Phylogenetic Methods for Computer-Aided Language Classification

Gerhard Jäger

gerhard.jaeger@uni-tuebingen.de

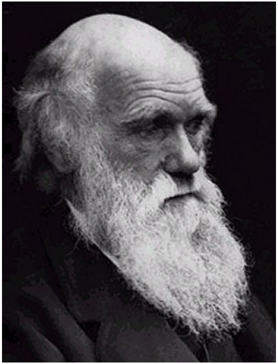
August 6, 2012

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Sprachwandel und Evolution

"The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. [...] We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation. The manner in which certain letters or sounds change when others change is very like correlated growth. [...] The frequent presence of rudiments, both in languages and in species, is still more remarkable. [...]
Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely, and lead to the gradual extinction of other tongues."

(Darwin, The Descent of Man)



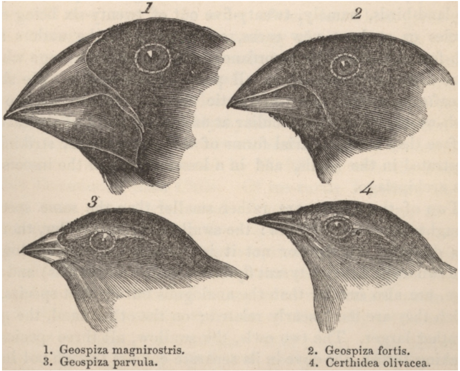
Language change and evolution

Vater Unser im Himmel, geheiligt werde Dein Name

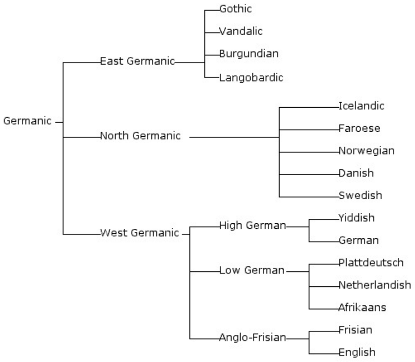
Onze Vader in de Hemel, laat Uw Naam geheiligd worden

Our Father in heaven, hallowed be your name

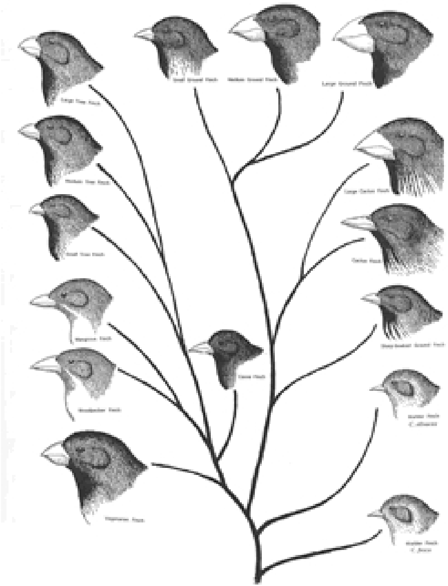
Fader Vor, du som er i himlene! Helliget vorde dit navn



Language change and evolution



- Germanic
  - East Germanic
    - Gothic
    - Vandalic
    - Burgundian
    - Langobardic
  - North Germanic
    - Icelandic
    - Faroese
    - Norwegian
    - Danish
    - Swedish
  - West Germanic
    - High German
      - Yiddish
      - German
    - Low German
      - Plattdeutsch
      - Netherlandish
      - Afrikaans
    - Anglo-Frisian
      - Frisian
      - English

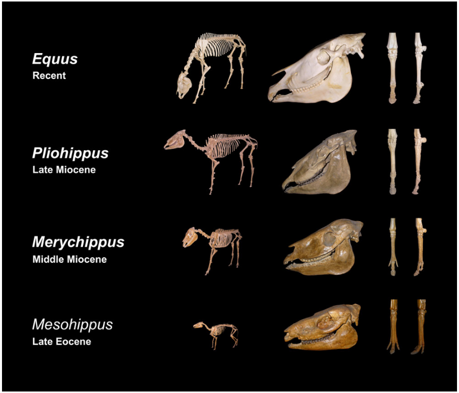


Language change and evolution

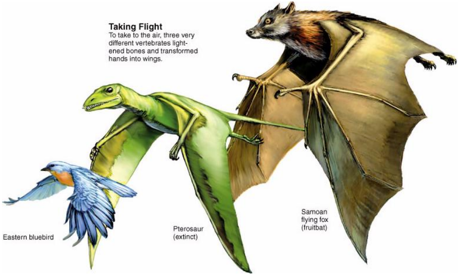
Mittelhochdeutsch:
Got vater unser, dâ du bist in dem himelrîche gewaltic alles des dir ist, geheiliget sô werde dîn nam

