Adding more Linguistics into Automated Reconstruction

Michael Cysouw

Philipps Universität Marburg

LOEWE Exzellente Forschung für Hessens Zukunft

DFG Deutsche Forschungsgemeinschaft

erc

The Problem:

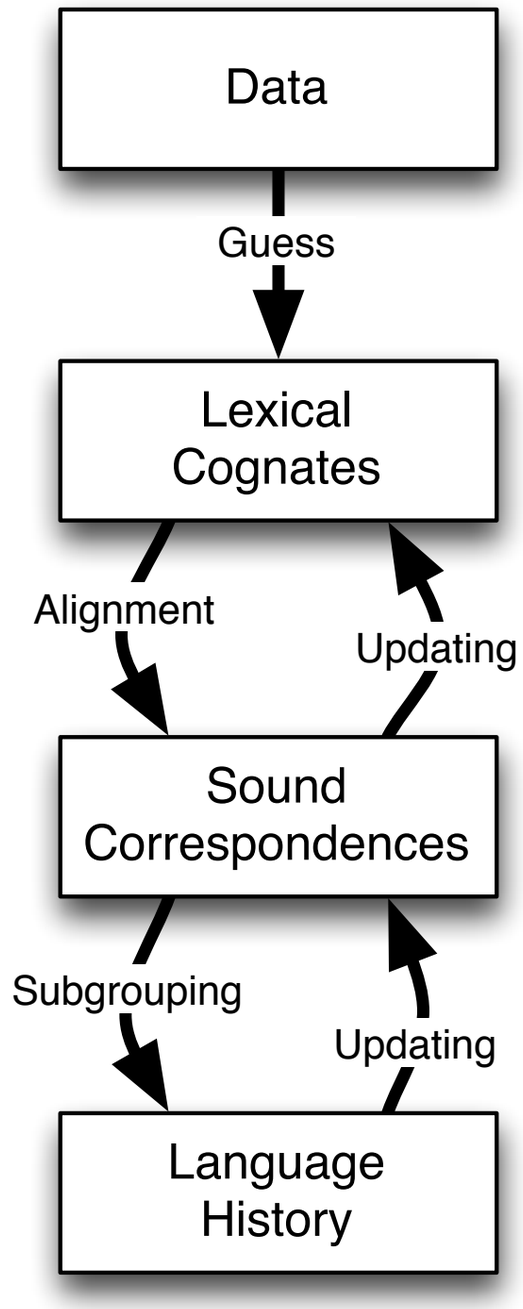
A method that only results in a tree of languages can lead to just three different responses from a linguistic specialist:

- the tree does not correspond to my research,
  so the tree is wrong
- the tree is the same as in my research,
  so the method does not tell us anything new
- I never investigated these languages,
  so show me why this should be true

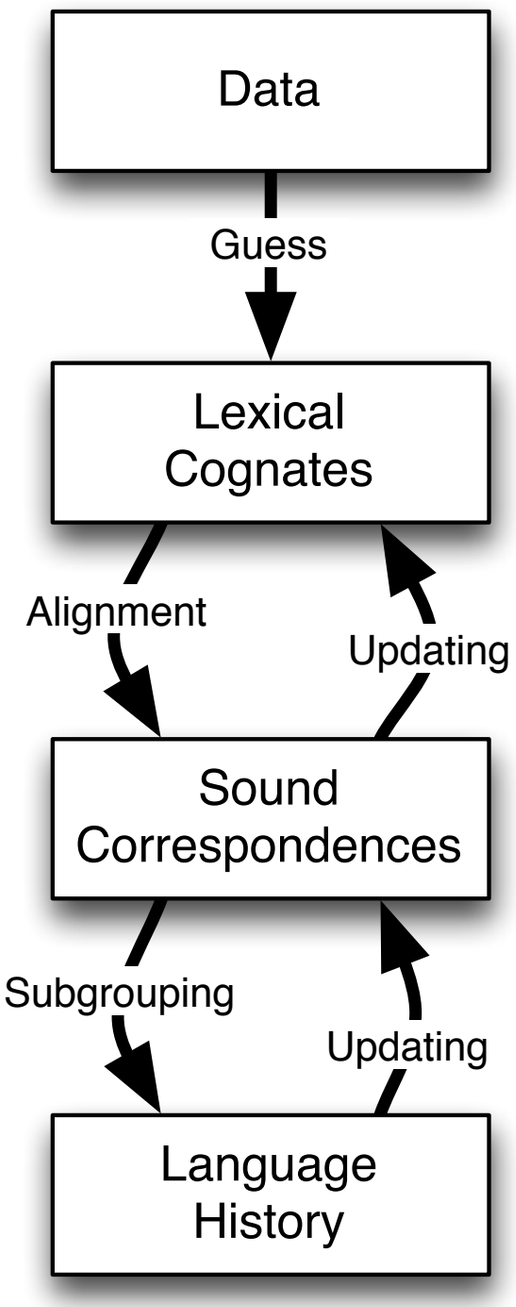
Crucial steps of the linguistic *Comparative Method*

- **Regularity of Correspondences**
  ("co-occurrence statistics")
- **Correspondence Sets**
  ("multiple alignment")
- **Reconstruction**
  ("summary of correspondences")

