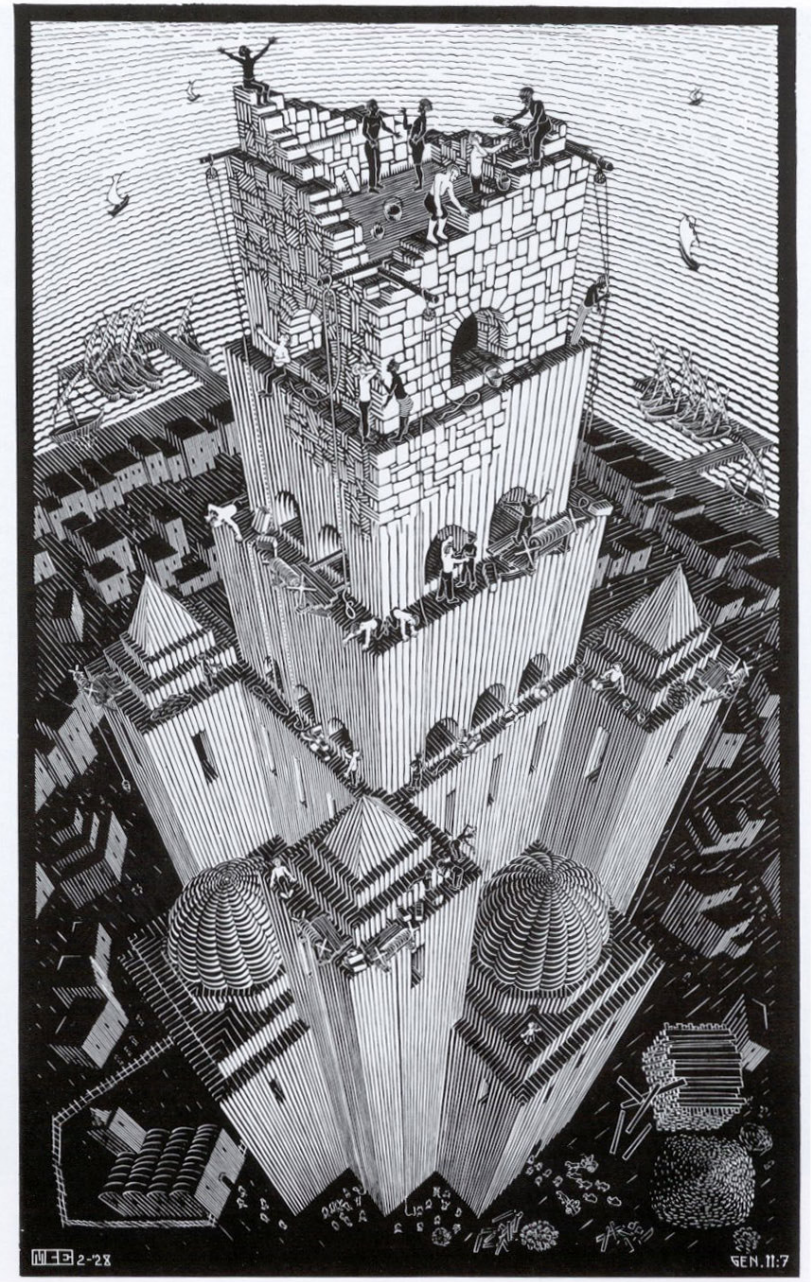
Network models of sound change

Dan Hruschka
Santa Fe Institute
dhrusch@santafe.edu

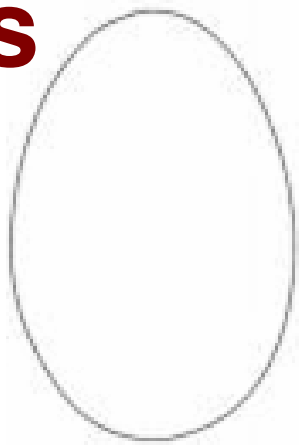
in collaboration with Tanmoy Bhattacharya, Eric Smith & Jon Wilkins

"Statistical Inference for Complex Networks"
December 2008



“Egg” in 24 Turkic languages

- jumurtGa
- simit
- umurxa



j	j	j	j	j	j		j	j	j	n	n	d	j	s	s	h	č		ǯ	ž	j	j	z
u	u	u	o	u	u		o	u	u	ɨ	ɨ	ɨ	u	ы	ɨ	ɨ	u	u	u	u	u	o	u
m	m	m	m	m	m		m	m	m	m	b	m	m	m	m	m		m	m	m	m	m	n
u	u	u	ɨ	u	u		u	u	u	ɨ	ɨ	ɨ	u	a	ī	ī	u	u	u	ɨ	ɨ	o	u
r	r	r	r	r	r			r	r	r	r	r	r	r			r	r	r	r	r	r	r
t	t	t		t	t		t	t	t		t	t		d	t	t			t	t	t	t	t
ɣ	ɣ		q	ɣ	q				G	x	q	q	q				ɣ	x	q	q	q	q	q
a	a	a	a	a	a			a	a	a	a	a	a	a			a	a	a	a	a	a	a

