Bioinformatische Methoden in der Historischen Linguistik
Aggregating word alignments

Gerhard Jäger

8. Februar 2013
Forum Scientiarum

From words to languages

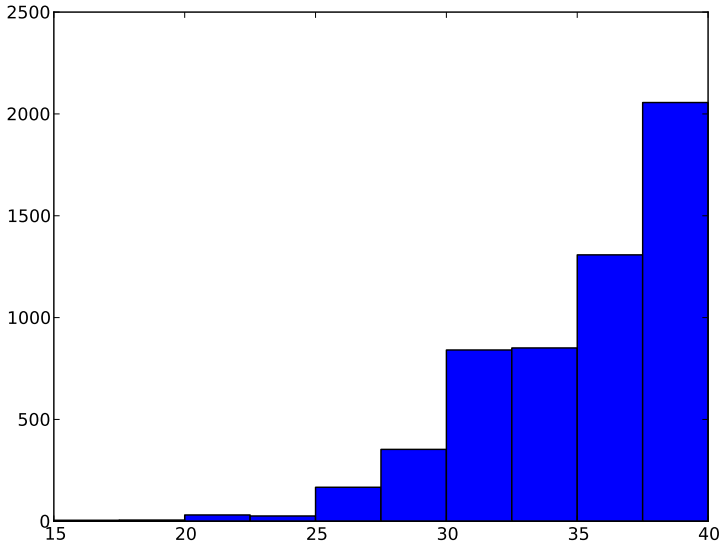
- alignment methods give us a measure of distance/similarity between individual words
- these need to be aggregated to get a distance measure between languages
- baseline approach to compute distance between L_1 and L_2:
  - compute Levenshtein distance between all 40 translation pairs from L_1 and L_2

$$d(L_1, L_2) = \sum_{i=1}^{N} \frac{d(w_i^{L_1}, w_i^{L_2})}{N}$$

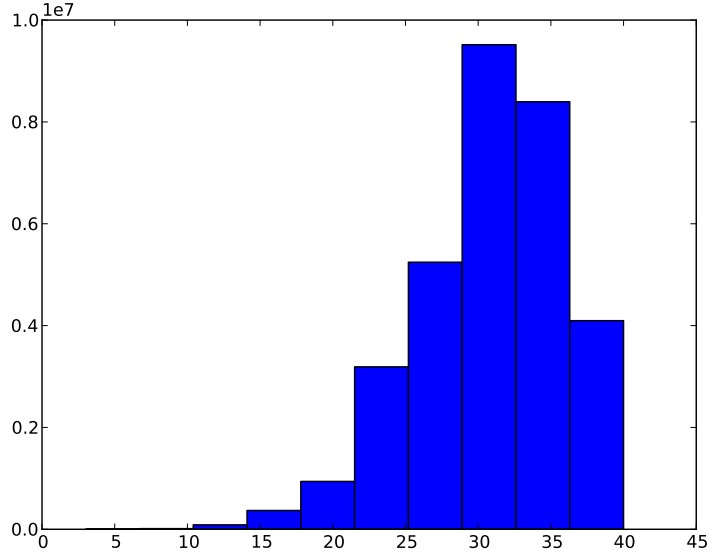
  where N is the number of concepts where we have a word from both languages
  - substantial number of missing data; N is often much smaller than 40

Missing data

- on average we actually only have 35.1 words per language
- if attested loans are excluded, the number goes down to 34.8

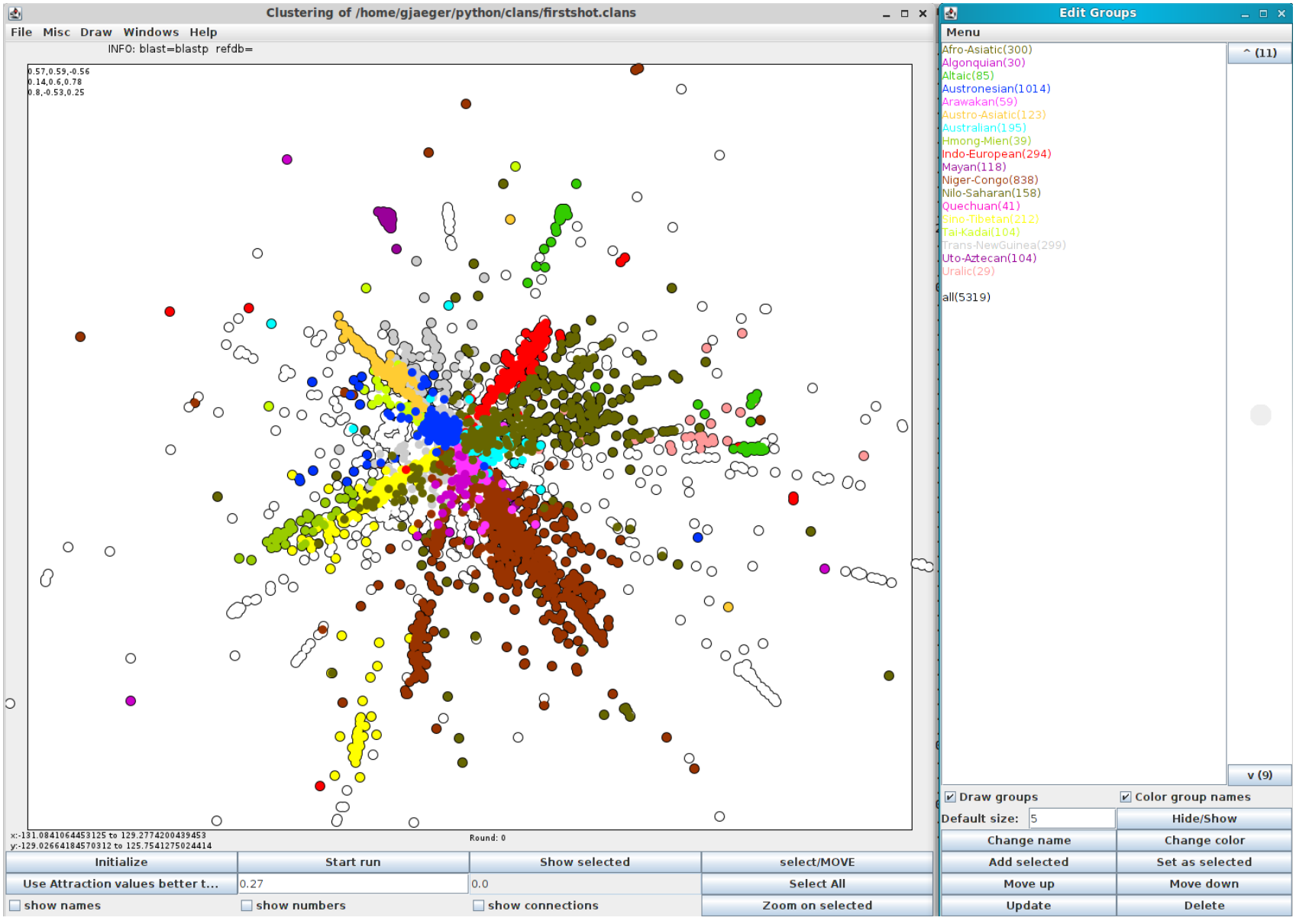


number of items per language



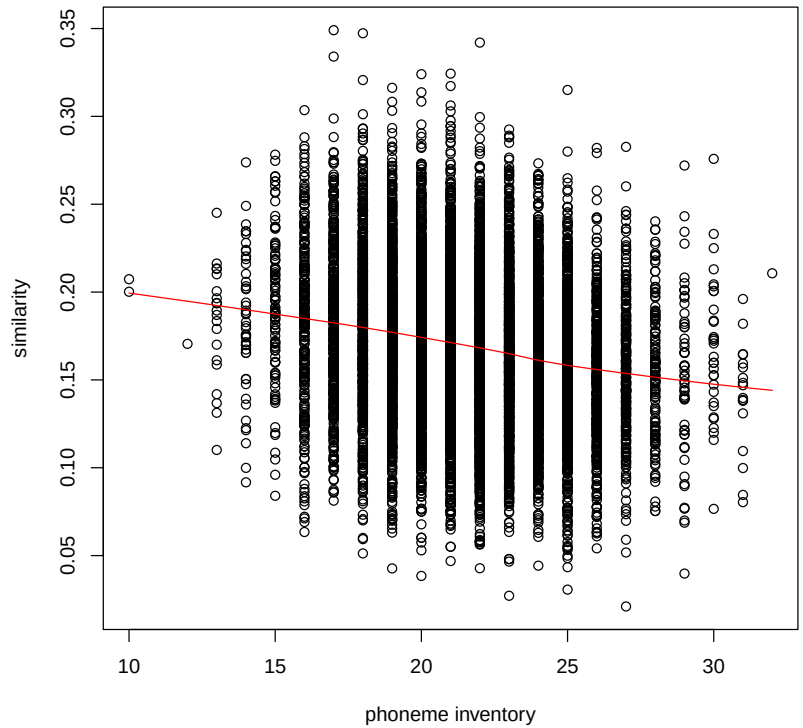
number of shared items per language pair

Evaluation


Clustering of /home/gjaeger/python/clans/firstshot.clans
File Misc Draw Windows Help
INFO: blast=blastp refdb=
0.57,0.59,-0.56
0.14,0.6,0.78
0.8,-0.53,0.25
x:-131.0841064453125 to 129.2774200439453
y:-129.02664184570312 to 125.7541275024414
Round: 0
Initialize
Start run
Show selected
select/MOVE
Use Attraction values better t...
0.27
0.0
Select All
show names
show numbers
show connections
Zoom on selected
Edit Groups
Menu
Afro-Asiatic(300)
Algonquian(30)
Altaic(85)
Austronesian(1014)
Arawakan(59)
Austro-Asiatic(123)
Australian(195)
Hmong-Mien(39)
Indo-European(294)
Mayan(118)
Niger-Congo(838)
Nilo-Saharan(158)
Quechuan(41)
Sino-Tibetan(212)
Tai-Kadai(104)
Trans-NewGuinea(299)
Uto-Aztecan(104)
Uralic(29)
all(5319)
^ (11)
v (9)
Draw groups
Color group names
Default size:
5
Hide/Show
Change name
Change color
Add selected
Set as selected
Move up
Move down
Update
Delete

Evaluation

- basic problem here: the smaller the sound inventories of the languages compared, the higher is the probability of false positives