Althochdeutsch:
Fater unser thû thâr bist in himile, si giheilagôt thîn namo

Gotisch:
Atta unsar þu in himinam, weihnai namo þein



Convergent evolution

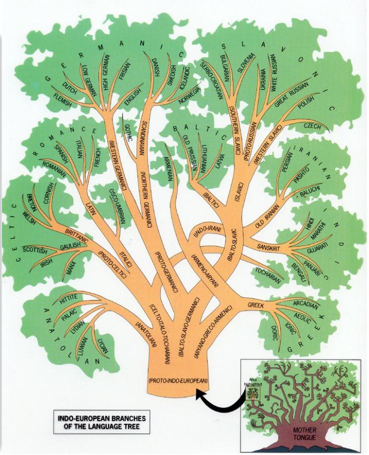

Taking Flight
To take to the air, three very different vertebrates lightened bones and transformed hands into wings.
Samoan flying fox (fruitbat)
Pterosaur (extinct)
Eastern bluebird

- Old English *docga* > English *dog*
- Proto-Paman **gudaga* > Mbabaram *dog* ('dog')

Language phylogeny

Comparative method

1. identifying *cognates*, i.e. obviously related morphemes in different languages, such as *new/nowy*, *two/dwa*, or *water/voda*
2. reconstruction of *common ancestor* and *sound laws* that explain the change from reconstructed to observed forms
3. applying this iteratively leads to phylogenetic language trees



INDO-EUROPEAN BRANCHES OF THE LANGUAGE TREE

Language phylogeny

Scope of the method

- reconstructed vocabulary shrinks with growing time depth
- maximal time horizon seems to be about 8,000 years
- grammatical morphemes and categories arguably more stable and less apt to borrowing
- problem here: limited number of features, cross-linguistic variation constrained by language universals, frequently convergent evolution
- comparative method is hard to apply in regions with high linguistic diversity and without written documents (Paleo-America, Papua)
- tree structure might be inappropriate if there is a significant effect of language contact (cf. Australia)

Computational Methods

- both cognate detection and tree construction lend themselves to algorithmic implementation
- Advantages:
  - easy to scale up
  - comparability of results
  - affords statistical evaluation
- Disadvantages:
  - cognacy judgments require lots of linguistic insight and experience
  - tree construction should be subject to historical (including archeological) and geographical plausibility

Computational Methods: Data

Electronic data sources

- **World Atlas of Language Structures** (Haspelmath et al.)
  - categorization of $>$ 2,000 language according to $>$ 140 discrete features relating to phonology, grammar, and vocabulary
- **Automated Similarity Judgment Program** (Wichmann et al.)
  - translation of a 40-item list of basic concepts into $>$ 5,000 languages
  - uniform phonetic transcription
- **Comparative Indoeuropean Database** (Dyen et al.)
  - 200-item list of basic concepts (*Swadesh list*), translated into 95 Indoeuropean languages
  - complete cognacy information based on expert judgments

Computational Methods: Phylogenetic Algorithms

Character-based methods

- input: vector of (usually binary) feature values for each data point (= language)
- background assumption: evolution proceeds by randomly changing features values
- **Maximum Parsimony**
  - non-parametric
  - reconstruction of a character vector for each non-leaf node of the reconstructed tree
  - evolutionary changes only along tree branches
  - best tree is the one which requires minimum number of changes
  - NP-hard; heuristic search required