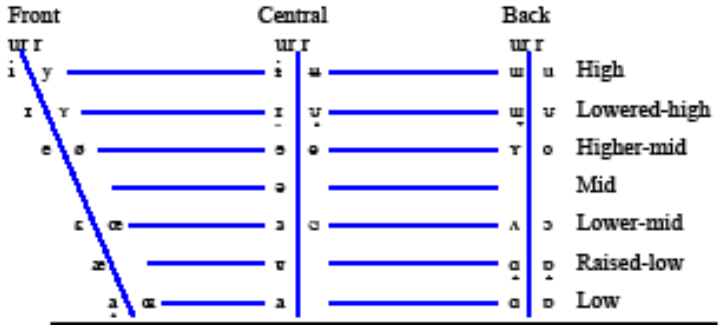
The Plan

Question—What theories of sound change best account for observed diversity in related languages?

- Brief introduction—concepts, units
- Specify one model in a likelihood framework
- Discuss plans for specifying other models and raise issues of inference and model selection

API Chart : UPSID 451 - BDPROTO (Dynamique Du Langage, Font IPALA) 2004, Christian Fressard, Ian Maddieson, Egidio Marsico

Vowels



Tones :

Extra high | High | Mid | Low | Extra low | Rising | Falling

Diacritics

Voiceless	n̥ e̥	Retroflex	ɵ̩	Advanced	t̟ u̟
Voiced	s̬	Nasal-release	dⁿ	Retracted	t̠ ɪ̠
Aspirated	tʰ eʰ	Lateral-release	dˡ	Centralized	ü
Breathy voiced	b̤ e̤	Breathy-release	dʱ	Raised	ɪ̝ ɑ̝
Creaky voiced	b̰ ḛ	No-audible-release	dˀ	Lowered	β̞ e̞
Labialized	tʷ eʷ	Lip-compressed	m̫	Retroflexed	aʴ
Palatalized	tʲ eʲ	Syllabic	n̩	Long	tː eː
Velarized	tˠ eˠ	Preglottalized	ʔd ʔe	Overshort	e̽
Pharyngealized	tˤ eˤ	Preaspirated	ʰt ʰe	ATR	e̘
Nasalized	ẽ	Prenasalized	nt	Dental	t̪
Ejective / Glottalized	tʼ eʼ	Prestopped	dl	Apical	t̺

Other symbols

Alveolar lateral flap	ɺ
Simultaneous ʃ and x	ɧ
Voiced alveolar r-sound	r͡r
Voiced alveolar fricative trill	r̝
Voiced alveolar affricate trill	d͡r
Voiceless alveolar affricate trill	t͡r
Voiced alveolar affricate lateral	d͡ɮ
Voiceless alveolar affricate lateral	t͡ɬ
Voiceless velar alveolar affricate lateral	k͡ɬ
Voiced alveolar fricative flap	ɾ̝
Voiceless alveolar affricate-lateral click	ǁ

Consonants

		Labial				Coronal					Dorsal			Guttural			
		Bilabial	Labiodental	Labial-palatal	Labial-velar	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Velar-alveolar	Uvular	Pharyngeal	Laryngeal	Epiglottal	Glottal
Obstruents	Stop	p b	p̪ b̪		k͡p g͡b	t̪ d̪	t d	t̠ d̠	ʈ ɖ	c ɟ	k g		q ɢ			ʡ ʢ	ʔ ʔ̬
	Implosive	ɓ ɓ			ɠ̊ ɠ		ɗ ᶑ			ʄ ʄ	ɠ̊ ɠ		ɗ ᶑ				
	Fricative sib.					s̪ z̪	s z	ʃ ʒ	ʂ ʐ	ɕ ʑ							
	Fricative ~ sib.	ɸ β	f v			θ ð	θ ð	θ ð	θ ð	ç ʝ	x ɣ		χ ʁ	ħ ʕ	ʜ ʢ		h ɦ
	Affricate sib.		p̪f b̪v			t̪s̪ d̪z̪	ts dz	tʃ dʒ	ʈʂ ɖʐ	cɕ ɟʑ							
	Affricate ~ sib.					t̪θ d̪ð	tθ dð	tɕ dʝ	ʈθ ɖð	cç ɟʝ	kx gɣ		qχ ɢʁ				
	Click	ʘ				ǀ	ǃ	ǂ									
Sonorants Liquids	Trill		ʙ				r		ɾ				ʀ				
	Tap						ɾ	ɾ	ɽ								
	Flap		ⱱ				ɾ	ɾ	ɽ								
	Lat. Fricative					ɬ ɮ	ɬ ɮ		ɬ ɮ		ʟ̝ ʟ̝	ɧ					
	Lat. Approximant					l	l	l	l	ʎ	ʟ ʟ						
	Nasal	m	m̪		ŋ͡m	n̪	n	n̠	ɳ	ɲ	ŋ		N				
	Approximant	ɸ̞ β̞	ʋ	ɥ	w ɰ		ɹ	ɻ	ɻ	j	ɰ		ʁ̞				