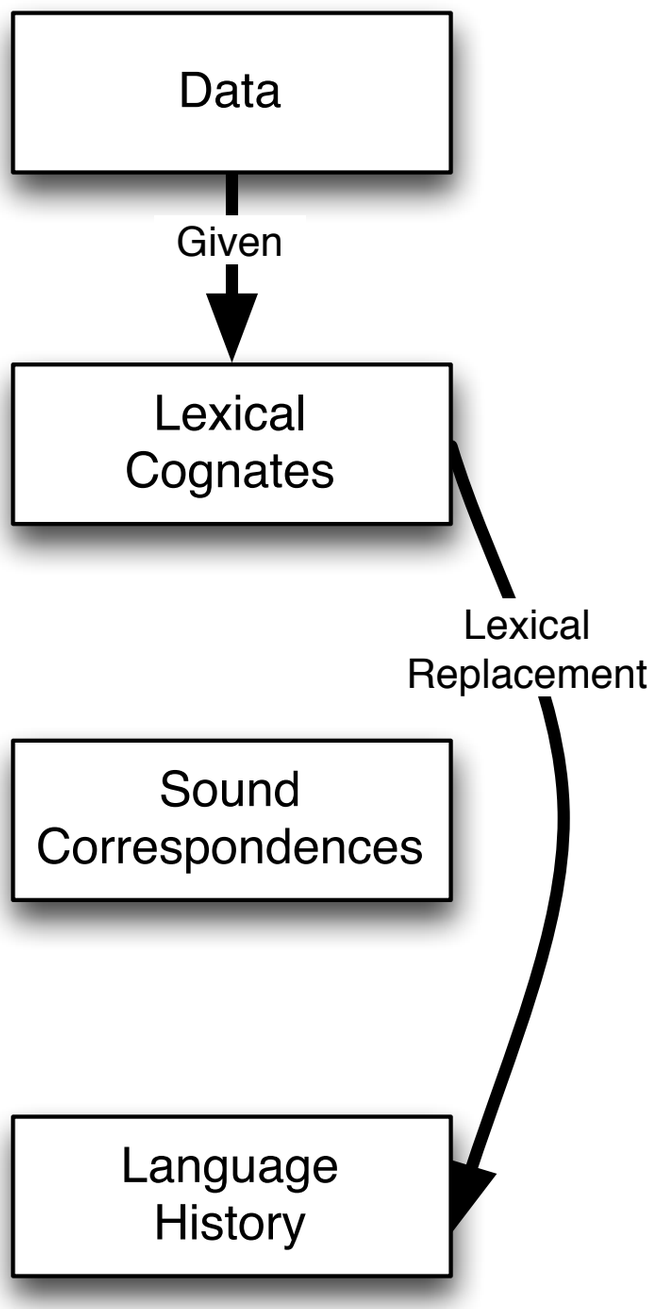
Comparative Method



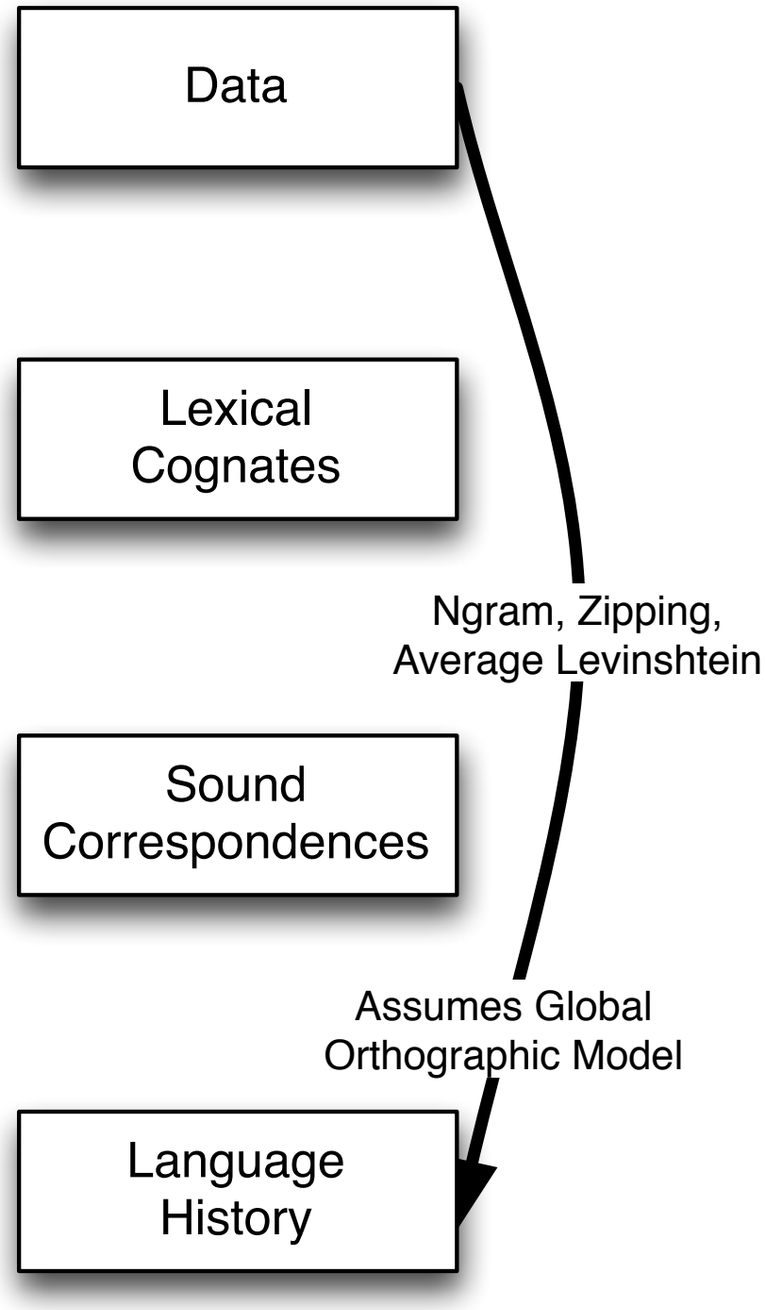
Comparative Method



Swadesh Method



Black Box Method



Swadesh Method

- **Necessary prerequisite:** known cognacy in parallel wordlist
- **Method:** Reconstruct history from distribution of cognates
- **Distances between language:** Swadesh (1952)
- **Stochastic model with first tree:** Sankoff (1969), Dobson (1969)
- **Modern replacement model:** Gray & Jordan (2000), (Gleason 1959)