Benchmark: LDND measure

- Wichmann et al.: doubly normalized Levenshtein distance (**L**evenshtein **D**istance **N**ormalized and **D**ivided)
- normalization for word length

$$\mathrm{nld}(x, y) \doteq \frac{d_{Lev}(x, y)}{\max(l(x), l(y))} \tag{1}$$

- normalization for language specific patterns (including sound inventory size):
  - normalization factor $1/\mu$
  - μ_{L_1,L_2}: mean of $\{nld(x, y) | x \in L_1, y \in L_1, \|x\| \neq \|y\|\}$

$$\begin{aligned} \mathrm{ldnd}(x, y, L_1, L_2) &\doteq \frac{nld(x, y)}{\mu_{L_1,L_2}} \\ \mathrm{ldnd}(L_1, L_1) &\doteq \frac{\sum_{x \in L_1, y \in L_2} \{\mathrm{ldnd}(x, y, L_1, L_2) : \|x\| = \|y\|\}}{\#\{x, y : \|x\| = \|y\|\}} \end{aligned}$$

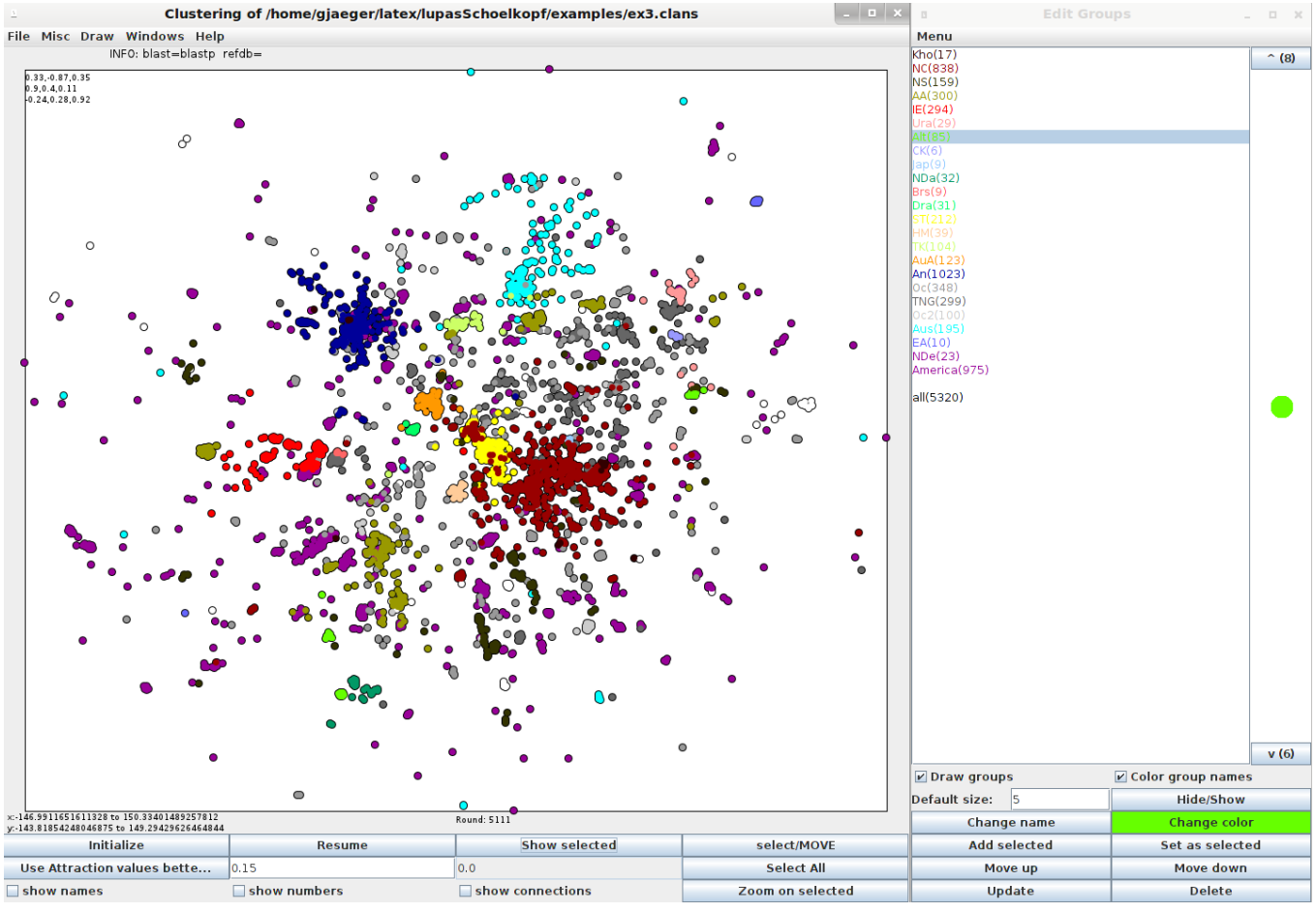
Benchmark: LDND measure

English / Swedish

	Ei	yu	wi	w3n	tu	fiS	...
yog	1	$^2/_3$	1	1	1	1	
du	1	$^1/_2$	1	1	$^1/_2$	1	
vi	$^1/_2$	1	$^1/_2$	1	1	$^2/_3$	
et	1	1	1	1	1	1	
tvo	1	1	1	1	$^2/_3$	1	
fisk	$^3/_4$	1	$^3/_4$	1	1	$^1/_2$	
⋮							

- average LDN along diagonal: 0.56
- average LDN off diagonal: 0.91
- LDND: $^{0.56}/_{0.91} = 0.61$

Benchmark: LDND measure

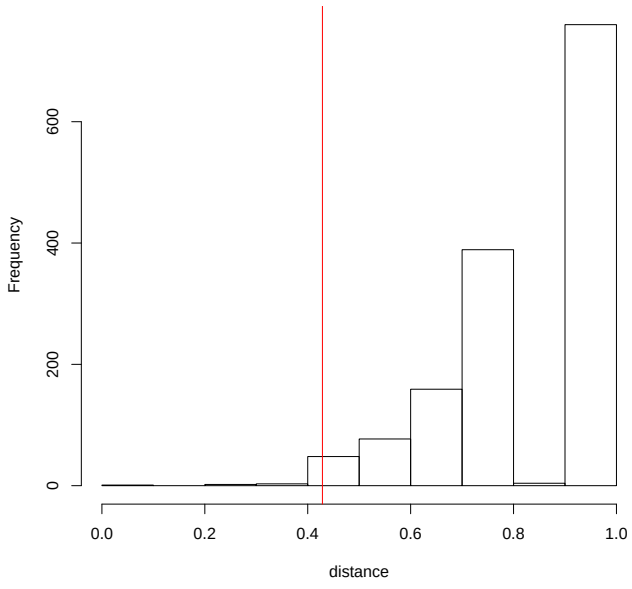


A bit of information theory

Swedish *fisk* = English *fish*?
Turkish *dört* = English *dirt*?

- first guess is good because the words sound similars **and the languages are closely related**
- second guess is bad (and wrong) even though the words sound similar **because the languages are not related**
- If two languages are related, knowing a word from one language reduces the uncertainty about its form in the other language
- *Hypothesis: degree of similarity between two languages* $\approx$ *average amount of information that the form of a word in one language carries about the form of its translation into the other language*

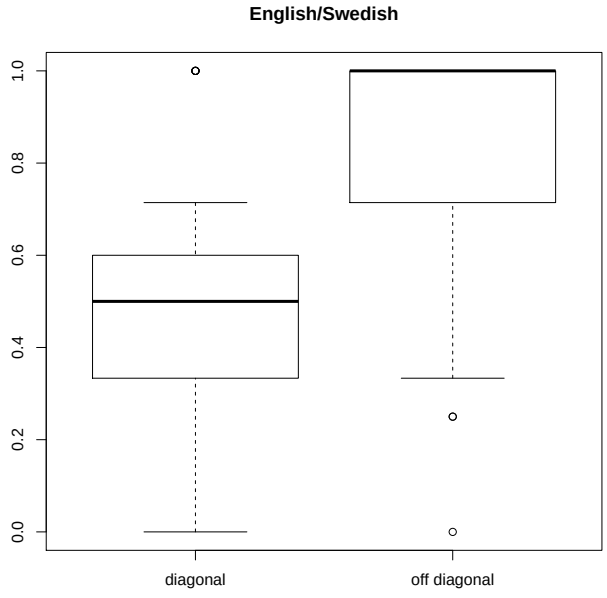
English and Swedish again

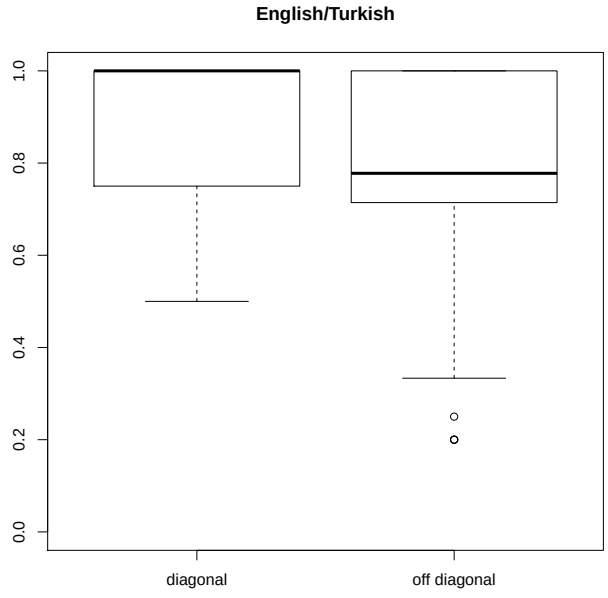


- Histogramm: off-diagonal distances
- red line: distance *fiS* $\sim$ *fisk* ($= 4.3$)
- relative frequency of off-diagonal entries ≤ 4.3: 0.004
- can be interpreted as *p*-value for the null hypothesis that the two words are not cognates
- $-log_2(0.004) = 7.9$ bit: amout of information that [fisk] carries about [fiS], given the general pattern of phonotactic similarities between unrelated English and Swedish words

Information theoretic estimate of language similarity

- similarity between two languages: average amount of information that a word from one language carries about its translation
- formally: average binary logarithm of the p-values for all Swadesh items in the data base