Many Theories

- No expectation of what changes to what
- Regularity, but no expectation of what changes to what
- Random walk through feature space
- Lenition (t → ts → s → h)
- Articulatory Reduction
- Other accounts
  - “An eft” becomes “a newt”
  - Sound symbolism—'stamp', stomp', 'tamp', 'tromp', 'tramp'
  - Constraints on language design (e.g. optimality theory)

An Opportunity

- “Although no comprehensive study of sound change that would allow us to distinguish common from uncommon innovations has ever been undertaken, historical linguists have acquired a sense of what kinds of change are likely to occur” (Blust 2004)

Common Methods

- **Sociolinguistics**—Study dialectical variation and change over short periods of time
- **Psycholinguistics**—Study variation in production and perception

Advantage

- Close view of change in action

Disadvantage

- Limited to frequent or fast changes

Historical Linguistics

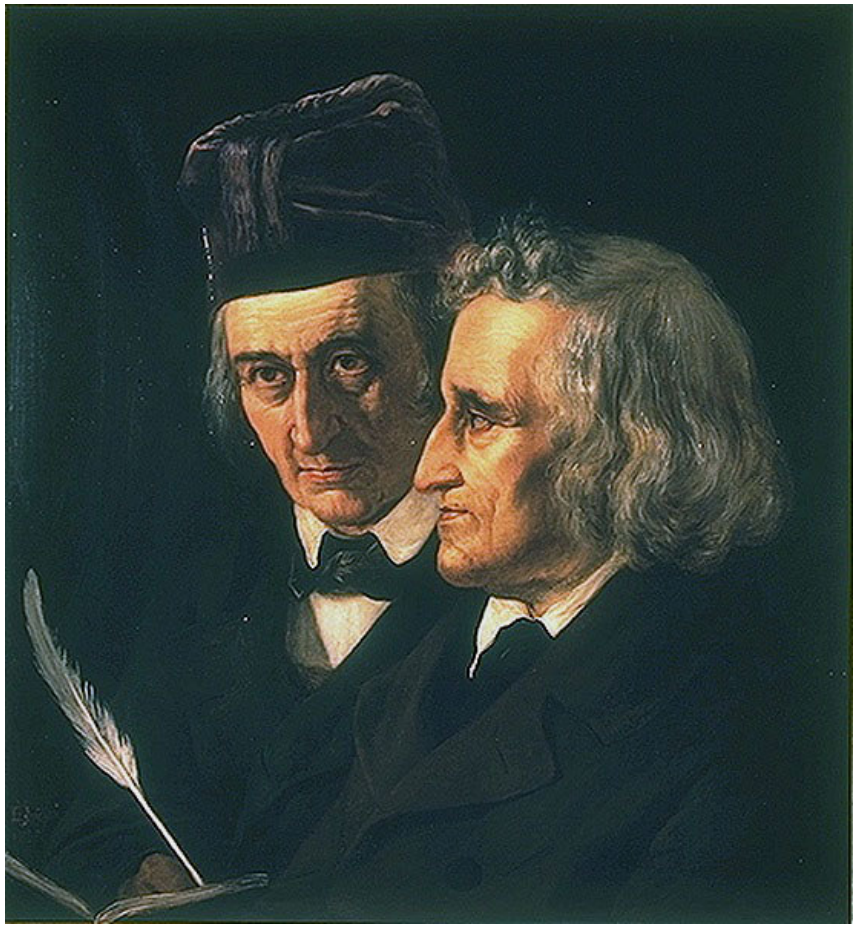
Study extant between-language variation and ancient recorded languages (Greek, Sanskrit)

