Estimating and Visualizing Language Similarities Using Weighted Alignment and Force-Directed Graph Layout

Gerhard Jäger

April 24, 2012, Avignon

joint work with Armin Buch, David Erschler & Andrei Lupas

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

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

CLANS

- **Cl**uster **An**alysis of **S**equences
- developed by bioinformaticians *Tancred Frickey* and *Andrei Lupas* as exploratory tool to explore evolutionary relationships among protein sequences (Frickey and Lupas 2004)
- similarities of proteins is determined via sequence alignment; resulting matrix is visualized using CLANS
- advantages in comparison to tree-based algorithms:
  - does not *a priori* assume a tree like signal (useful when lateral transfer plays a role)
  - fast (esp. in comparison to character based algorithms)
  - robust (noise in data items does not accumulate)
- general impression so far (Lupas, p.c.):
  - tree algorithms are more precise when evolutionary distances are small; CLANS is more sensitive to weak evolutionary signals

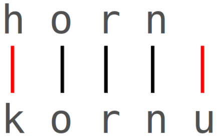
The Automated Similarity Judgment Program

- Project at MPI EVA in Leipzig around Sören Wichmann
- covers more than 5,000 languages
- basic vocabulary of 40 words for each language, in uniform phonetic transcription
- freely available

used concepts: *I, you, we, one, two, person, fish, dog, louse, tree, leaf, skin, blood, bone, horn, ear, eye, nose, tooth, tongue, knee, hand, breast, liver, drink, see, hear, die, come, sun, star, water, stone, fire, path, mountain, night, full, new, name*

First shot: Levenshtein Distance

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



h o r n

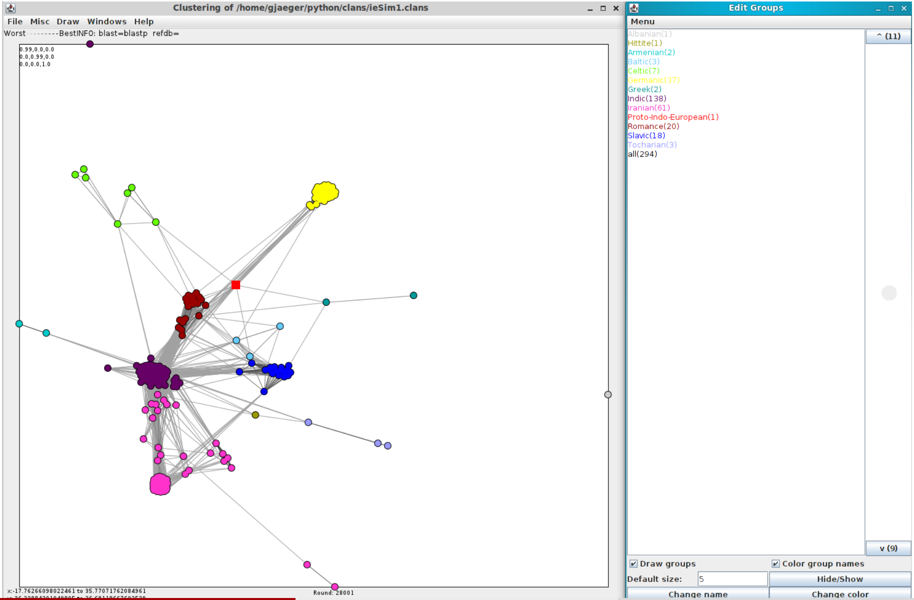
k o r n u

- edit distance $= 2$
- transformation into similarity measure

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

- similarity between L1 and L2: average similarity of translation pairs between L1 and L2