Computational Methods: Phylogenetic Algorithms

Character-based methods

- input: vector of (usually binary) feature values for each data point (= language)
- background assumption: evolution proceeds by randomly changing features values
- **Maximum Likelihood**
  - parametric model
  - assumes an explicit stochastic model of the likelihood of an evolutionary change in a given position of a tree
  - algorithm(s) search for tree that maximizes overall likelihood
- **Bayesian Phylogenetic Inference**
  - conceptually similar to Maximum Likelihood
  - incorporates prior knowledge on the probability of tree topologies
- both methods are computationally even more demanding than Maximum Parsimony

Computational Methods: Phylogenetic Algorithms

Distance-based methods

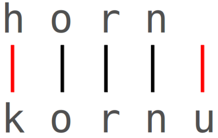
- input: matrix of distances between data points
- background assumption: evolution increases distance between species/languages
- no reconstruction of ancestral forms
- **Neighbor Joining**
  - hierarchical clustering
  - bottom-up
  - the two points with the smallest normalized distance are joined into a cluster, the distance of this new point to the other points is computed, and this procedure is repeated until only one cluster is left
  - fast: $\mathcal{O}(n^3)$ in the number of data points
- NJ consistently outperforms other distance based methods (such as UPGMA) when applied to language phylogeny

Estimating distances

- basically two approaches to estimating language distances:
    - count the number of shared cognates in a given Swadesh list of of core concepts
    - compute the pairwise phonetic distance between synonymous concepts, and aggregate the results
- latter approach is closely connected to computational dialectometry

Normalized Levenshtein Distance

- first step: finde minmal edit distance between all translation pairs of the languages to be compared
- e.g. German ↔ Latin

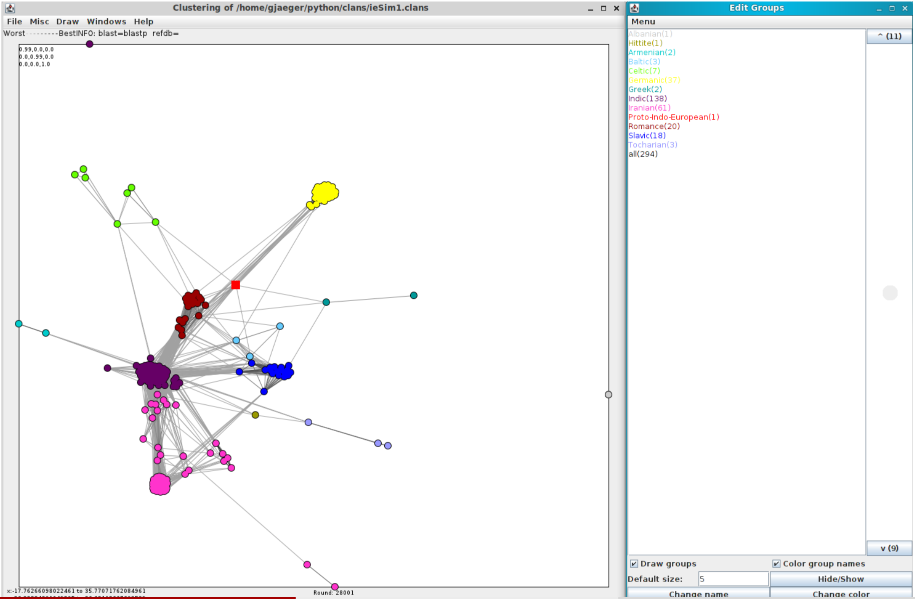


- edit distance = 2
- transformation into similarity measure to correct for varying word lengths:

$$\mathrm{sim}(x,y) \doteq \frac{2(\max(l(x),l(y)) - d_{Lev}(x,y))}{l(x)+l(y)}$$

- similarity between L_1 and L_2: average similarity of translation pairs between L_1 and L_2

Normalized Levenshtein Distance


Clustering of /home/gjaeger/python/clans/ieSim1.clans
File Misc Draw Windows Help
Edit Groups
Menu
^ (11)
v (9)
Hide/Show
Change name
Change color

Worst--------BestINFO: blast=blastp refdb=

```
0.99,0.0,0.0
0.0,0.99,0.0
0.0,0.0,1.0
```

