Opportunities in quantitative historical linguistiics

Eric Smith

work with

Tanmoy, Jon, Bill, Ian, Dan Hruschka, Logan Sutton, Hyejin Youn

collaborative with

Murray Gell-Mann, Ilia Peiros, George Starostin

Outline

- Brief history of languages at SFI
- The steps toward quantitative phylogenetics
  - Example I: lexical reconstruction
  - Example II: inferring models of word-order change
- New opportunities for empirical discovery

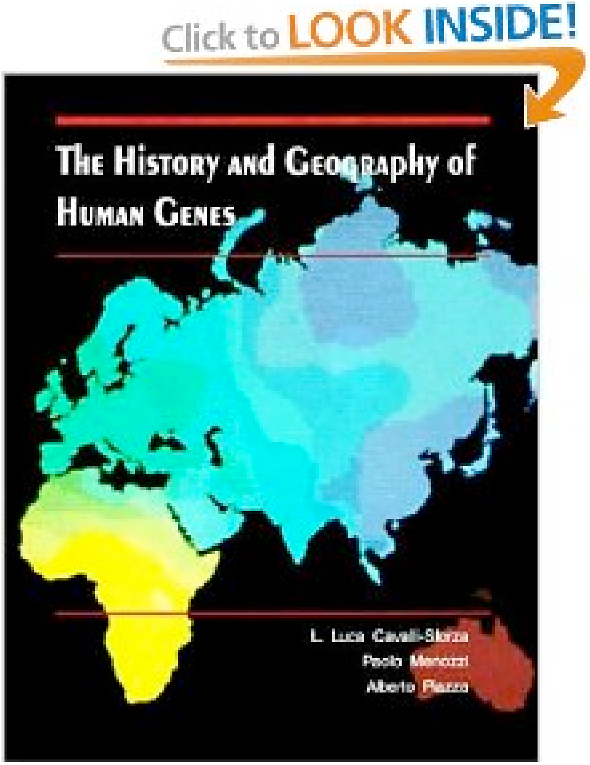
The Evolution of Human Languages Program at SFI

<http://ehl.santafe.edu/>
<http://starling.rinet.ru/main.html>

Goals: Deep reconstruction of language history and connection with genes and the archaeological record

Murray Gell-Mann

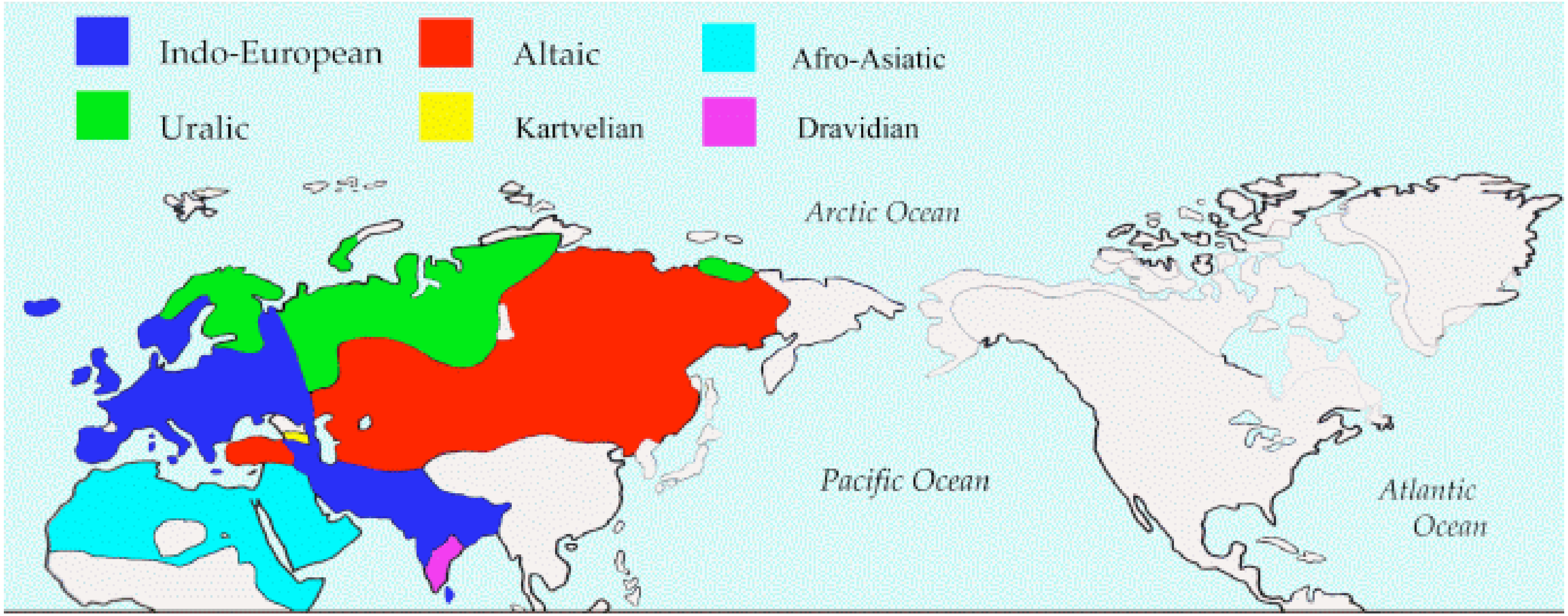
George Starostin



Ilia Peiros

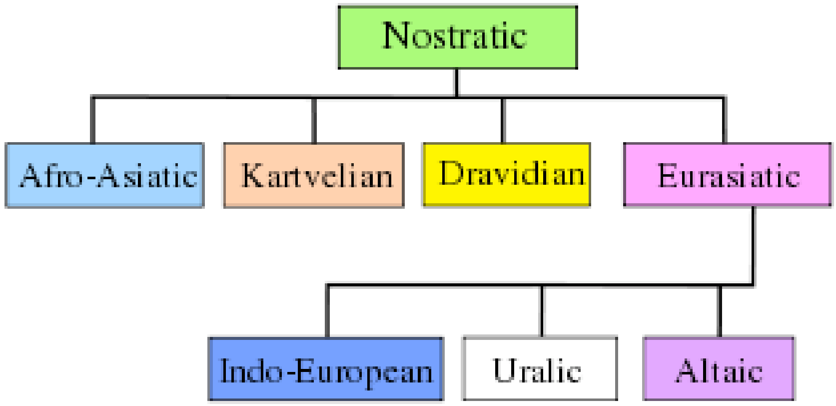
Sergei Starostin

Deep reconstruction: 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

Methods: 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)

eins	one
zwei	two
drei	three
vier	four
fünf	five
…	…
Tier	animal
Reh	deer
Hund	dog
Vater	father
Mutter	mother
Milch	milk
Finger	finger
Arm	arm
Bein	leg
…	…

Sound Map

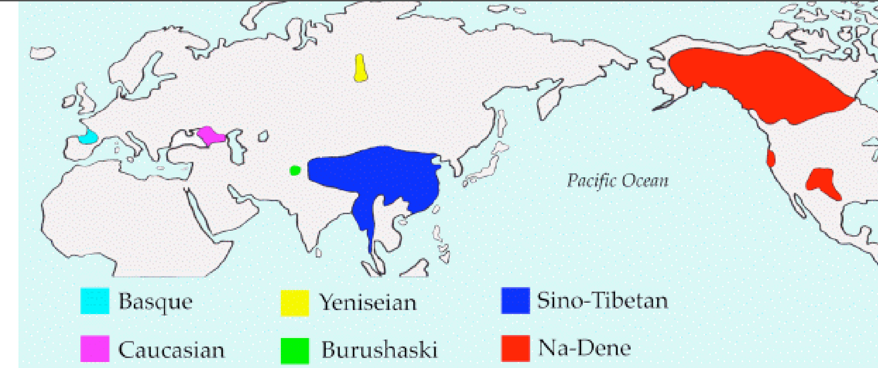
$$P_{\mathrm{Preserve}}(\mathrm{word}) = e^{-t/\tau}$$

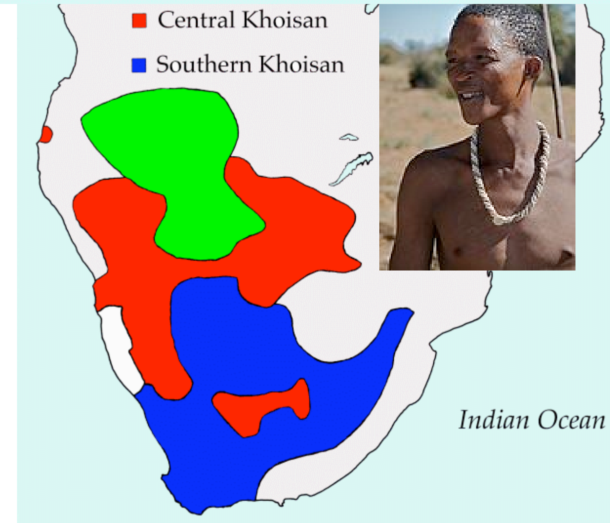
$$\Delta t_{\mathrm{sep}} = -\tau \log\left(\mathrm{frac.\ \ preserved}\right)$$

EHL today

- The Dene-Caucasion hypothesis
- Borean? Relic of the last ice-age?
- Reconstructions of Dravidian, Khoisan, ...