Advantages

- Longer time scales

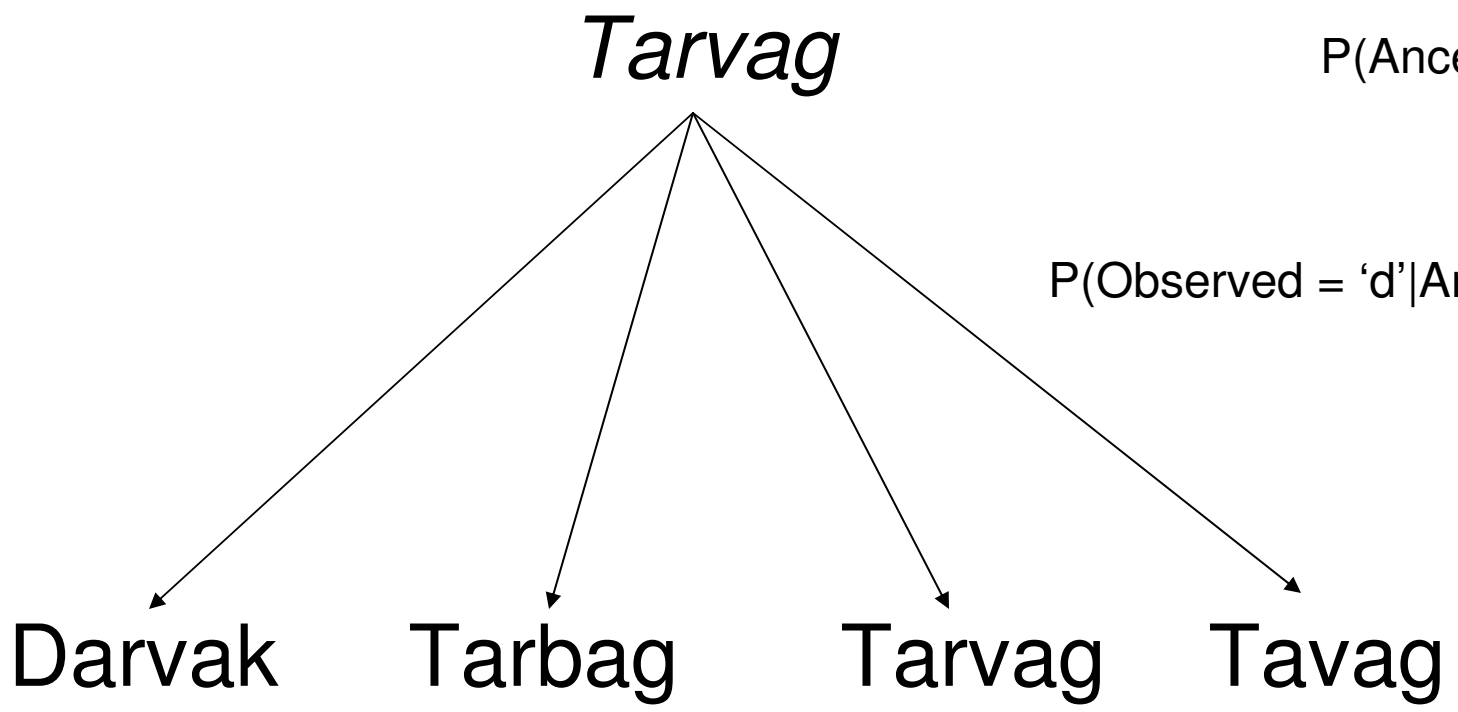
Disadvantages

- Can't see process
- **Requires some way of making inferences about generative models**



Jacob and Wilhelm Grimm (1855)

A Model



1. **Align words**
2. **Estimate probabilities of sound change**
3. **Estimate probabilities of ancestral sounds**
4. Identify cognates
5. Infer historical relationships

Likelihood Function

$$p(L_1, L_2, ..., L_N \mid p(A), p(s \mid A), alignment) =$$

$$\prod_{CognateClasses} \prod_{positions} \sum_{Ancestors} p(A) \prod_{L} p(s \mid A)$$

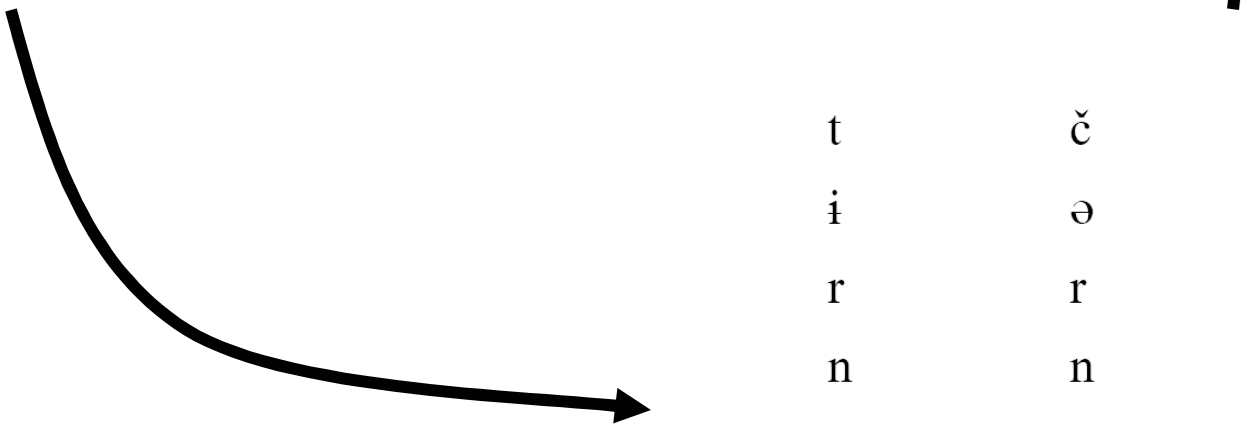
P(s|A) = probability of observed sound given an ancestor
P(A) = probability of ancestral sound

