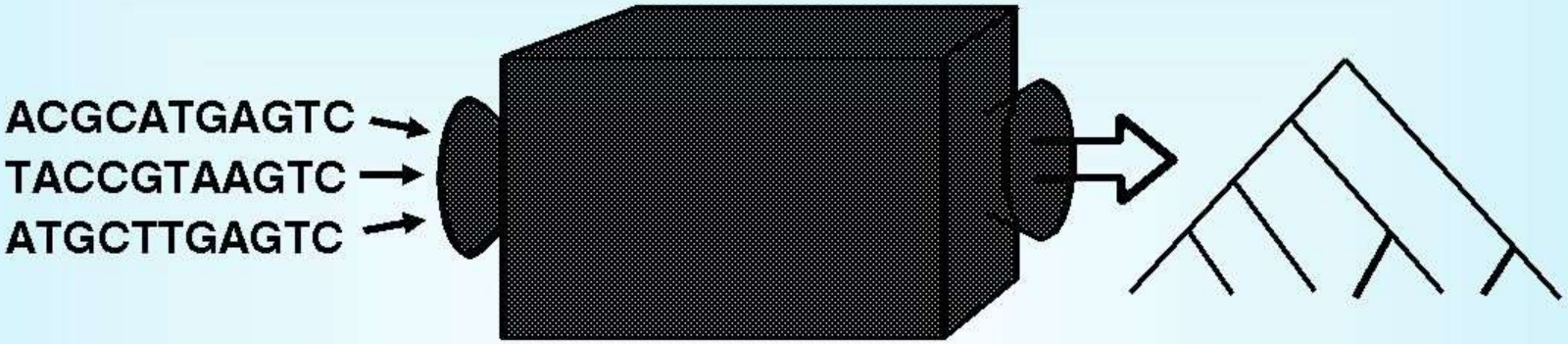
Motto:



"Unfortunately, phylogenetic analysis is frequently treated as a black box into which the data are fed and out of which ' The Tree ' springs."

Introduction to chapter 11 of *Molecular Systematics*, 2nd ed., edited by Hillis/Moritz/Mable, 1996:407.

We shall not follow this rule -

instead ...

The Distribution of Word Lists and its Impact on the Subgrouping of Languages

Hans J. Holm

هنس هولم

Հանս Հոլմ

한스 홀음

1.1 Linguistic Approaches

- Traditional linguistic methods for 200 years only poor results

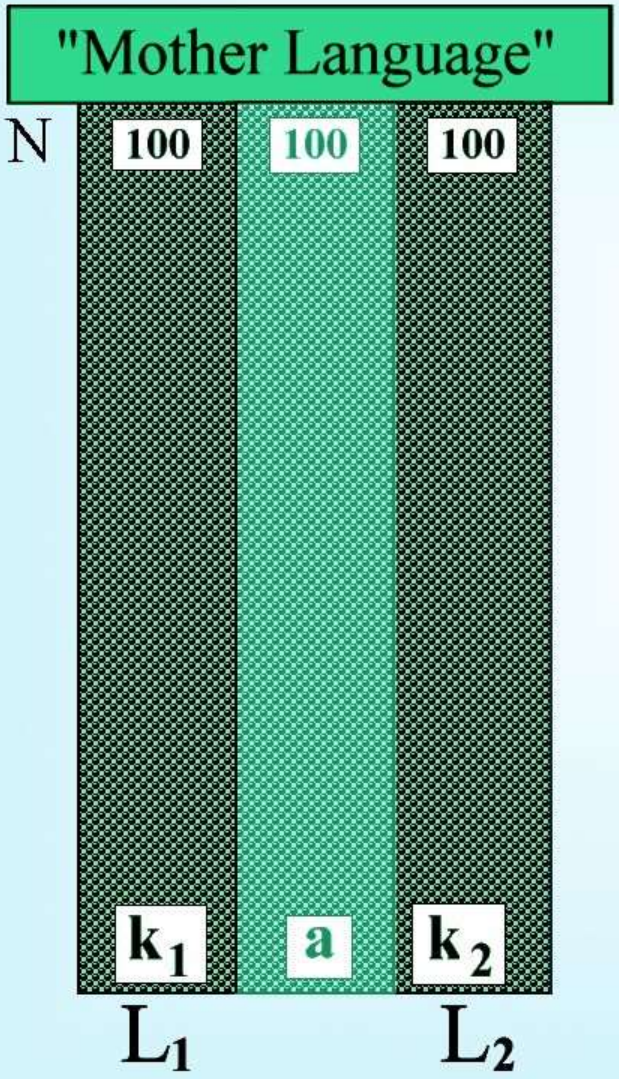
However ...

1.2 Quantitative Approaches

- Traditional linguistic methods for 200 years only poor results

- Quantitative attempts often no better:
  - proud of identifying 'Greek' vs. 'Germanic' (!)
  - often fixated on mechanistic rate assumption
  - confuse surface resemblance with genealogical relationship.

2.1 Stochastic Model of Language Change

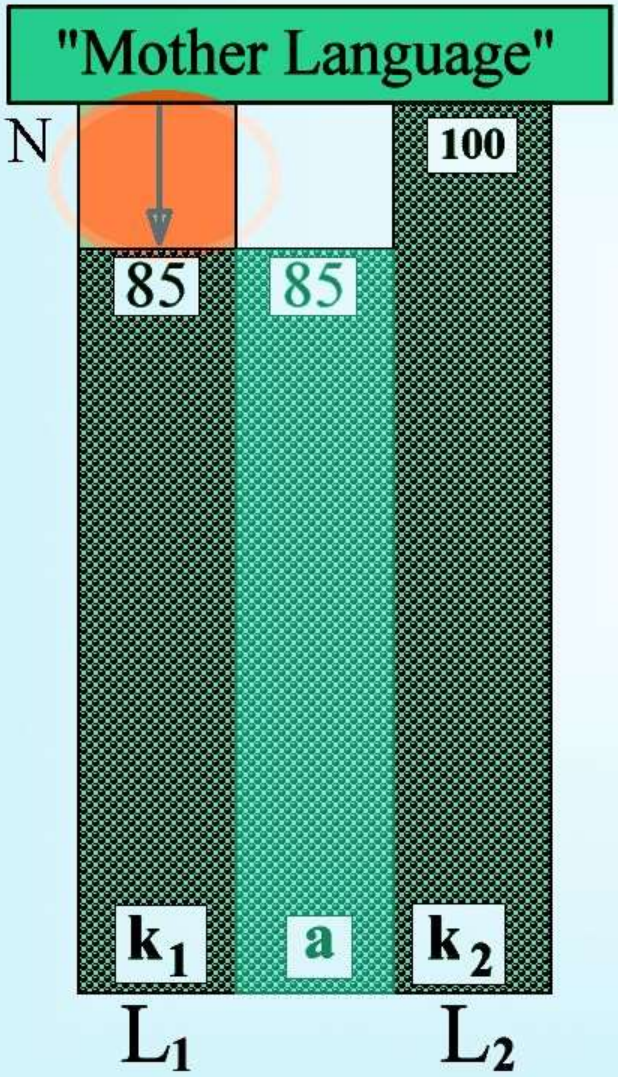


Step 1 - 'Era of Separation'

Mother language L_X splits into two daughter languages, both starting with

- 'k' = 100% inherited features,
- 'a' = 100% agreements.