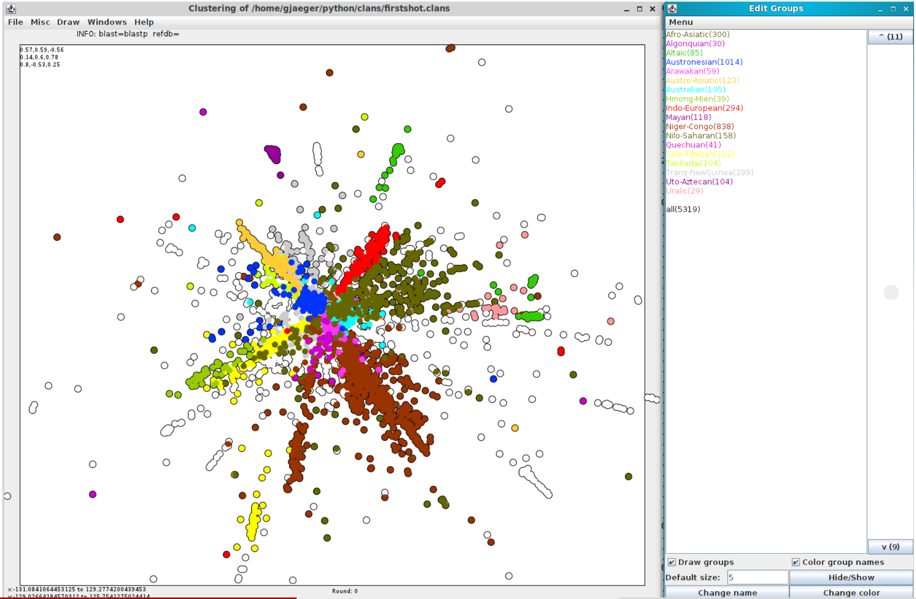
x:-17.76266098022461 to 35.77071762084961

Round: 28001

Albanian(1)
Hittite(1)
Armenian(2)
Baltic(3)
Celtic(7)
Germanic(37)
Greek(2)
Indic(138)
Iranian(61)
Proto-Indo-European(1)
Romance(20)
Slavic(18)
Tocharian(3)
all(294)

☑ Draw groups
☑ Color group names
Default size: 5

Normalized Levenshtein Distance



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

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)

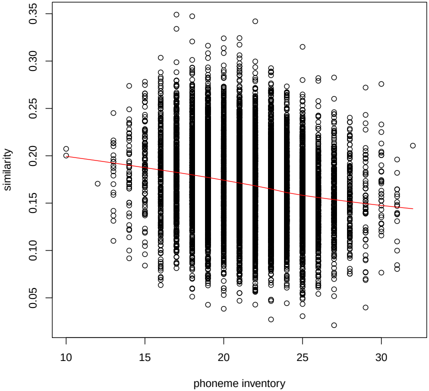
all(5319)

☑ Draw groups ☑ Color group names

Default size: 5

Normalized Levenshtein Distance

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

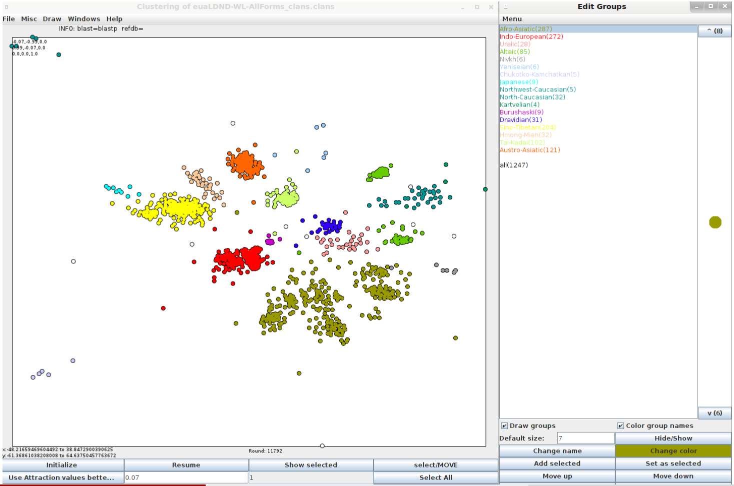


Levenshtein Distance, Normalized and Divided (LDND)

- Wichmann et al.:
  1. correct for word length by dividing by the length of the longer word
  2. divide by the average normalized distance between words from L_1 and words from L_2 that are not translation pairs