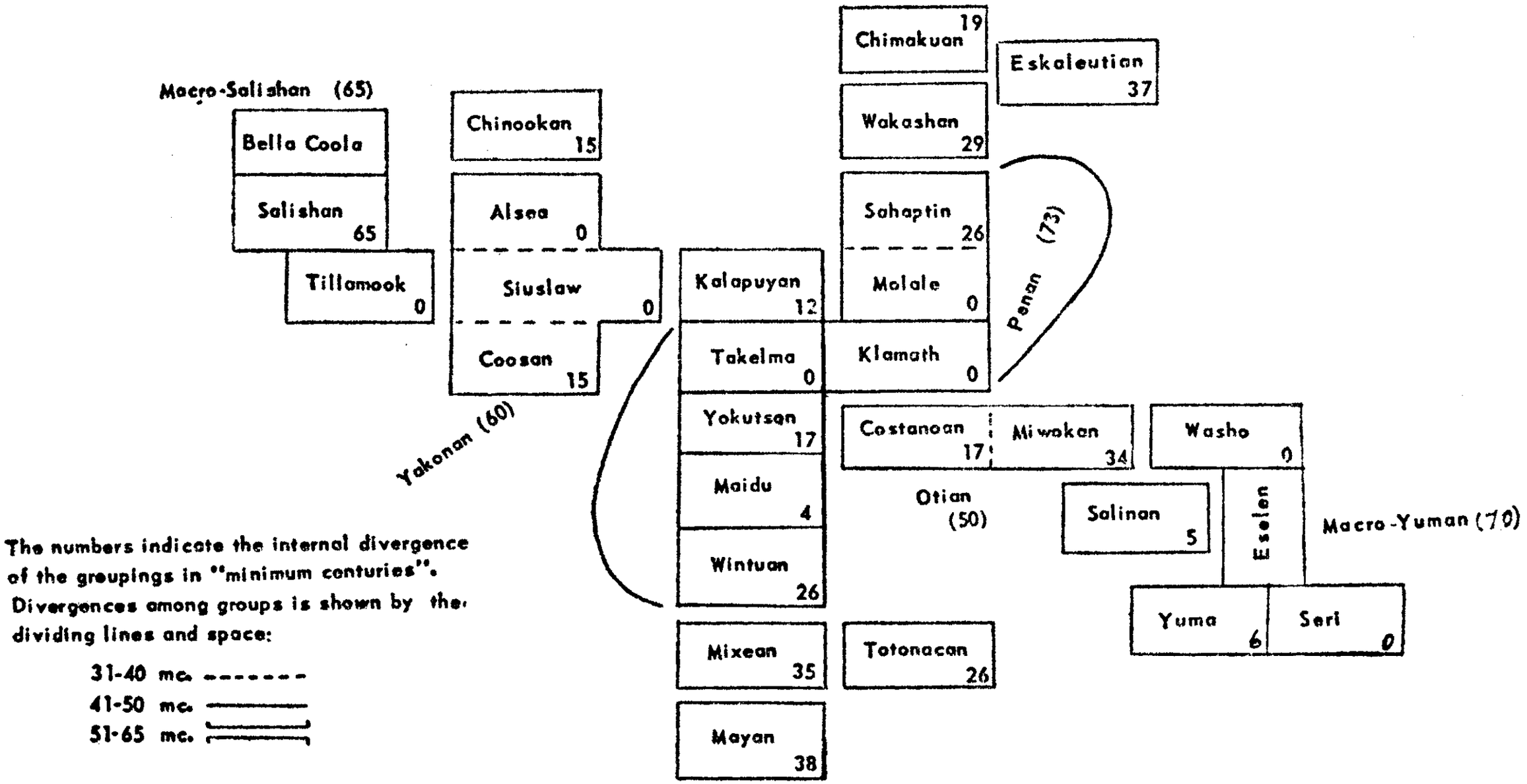


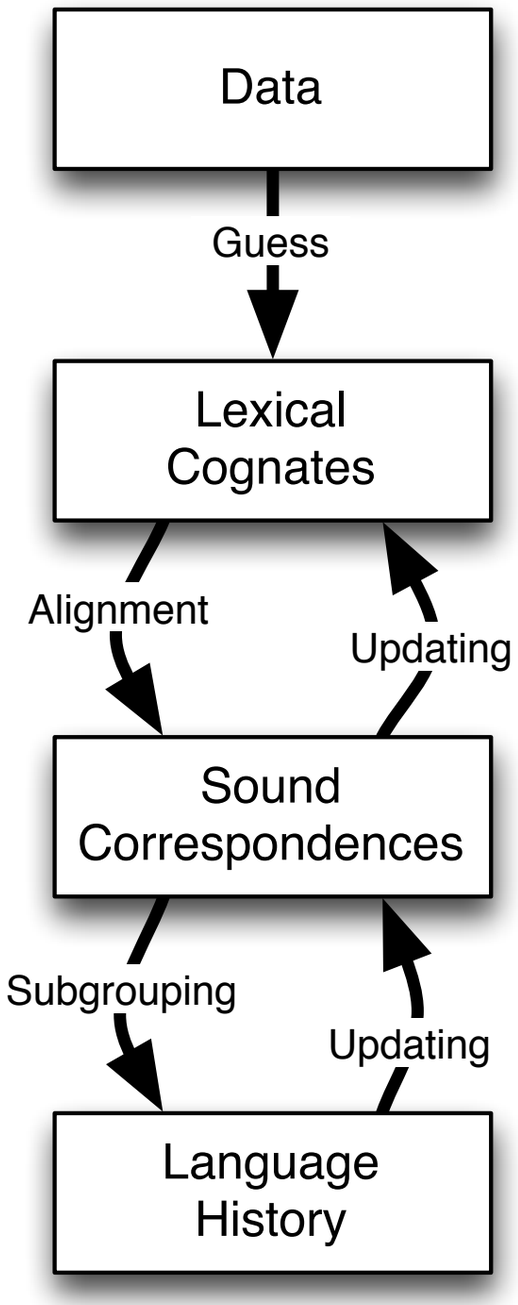
FIG. 2. Internal and external relationships of the Penan micro-phylum.

Swadesh, Morris. 1952. Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos. Proceedings of the American Philosophical Society 96(4). 452-463.
Swadesh, Morris. 1959, "Linguistics As An Instrument of Prehistory. Southwestern Journal of Anthropology 15, no. 1: 20-35.