2.1 Stochastic Model of Language Change



Step 2:

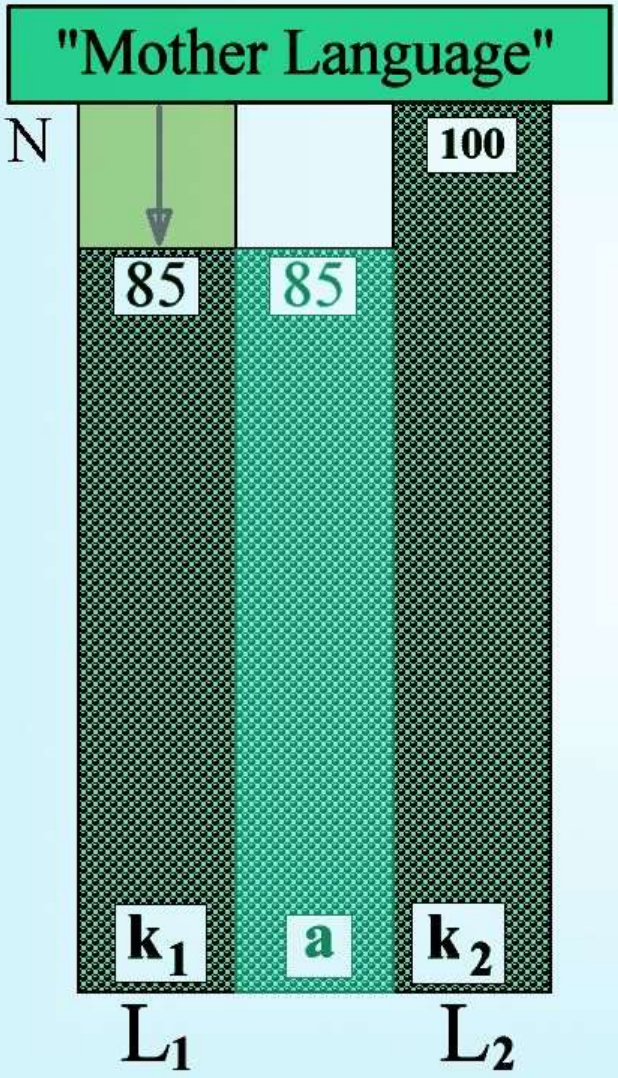
Only L_1 changes 15%

$\rightarrow$ a = 85 %

Nature of change:

?

2.1 Stochastic Model of Language Change



Step 2:

Only L_1 changes 15%

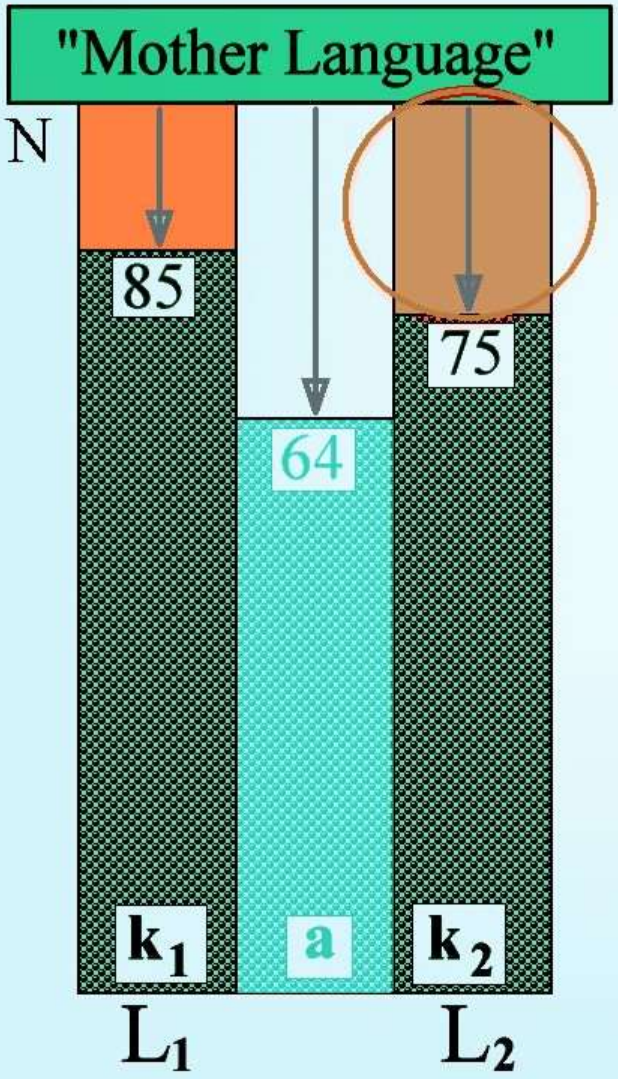
$\rightarrow a = 85\ \%$

Nature of change:

Loss of inherited features

by
- independent
- irregular
- irreversible influences.

2.1 Stochastic Model of Language Change



Step 3:

Also L_2 changes: 25 %

→ a = 64 %

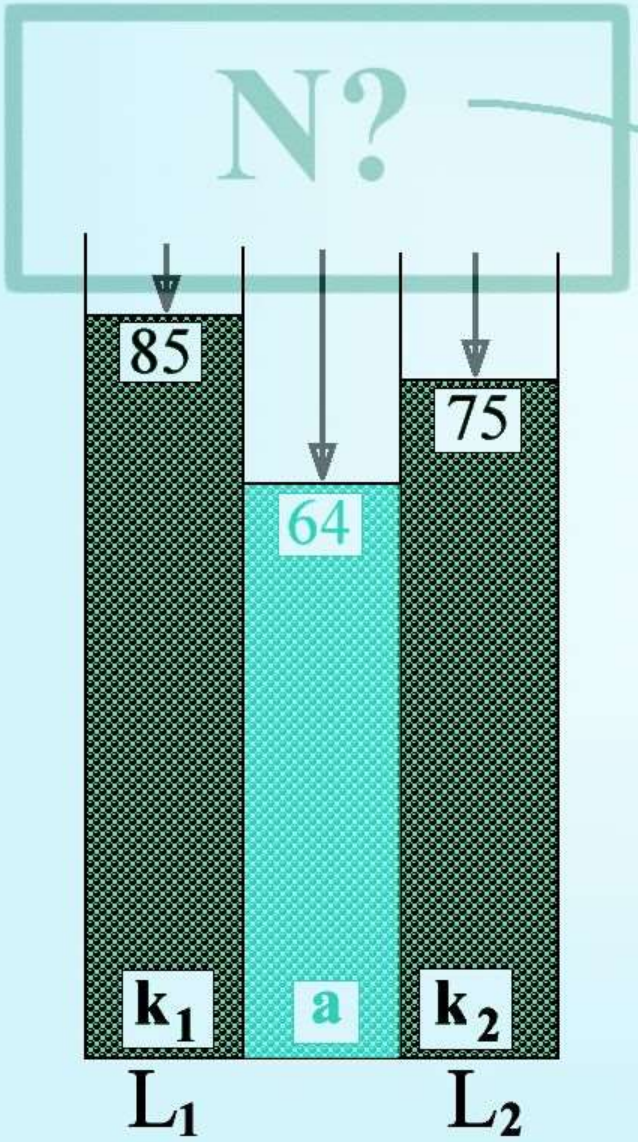
"Hypergeometric process"

with parameters

- k_1 and k_2 preserved cognates
- $a_{1,2}$ agreements

However:

2.2 Estimation of Universe



Universe 'N' at era of separation in fact unknown!