- The EHL database and Starling





Turkic etymology :

New query

Total of 2017 records 101 page

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

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
*Āb ⊞	Altaic etymology ⊞	hunt, chase	охота	*ab* (Orkh.), *av* (OUygh.)	*av* (MK)	*av*	*aw*	*aw* (Pav. C.)	*ɔv*	*aw*, dial. *ɒ*		*ov*	*āv*
*ab- ⊞	Altaic etymology ⊞	to crowd, come together	собираться, встречаться	av- (OUygh.)	av- (MK, KB)								
					avut (MK),								

Concepts and steps in a quantitative phylogenetics

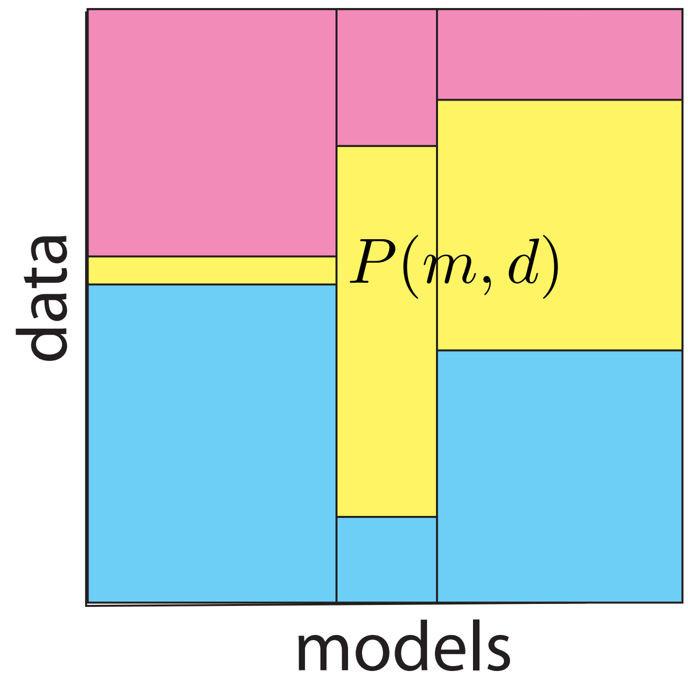
- Roles of Likelihood and Bayesian methods
  - Frequent-pattern versus rare-feature innovations
  - Bayes's theorem and prior prejudice
  - Typological constraints and Bayesian priors
  - Information criteria and significance of parameters
- The likelihood part of a phylogenetic algorithm
  - Overall structure of sound and meaning change
  - Alignment, sound correspondence, and errors
  - Context discovery, detection of borrowings
  - Classification and reconstruction

Rare innovations versus clusters of common innovations

- Rare innovations: single features with ~0 probability to occur by chance (go in Bayesian priors)
  - 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: (estimate with likelihood)
  - 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

Bayes's Theorem and model comparison

$$\sum_d P(m,d) \quad \equiv \quad P(m)$$



$$\sum_m P(m,d) \quad \equiv \quad P(d)$$

- Represent both data *and models* with a joint probability
- Split joint probability into conditionals either of two ways

$$\begin{aligned} P(m,d) &= P(m \mid d)\, P(d) \\ &= P(d \mid m)\, P(m) \end{aligned}$$

- Bayes's theorem: priors and likelihoods

$$\begin{aligned} P(m \mid d) &= \frac{P(d \mid m)\, P(m)}{P(d)} \\ &= \frac{P(d \mid m)\, P(m)}{\sum_{m'} P(d \mid m')\, P(m')} \end{aligned}$$

Bayesian prior

Bayesian posterior

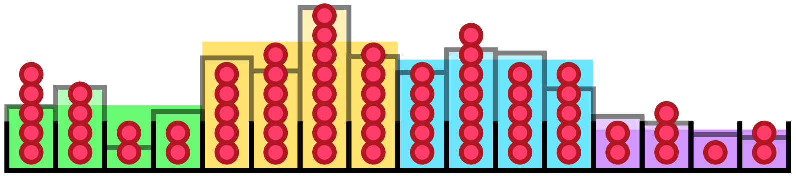
Typological constraints are a natural domain for priors (Ian's lecture)

- Three roles for priors
  - Include frequency evidence from outside this sample
  - Include non-frequency evidence (rare innovations)
  - Represent out-of-field evidence (molecular phylogenies)
- On states
  - phoneme inventory, word order, ...
  - implicational relations (pronouns, time, color, aspect)
- On transitions
  - phoneme contexts, intermediate word-order states (NDO)
  - geometric models of phonology or semantics?

Akaike and Bayesian Information Criteria

$$\mathrm{AIC} \equiv -2\log(L) + 2k$$

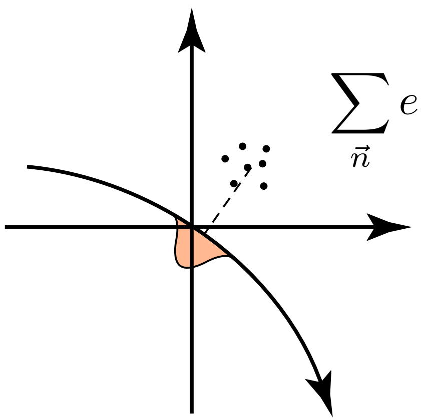
$$\mathrm{BIC} \equiv -2\log(L) + k\log(n)$$



$$P(n_1,\ldots n_K) = \prod_{i=1}^{K} p_i^{n_i}\left(\frac{N!}{n_1!,\ldots,n_K!}\right) = e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)}$$

$$D\left(\frac{\vec{n}}{N}\,\|\,\vec{p}\right) \equiv \sum_{i=1}^{K}\frac{n_i}{N}\log\frac{\frac{n_i}{N}}{p_i}$$

Kullback-Leibler divergence,
or Relative Entropy



$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)}\delta\left(\frac{\tilde{n}}{N}-\tilde{p}\right) = \rho(\tilde{p})$$

$$\mathcal{L}(\tilde{n}_1,\ldots,\tilde{n}_k) = \prod_{j=1}^{k}\tilde{p}_j^{\tilde{n}_j}$$

Likelihoods & their maxima

$$\tilde{p}_i \to \frac{\tilde{n}_i}{N} \qquad \max_{\tilde{p}}\log\mathcal{L}(\tilde{n}_1,\ldots,\tilde{n}_k) = N\sum_{j=1}^{k}\frac{\tilde{n}_j}{N}\log\frac{\tilde{n}_j}{N}$$

$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)} N\sum_i \tilde{p}_i\log\tilde{p}_i \approx N\sum_i \tilde{p}_i(\vec{p})\log\tilde{p}_i(\vec{p}) + \frac{k}{2}$$

(Distribution of max-likelihoods)

$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)}\sum_i \tilde{n}_i(\vec{p})\log\tilde{p}_i \approx N\sum_i \tilde{p}_i(\vec{p})\log\tilde{p}_i(\vec{p}) - \sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)}\frac{N}{2}\sum_{j=1}^{k}\frac{\left(\tilde{p}(\vec{n}/N)-\tilde{p}(\vec{p})\right)^2}{\tilde{p}_i(\vec{p})}$$

$$= N\sum_i \tilde{p}_i(\vec{p})\log\tilde{p}_i(\vec{p}) - \frac{k}{2}$$

$$\sim N\sum_i \tilde{p}_i\log\tilde{p}_i - k$$

Maximize average likelihood of real samples over average models

Unbiased sample estimator!