Some assumptions were required to get here

Estimation Strategy

P(Observed sound|Ancestral sound)

P(Ancestral sound)



t	č	t
ɨ	ə	ɨ
r	r	r
n	n	
a	e	a
q		x

Find alignment that maximizes L

Data

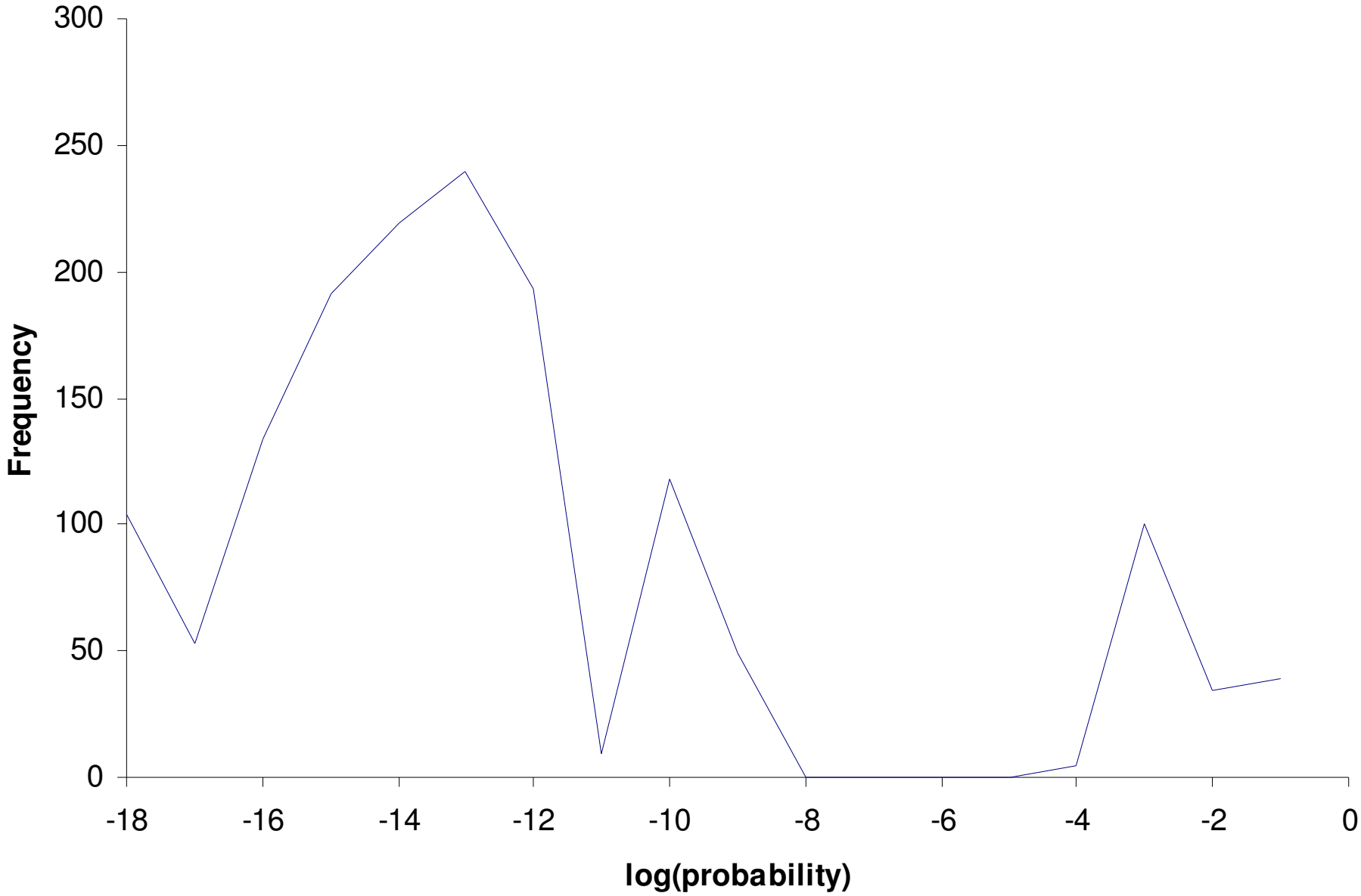
- Word lists (350 to 1400 words each) for 29 Turkic languages
- “Phonetic” transcriptions