Only Hg allows to compute expected value by

$$\hat{N} = \frac{k_1 \bullet k_2}{a_{1,2}}$$

3.1 Applications up to now

N defines state at era of split:
= ranked nodes of departure

Only few applications ...

3.1 Applications up to now

N defines state at era of split:
= ranked nodes of departure

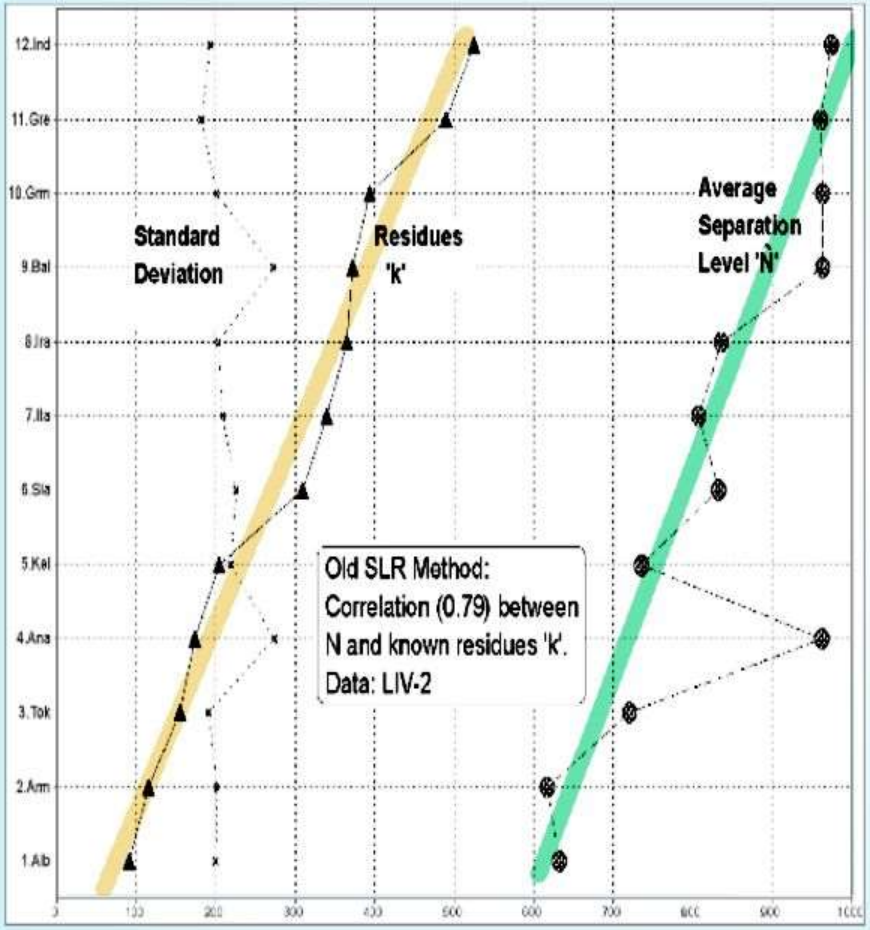
---

applied to data of
- Pokorny 1959 by Holm (2000)
- Mixe-Zoquean by Cysouw et al. (2006)
- Lexikon der idg. Verben = LIV (Rix et al. 2001)

in this paper

However:

3.2 Unwanted Dependence



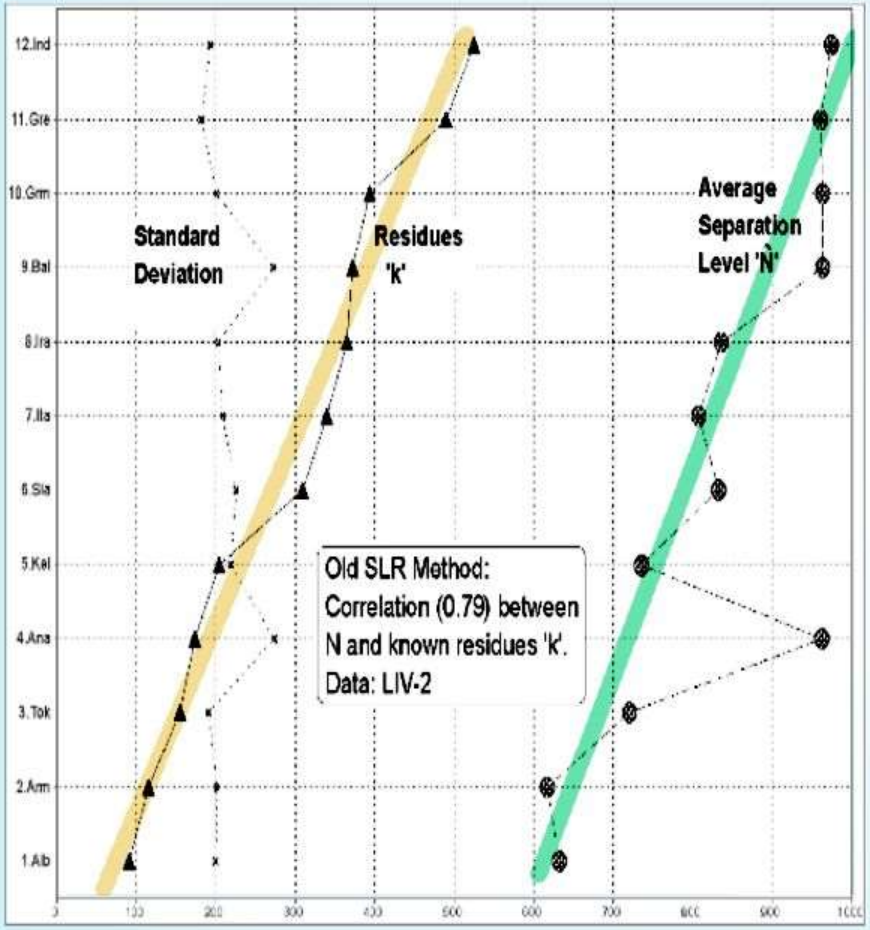
Separation level N

depends on residues 'k'

= Bias

But why ??

3.2 Unwanted Dependence



Separation level N depends on residues 'k'

= Bias

logically not due to
- algorithm
- poor knowledge
- scatter

4.1 Revisiting the Properties of Word Lists

Requirements of Hg fulfilled?

- draws independent? - yes
- probability equal for every word? - no !!!

How can we measure this?

4.1 Revisiting the Properties of Word Lists

Requirements of Hg fulfilled?