More detail on Akaike derivation

$$\mathcal{L}(\tilde{n}_1, \ldots, \tilde{n}_k) = \prod_{j=1}^{k} \tilde{p}_j^{\tilde{n}_j}$$

Likelihood of *any* data given a particular model

Average log-likelihood of actual data, from a model produced with fixed estimated parameters p~

$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)} \log \mathcal{L}(\tilde{n}(\vec{n}) \mid \tilde{p}) = \sum_i \tilde{n}_i(\vec{p}) \log \tilde{p}_i$$

(average) (data) (model)

Now this averaged log-likelihood, averaged over estimated *models*; 2nd-order Taylor exp'n

$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)} \sum_i \tilde{n}_i(\vec{p}) \log \tilde{p}_i(\vec{n}) \approx N \sum_i \tilde{p}_i(\vec{p}) \log \tilde{p}_i(\vec{p}) - \sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)} \frac{N}{2} \sum_{j=1}^{k} \frac{\left(\tilde{p}(\vec{n}) - \tilde{p}(\vec{p})\right)^2}{\tilde{p}_i(\vec{p})}$$

$$= N \sum_i \tilde{p}_i(\vec{p}) \log \tilde{p}_i(\vec{p}) - \frac{k}{2}$$

(average) (estimated models)

But we don't have the ideal p~s; the best we can do is obtain an unbiased estimator from any single sample; for this we need to identify the bias typical of samples

$$\sum_{\vec{n}} e^{-ND\left(\frac{\vec{n}}{N}\|\vec{p}\right)} N \sum_i \tilde{p}_i \log \tilde{p}_i \approx N \sum_i \tilde{p}_i(\vec{p}) \log \tilde{p}_i(\vec{p}) + \frac{k}{2}$$

(sample ML) (ideal ML) (variance correction)

Use this to replace the ideal ML with an unbiased estimator from samples, get previous slide

Word lists are the starting point for lexical (= phonological / semantic) reconstruction

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

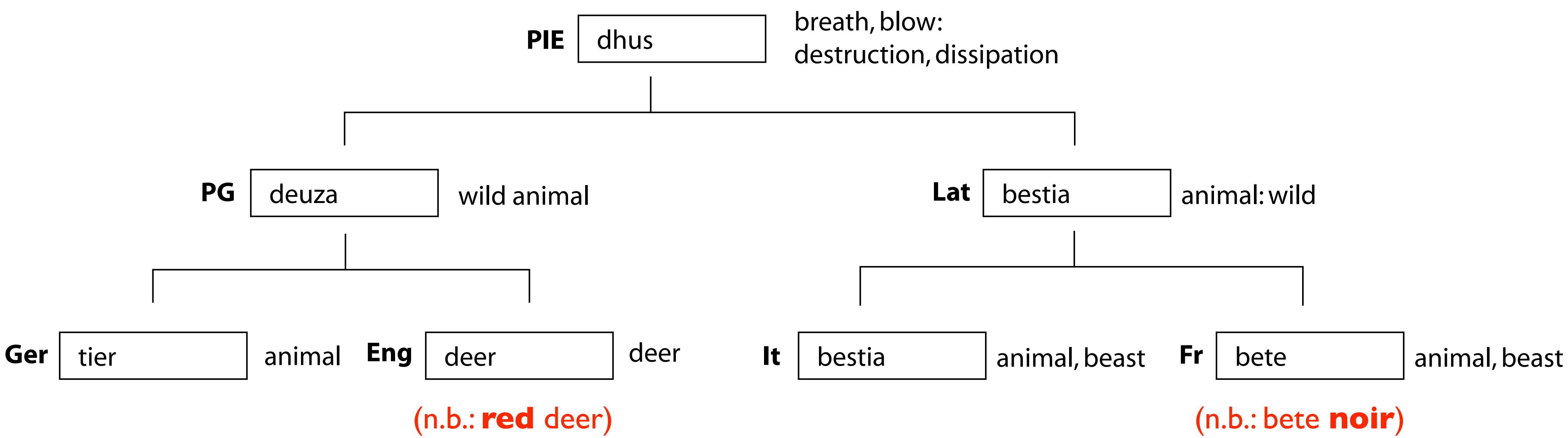
Turkic etymology :

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
*Āb ⊞	Altaic etymology ⊞	hunt, chase	охота	*ab* (Orkh.), *av* (OUygh.)	*av* (MK)	*av*	*aw*	*aw* (Pav. C.)	ɔv	*aw*, dial. *ɵ*		*ov*	āv
*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.)	xɔwuč	*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	avičγa, abučγa 1, abučqa 2 (OUygh.)	avičγa 1 (MK, KB)	abuš 3 dial.	abušqa, awucqa 1 dial. (Sib.)	abušqa, avušqa 4 (Abush., Sangl.)					

Objects that must be modeled are joint histories of sound and meaning for a collection of words

- 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 (Bill's lecture)

- 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

eins	one
zwei	two
drei	three
vier	four
fünf	five
…	…
Tier	animal
Reh	deer
Hund	dog
Vater	father
Mutter	mother
Milch	milk
Finger	finger
Arm	arm
Bein	leg
…	…

Sound Map

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
vier
funf

Eng

one
two
three
four
five

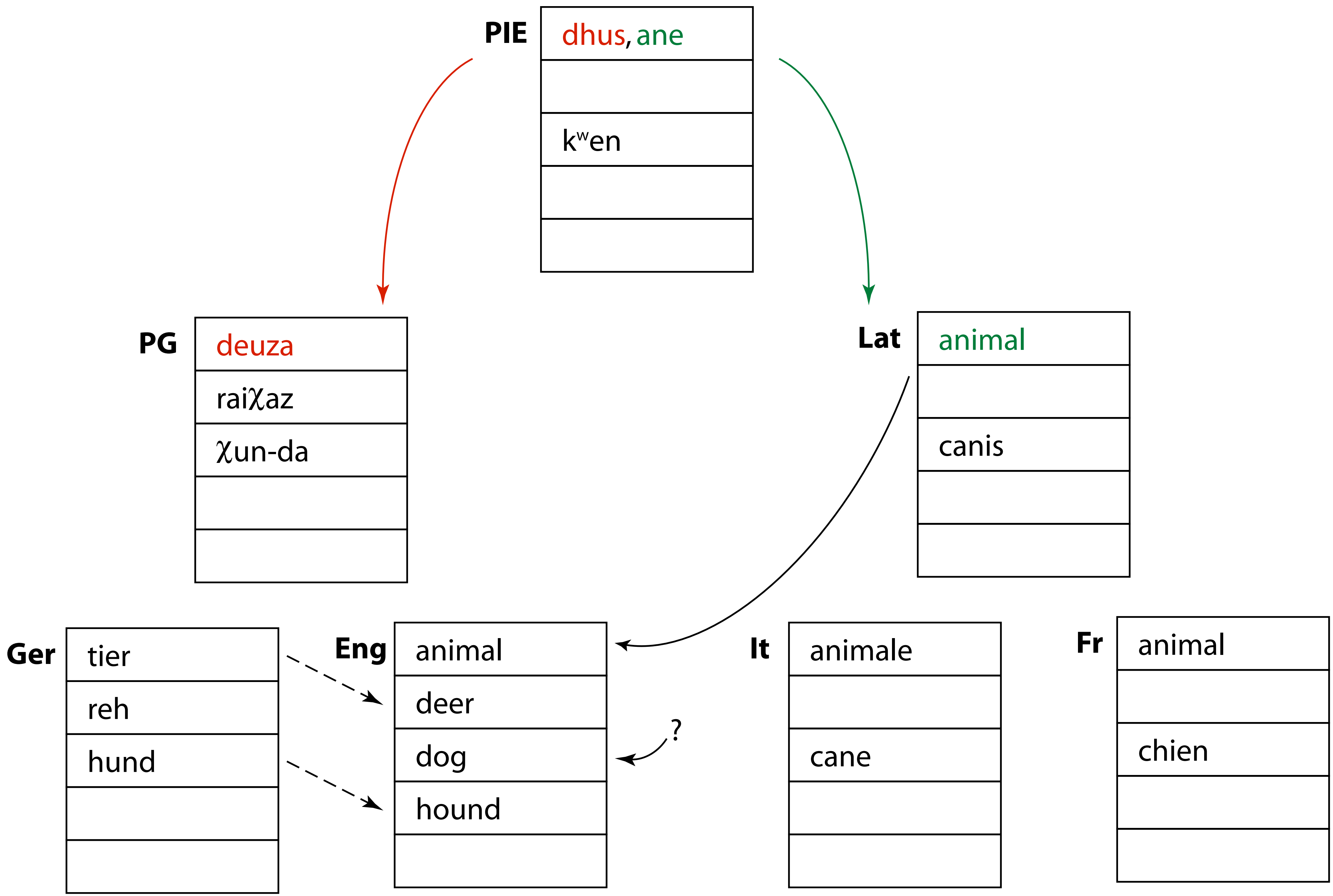
It

uno
due
tre
quattro
cinque

Fr

un
deux
trois
quatre
cinq

Sound maps help identify meaning shifts



The "alignment problem": what to compare w/ what?