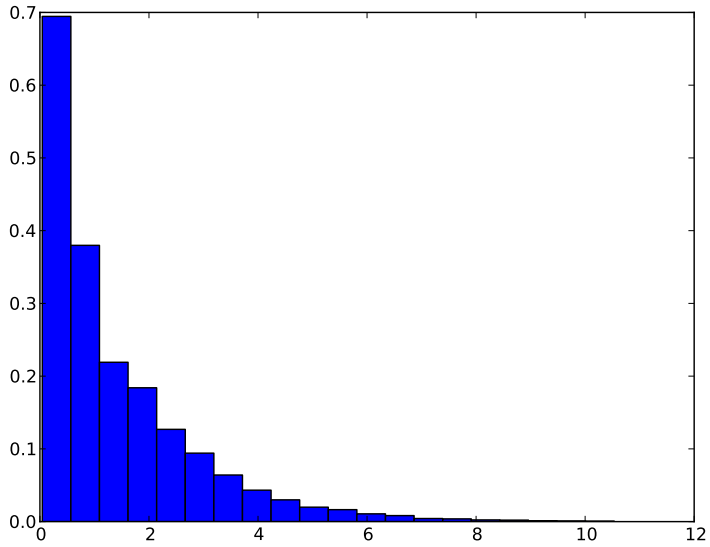
Information theoretic estimate of language similarity

- formally:
  - let $d(w_i^{L_1}, w_j^{L_2})$ be the normalized Levenshtein distance of the i-th word from L_1 and the j-th word from L_2 and N the number of shared concepts of L_1 and L_2.

$$pv(w_i^{L_1}, w_i^{L_2}) = \frac{|\{(j,k)|i \neq j, d(w_j^{L_1}, w_k^{L_2}) < d(w_i^{L_1}, w_i^{L_2})\}|}{N(N-1)}$$

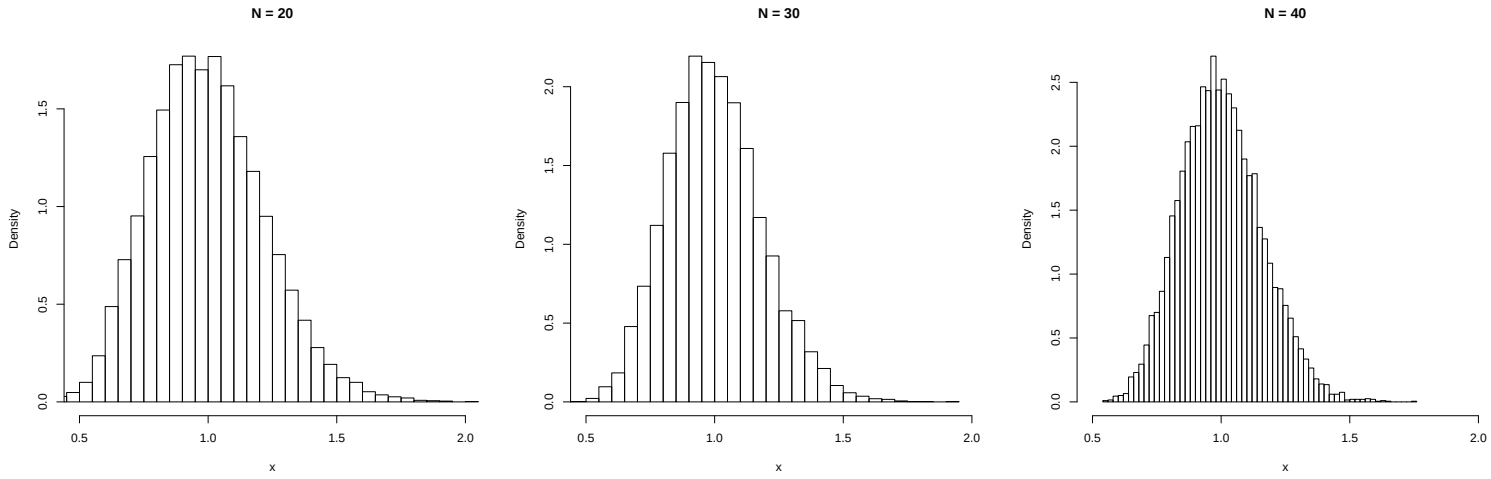
How to deal with missing data

- for unrelated languages, $pv(w_i^{L_1}, w_i^{L_2})$ is just a random variable
- approximately exponentially distributed (with mean $= \frac{1}{\log 2}$):



How to deal with missing data

- the mean of N exponentially distributed variables is approximately normally distributed[1]
- variance depends on N though



[1] Strictly speaking, it is a Erlang distribution, but for $N > 10$ or so, a normal distribution is a reasonable approximation.

How to deal with missing data

- let $x_1, \ldots, x_N$ be independent identically distributed random variables with standard deviation σ

$$sd(\frac{1}{N}\sum_{i=1}^{N} x_n) = \frac{\sigma}{\sqrt{N}}$$

- both mean and variance of the negative (binary) logarithms of the individual p-values are $\frac{1}{\log 2}$
- so the following function is standard normally distributed for unrelated languages

$$log2\sqrt{N}(\sum_{i=1}^{N} pv(w_i^{L_1}, w_i^{L_2}) - \frac{1}{\log 2})$$

How to deal with missing data

- the following function gives the probability that the degree of similarity that we find between L_1 and L_2 is due to chance:

$$d(L_1, L_2) = \text{erfc}(\log 2\sqrt{N}(\sum_{i=1}^{N} pv(w_i^{L_1}, w_i^{L_2}) - \frac{1}{\log 2}))$$

- as the *complementary error function* erfc is monotonically decreasing, we can define the similarity betwen L_1 and L_2 as

$$\text{sim}(L_1, L_2) = \sqrt{N}(\sum_{i=1}^{N} pv(w_i^{L_1}, w_i^{L_2}) - \frac{1}{\log 2})$$

Comparing unweighted and weighted alignment

- same procedure for aggregating word-alignments to language similarities can be applied to weighted alignments
- some results

Levensthein:

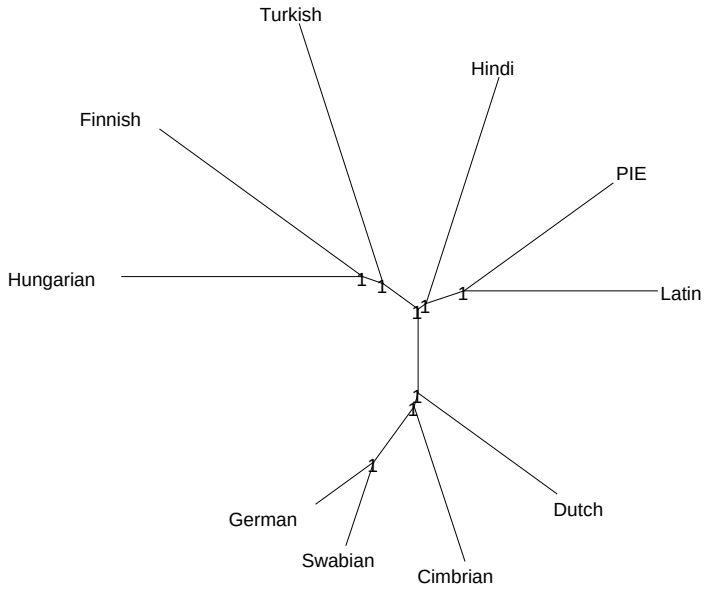
	German	Swabian	Cimbrian	Dutch	Hindi	PIE	Latin	Hungarian	Finnish	Turkish
German	45.7	35.2	26.7	25.8	10.1	14.9	10.9	6.8	6.0	6.9
Swabian	35.2	46.5	22.0	21.8	10.0	13.0	11.5	7.2	6.8	6.1
Cimbrian	26.7	22.0	42.3	20.7	11.8	10.7	9.8	5.9	6.7	6.1
Dutch	25.8	21.8	20.7	45.7	9.5	14.0	11.2	6.9	5.7	5.1
Hindi	10.1	10.0	11.8	9.5	45.7	14.4	12.1	6.5	7.1	7.5
PIE	14.9	13.0	10.7	14.0	14.4	46.5	19.6	8.1	6.4	5.2
Latin	10.9	11.5	9.8	11.2	12.1	19.6	46.5	8.0	6.1	6.9
Hungarian	6.8	7.2	5.9	6.9	6.5	8.1	8.0	42.7	11.5	8.5
Finnish	6.0	6.8	6.7	5.7	7.1	6.4	6.1	11.5	37.9	7.4
Turkish	6.9	6.1	6.1	5.1	7.5	5.2	6.9	8.5	7.4	45.7

weighted alignment:

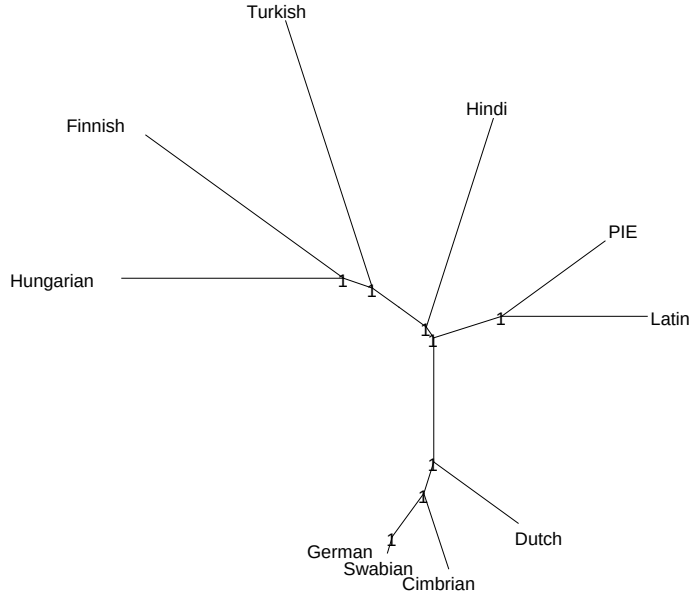
	German	Swabian	Cimbrian	Dutch	Hindi	PIE	Latin	Hungarian	Finnish	Turkish
German	42.3	36.3	31.6	29.0	12.0	16.9	12.2	7.5	7.2	6.3
Swabian	36.3	42.3	27.8	26.0	12.4	15.5	12.4	8.2	7.1	6.2
Cimbrian	31.6	27.8	40.8	24.8	13.0	12.8	10.9	7.5	7.5	6.4
Dutch	29.0	26.0	24.8	41.4	11.8	16.7	12.7	7.5	5.9	5.2
Hindi	12.0	12.4	13.0	11.8	42.9	14.6	13.3	8.1	6.9	7.2
PIE	16.9	15.5	12.8	16.7	14.6	45.8	22.6	8.3	7.7	5.2
Latin	12.2	12.4	10.9	12.7	13.3	22.6	44.2	7.4	6.3	7.5
Hungarian	7.5	8.2	7.5	7.5	8.1	8.3	7.4	42.9	12.3	9.0
Finnish	7.2	7.1	7.5	5.9	6.9	7.7	6.3	12.3	34.3	7.0
Turkish	6.3	6.2	6.4	5.2	7.2	5.2	7.5	9.0	7.0	44.2