- draws independent? - yes

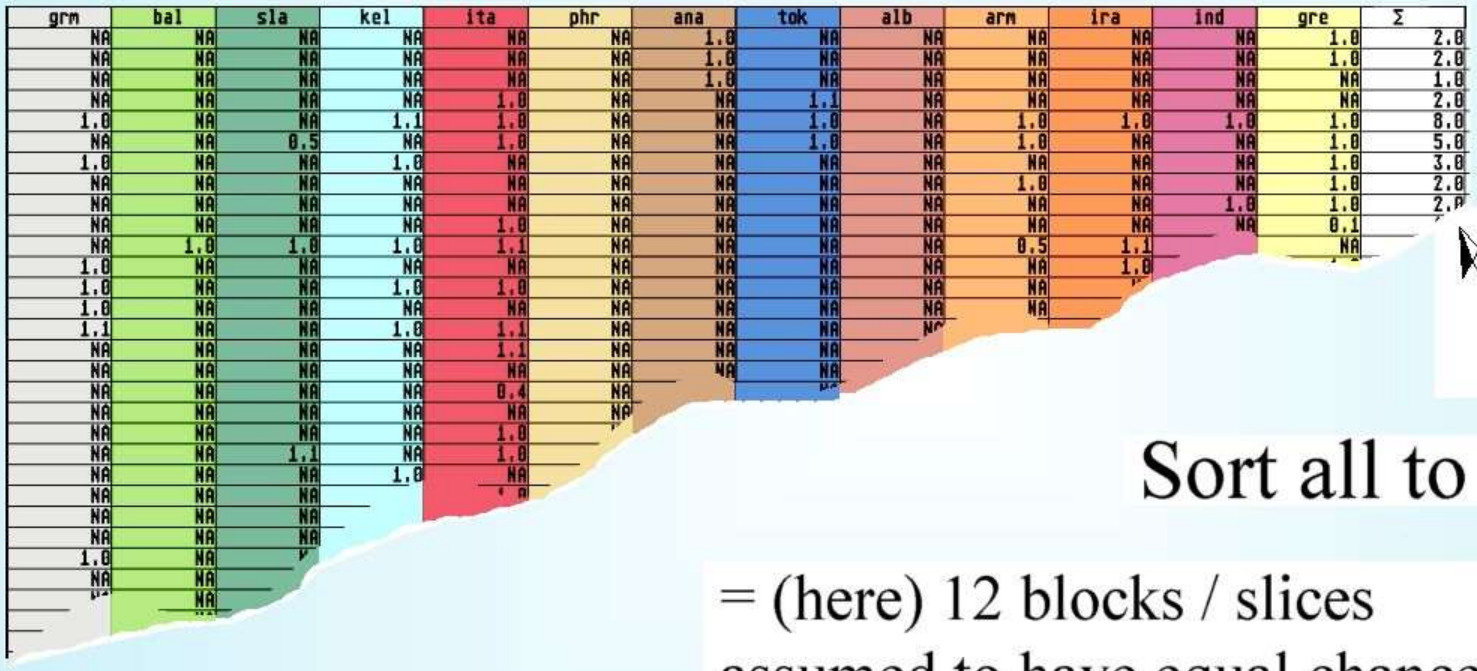
- probability equal for every word? - no !!!

Necessary to do worldwide tests?

No - only distribution of concrete list needed!

4.2 Detecting the Distribution

Spreadsheet with 12 IE branches



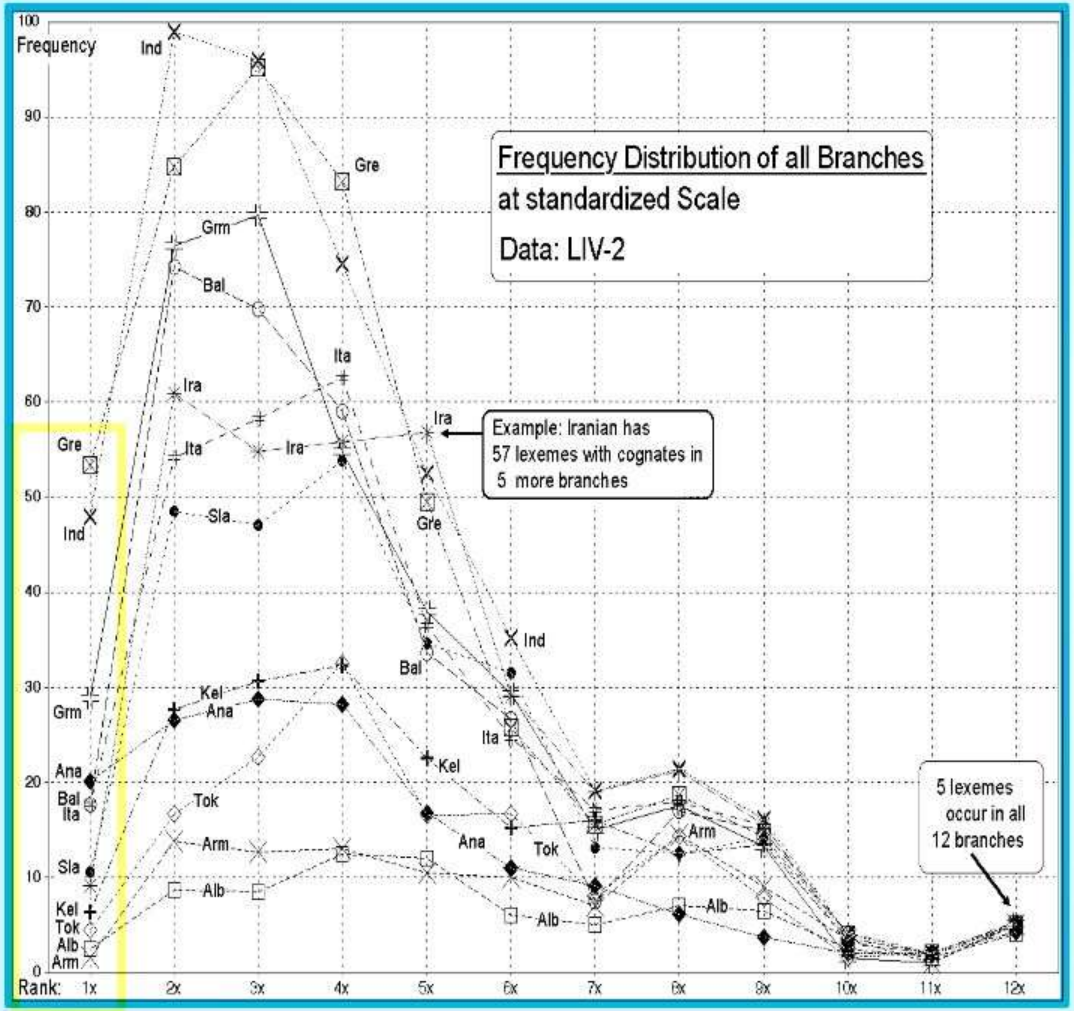
grm	bal	sla	kel	ita	phr	ana	tok	alb	arm	ira	ind	gre	Σ
NA	NA	NA	NA	NA	NA	1.0	NA	NA	NA	NA	NA	1.0	2.0
NA	NA	NA	NA	NA	NA	1.0	NA	NA	NA	NA	NA	1.0	2.0
NA	NA	NA	NA	NA	NA	1.0	NA	NA	NA	NA	NA	NA	1.0
NA	NA	NA	NA	1.0	NA	NA	1.1	NA	NA	NA	NA	NA	2.0
1.0	NA	NA	1.1	1.0	NA	NA	1.0	NA	1.0	1.0	1.0	1.0	8.0
NA	NA	0.5	NA	1.0	NA	NA	1.0	NA	1.0	NA	NA	1.0	5.0
1.0	NA	NA	1.0	NA	NA	NA	NA	NA	NA	NA	NA	1.0	3.0
NA	NA	NA	NA	NA	NA	NA	NA	NA	1.0	NA	NA	1.0	2.0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.0	1.0	2.0
NA	NA	NA	NA	1.0	NA	NA	NA	NA	NA	NA	NA	0.1	
NA	1.0	1.0	1.0	1.1	NA	NA	NA	NA	0.5	1.1		NA	
1.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.0			
1.0	NA	NA	1.0	1.0	NA	NA	NA	NA	NA				
1.0	NA	NA	NA	NA	NA	NA	NA	NA	NA				
1.1	NA	NA	1.0	1.1	NA	NA	NA	NA					
NA	NA	NA	NA	1.1	NA	NA	NA						
NA	NA	NA	NA	NA	NA	NA	NA						
NA	NA	NA	NA	0.4	NA								
NA	NA	NA	NA	NA	NA								
NA	NA	NA	NA	1.0									
NA	NA	1.1	NA	1.0									
NA	NA	NA	1.0	NA									
NA	NA	NA											
NA	NA	NA											
NA	NA	NA											
1.0	NA												
NA	NA												
	NA												