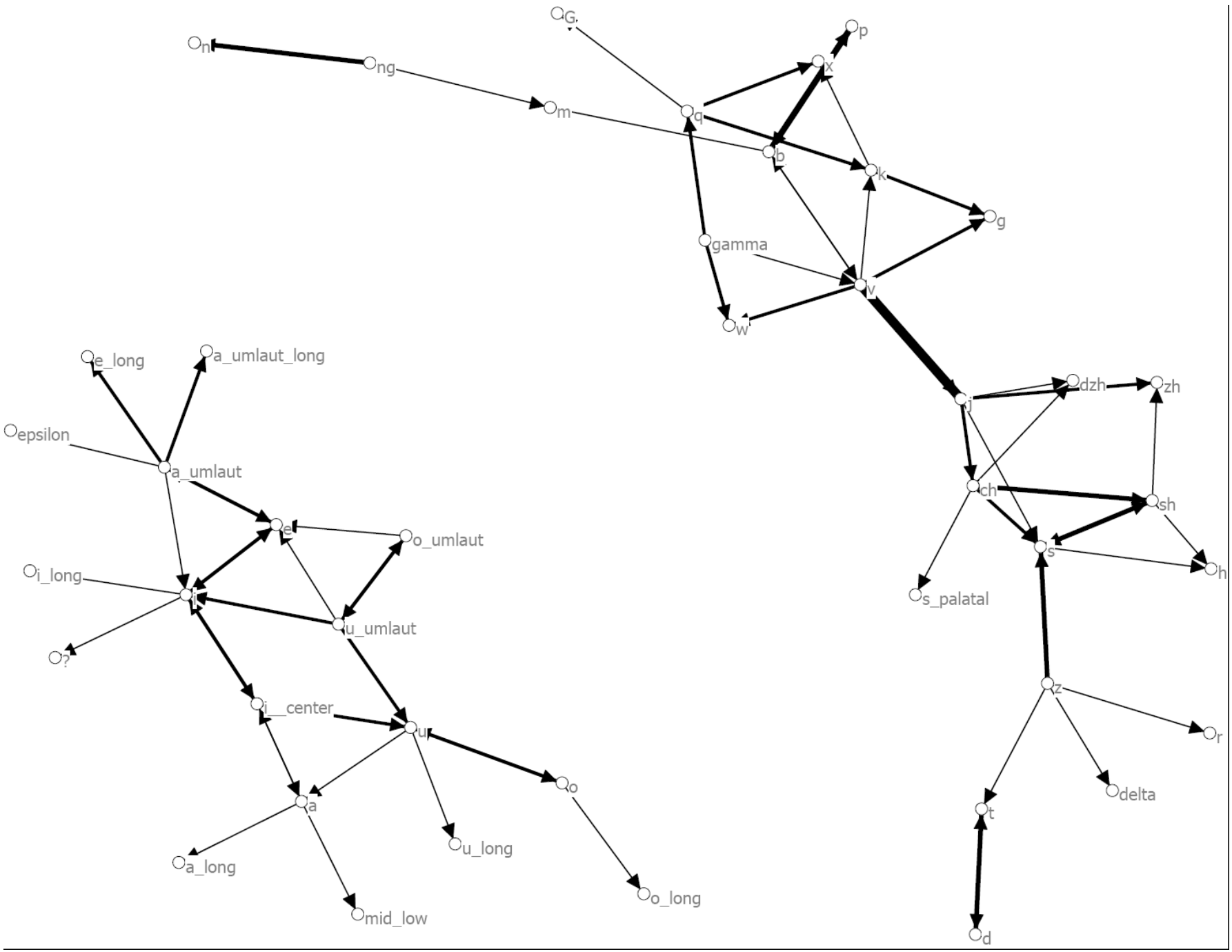
- Advantages
  - Shallow time depth
- Weaknesses
  - Idealization
  - Synchronic