Comparison

- applying Neighbor Joining phylogeny induction

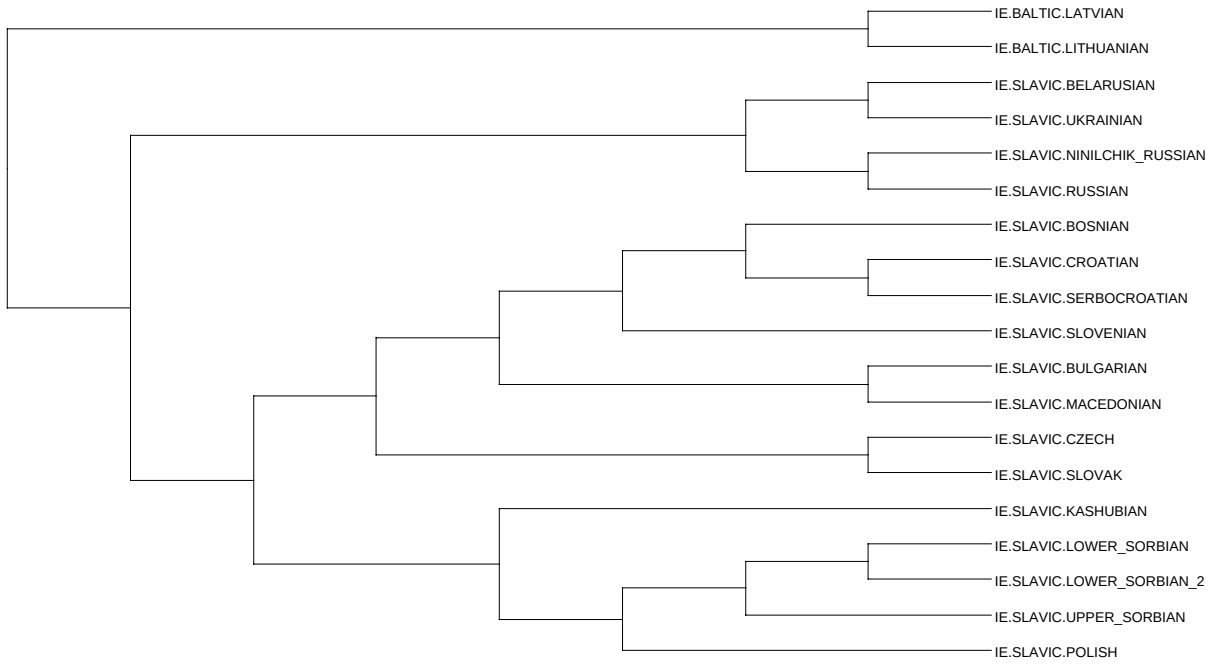


Levenshtein alignment

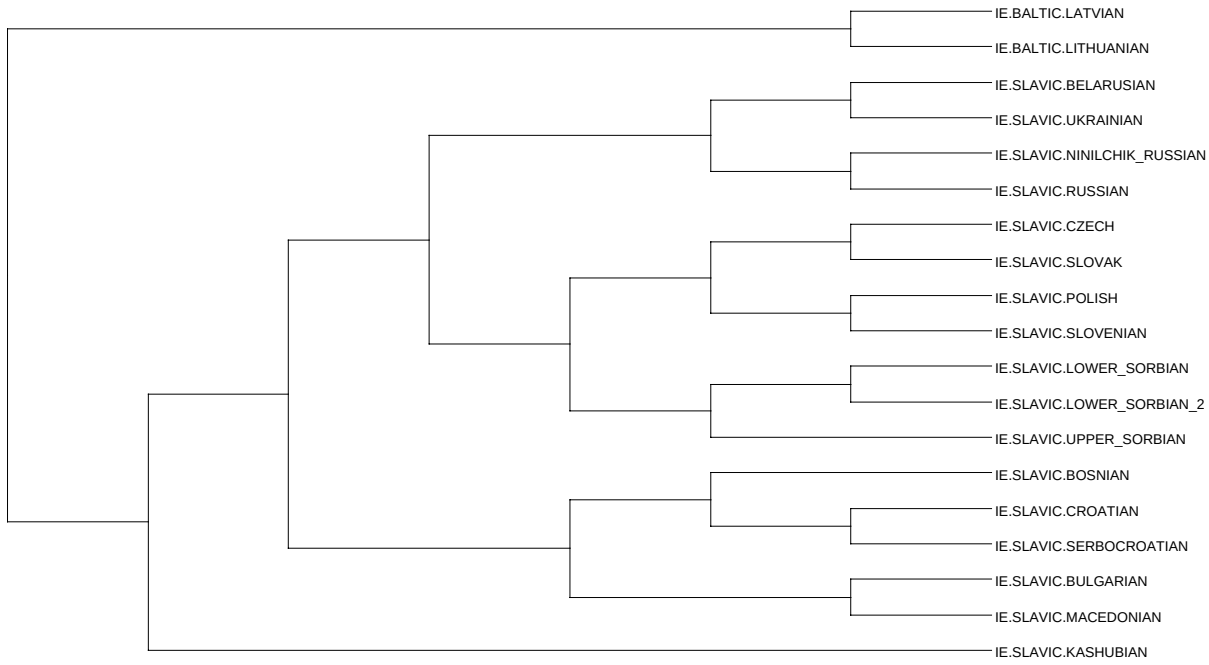


weighted alignment

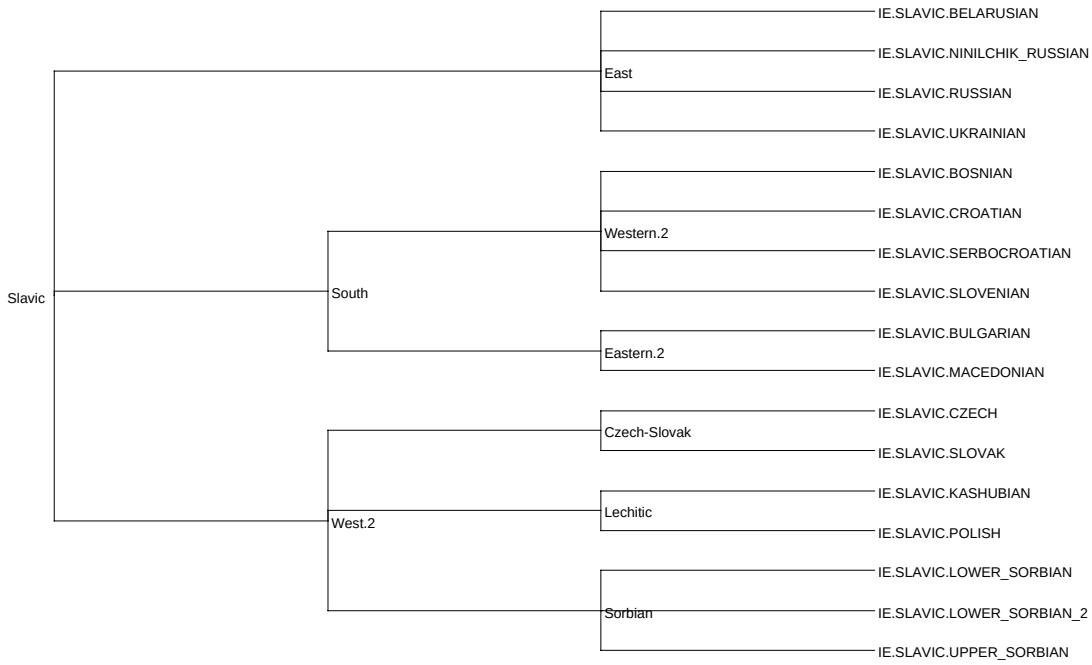
The Balto-Slavic languages: Levenshtein alignment



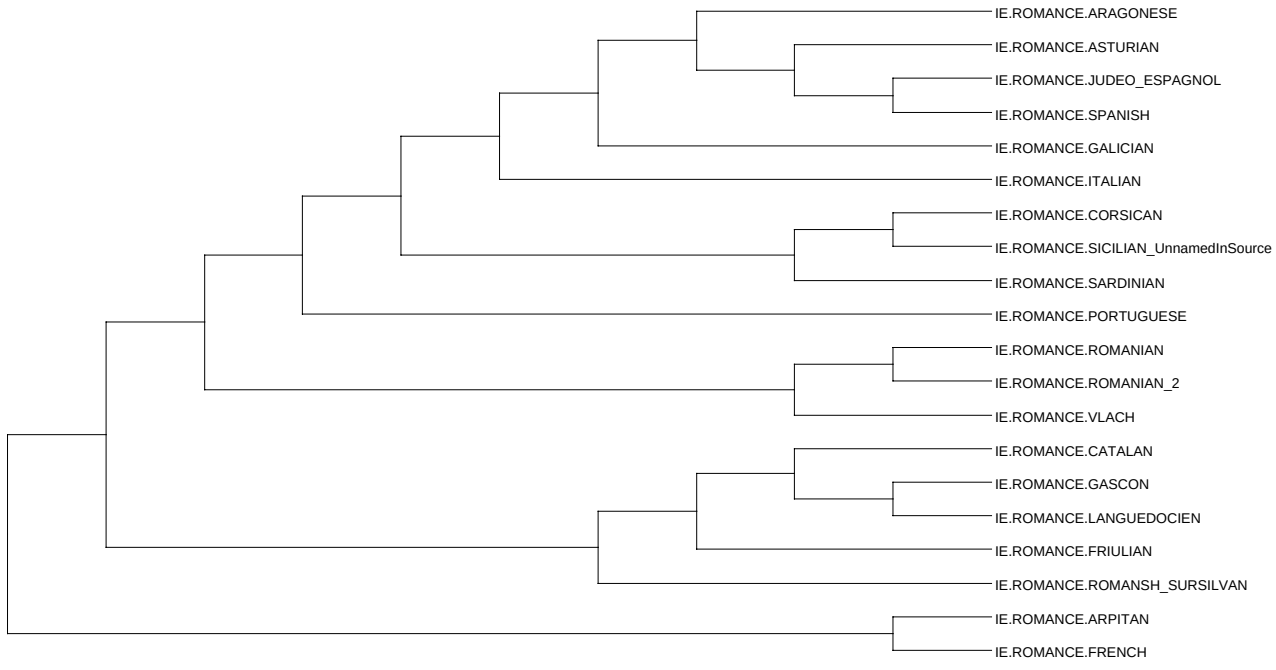
The Balto-Slavic languages: Weighted alignment