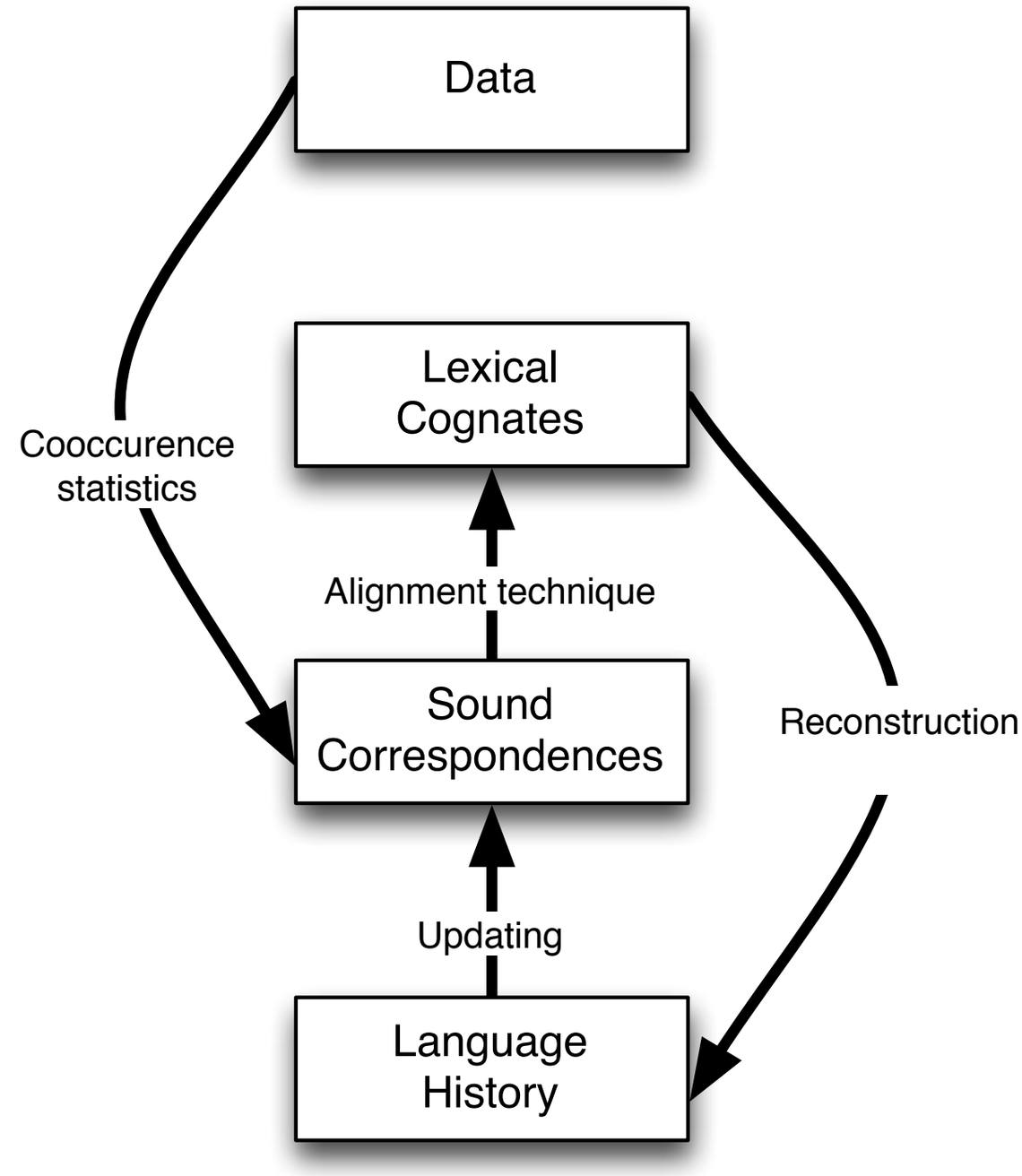
Black Box Methods

- **Necessary prerequisite:** Global orthographic model ("same alphabet for all data")
- **Method:** Compare global orthographic similarity
- **Aggregate Levenshtein Distance:** Batagelj et al. (1992), Kessler (1995)
- **N-Gram Similarity:** Huffman (1998)
- **Zipping Distance:** Benedetto et al. (2002)

Comparative Method



Sound Change Method



Regular symbol correspondences

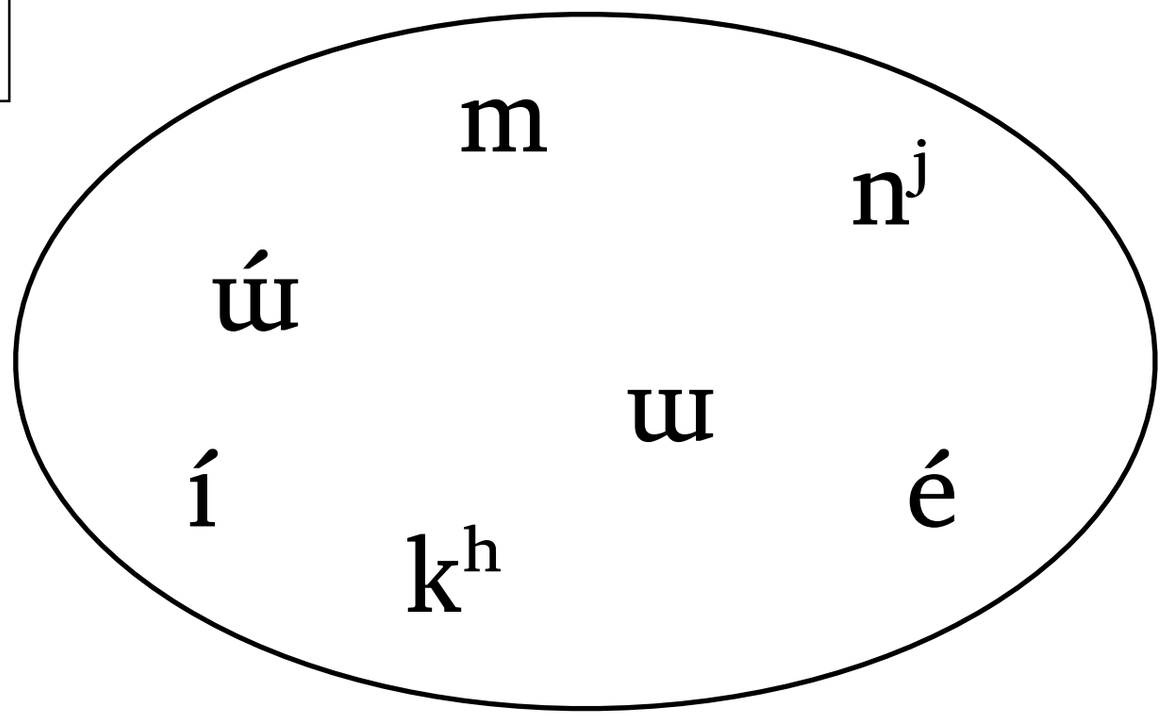
Parts/Length Ratio

("informativity of the parts")

- **DNA 10^{-3}** viz. 4 / 1000s
  (nucleotides vs. length of strings)
- **Protein 10^{-1}** viz. 20 / 100s
  (amino acids vs. length of protein)
- **Words 10^{+1}** viz. 40 / 6
  (sounds vs. length of word)
- **Sentences 10^{+2}** viz. 1000s / 10
  (number of words vs. length of sentence