Levenshtein Distance, Normalized and Divided (LDND)

- Application to all Eurasiatic languages in ASJP:



Clustering of euaLDND-WL-AllForms_clans.clans

INFO: blast=blastp refdb=

-0.07,-0.99,0.0
0.99,-0.07,0.0
0.0,0.0,1.0

x:-48.21659469604492 to 38.8472900390625
y:-61.36861038208008 to 64.63750457763672

Round: 11792

Use Attraction values bette... 0.07

1

Edit Groups

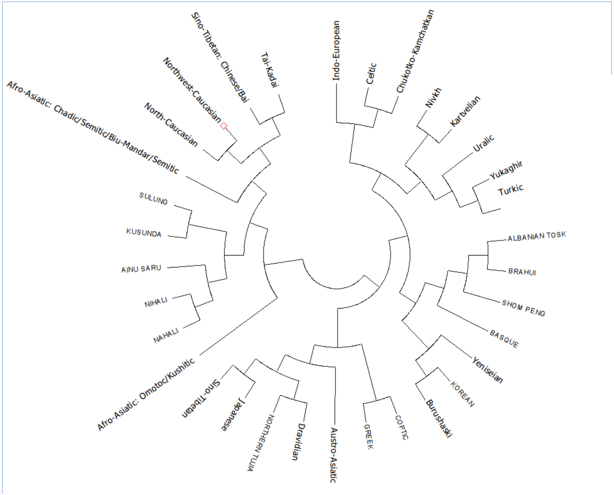
Afro-Asiatic(287)
Indo-European(272)
Uralic(28)
Altaic(85)
Nivkh(6)
Yeniseian(6)
Chukotko-Kamchatkan(5)
Japanese(9)
Northwest-Caucasian(5)
North-Caucasian(32)
Kartvelian(4)
Burushaski(9)
Dravidian(31)
Sino-Tibetan(204)
Hmong-Mien(32)
Tai-Kadai(102)
Austro-Asiatic(121)

all(1247)

☑ Draw groups
☑ Color group names
Default size: 7

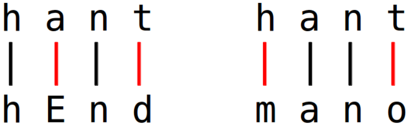
Levenshtein Distance, Normalized and Divided (LDND)

- Application to all Eurasiatic languages in ASJP; NJ-tree:



Weighted Alignment

- Levenshtein distance is very coarse grained



h a n t   h a n t

h E n d   m a n o

- similarity is 0.5 in both cases
- correspondences *a*∼*E*, *t*∼*d* are (according to linguistic criteria like place of articulation) much more natural than *h*∼*m* or *t*∼*o*
- German appears equidistant to English and Spanish here, even though the distance to English is clearly smaller

Weighted Alignment

- **Needleman Wunsch Algorithm**
  - similar to computation of Levenshtein distance
  - edit operations are *weighted*: algorithm finds optimal alignment, that minimizes total weight
  - can be done in $\mathcal{O}(nm)$ (with n and m being the lengths of the strings to be aligned) via dynamic programming
  - $a \sim E$, $d \sim t$ should have lower weight than $t \sim o$
- How to determine these weights?
  - bioinformatics: **pointwise mutual information** (a.k.a. **log-odds**)
  - logarithm of the probability of a replacement, divided by probability of chance co-occurrence of molecula pair in question

Pointwise Mutual Distance

- **pointwise mutual information** between event x and event y:

$$pmi(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

- Prokic (2010):
  - let x and y be two segments
  - $p(x,y)$: probabilty that x and y are aligned to each other in a pair of cognates
  - $p(x), p(y)$: probability of x/y to occur individually
- iterative procedure:
  - perform Levenshtein alignment between a given training set of cognate pairs
  - estimate $pmi(x,y)$ for all segment pairs by using relative frequencies as proxy for probabilites
  - perform Needleman-Wunsch alignment with $pmi(x,y)$ as weight of $x \sim y$
  - repeat this procedure until weights do not change anymore