Cross-totals of digits

Sort all to Σ

= (here) 12 blocks / slices
assumed to have equal chance of survival

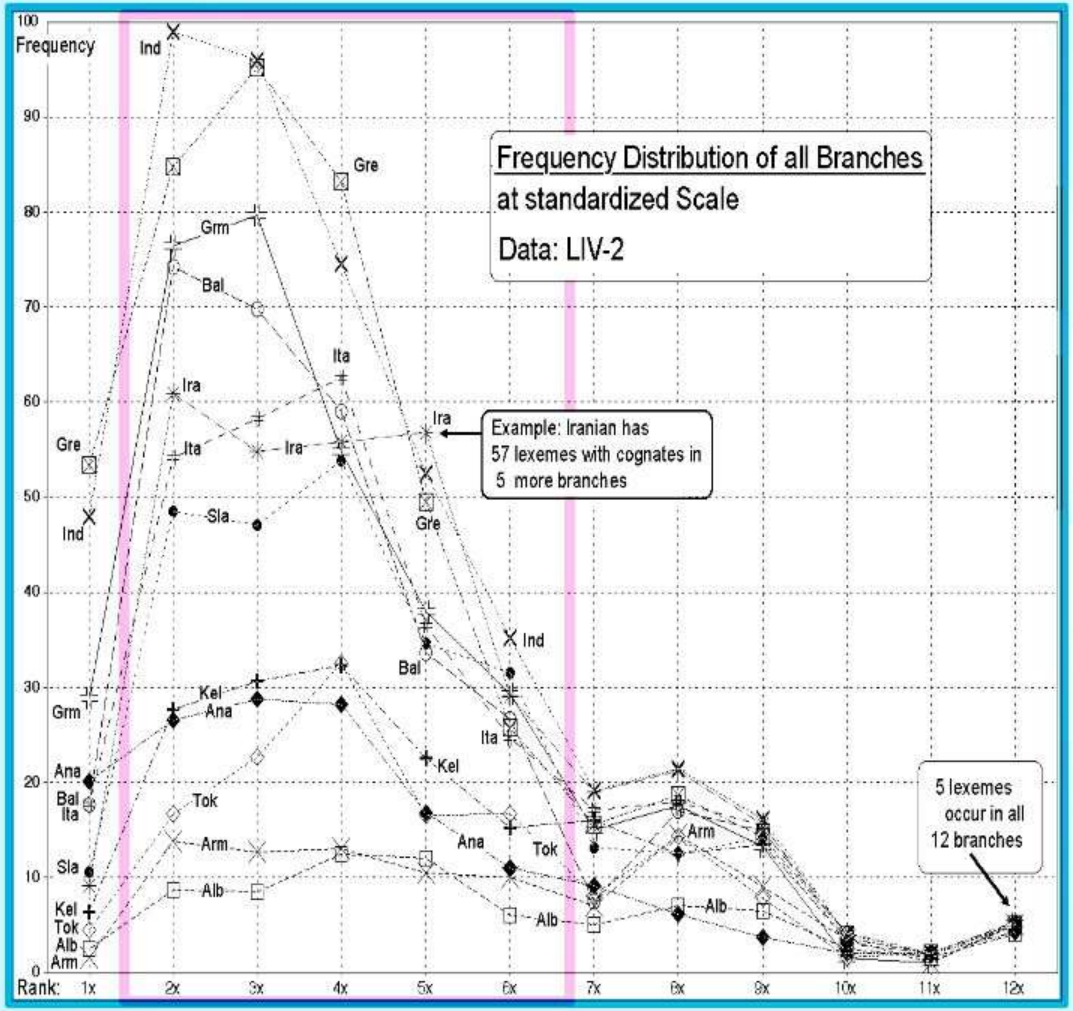
4.3 Analysis of the Distribution



Frequency-ordered data display:

- extreme left: some verbs in one language only

4.3 Analysis of the Distribution

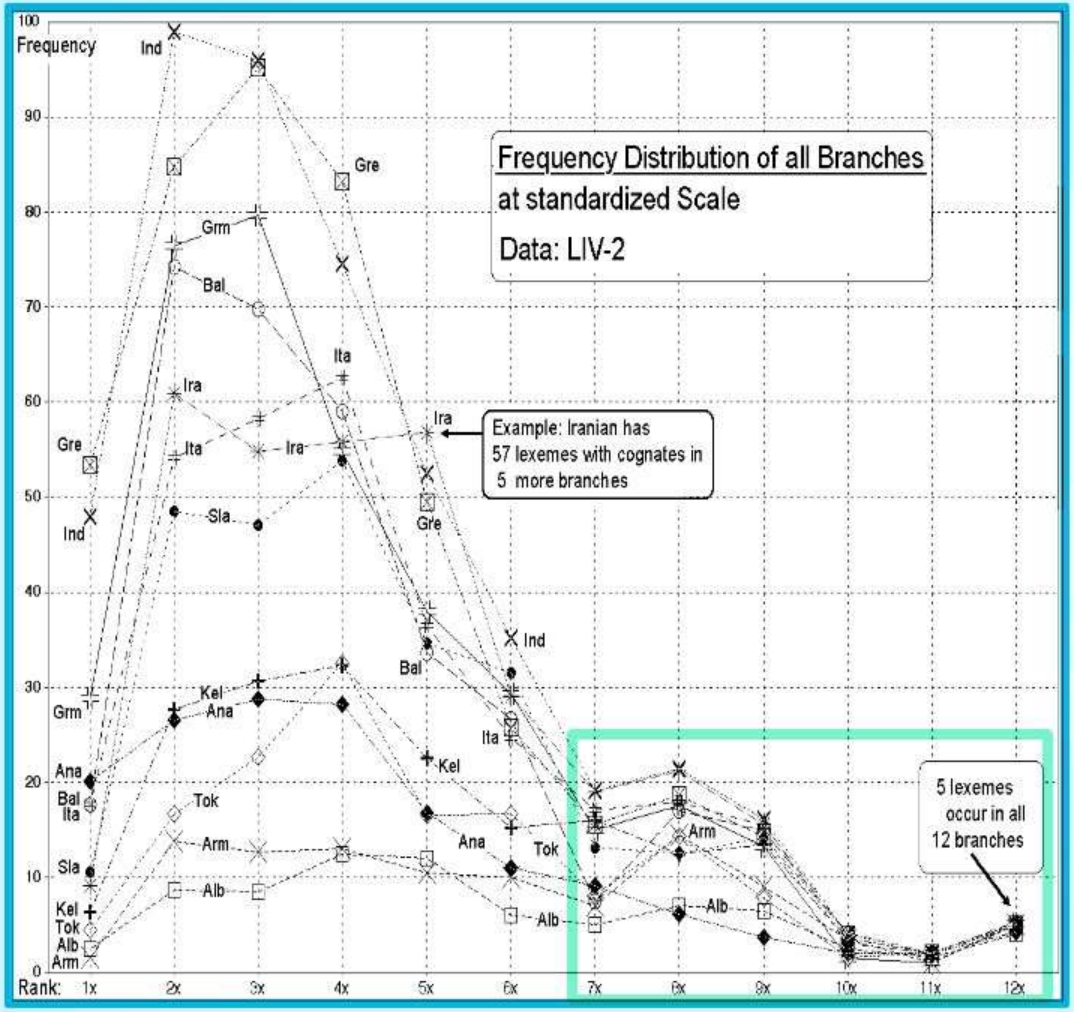


Frequency-ordered data display:

- extreme left: some verbs in one language only

- left hand: many verbs in few languages

4.3 Analysis of the Distribution



Frequency-ordered data display:

- extreme left: some verbs in one language only

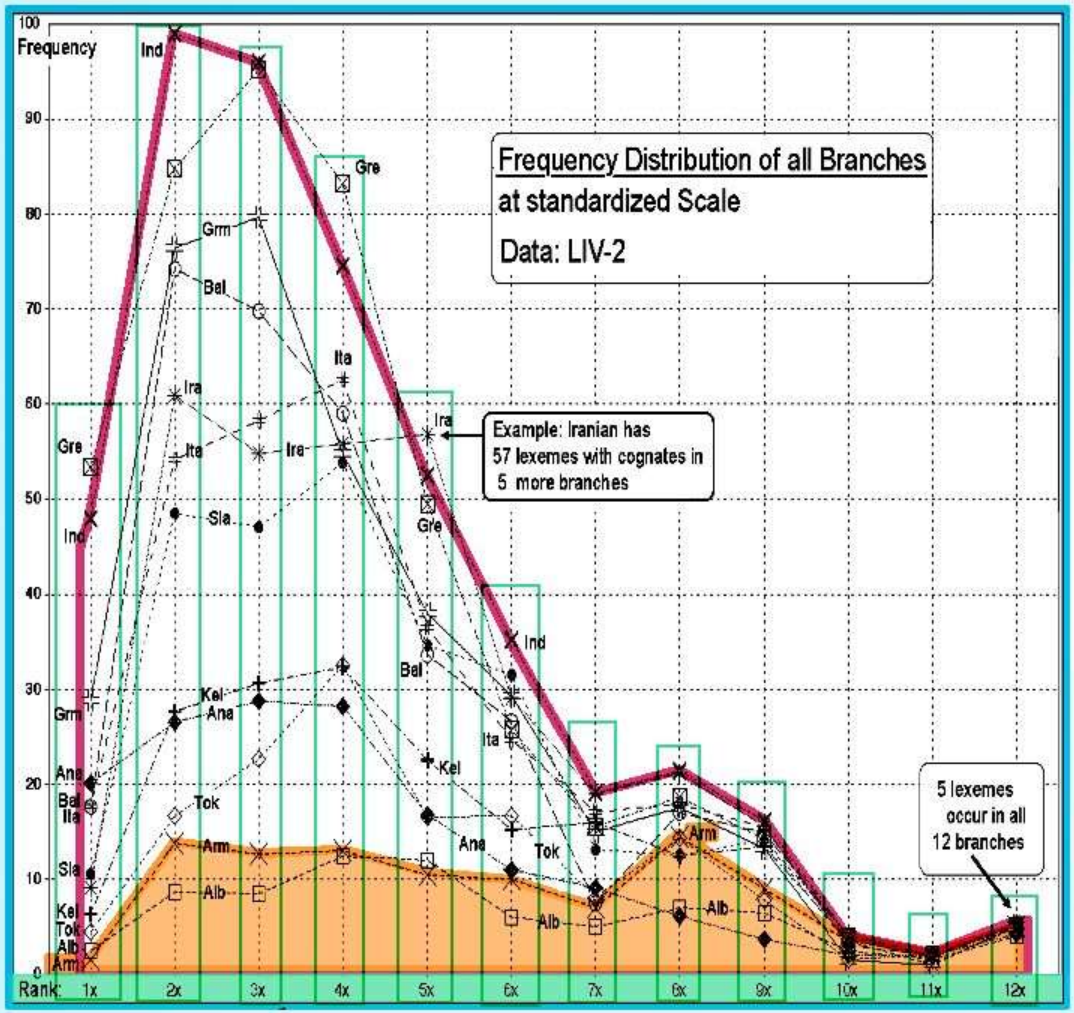
- left hand: many verbs in few languages

- right hand: fewer verbs in many languages

Question:

Where are connections with our formula ??