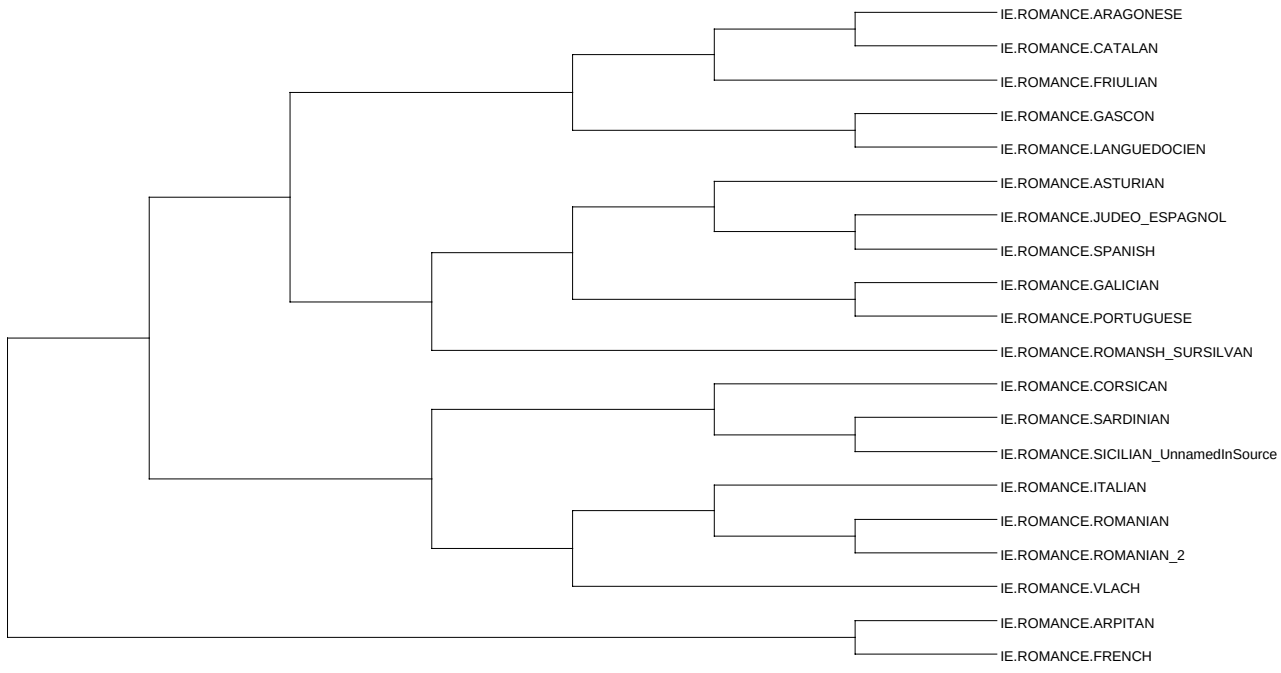
The Slavic languages: Ethnologue classification



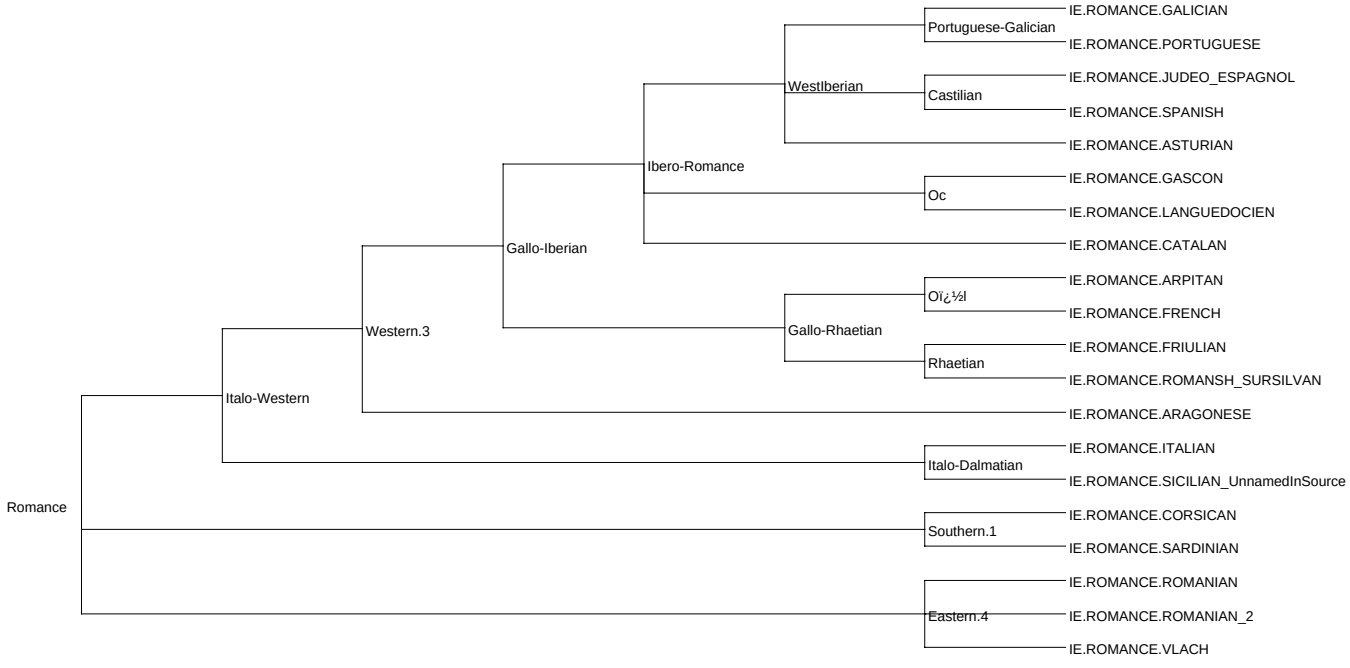
The Romance languages: Levenshtein alignment



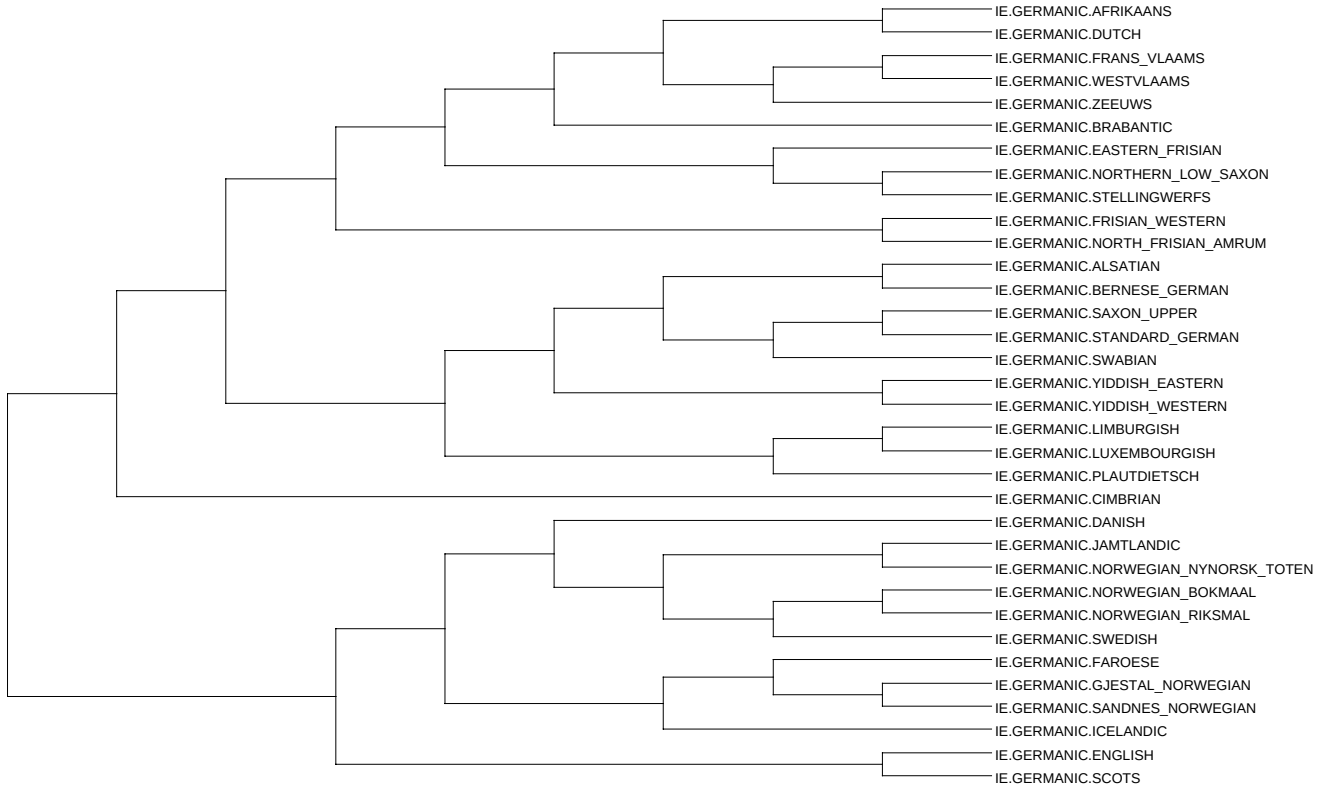
The Romance languages: Weighted alignment



