		b	b	b	b	b	b	m	b	b	b		b
e	e	ü	i	e	u	ü	e	ö	e	ö	ö	i	i		e	e	i	i	i	e	i	ū	ü	i		i
d	ð	j	j	j	j	j	z	j	j	z	z	j	d		d	d	j	j	j	j	j		j	j		j
ü	ü	ü	e	i	u	ü	ɨ	ü	i	ə	ü	i	i		i	i	i	i	i	e	i		ü	i		i
k	k	k	k	k	k	k	k	k	k	k	k	k	k		k	k	k	k	k	k	k	k	k	k		k

Etymology 12: "Breast, nipple"

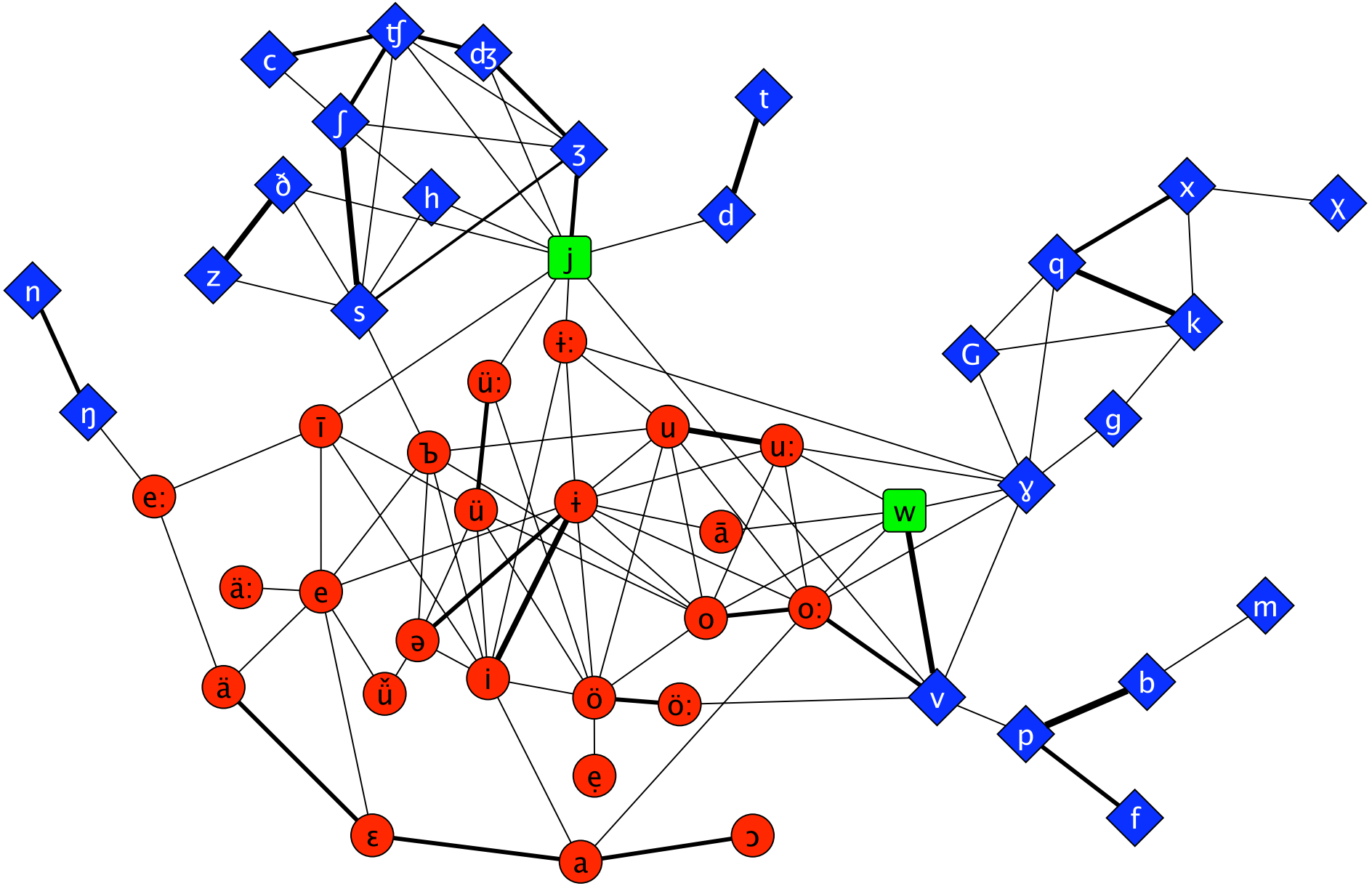
	m	m		m		m	m			m	m					m	m		m	m	m						
	e	ɛ		ä		ä	ā			ā	ē					ä	ä		ä	ä	ä						
	m	m		m		m	m			m	m					m	m		m	m	m						
																					m						
	e	i		ä		ä	e			ä	ē					ä	ä		ä	ä	ä						
										k						j	j										