4.4 Detecting the Reason

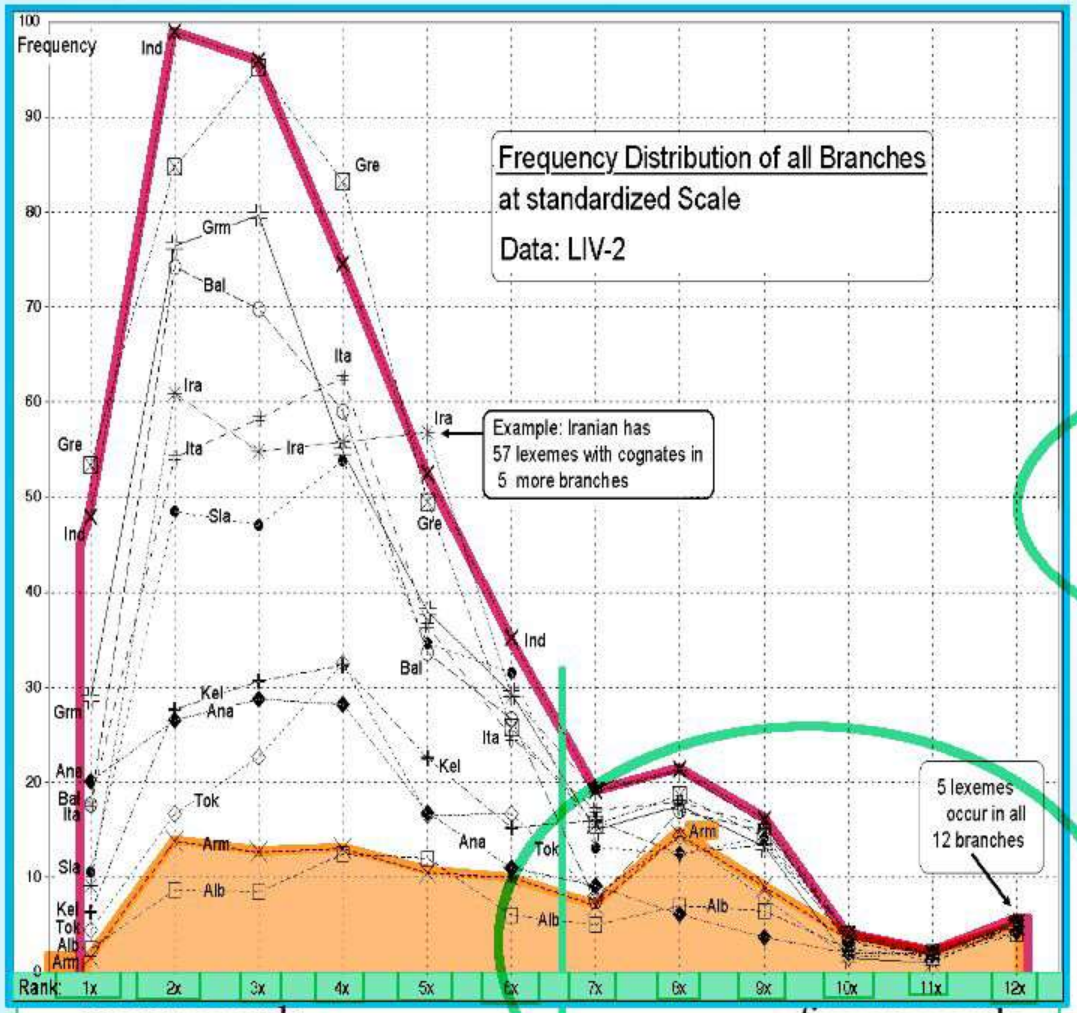


'k' preserved cognates?
= area below curve!

a agreements?
= frequency / rank slices!

What, then is wrong here?

4.4 Detecting the Reason



'k' preserved cognates?
= area below curve!
a agreements?
= frequency / rank slices!

we perceive: languages with low 'k' own relatively higher proportion of 'a'.

Since a is denominator in

$$\hat{N} = \frac{k_1 \bullet k_2}{a_{1,2}}$$

Result = false earlier split

5.1 Operationalization

- Calculate only data with same chance of being changed, >

never use total numbers, but

each slice at a time

But what about the scatter?

5.1 Operationalization

- Calculate only data with same chance of being changed, >

never use total numbers, but

each slice at a time

- Slices must be big enough to avoid unacceptable scatter, > **use not**
  - low frequency (left hand) slices alone, because low agreements > extreme scatter
  - high frequency (right hand) slices alone, because insignificant = uninformative

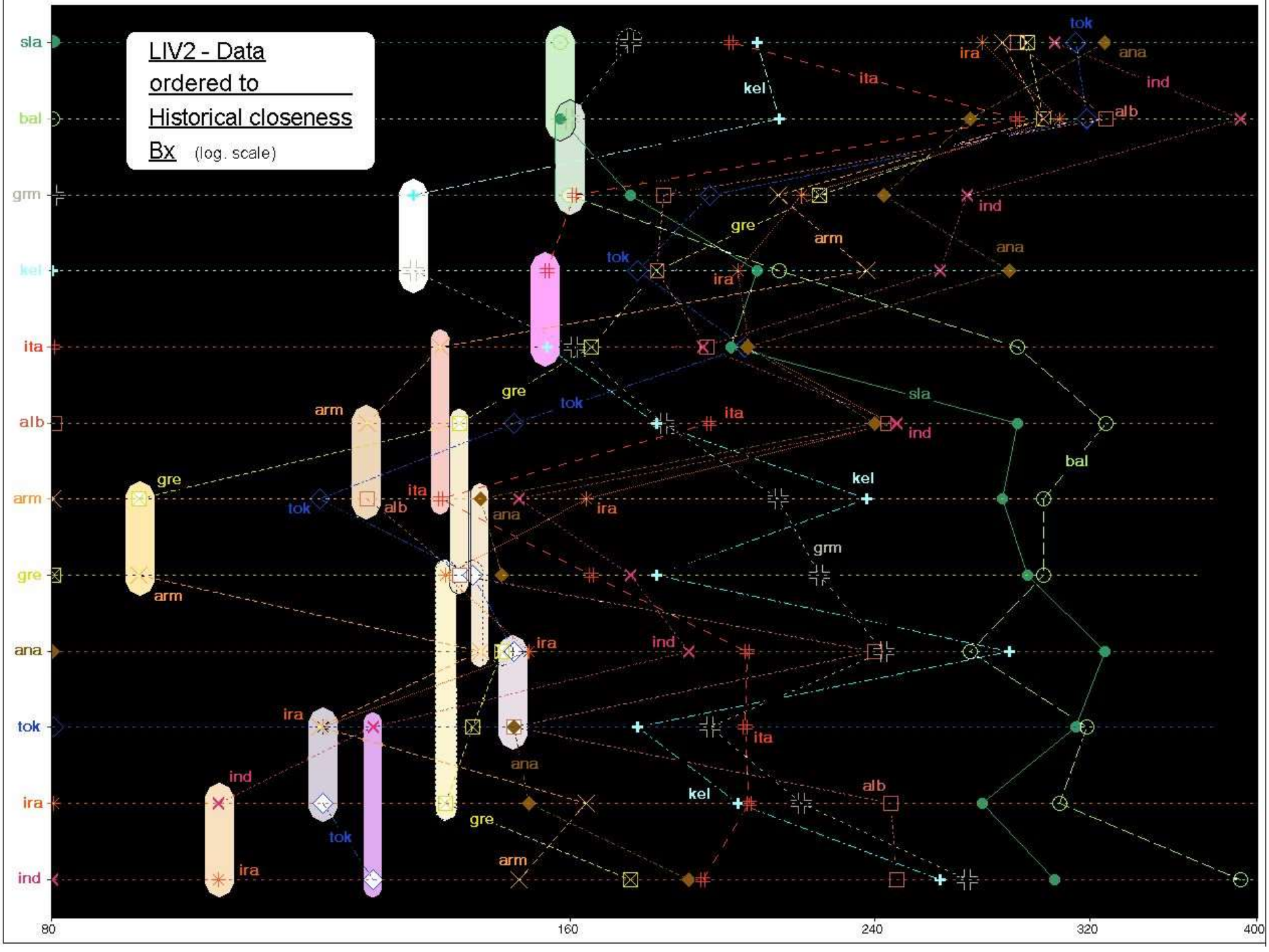
Best: Use all slices

6.1 From Final Matrix to New Subgrouping

- take arithmetic mean of all slices per language
  (eventually standardize to 100)
  → final matrix of 11·12 / 2 = 66 nodes N between every pair of languages

---

Useful: Flatten the unsorted sequences according to prior knowledge or Bx-method (Holm 2005:640)


LIV2 - Data ordered to Historical closeness Bx (log. scale)
sla
bal
grm
kel
ita
alb
arm
gre
ana
tok
ira
ind
80
160
240
320
400

6.1 From Final Matrix to New Subgrouping

- take arithmetic mean of all slices per language
  (eventually standardize to 100)
  → final matrix of 11·12 / 2 = 66 nodes N between every pair of languages

---

- Building the tree
  Not one way hill-climbing - since no clusters, but proceed on 'broad front', first finding next node for every language separately!