‘bag of symbol” approach

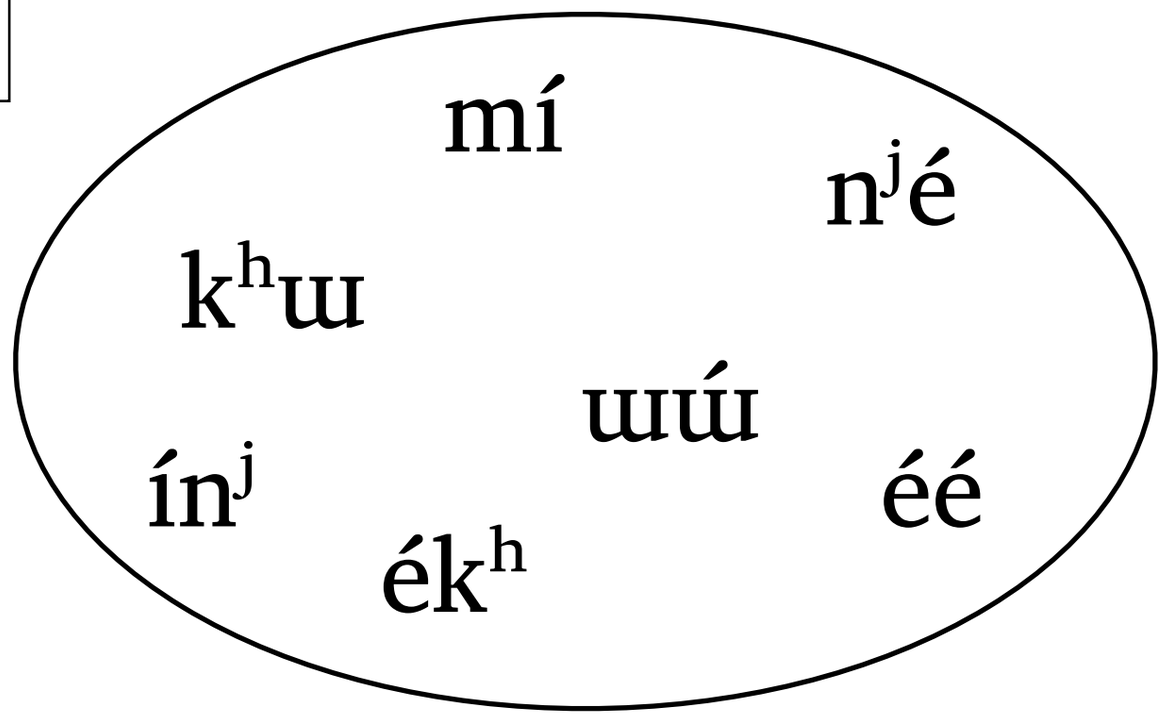
mínʲéékʰɯúú



(1-grams)

‘bag of symbol” approach

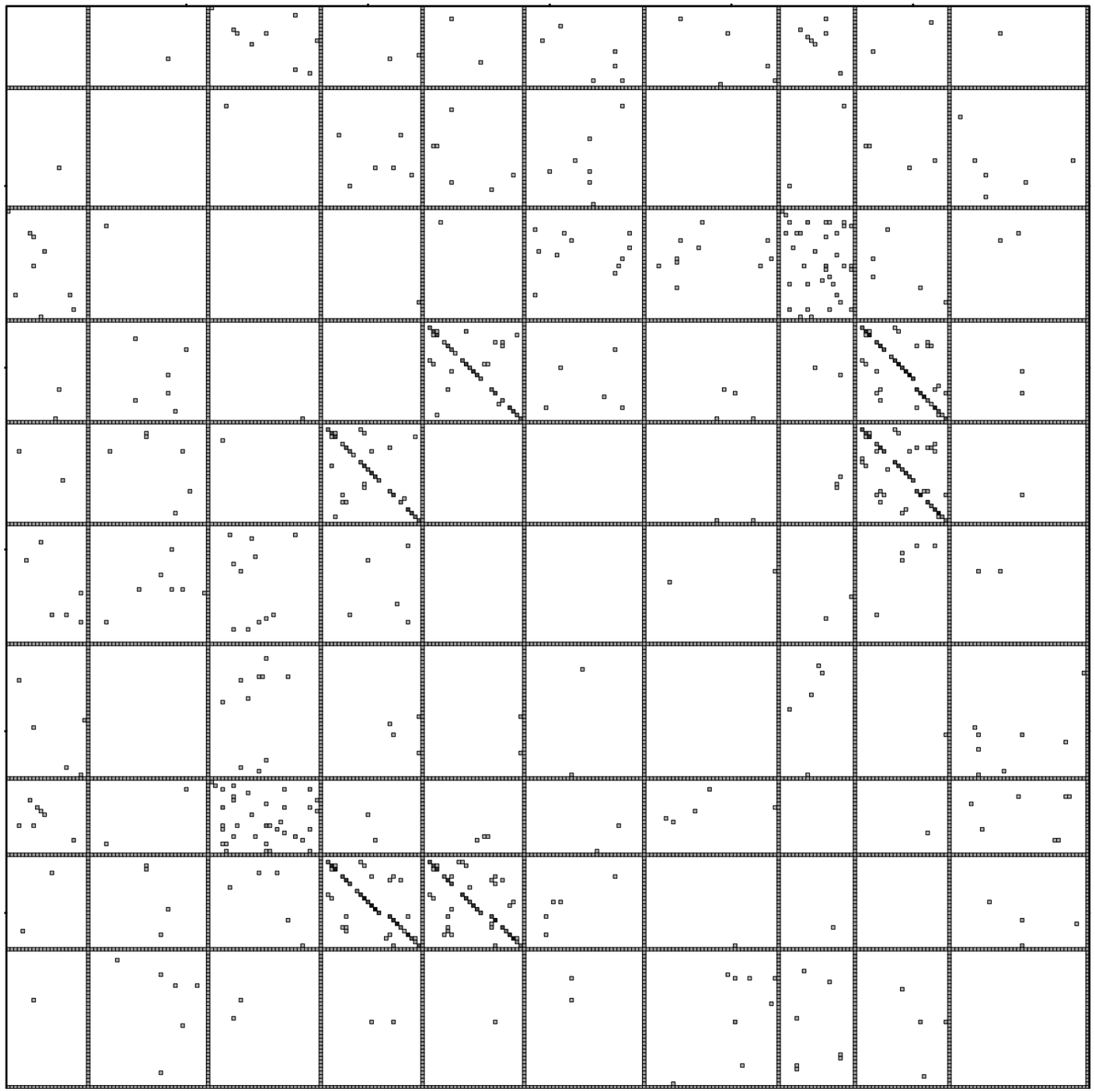
mínʲéékʰɯúú



(2-grams)

	Bora	Muinane
down	tʃɨ́nʲe, paári	báari, gíino
bee	íímúʔóexpʰi, téʔtsʰipa	nɨ́ɨ́biɨri, mɨ́ɨbiɨriʔɨ
sharp	tsʰúʔxɨβáne	sɨ́ɨ́xéβano
…	…	…

Bora	Muinane	Bora	Muinane	Bora	Muinane
#k	#k^h	#i	#i	#n	#n
kɨ	k^hɯ	#a	#a	#m	#m
se	ts^hɨ	di	ti	mɨ	mɯ
xe	xɨ	du	to	nɨ	nɯ
ga	k^wa	#d	#t	us	ts^hɨ
ba	pa	#s	#ts^h	#t	#t^h
#b	#p	gi	ʧi	ɨg	ɯk^w
e#	ɨ#	ni	ni	#ɸ	#p^h