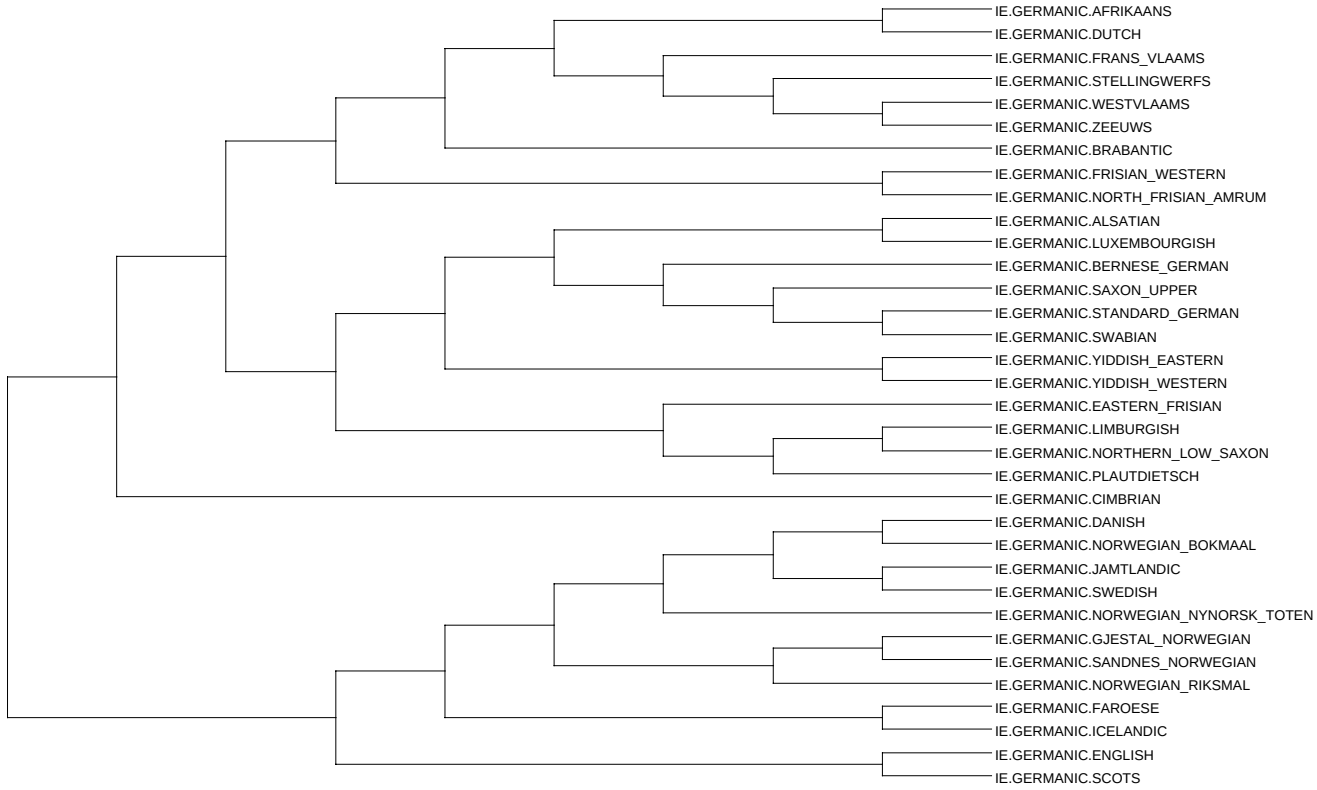
The Romance languages: Ethnologue classification



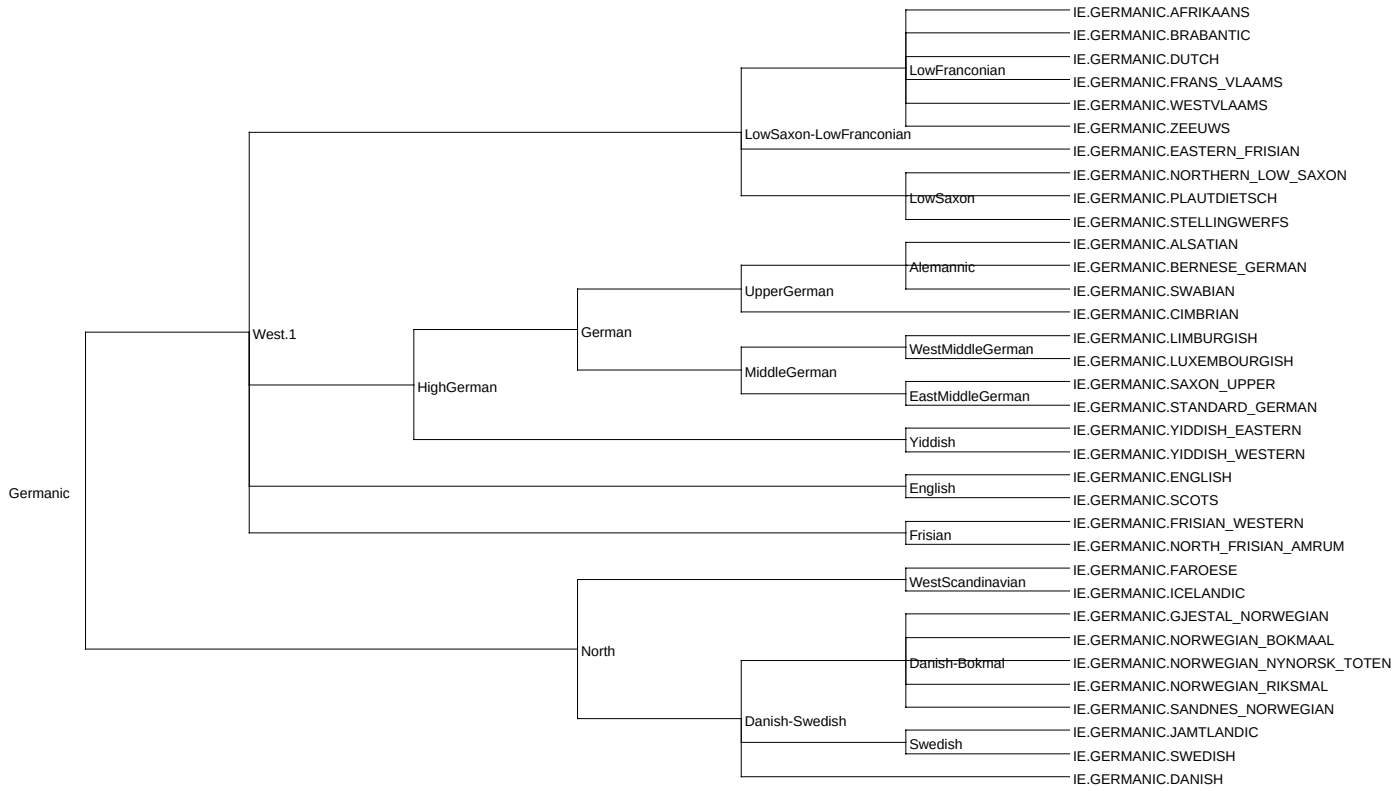
The Germanic languages: Levenshtein alignment



The Germanic languages: Weighted alignment

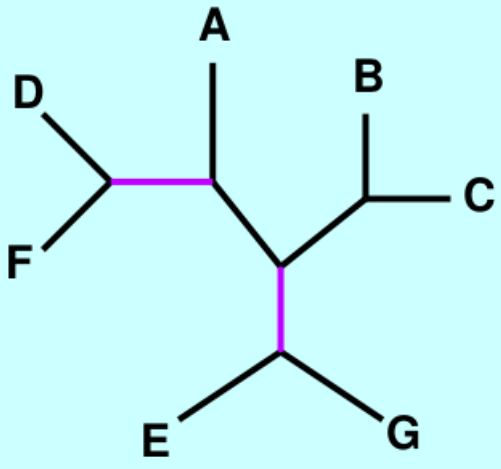


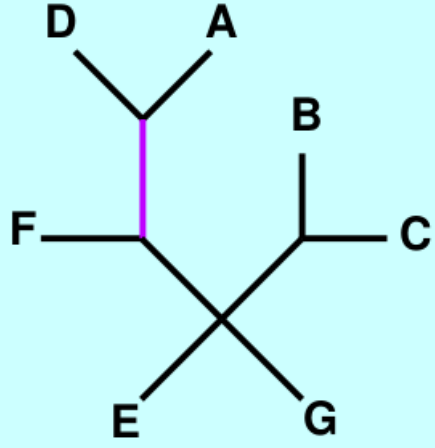
The Germanic languages: Ethnologue classification



Tree distances: Robinson-Fould

The symmetric difference metric





Partitions

{ADF | BCEG}
{DF | ABCEG}
{BC | ADEFG}
{EG | ABCDF}

Partitions

{ADF | BCEG}
{AD | BCEFG}
{BC | ADEFG}

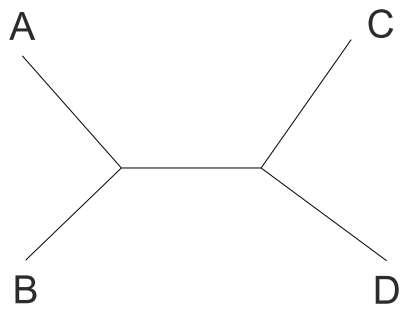
The symmetric difference is the number of partitions that are in one but not both of these lists, in this case 3.

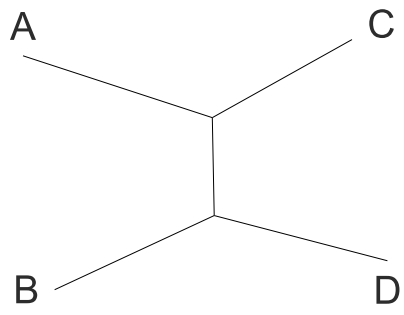
Tree distances: Robinson-Fould

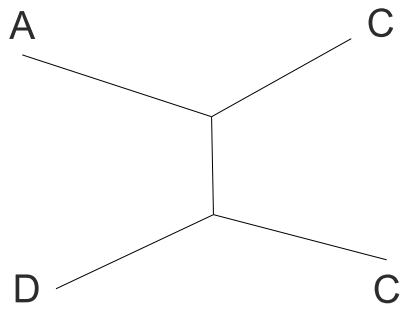
- normalized RF-distance: number of different partitions, divided by the total number of partitions in tree 1 + total number of partitions in tree 2
- in the example: $\frac{3}{4+3}$

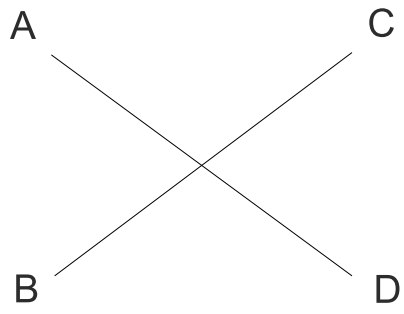
Tree distances: Quartet distance

- for a quartet of species, there are four possible tree topologies, 3 **butterflies** and 1 **star**