Weighted Alignment

- dialectometry: cognate pairs are known
- language classification: only limited set of expert-compiled cognate lists are available, and only for small number of language families
- quick and dirty procedure to obtain cognate lists from ASJP data:
  - p-value of translation pair (w_1, w_2) from languages L_1, L_2:

    $$P(\mathrm{sim}(w_1, w_2) > P(w_1, w_w)|(w_1', w_2') \in L_1 \times L_2; \|w_1\| \neq \|w_2\|)$$

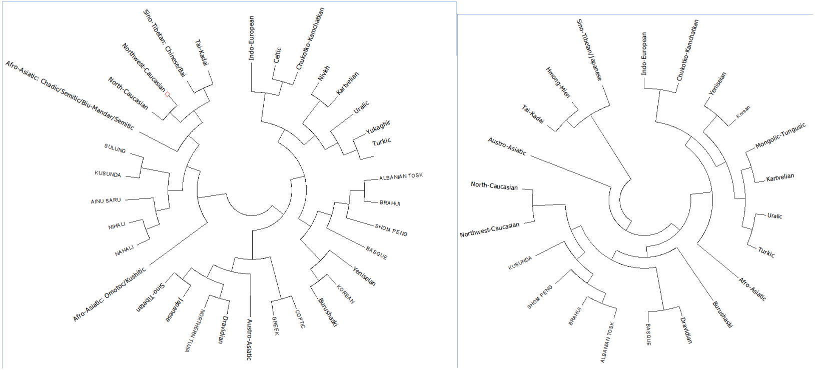
  - intuitively: probability that the level of similarity between w_1 and w_2 is as high as it is if the two are not cognates
  - two words are considered cognates if
    - L_1 and L_2 have a normalized LDND-distance of ≤ 0.2, and
    - p-value of (w_1, w_2) is < 0.01
  - further pairs are added by forming the transitive closure of the cognacy relation

Weighted Alignment

- calibrated similarity between languages:
  - compute weighted alignemnt score for all translation pairs
  - compute weighted alignemnt score for all non-translation pairs
  - determine p-value for each translation pair ($=$ relative frequency of non-translation pairs with higher weighted alignment score)
  - similarity between L_1 and L_2 is average logarithm of p-values of all translation pairs

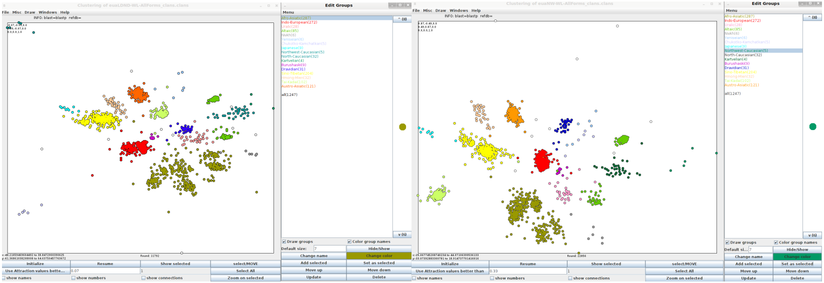
Qualitative comparison

- NJ trees for the languages of Eurasia (left: LDND; right: NWPV)



Qualitative comparison

- CLANS visualization for the languages of Eurasia (LDND left, NWPV right)



```
x:-48.216594696044492 to 38.8472900390625
y:-41.36861038208008 to 64.63750457763672
Round: 11792
```

Afro-Asiatic(287)
Indo-European(272)
Uralic(28)
Altaic(85)
Nivkh(6)
Yeniseian(6)
Chukotko-Kamchatkan(5)
Japanese(9)
Northwest-Caucasian(5)
North-Caucasian(32)
Kartvelian(4)
Burushaski(9)
Dravidian(31)
Sino-Tibetan(204)
Hmong-Mien(32)
Tai-Kadai(102)
Austro-Asiatic(121)

all(1247)

```
x:-25.067745208740234 to 44.07106399536133
y:-33.07392883300781 to 18.914737701416016
Round: 22856
```

Afro-Asiatic(287)
Indo-European(272)
Uralic(28)
Altaic(85)
Nivkh(6)
Yeniseian(6)
Chukotko-Kamchatkan(5)
Japanese(9)
Northwest-Caucasian(5)
North-Caucasian(32)
Kartvelian(4)
Burushaski(9)
Dravidian(31)
Sino-Tibetan(204)
Hmong-Mien(32)
Tai-Kadai(102)
Austro-Asiatic(121)

all(1247)