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"

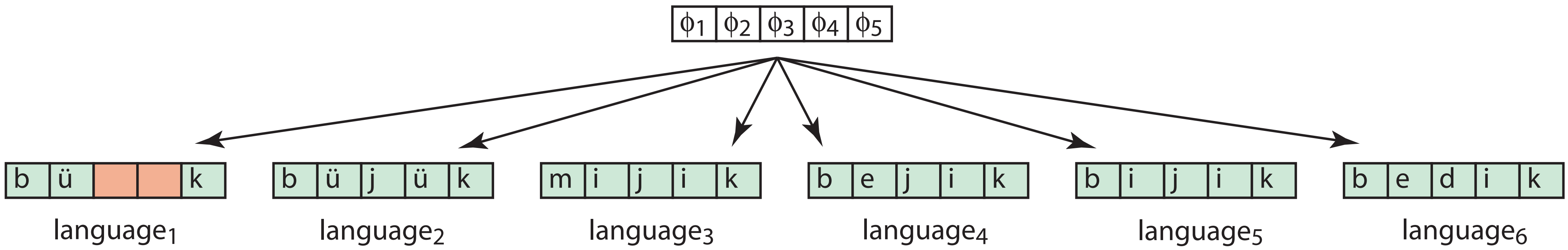
b	b	b	b	b	b	b	b	b	b	p	m	b	b	b	b	b	b	b	b	m	b	b	b		b
e	e	ü	i	e	u	ü	e	ö	e	ö	ö	i	i	e	e	i	i	i	e	i	ǖ	ü	i		i
d	δ	j	j	j	j	j	z	j	j	z	z	j	d	d	d	j	j	j	j	j		j	j		j
ü	ü	ü	e	i	u	ü	ɨ	ü	i	ə	ü	i	i	i	i	i	i	i	e	i		ü	i		i
k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k	k		k

Etymology 12: "Breast, nipple"

m	m	m	m	m	m	m	m	m	m	m	m
e	ɛ	ä	ä	ǟ	ǟ	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			

Maximum-likelihood estimation of history and process (phonological only, here)

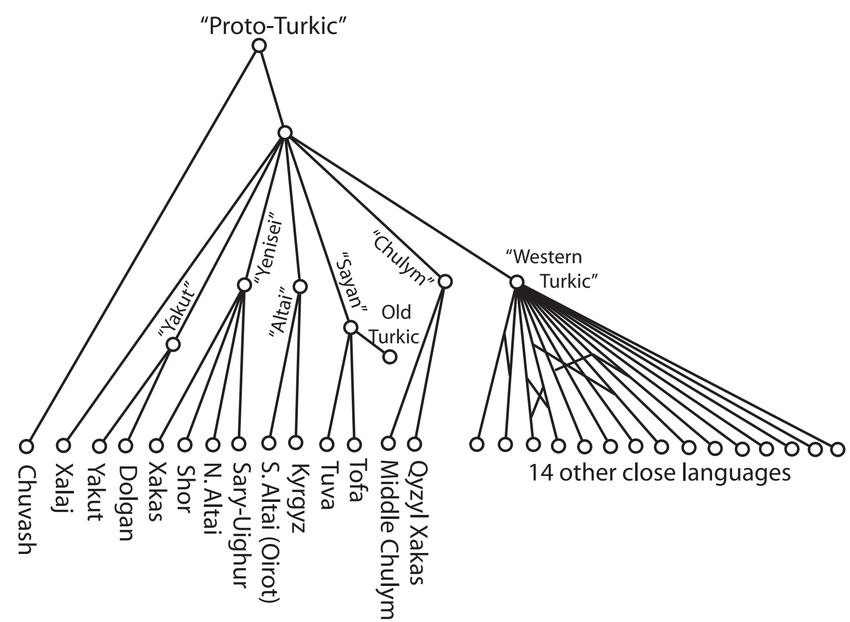
- Suppose we have proposed an alignment of positions in the daughter languages
- Propose phoneme assignments to aligned positions in the ancestor (with probabilities)
- Estimate regular correspondence of ancestor to daughter phonemes (w/ or w/o probabilities)
- Estimate random violations (with probabilities)



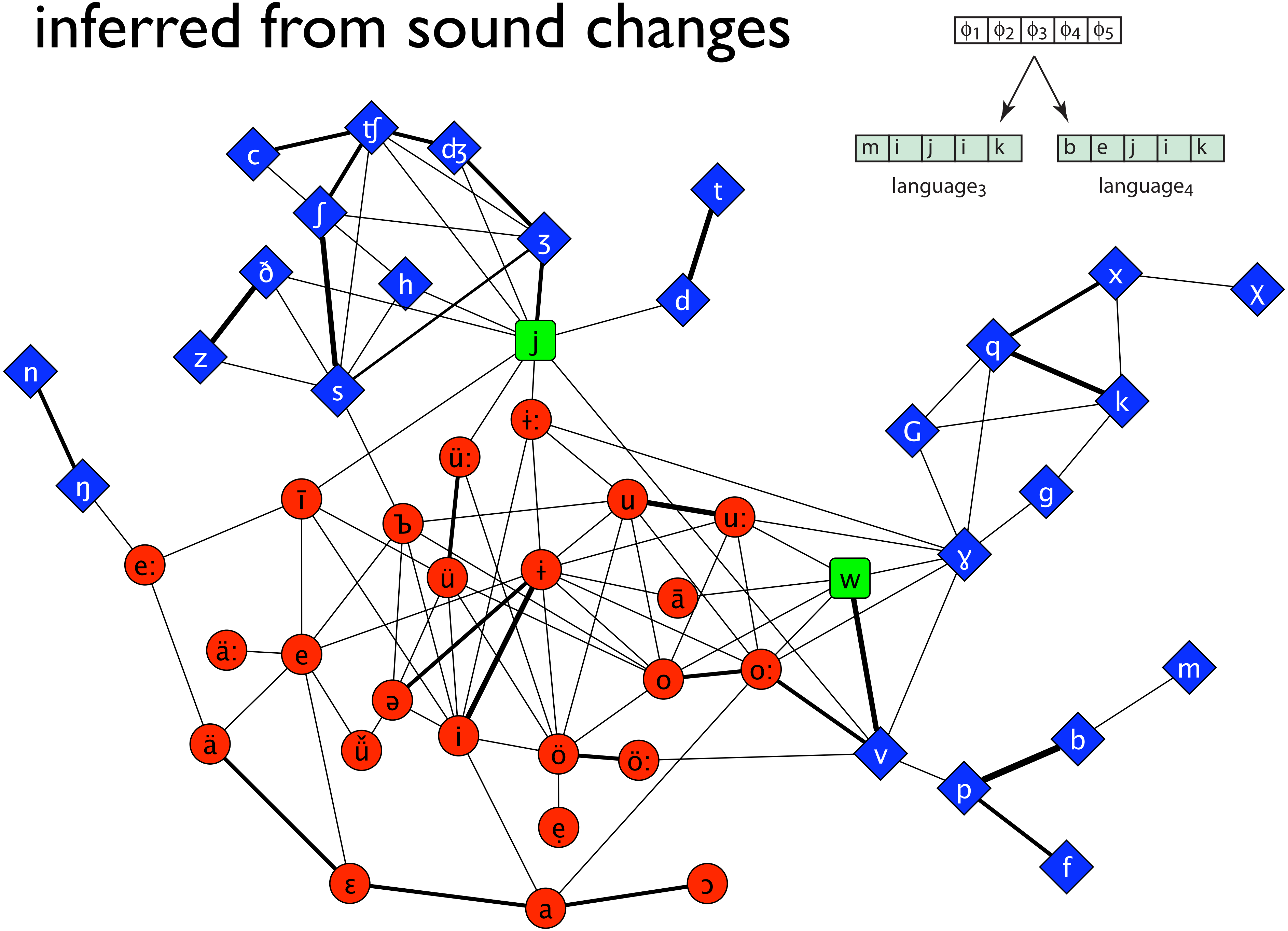
Sound correspondences among the languages

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	$ə^w$
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

SA Starostin, AV Dybo and OA Mudrak 2003. An Etymological Dictionary of Altaic Languages. Leiden: Brill