6.1 From Final Matrix to New Subgrouping

Bx-flattening helped us to pre-order the data

Now we can reconstruct the tree, proceeding at broad front:

Lang	Sla	Bal	Grm	Kel	Ita	Alb	Arm	Gre	Ana	Tok	Ira	Ind
Sla	Sla	74	104	120	157	110	140	135	135	121	131	163
Bal		Bal	105	115	120	103	141	143	185	140	141	143
Grm			Grm	110	113	111	124	115	141	132	137	135
Kel				Kel	93	126	115	138	122	136	130	142
Ita					Ita	103	118	106	129	117	124	122
Alb						Alb	93	94	98	124	115	113
Arm							Arm	96	106	119	116	113
Gre								Gre	116	114	134	108
Ana									Ana	100	124	108
Tok										Tok	112	109
Ira											Ira	80
Ind												Ind
K :	265.4	308.1	332.1	181.5	299.1	74.8	100	398.7	139.6	145.4	323.8	424.9

highest

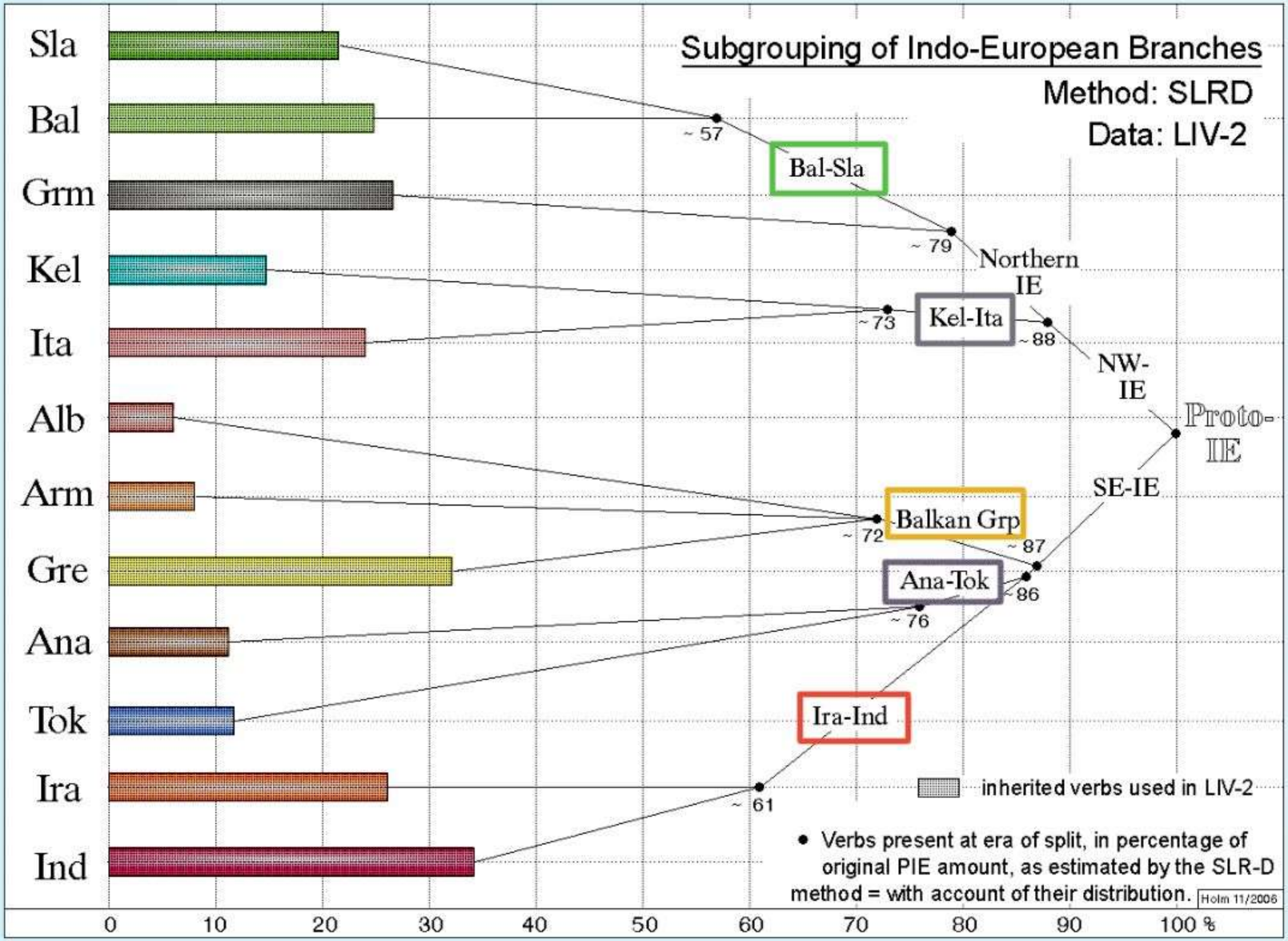
median

Latest

splits

Data: Arithmetic Mean of Rank Slices; LIV-2; Method: SLRD

7.1 The Tree



(by re-itering this process ..)

7.2 Discussion

Found: bias in distribution, but
Possibly more hidden bias in data from:

- Extremely different **cultural background**, e.g.
  - hunter- & gatherer communities in the north vs.
  - advaned civilisations in Anatolia
- Differences in **reliability of research** itself
- Peripheral (=conservative) vs. central (innovative) **position** of languages.
  (Opposite to Nichols and MDS, which hold that changes must increase with distance)

7.3 Conclusion

Linguistically:

Result refutes early split of / from Anatolian !

Methodologically:

Distributional bias considered in now

"Separation Level Recovery accounting for Distribution" (SLRD).

- Regrettably: Large amount of data needed -

7.3 Conclusion

Linguistically:

Result refutes early split of / from Anatolian !

Methodologically:

Distributional bias considered in now

"Separation Level Recovery accounting for Distribution" (SLRD).

- Regrettably: Large amount of data needed -

What should we have learnt?

7.3 Conclusion

Never trust methods that only crank data through parsimony, compatibility, or MrBayes packages without regarding their hypergeometric behavior.

Note that even apparently good results regularly appear, due to

- very strong signals, or
- simply chance.

7.4 Outlook and Test in Real Environment

Any subgrouping result must be projectable into real geography!!