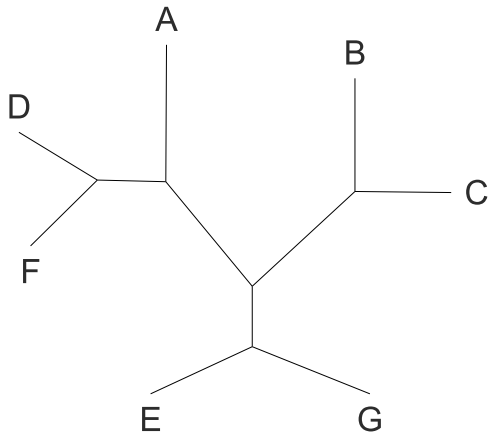


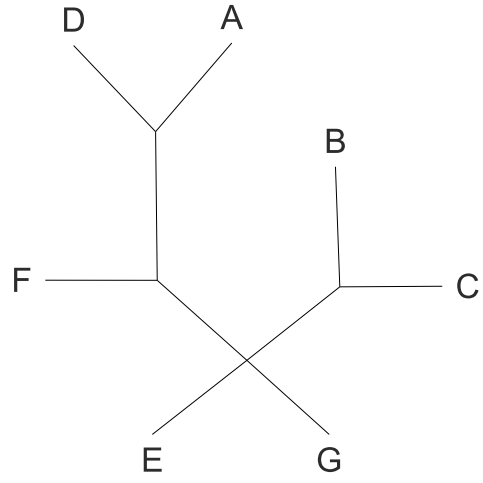


Tree distances: Quartet distance

- **quartet distance** between two unrooted trees is the number quartets that have a different topology in the two trees

Tree distances: Quartet distance

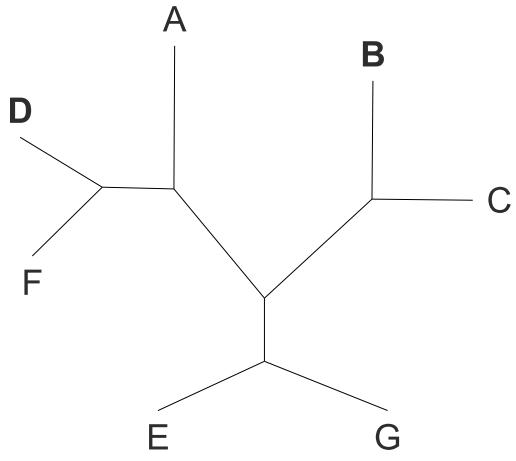


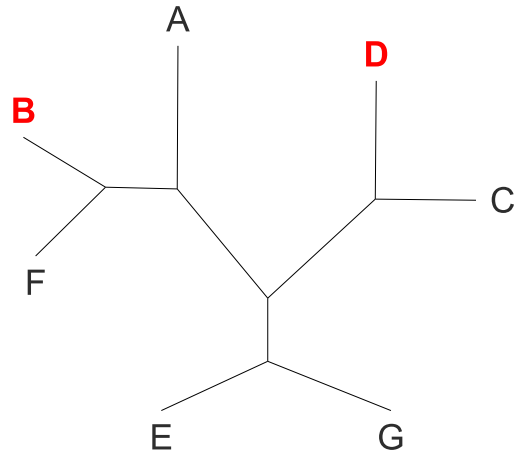


- $\binom{7}{4} = 35$ quartets in total
- 25 are shared, 10 are different
- normalized qdist: $\frac{10}{35}$

Tree distances

- Robinson-Fould distance is more intuitive, but quartet distance is more robust





- Robinson-Fould distance: 6; normalized $\frac{6}{8} = 0.75$
- quartet distance: 23; normalized $\frac{23}{35} \approx 0.66$

Expert trees

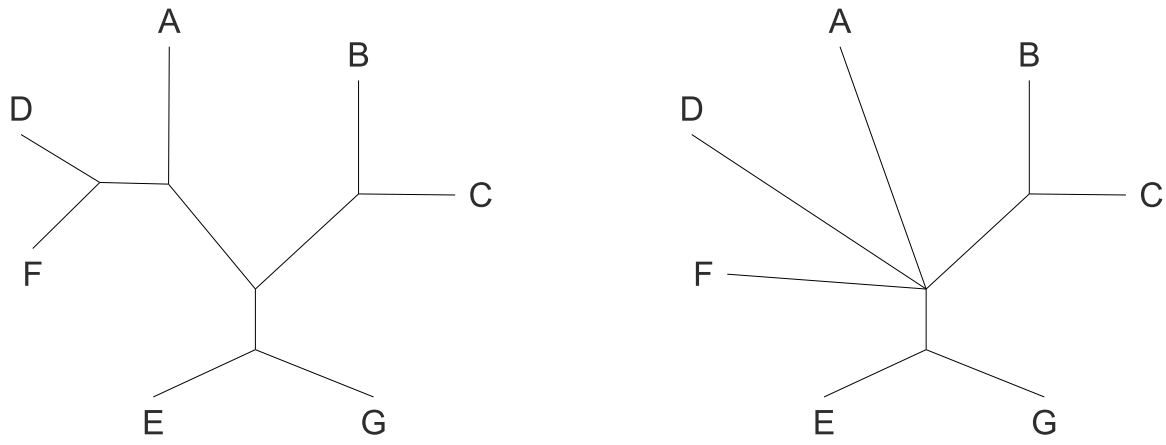
- quality of phylogenetic inference can be evaluated by comparison to **expert classifications**
- three commonly used classification systems:
  1. two-level taxonomy from WALS (World Atlas of Language Structure)
  2. multi-level taxonomy from Ethnologue
  3. more conservative multi-level taxonomy according to Harald Hammarström
- all three are part of the meta-data in ASJP

Expert trees and tree distances

- most nodes in the expert trees are multiple branching
- trees that are produced by phylogenetic software are always binary branching
- this leads to misleadingly high tree distances

Expert trees and tree distances

- suppose the left tree is extracted from the data and the right one is an expert tree



- as the left tree correctly captures all taxa in the right tree, this seems to be a perfect fit
- however:
  - normalized Robinson-Fould distance: 0.33
  - normalized quartet distance: 0.11

Expert trees and tree distances

- in practice
  - 5,644 languages in ASJP (excluding creoles etc.)
  - there 5,641 partitions in every inferred tree
  - Ethnologue: 1,803 partitions
  - WALS: 391 partitions
  - Hammarström: 1,735 partitions
- Robinson-Fould distance to WALS tree will be at least 0.68, no matter how well the algorithm performs
- minimum quartet distance: not easy to calculate, but also substantial

Measures of fit

- more realistic measures of goodness of fit:
  - **Robinson-Fould fit**:

$$\frac{\text{number of shared partitions}}{\text{total number of partitions in the expert tree}}$$

  - **quartet fit**:

$$\frac{\text{number of shared butterflies}}{\text{total number of butterflies in the expert tree}}$$

- these measures are always between 0 and 1
- 1 means that all groupings from the expert classification are correctly recovered

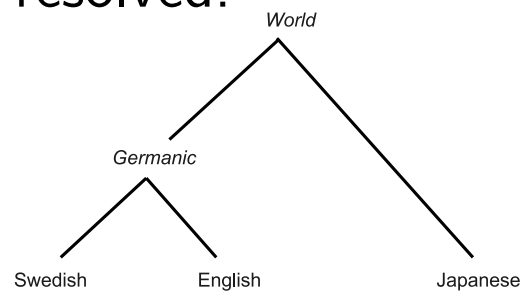
Triplet fit

- pick a triplet of languages A, B, C which has a resolved tree structure $((A, B), C)$ according to the expert tree
- determine predicted distances:

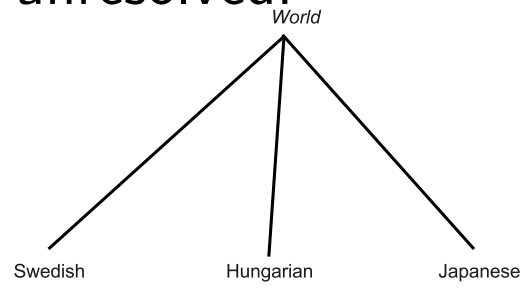
$$\begin{aligned} d(\text{Swedish}, \text{English}) &= 0.486 \\ d(\text{Swedish}, \text{Japanese}) &= 0.905 \\ d(\text{English}, \text{Japanese}) &= 0.897 \end{aligned}$$

- - $d(A, B) < \min(d(A, C), d(B, C)) \mapsto$ **correct**
  - otherwise $\mapsto$ **incorrect**
- triplet fit of a distance measure to an expert tree: proportion of resolved triplets that come out correct

resolved:

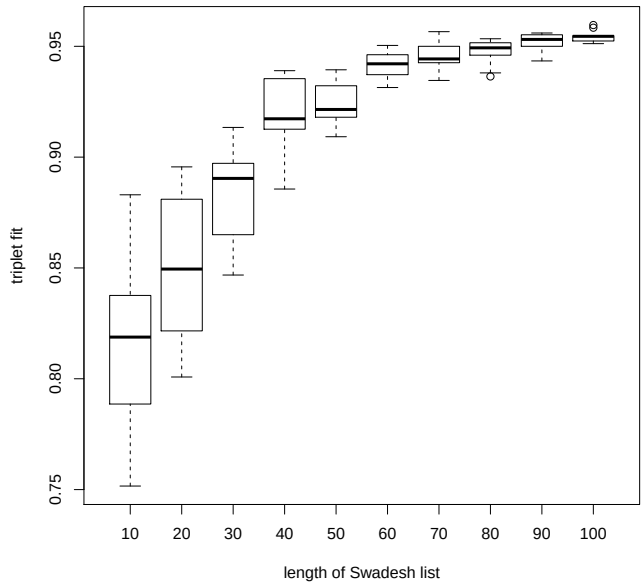


unresolved:



Triplet fit

- Wichmann et al. claim that 40 Swadesh items are enough; longer Swadesh lists lead to a decrease in quality according to their methods



- Note: More data is better than less data.