Sound correspondences among the languages

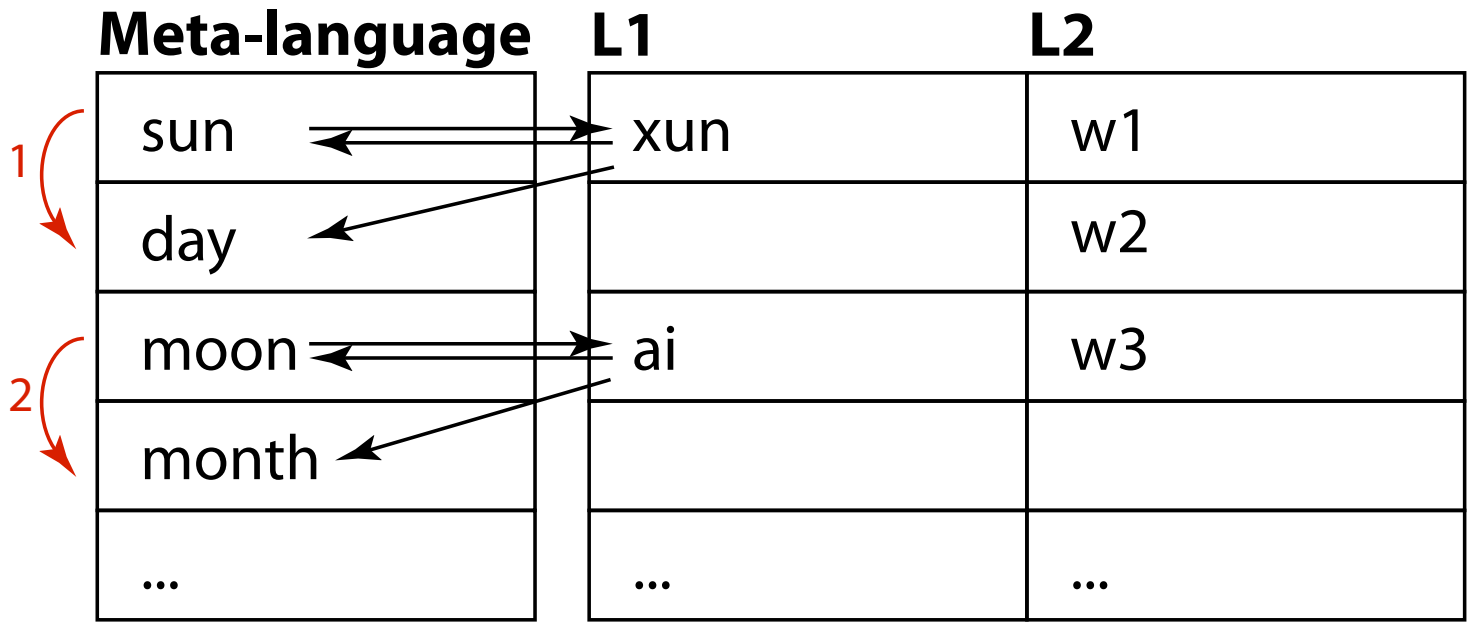
ancestor	probAncestor	ATU	KRH	JAK	DOLG	TUV	TOF	HAK	SHR	ALT	KRG	UIG	UZB	KAZ	KLPX	NOGX	BAS	TAT	QUM	BLKX	KRMX	TRM	AZB	GAGX	TRK	KHAL	CHV
97	0.097	a	a	a	a	a	a	a	a	a	a	a	ɔ	a	a	a	a	a	a	a	a	a	a	a	a	a	o
616	0.042	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	i	i	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ə
101	0.042	e	e	e	e	e	e	i	e	e	e	ä	e	e	e	e	i	i	e	e	e	e	ä	e	e	ä	ъ
117	0.041	u	u	u	u	u	u	u	u	u	u	u	u	u	u	u	o	o	u	u	u	u	u	u	u	u	ъ
105	0.029	i	i	i	i	i	i	ə	i	i	i	i	i	i	i	i	e	e	i	i	i	i	i	i	i	i	ə
61948	0.026	ü	ü	ü	ü	ü	ü	ü	ü	ü	ü	ü	u	ü	ü	ü	ö	ö	ü	ü	ü	ü	ü	ü	ü	i	əʷ
111	0.018	o	o	o	o	o	o	o	o	o	o	o	u̇	o	o	o	u	u	o	o	o	o	o	o	o	o	ъ
61942	0.016	ö	ö	ö	ö	ö	ö	ö	ö	ö	ö	ö	u̇	ö	ö	ö	ü	ü	ö	ö	ö	ö	ö	ö	ö	e	ъ
113	0.058	q	q	x	k	q	q	x	q	q	q	q	q	q	q	q	q	q	q	q	q	G	G	q	k	q	x
106	0.048	j	j	s	h	č	č	č	č	j	ǯ	j	j	ž	ž	j	j	j	j	j	j	j	j	j	j	j	s
116	0.044	t	t	t	t	d	d	t	t	t	t	t	t	t	t	t	t	t	t	t	t	d	d	d	d	t	t