Typology I: sound relations and features inferred from sound changes



Context dependence: an Artificial Intelligence problem

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

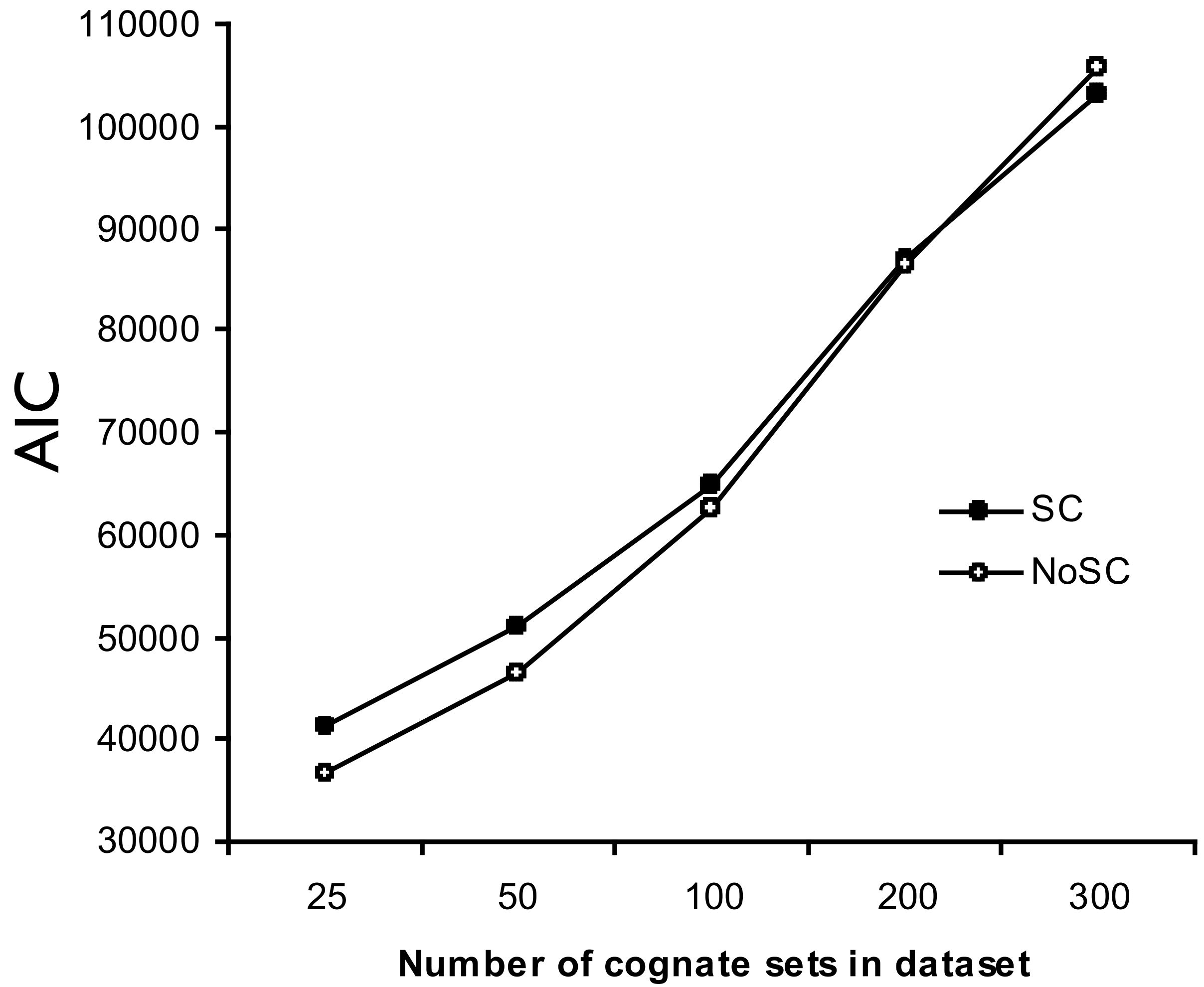
From R.L. Trask, "Language change"

- Sound change can be regular, but not at single-phoneme level
- Conditioning contexts can be lost; must be guessed
- AIC and BIC can be used to judge guesses

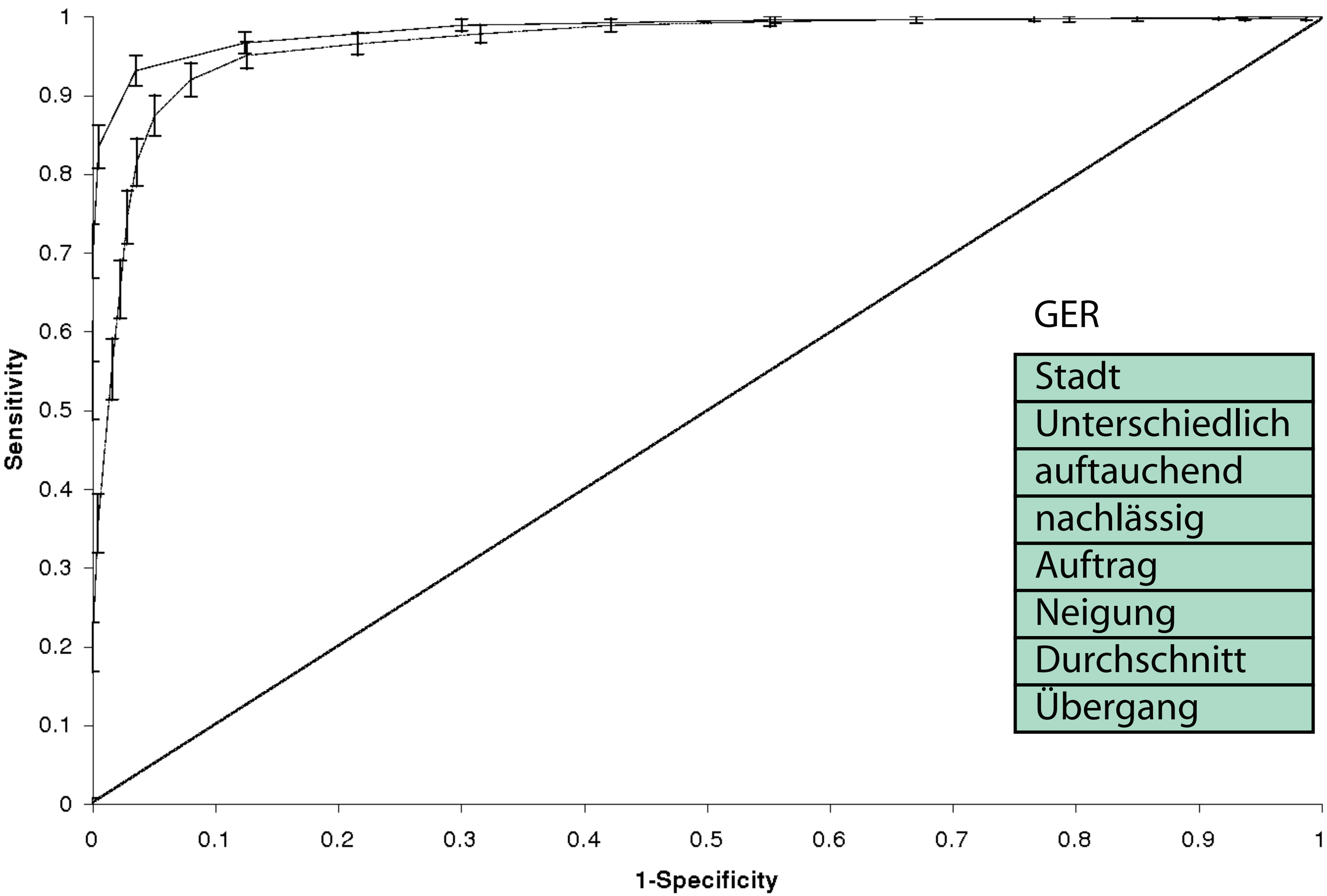
Vowel harmony in the Turkic language family: a predictive supra-segmental context

Front/back agreement in roots and affixes

nom	a book
inek	a cow
nomnar	books
inekter	cows



Identification of borrowings: words that do not fit the system pattern

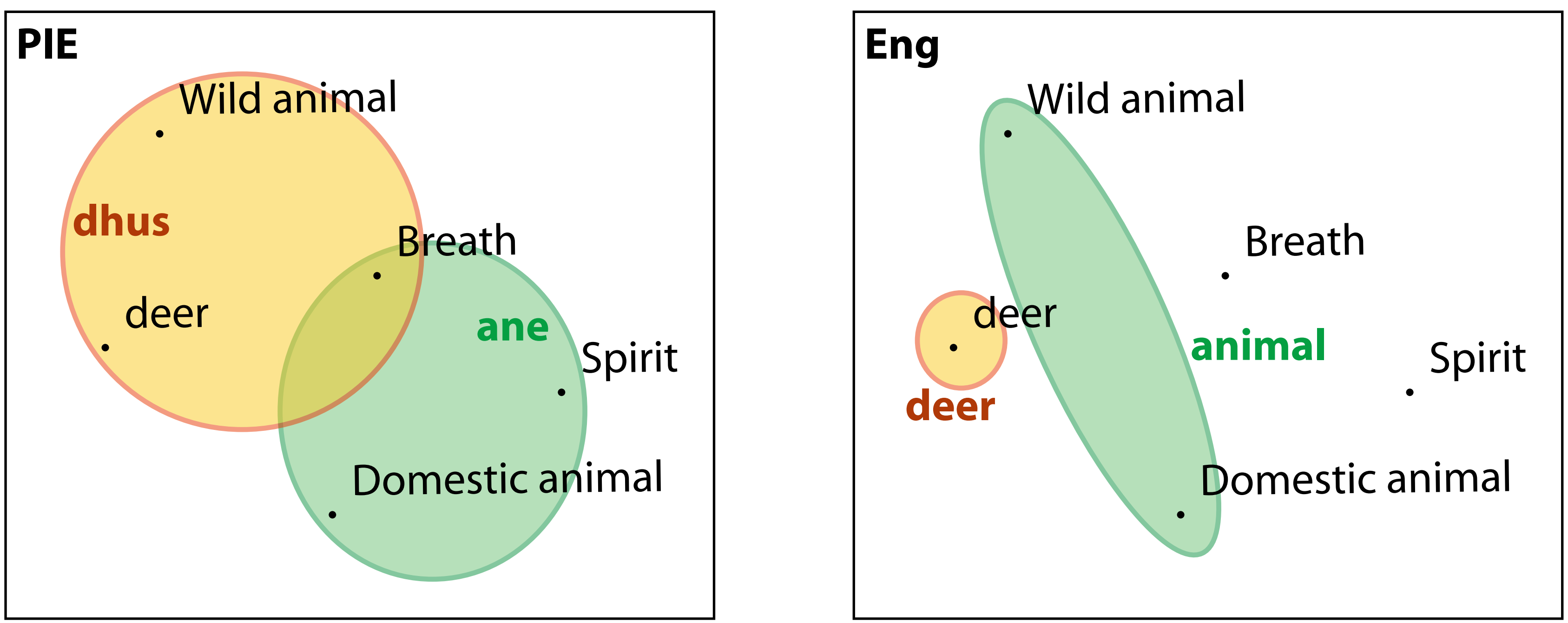


GER	ENG
Stadt	city
Unterschiedlich	different
auftauchend	emergent
nachlässig	negligent
Auftrag	mission
Neigung	passion
Durchschnitt	intersection
Übergang	transition