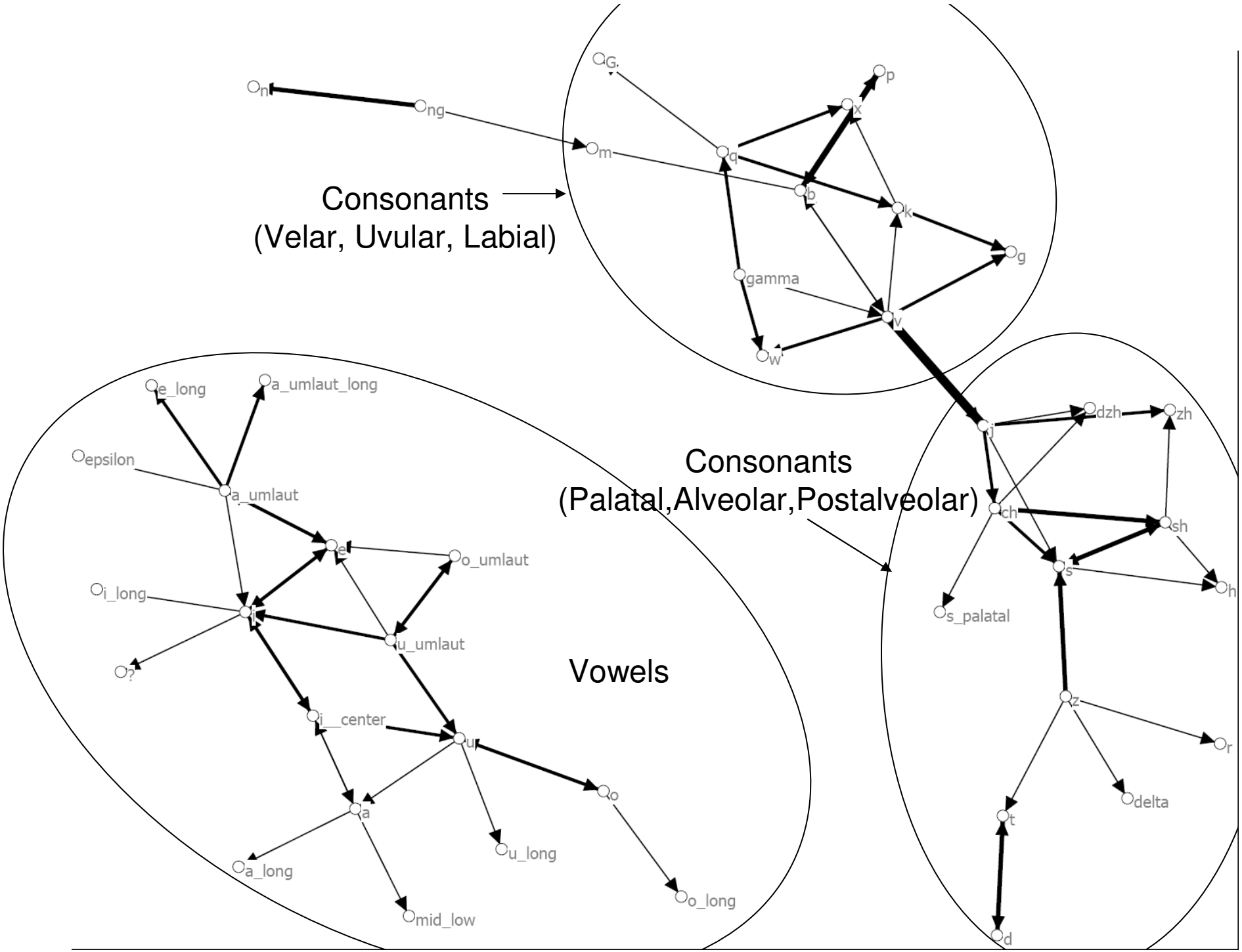
Finding ML estimates

- 100 words
  - 9635 observations
  - 24 hours
  - loglikelihood = -8239
  - Ancestral sounds—15 consonants and 9 vowels with probability > 10^{-17}
- 200 words
  - 16127 Observations
  - 118 hours
  - loglikelihood = -16903
  - Ancestral sounds– 18 consonants and 9 vowels with probability > 10^{-17}

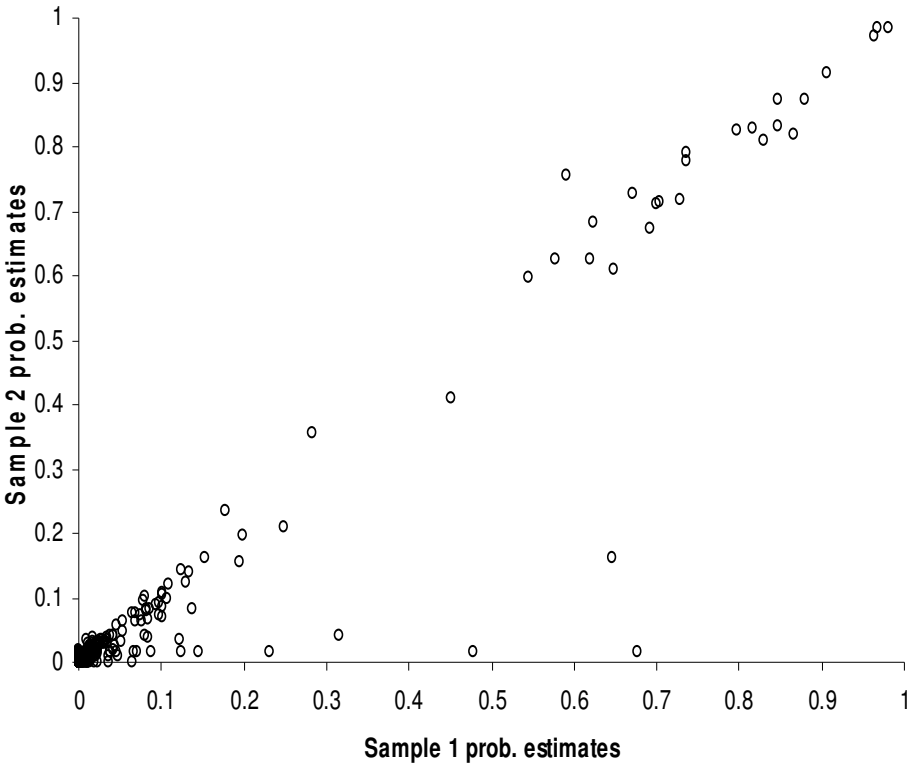
Distribution of Probabilities in the Sound Change Matrix