114	0.043	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r
108	0.032	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l
107	0.031	k	k	k	k	k	x	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k
98	0.023	b	b	b	b	b	b	p	p	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	b	p
110	0.021	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n
61850	0.019	š	š	s	s	š	š	s	š	š	š	š	š	s	s	s	š	š	š	š	š	š	š	š	š	š	l
109	0.019	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m	m
122	0.018	z	z	s	s	s	s	s	s	s	z	z	z	z	z	z	δ	z	z	z	z	z	z	z	z	z	r
8240	0.017	č	č	s	s	š	š	s	š	č	č	č	č	š	š	š	s	č	č	č	č	č	č	č	č	č	s
611	0.016	ɣ	ɣ	0	0	ɣ	ɣ	ɣ	ɣ	0	ō	ɣ	ɣ	w	w	w	w	w	w	w	ɣ	G	ɣ	0	ɣ	ɣ	v
115	0.015	s	s	0	0	s	s	s	s	s	s	s	s	s	s	s	h	s	s	s	s	s	s	s	s	s	s
103	0.012	g	g	ɣ	g	g	g	g	ɣ	ɣ	g	g	k	g	k	k	g	g	k	g	g	g	g	g	g	g	g
112	0.01	p	p	b	b	v	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p	p
62111	0.009	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	ŋ	n	ŋ	n	n	n	ŋ	n
118	0.008	b	v	b	s	g	g	v	b	v	j	v	v	j	j	j	j	v	j	j	v	v	v	v	v	v	v