Receiver **O**perating **C**haracteristic curve

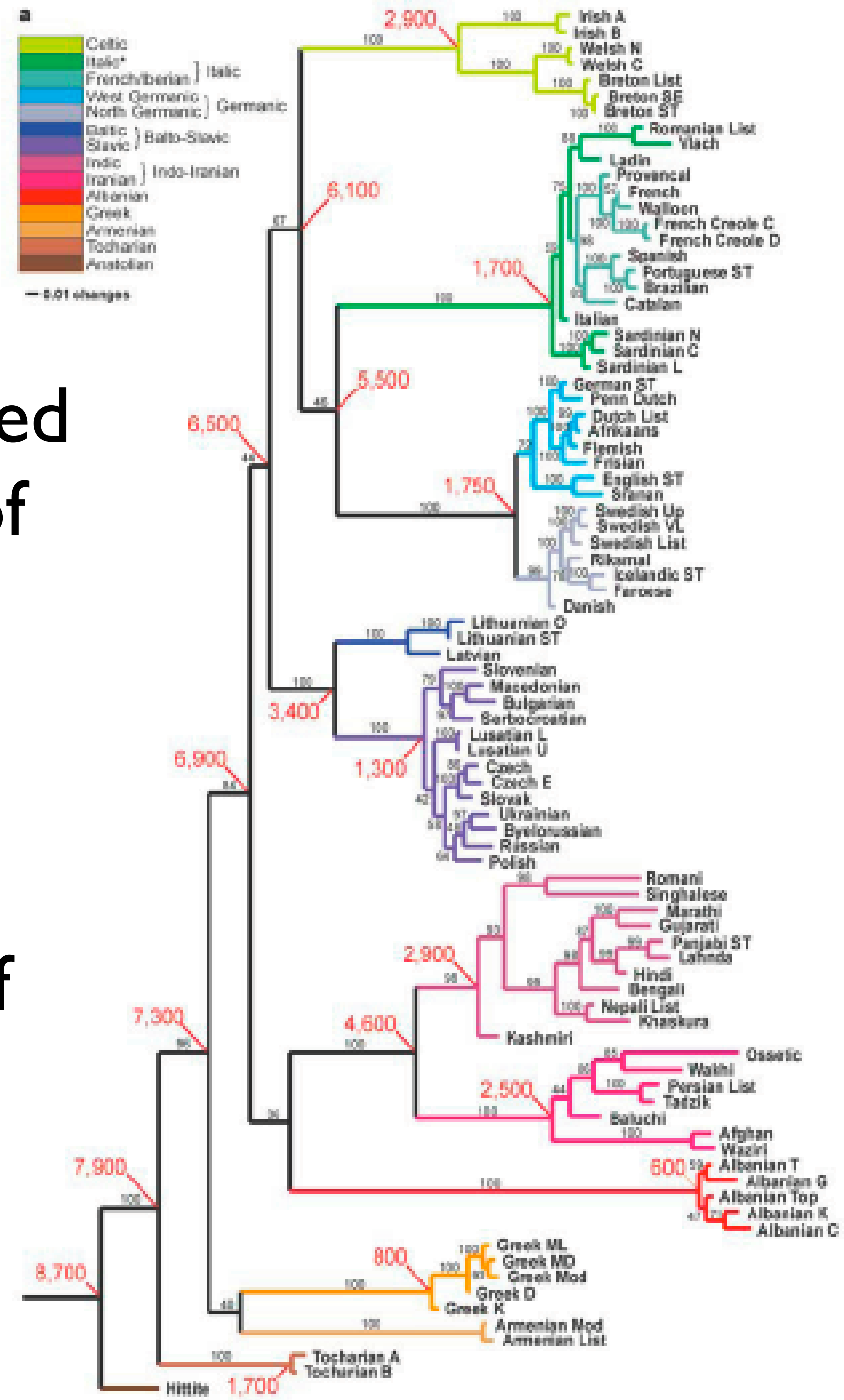
New conceptual domain: mathematical likelihood modeling of semantic shift

- 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



Phylogeny and reconstruction

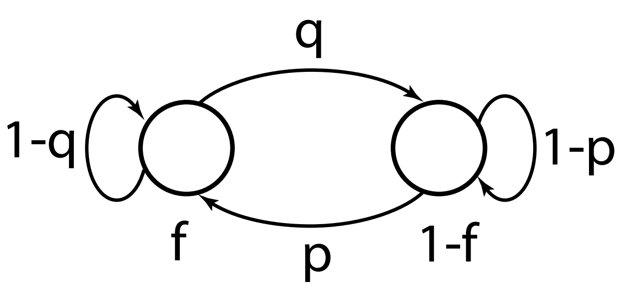
- Modern glottochronology applied to expert linguists’ judgments of cognate classifications
- Presence/absence data format modeled after genes
- *Not* yet a model of processes of sound and meaning change



Russell D. Gray & Quentin D. Atkinson

NATURE | VOL 426 | 27 NOVEMBER 2003 | www.nature.com/nature

Maslova: how much can you do with incomplete reconstructions of the past?

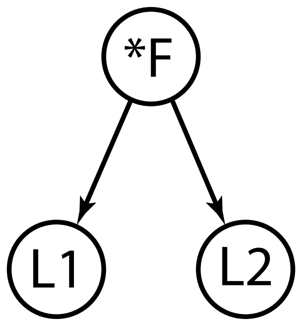


Two quantities: stationary frequency
typical number of pairs

$$\bar{f} = \frac{p}{p+q}$$

$$\bar{h} = (p+q)^2\,[2-(p+q)]\,2\bar{f}\left(1-\bar{f}\right)$$

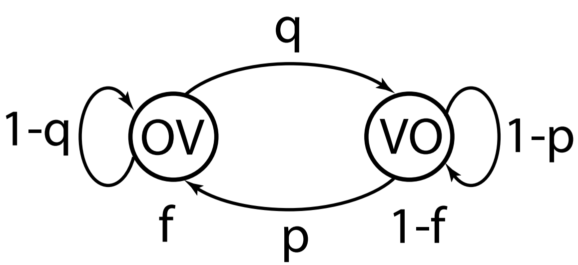
Regression model for observed frequencies estimate (p/q): $f_i = \bar{f} + \epsilon_i$



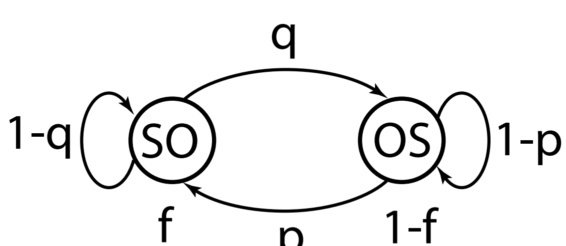
Fit of number of pairs against mean,
in children of a common family ancestor:
put bounds on (p+q), p/q

$$h_i - \bar{h} = \left(f_i - \bar{f}\right) 2\,(p+q)\left(1-2\bar{f}\right) + \varepsilon_i$$

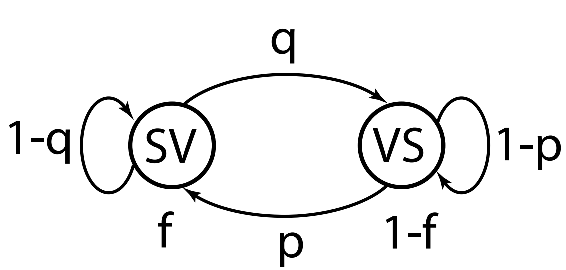
Three partitions of standard word order



$$\begin{aligned} f(OV) &\approx 0.53 \\ \frac{1}{p+q} &\approx 24\text{ky} \end{aligned}$$



$$\begin{aligned} f(SO) &\approx 0.96 \\ \frac{1}{p+q} &\leq 11\text{ky} \\ p/q &\geq 15 \end{aligned}$$



$$\begin{aligned} f(SV) &\approx 0.86 \\ \frac{1}{p+q} &\in (10\text{ky}, 30\text{ky}) \\ p/q &\geq 3 \end{aligned}$$