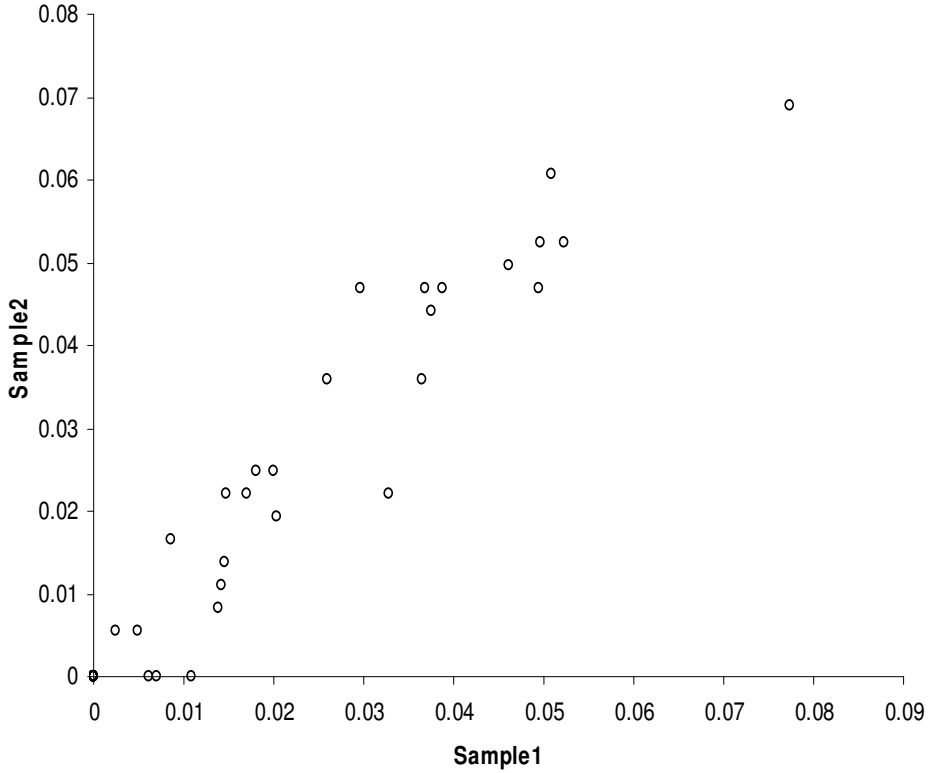

n
ng
m
G
p
x
q
b
k
g
gamma
v
w
dzh
zh
j
ch
sh
s
h
s_palatal
z
r
delta
t
d
e_long
a_umlaut_long
epsilon
a_umlaut
e
o_umlaut
i_long
i
u_umlaut
?
i__center
u
o
a
u_long
a_long
o_long
mid_low


n
ng
m
G
p
x
q
b
k
g
gamma
v
w
Consonants
(Velar, Uvular, Labial)
e_long
a_umlaut_long
epsilon
a_umlaut
e
o_umlaut
i_long
i
u_umlaut
?
Vowels
i__center
u
o
a
u_long
a_long
o_long
mid_low
j
dzh
zh
Consonants
(Palatal,Alveolar,Postalveolar)
ch
sh
s
h
s_palatal
z
r
delta
t
d

Comparing 2 samples' estimates



Estimated Sound Change Probabilities



Estimated Ancestral Probabilities

Recap

- We have ML estimates for data given model
- They fit qualitative observations made by many linguists
- They are consistent for different samples

Questions

- How to compare with other models?
  - Feature-based changes
  - Natural classes (Bouchard-Cote et al. 2007)
- We will always have misspecification. What to do?
  - Historical relationships
  - Regularity of sound change (Foot-Pied, Five-Pent)
  - Idiosyncratic, but common kinds of change (“A newt” & “an eft”)
- Is there a way to estimate faster?