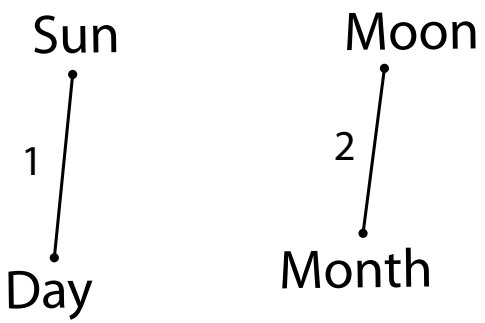


Sound relations inferred from sound changes



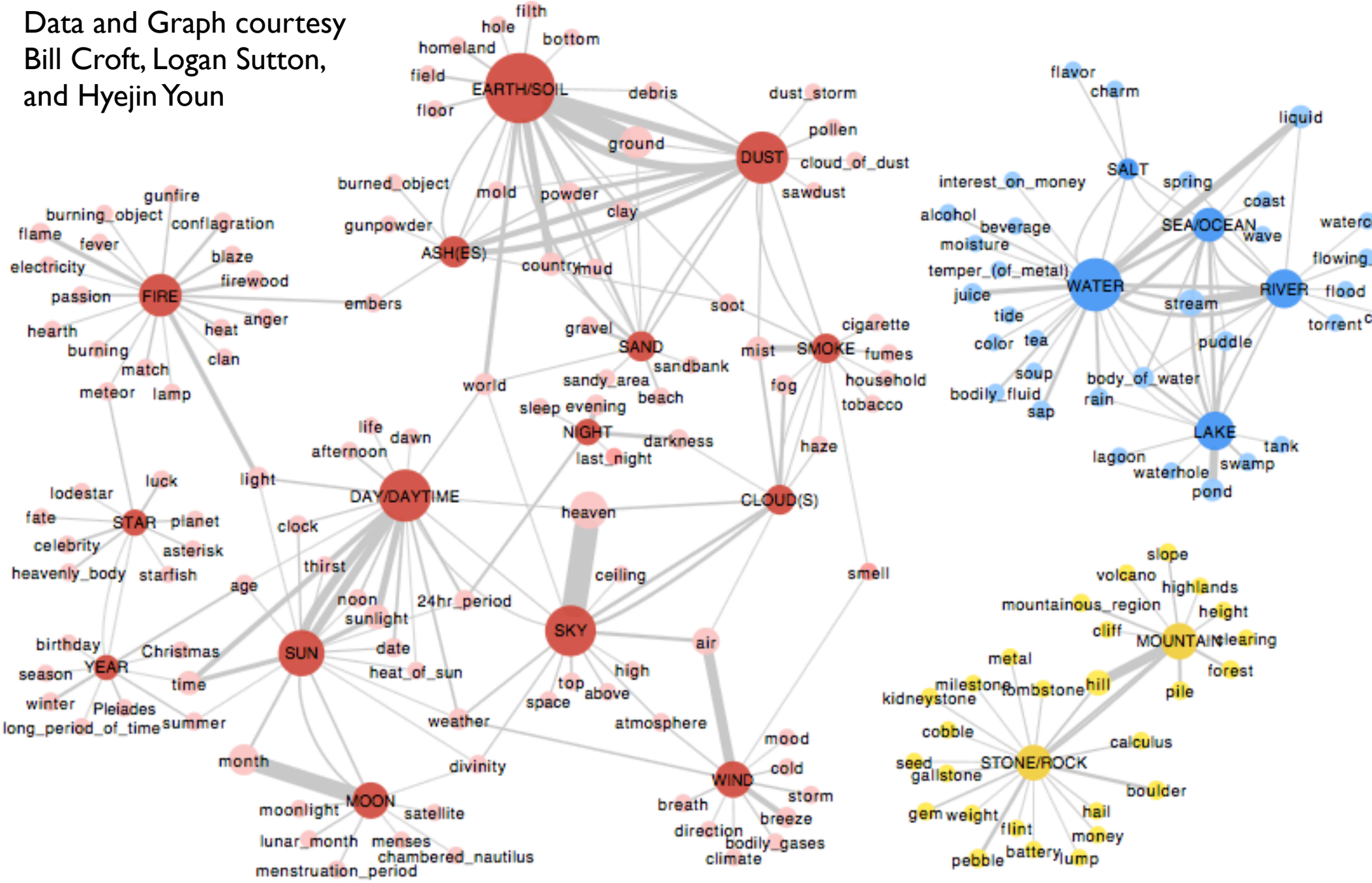
Inferring polysemy with English as a meta-language





Network of polysemes in 81 diverse languages

Data and Graph courtesy
Bill Croft, Logan Sutton,
and Hyejin Youn



Some concluding comments

- Historical linguistics has been a "rule-based" system, something like formal logic
- Can we re-derive such rule-based systems within principles of statistical inference?
- New high-volume data sources and new data types: a *huge* opportunity for computational analysis
- Languages provide another window on evolution and historical inference from molecular sequences
- BUT: *Good* quantitative linguistics will require collaboration, patience, and work