New opportunities for quantitative characterization of regularities

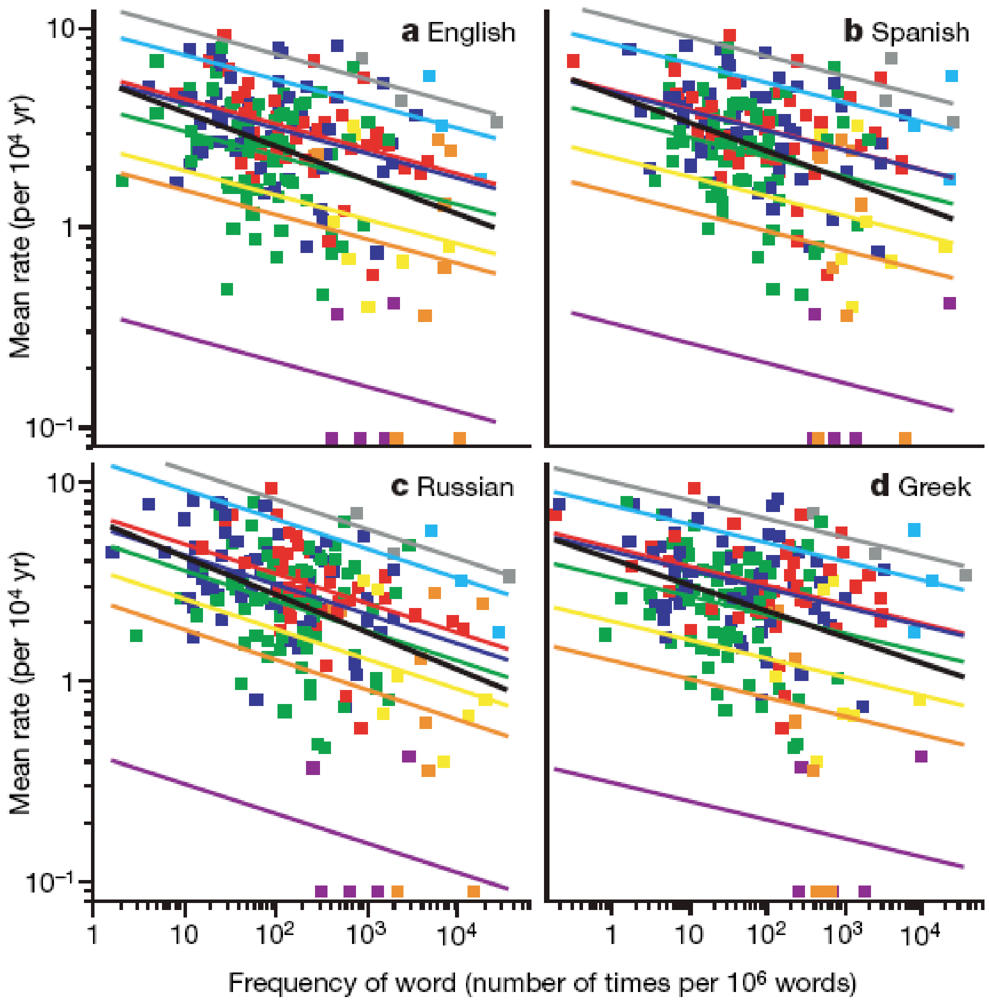
- Cross-linguistic regularities in frequency of use, and relations to rates of change
- Punctuated equilibrium and correlations of the "clock" of language change with culture
- Polysemy, synonymy, and semantics
- Full speaker-corpus archives (Norquist et al.), formant-based analysis (Labov), ...

Typology II: 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]

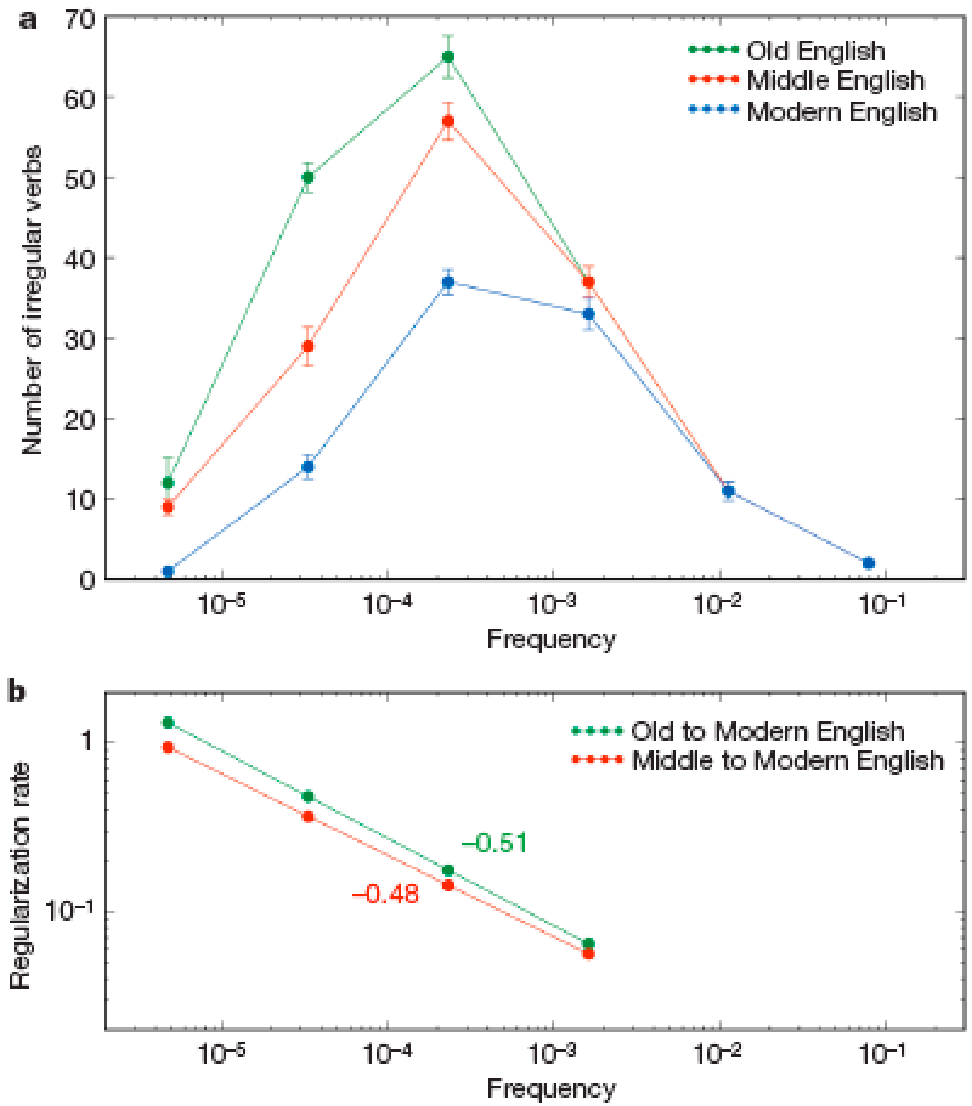
NATURE | Vol 449 | 11 October 2007



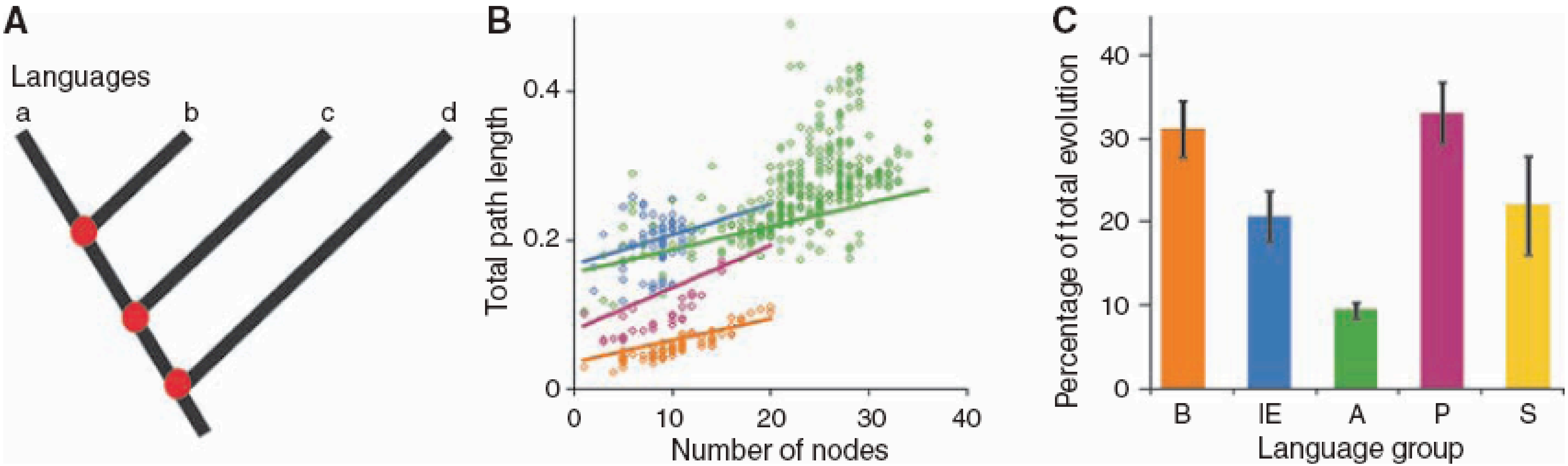
Quantifying the evolutionary dynamics of language

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: do cultural phenomena like splitting drive language divergence?