Pairwise symbol co-occurrences between 10 languages

Correspondence Sets
(“multiple Alignment”)

tongue

American English	tʰ	ʌ	ŋ	-	-
Canadian English	tʰ	ʌ	ŋ	-	-
Central German (Cologne)	ʦ	ʊ	ŋ	-	-
Central German (Honigberg)	ʦ	aː	-	-	-
Central German (Luxembourg)	ʦ	ɔ	ŋ	-	-
Central German (Murrhardt)	ʦ	ʊ	ŋ	-	-
Danish	tʰ	ʊ	ŋ	-	ə
Dutch (Antwerp)	t	ʌ	ŋ	-	-
Belgian Dutch	t	ʊ	ŋ	-	-
Dutch (Limburg)	t	ʊ	ŋ	-	-
Dutch (Ostend)	t	ʊ	ŋ	-	ə
Dutch	tʰ	ʌ	ŋ	-	-
New Zealand English (Auckland)	tʰ	ɐ	ŋ	-	-
English (Buckie)	tʰ	ʌ	ŋ	-	-
Indian English (Delhi)	t	ɐ	ŋ	-	-
Nigerian English (Igbo)	tʰ	ʌ	ŋ	ɡ	-
South African English (Johannisburg)	tʰ	ɜ	ŋ	-	-
English (Lindisfarne)	t	ɔ	ŋ	-	-
English (Liverpool)	t͡θʰ	ʊ	ŋ	ɡ	-
English (London)	tʰ	ɐ	ŋ	-	-
English (North Carolina)	tʰ	ɜː	ŋ	-	-
Australian English (Perth)	tʰ	ɐ	ŋ	-	-
English (Singapore)	t	ɑ	ŋ	-	-
English	tʰ	ɐ	ŋ	-	-
English (Tyrone)	tʰ	ɔ	ŋ	-	-
Faroese	tʰ	ʊ	ŋ	k	a
German	ʦ	ʊ	ŋ	-	ə
High German (North Alsace)	ʦ	ʊ	ŋ	-	-
High German (Biel)	ʦ	ʊ	ŋ	-	ə
High German (Bodensee)	ʦ	ʊ	ŋ	-	ə
High German (Graubuenden)	ʦ	ʊ	ŋ	ɡ	ɐ
High German (Herrlisheim)	ʦ	ʊ	ŋ	-	-
High German (Ortisei)	ʦ	ʊ	ŋ	ɡ	ɛ
High German (Tuebingen)	ʦ	u	ŋ	ɡ	-
High German (Walser)	ʦ	ʊ	ŋ	ɡ	ə
Icelandic	tʰ	ʊ	ŋ	k	a
Low German (Achterhoek)	tʰ	ʊ	ŋ	-	ə
Low German (Bargstedt)	tʰ	ʊ	ŋ	-	-
Norwegian (Stavanger)	tʰ	ʊ	ŋ	-	ə
Scottish	tʰ	ʌ	ŋ	-	-
Swedish (Skane)	tʰ	Øʏ	ŋ	j	ɐ
Swedish (Stockholm)	tʰ	ʊ	ŋ	-	ə
West Frisian (Grou)	t	ɔ	ŋ	-	ə
Yiddish (New York)	ʦ	u	ŋ	ɡ	-

(LingPy library, maintained by Mattis List)

foot

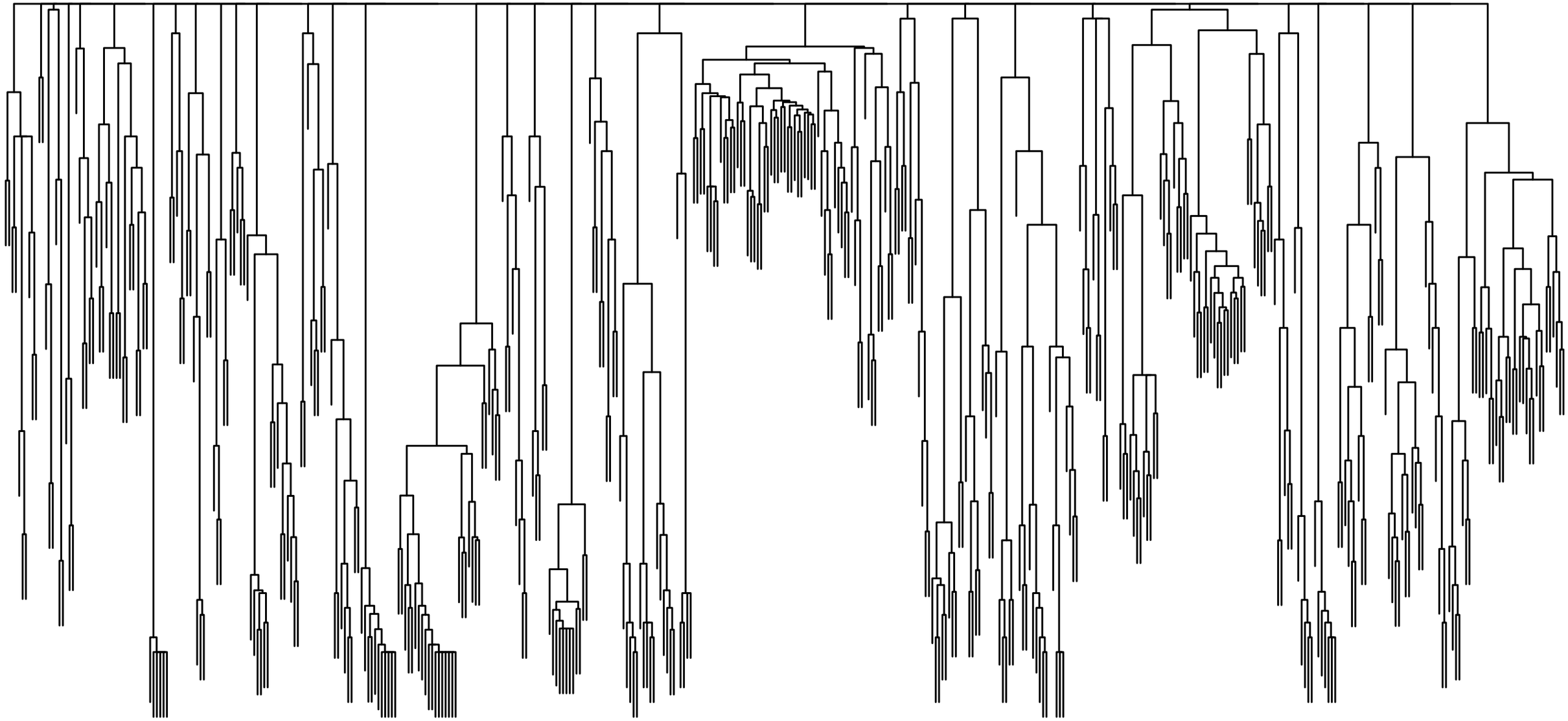
American English	f	ʊ	t
Canadian English	f	ʊ	t
Central German (Cologne)	f	ʊ	s
Central German (Honigberg)	f	eu	s
Central German (Luxembourg)	f	œʊ	s
Central German (Murrhardt)	f	uə	s
Danish	f	ʊ	lˀ
Dutch (Antwerp)	f	u	t
Belgian Dutch	v	u	t
Dutch (Limburg)	f	oː	t
Dutch (Ostend)	v	uə	t
Dutch	f	u	t
New Zealand English (Auckland)	f	ʊ	t
English (Buckie)	f	æ	tʰ
Indian English (Delhi)	f	ʊ	t
Nigerian English (Igbo)	f	u	t
South African English (Johannisburg)	f	ʊ	tʰ
English (Lindisfarne)	f	ɔ	t
English (Liverpool)	f	ʊ	θ
English (London)	f	ʉ	ʔ
English (North Carolina)	f	øː	t
Australian English (Perth)	f	ʊ	t
English (Singapore)	f	u	t
English	f	ʊ	tʰ
English (Tyrone)	f	ɔ	t
Faroese	f	ɒu	t
German	f	uː	s
High German (North Alsace)	ɸ	eː	s
High German (Biel)	f	oə	s
High German (Bodensee)	f	ʊɛ	s
High German (Graubuenden)	f	ʊɔ	s
High German (Herrlisheim)	ɸ	eː	s
High German (Ortisei)	f	ʊə	s
High German (Tuebingen)	f	uə	s
High German (Walser)	ɸ	oɛ	s
Icelandic	f	ɔʊ	t
Low German (Achterhoek)	β	oː	t
Low German (Bargstedt)	f	ɔɪ	t
Norwegian (Stavanger)	f	uː	t
Scottish	f	ʉ	tʰ
Swedish (Skane)	f	ʉu	t
Swedish (Stockholm)	f	o	tː
West Frisian (Grou)	f	oə	t
Yiddish (New York)	f	u	s