Benchmarking

- two reasonable Gold standards for comparing these two similarity/distance measures:
  - expert judgments on cognacy
  - expert judgments on language classification

Benchmarking: cognacy

- experiment:
  - extract those items from the Dyen-Kruskal database that occur in ASJP
  - define a cognacy estimator based on LDND by finding the optimal cutoff
  - do the same for NWPV
  - compare
- result
  - LDND: optimally achievable *Matthews Correlation Coefficient*: $\mathbf{0.547}$
  - NWPV: optimally achievable *Matthews Correlation Coefficient*: $\mathbf{0.574}$
    (+1 means perfect prediction, -1 perfect mis-prediction)

Benchmarking: language classification

- expert classifications of languages of the world:
  - WALS: two taxonomic levels for each language (family, genus)
  - Ethnologue: detailed classification
  - Hammerström (2010): equally detailed but more conservative classification
- measure for the quality of a NJ tree: number of internal nodes in the expert tree that have a counterpart in the NJ tree (i.e. a node that dominates the same set of leaves)

Benchmarking: language classification

Results for LDND vs. *NWPV* (weighted alignment + p-value estimation)

	Africa	Eurasia	Australia/Oceania	America

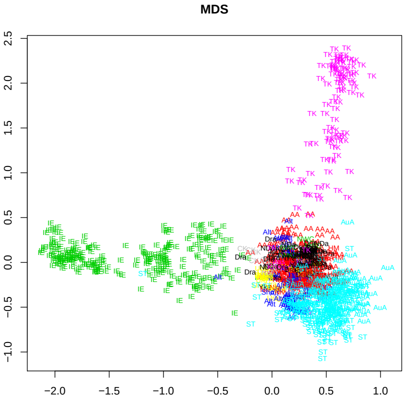
WALS	0.517(30/58)
	0.500(29/58)
		0.630(46/73)
		0.644(47/73)
			0.573(82/143)
			0.524(75/143)
				0.743(81/109)
				0.752(82/109)

Ethnologue	0.477(234/490)
	0.473(232/490)
		0.455(148/325)
		0.489(159/325)
			0.510(311/610)
			0.515(314/610)
				0.603(178/295)
				0.603(178/295)

Hammerström	0.508(248/488)
	0.5(244/488)
		0.467(162/347)
		0.493(171/347)
			0.519(318/613)
			0.524(321/613)
				0.669(195/292)
				0.682(199/292)

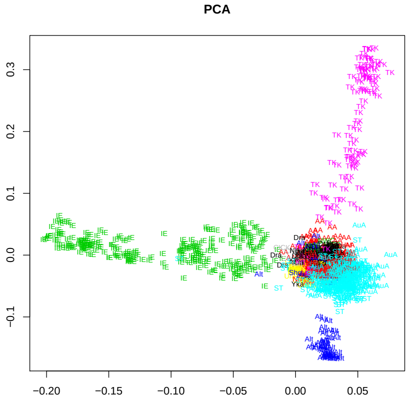
Visualization: Multi-Dimensional Scaling

- MDS applied to NWPV-matrix of the Eurasian languages



Visualization: Principal Component Analysis

- PCA applied to NWPV-matrix of the Eurasian languages



Visualization: CLANS

Force-Directed Graph Layout

- method to visualize graphs or similarity matrices in two or three dimensions
- simulation of a physical system:
    - data items ⇔ physical particles
    - pairwise attractive force between particles proportional to their similarity
    - constant repelling force between any pair of particles
    - *this is just one of many protocols to determine forces*
        - initially, all particles are placed at random
        - in each time step, each particle is move a small amount along the resulting force vector
        - last step is repeated until a stable state is reached
- tends to stabilize in a state where groups of mutually similar items form clusters

Prospects

- Multiple Sequence Alignment
- Pair Hidden Markov Models for sequence alignment and reconstruction of ancestral forms
- resampling methods (bootstrapping, jackknifing)
- simulations
- correlation with geographic and/or genetic and phenotypic anthropological distance measures