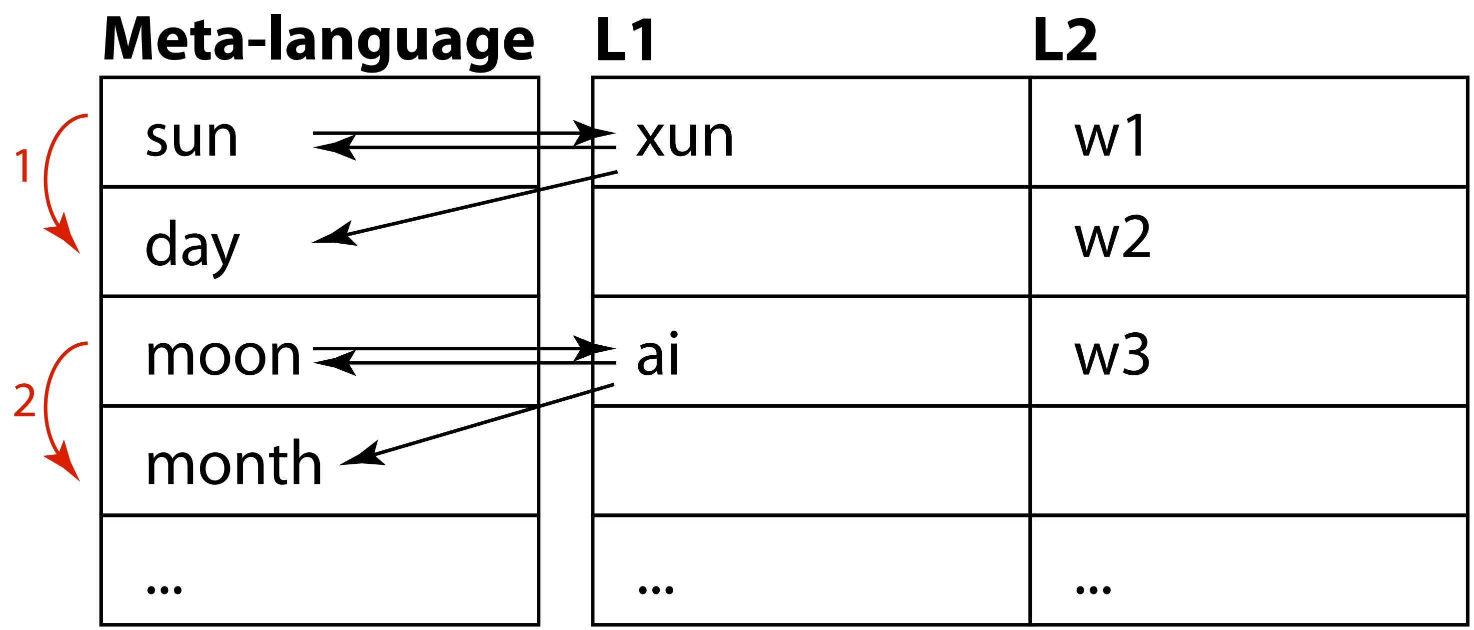


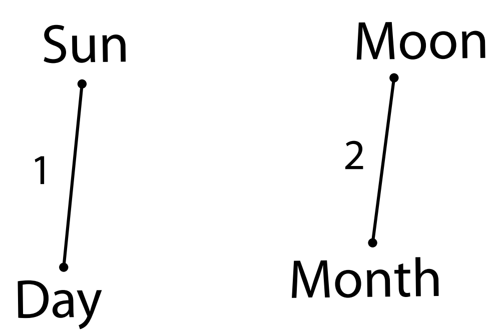
Languages Evolve in Punctuational Bursts

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

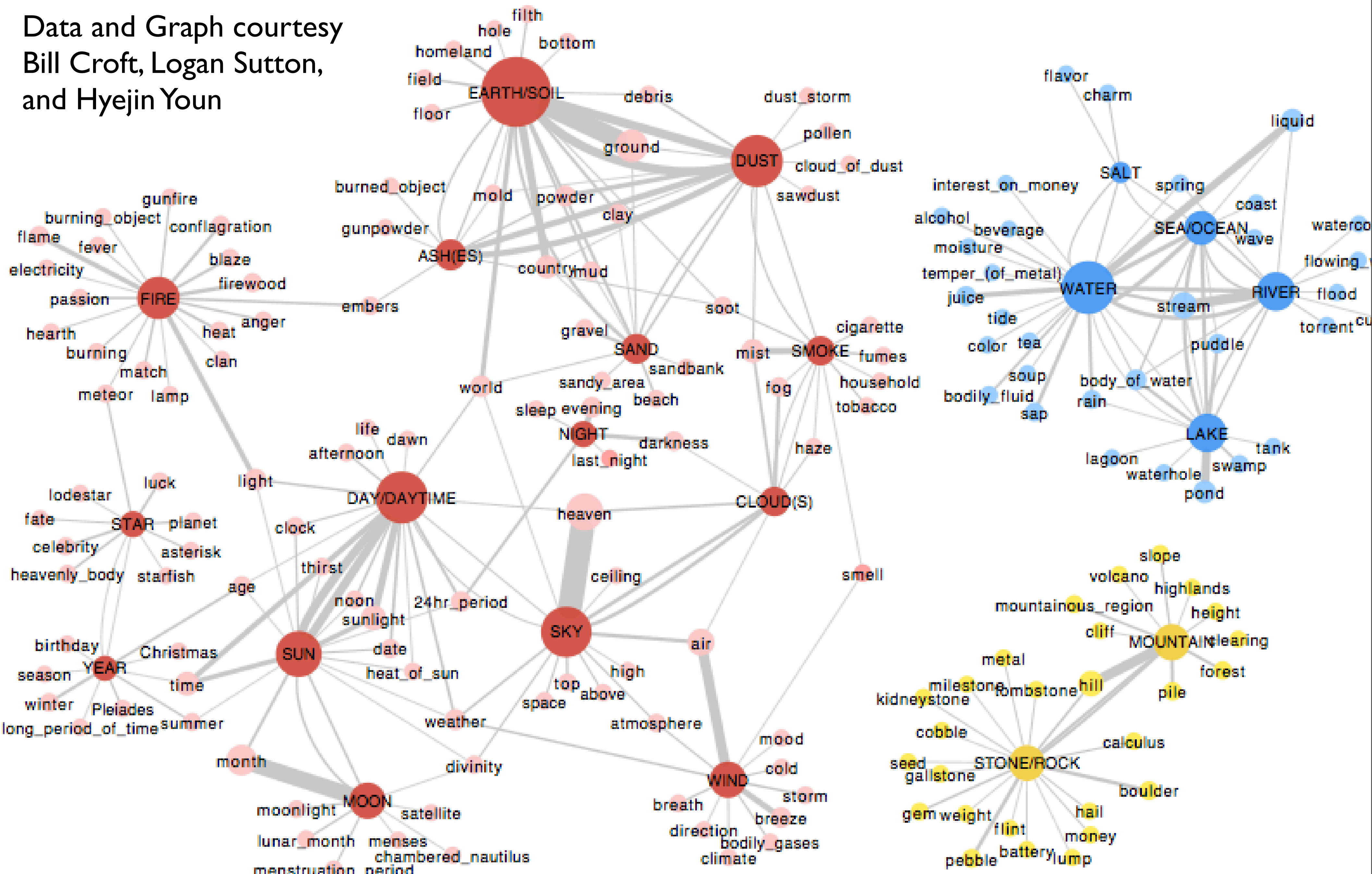
Toward a likelihood model of semantic shift: Inferring polysemy with English as a meta-language





Network of polysemes in 81 diverse languages

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



Summary comments

- Much linguistics traditionally modeled on logic;
  Can we root historical linguistics in probability?
- Language evolution is *not* molecular evolution;
  Same principles lead to different models
- New and refined quantitative signatures: Both methodological and conceptual opportunities