Reconstruction

("summary of correspondences")

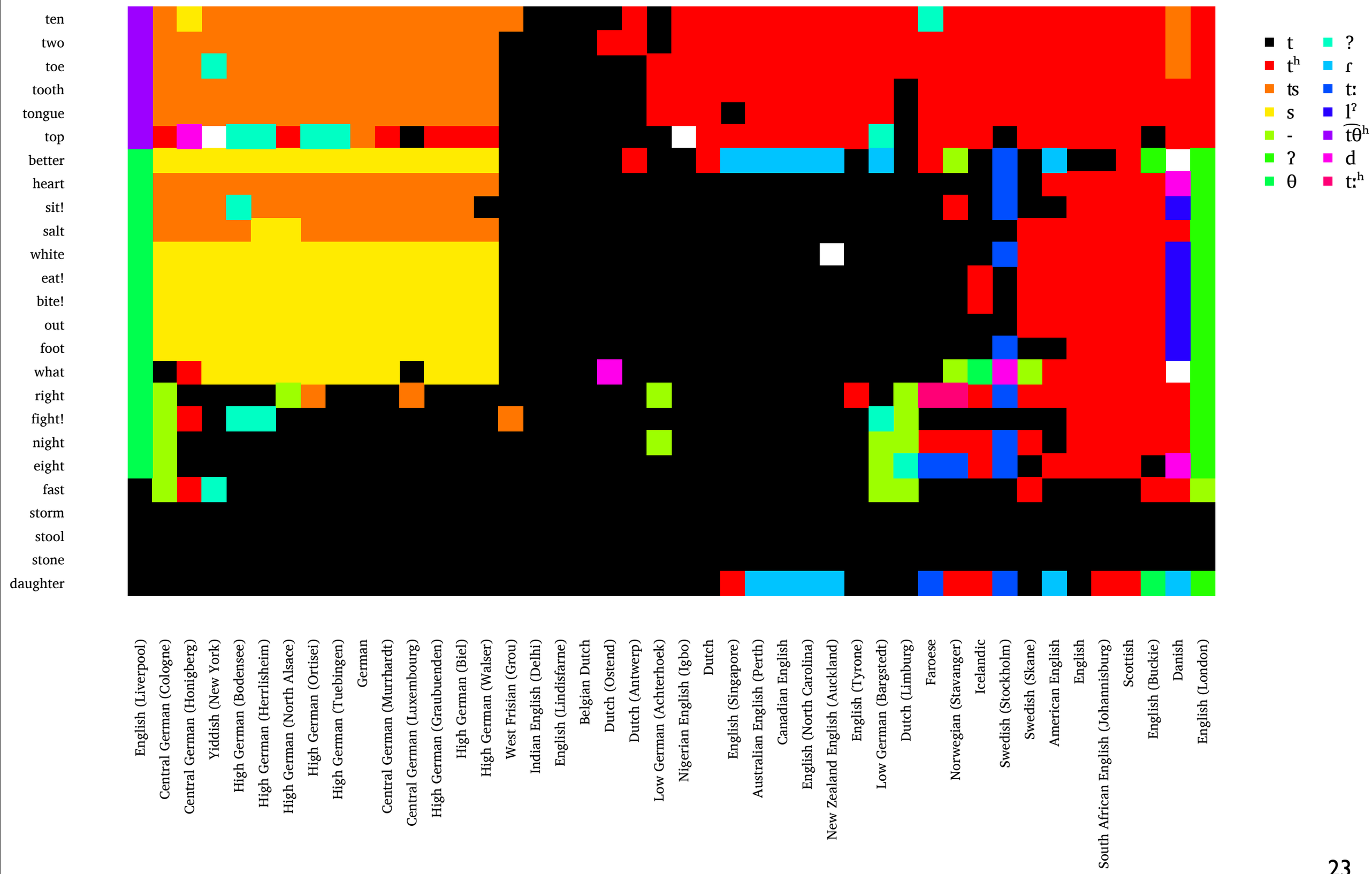
Reconstruction

- What should be the (computational) representation of reconstructions ?
- Linguists mostly think of them as actual sounds
  - that is a highly problematic point-estimate of the actual variation
- Proposal: a proto-sound is a cluster of correspondence sets
  - clusters of "columns in an alignment"