Some results

- compute 5,644 $\times$ 5,644 distance matrices based on Levenshtein alignment vs. weighted alignment
- perform Neighbor-Joining algorithm
- measure fit to expert trees
- Robinson-Fould fit:

	Levenshtein	weighted	Aline
WALS	0.624	0.639	0.622
Ethnologue	0.485	0.490	0.477
Hammarström	0.457	0.473	0.447

Some results

- quartet fit:

	Levenshtein	weighted	ALINE
WALS	0.869	0.886	0.855
Ethnologue	0.839	0.857	0.824
Hammarström	0.890	0.896	0.891

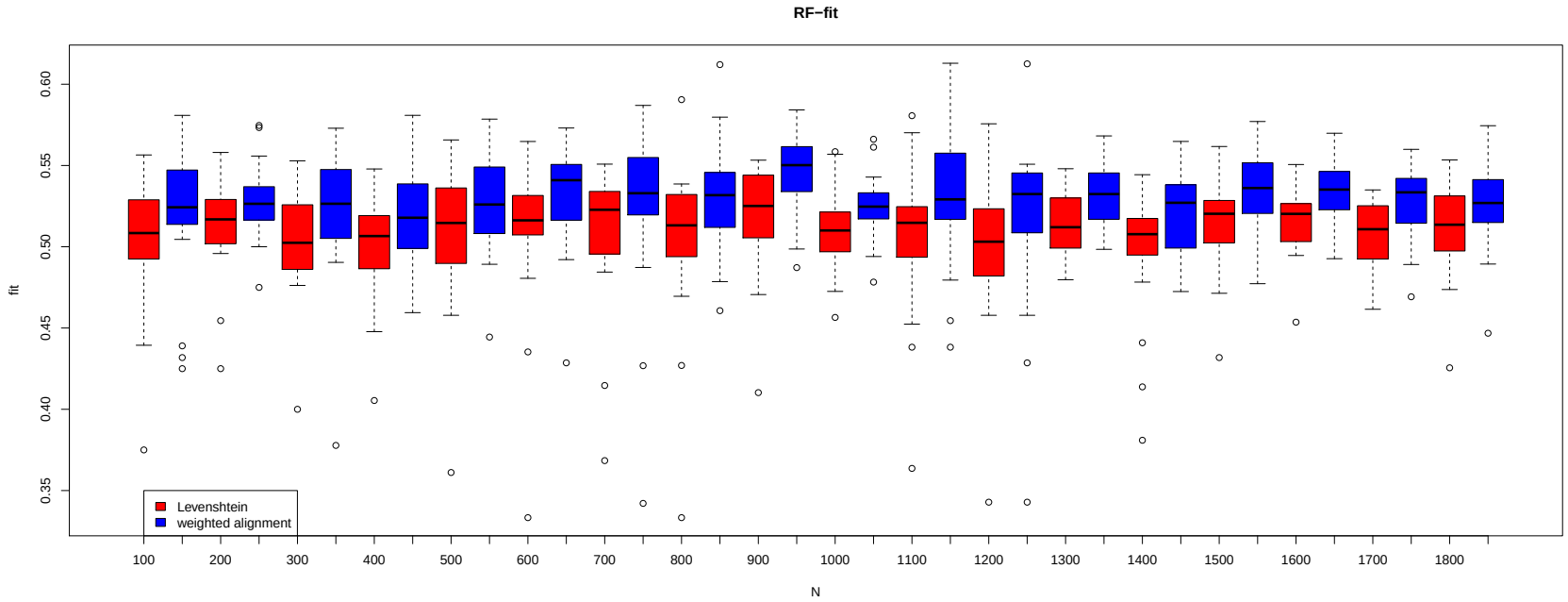
Some results

- triangle fit:

	Levenshtein	weighted	ALINE
WALS	0.8816	0.9055	0.8876
Ethnologue	0.7733	0.7980	0.7734
Hammarström	0.7670	0.7904	0.7699

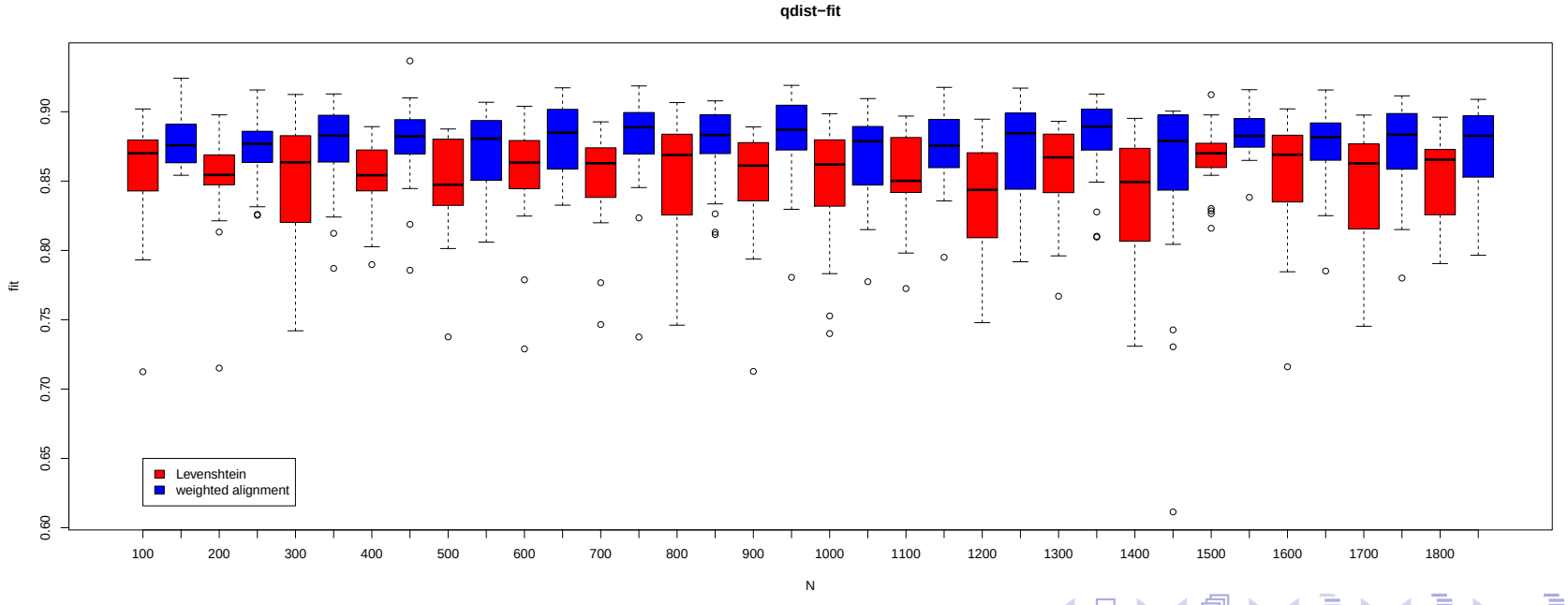
Some results

- procedure: for $N = 100, 200, 300, \ldots, 1{,}800$:
  - pick N languages at random
  - compute $N \times N$ distance matrices based on Levenshtein vs. weighted alignment
  - perform Neighbor Joining algorithm
  - measure fit to Hammarström expert tree
- Robinson-Fould fit:



Some results

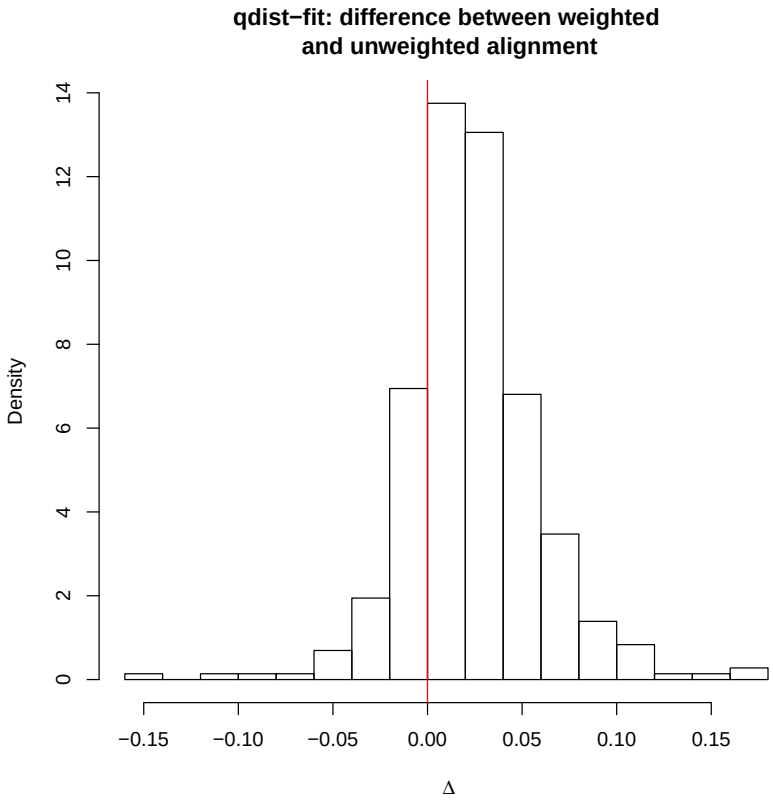
- procedure: for $N = 100, 200, 300, \ldots, 1{,}800$:
  - pick N languages at random
  - compute $N \times N$ distance matrices based on Levenshtein vs. weighted alignment
  - perform Neighbor Joining algorithm
  - measure fit to Hammarström expert tree
- quartet fit:



Some results

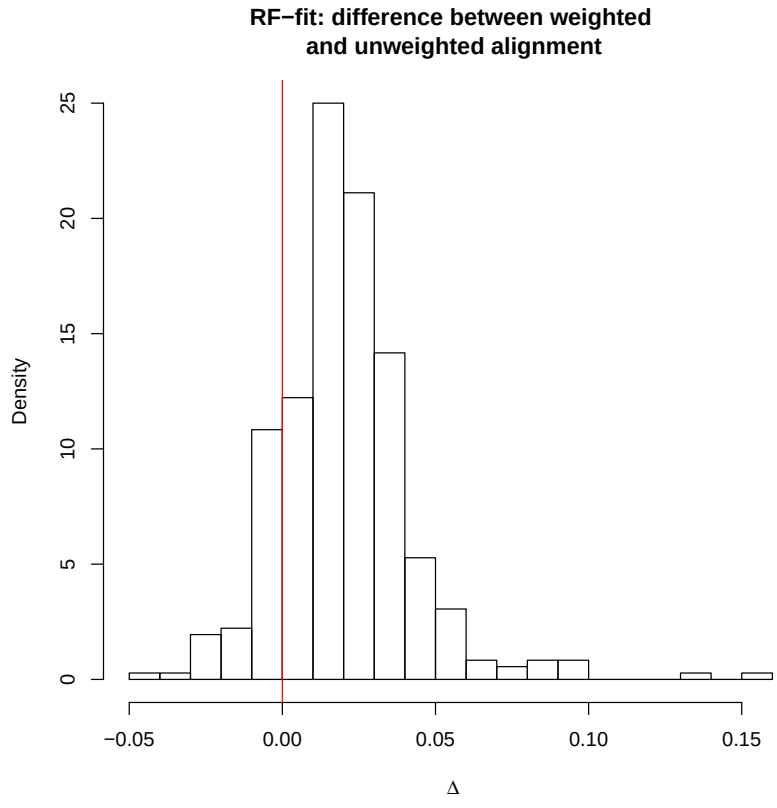
- difference between weighted alignment and Levenshtein alignment for the same data set

- Robinson-Fould fit



$\mu = 0.020$

- quartet fit



$\mu = 0.023$