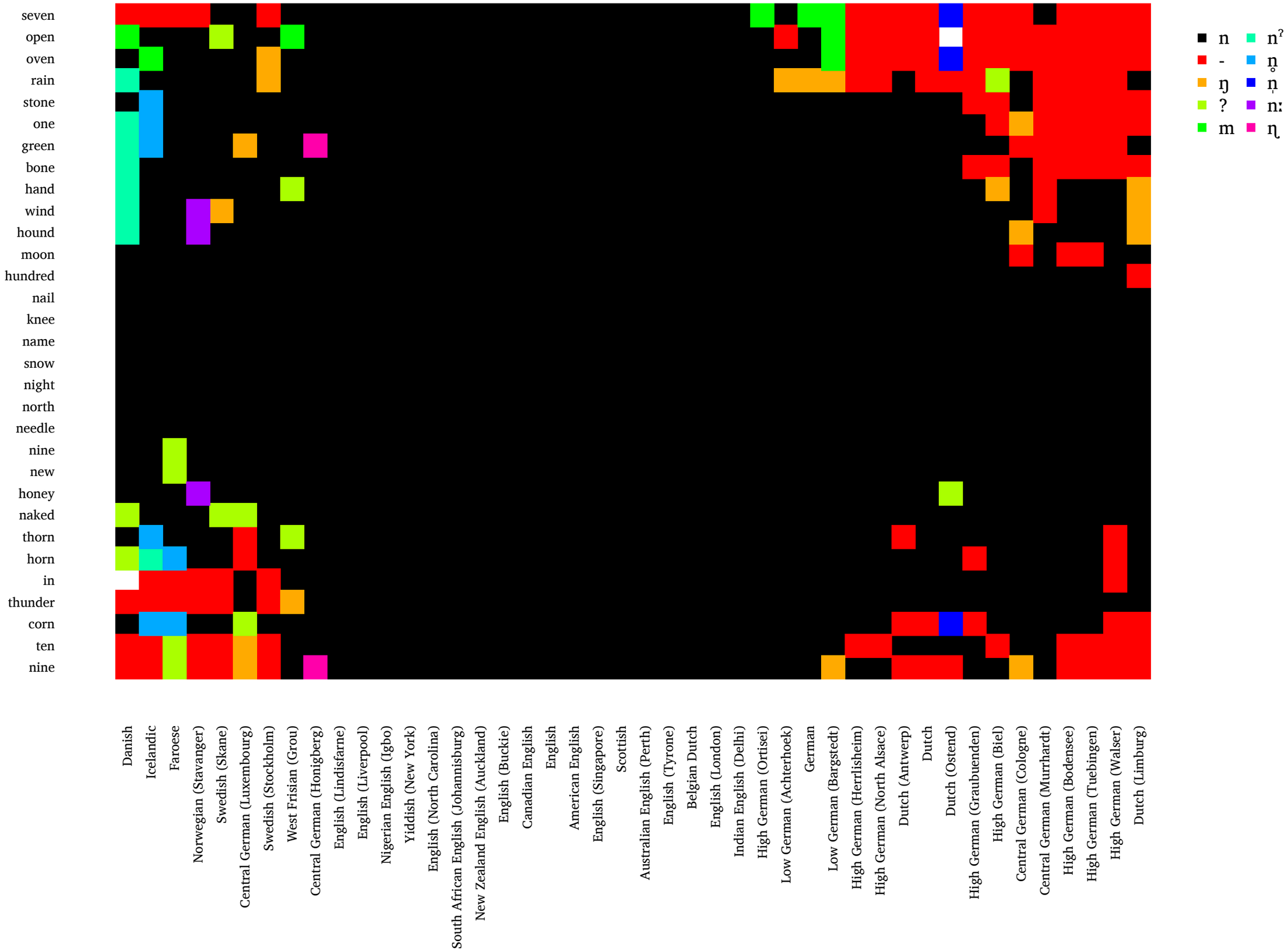


greedy clustering (“complete clustering”)
of all correspondences (“columns in multi-alignments”)
in a Germanic data set

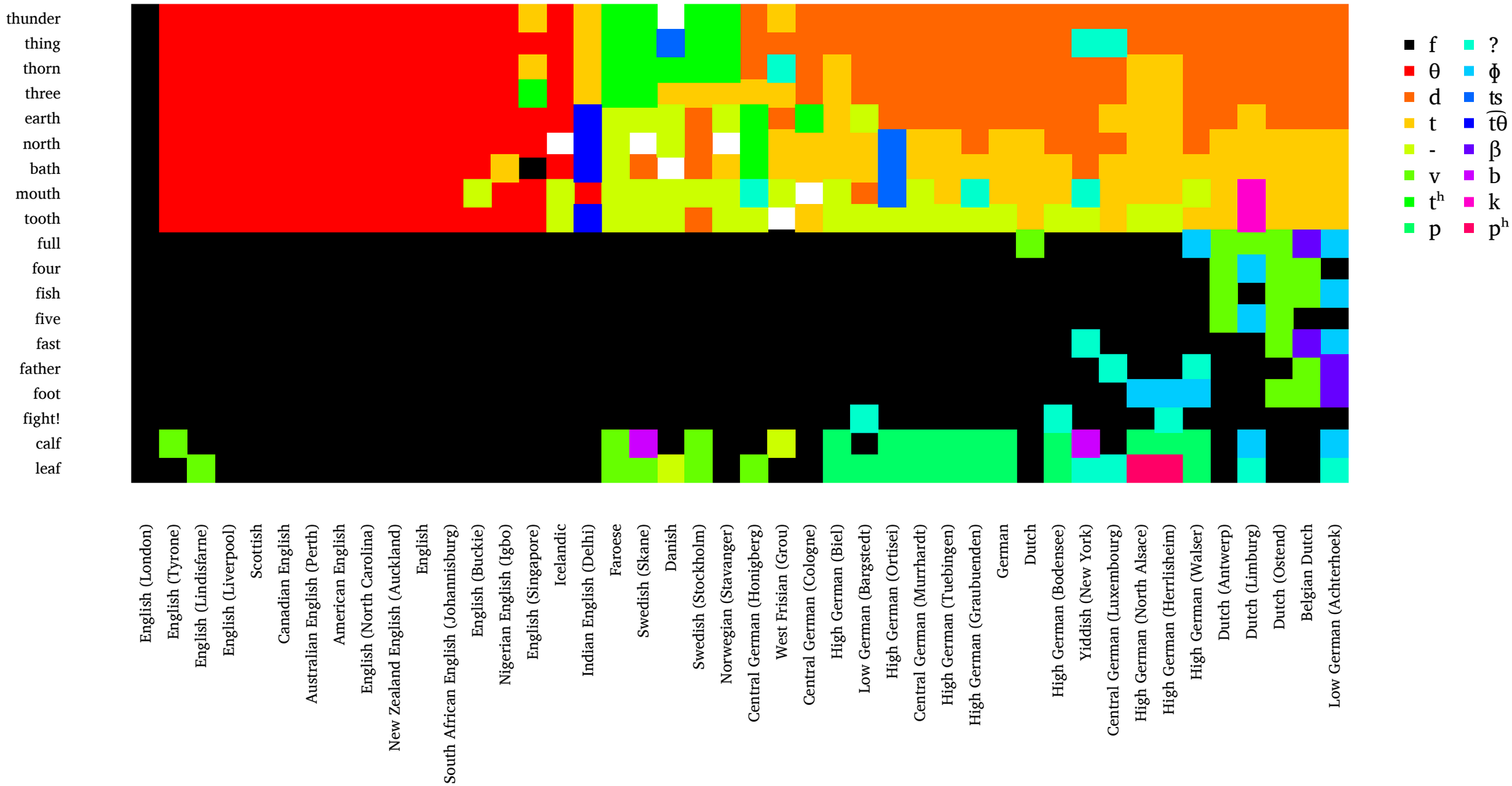
Correspondences with high frequency of [t]



Correspondences with high frequency of [n]



Correspondences with high frequency of [f]



Prospects

- Add more linguistics into reconstruction
- Retaining variation in the reconstruction process is important for deep phylogenies
- Visual UI for non-computational linguists seems highly important