First shot: normalized Levenshtein Distance



Clustering of /home/gjaeger/python/clans/ieSim1.clans

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

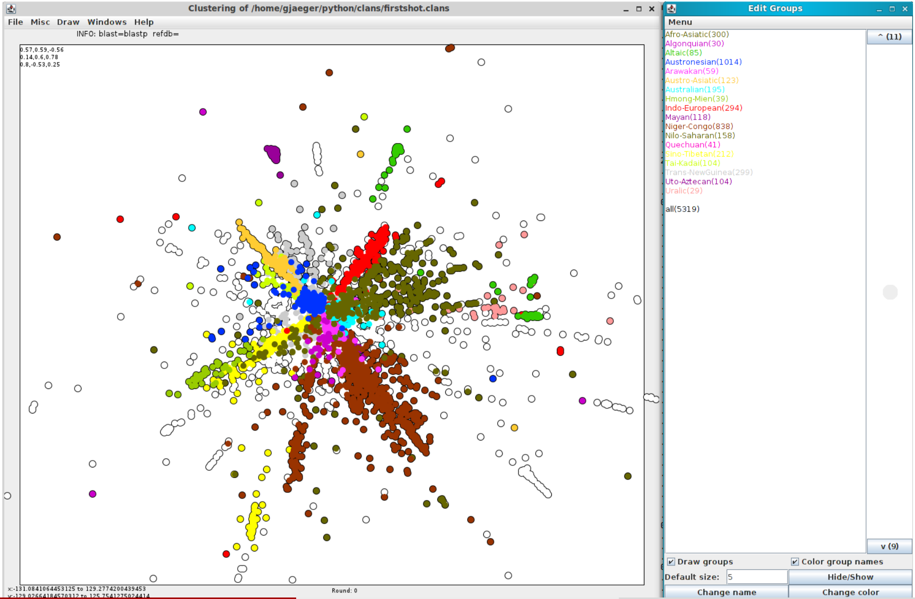
Round: 28001

Edit Groups

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

First shot: normalized Levenshtein Distance



Clustering of /home/gjaeger/python/clans/firstshot.clans

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

Edit Groups

- 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)

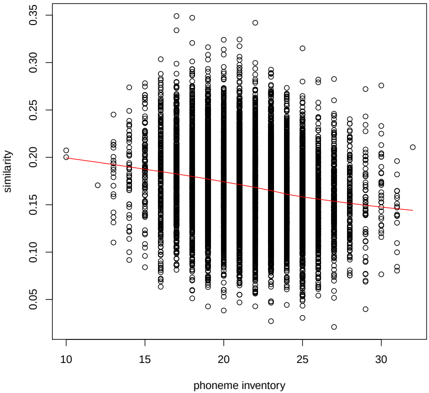
☑ Draw groups

☑ Color group names

Default size: 5

First shot: normalized Levenshtein Distance

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



Benchmark: LDND measure

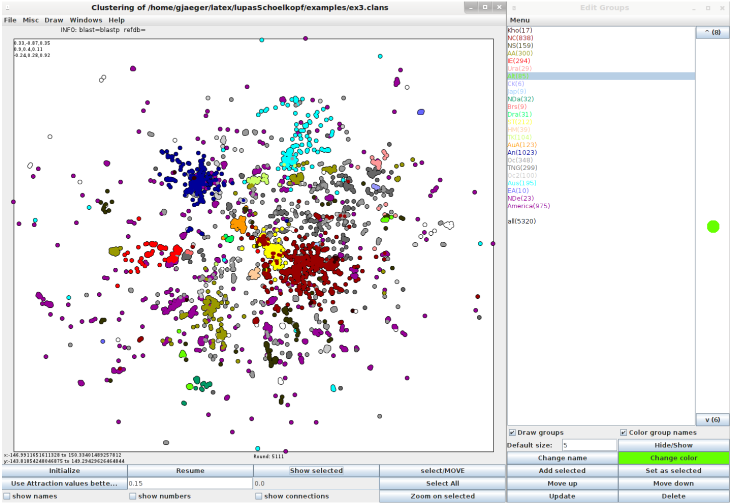
- Wichmann et al.: doubly normalized Levenshtein distance (**L**evenshtein **D**istance **N**ormalized and **D**ivided)
- normalization for word length

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

- 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



Clustering of /home/gjaeger/latex/lupasSchoelkopf/examples/ex3.clans

INFO: blast=blastp refdb=

```
0.33,-0.87,0.35
0.9,0.4,0.11
-0.24,0.28,0.92
```

```
x:-146.9911651611328 to 150.33401489257812
y:-143.81854248046875 to 149.29429626464844
```

Round: 5111

Use Attraction values bette... 0.15 0.0

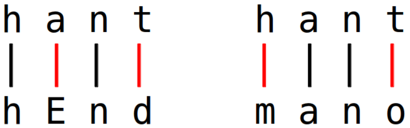
☐ show names ☐ show numbers ☐ show connections

☑ Draw groups ☑ Color group names

Default size: 5

Needleman-Wunsch-Algorithmus

- Levenshtein distance is somewhat coarse grained



```
h a n t    h a n t
| | | |    | | | |
h E n d    m a n o
```

- simply normalized distance is 0.5 in both cases
- after second normalization, [hant] even appears somewhat closer to [mano] ($ldnd = 0.54$) than to [hEnd] ($ldnd = 0.55$)
- 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

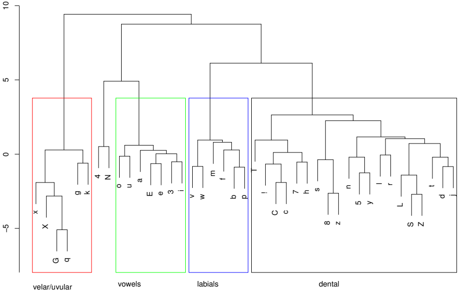
- **Needleman Wunsch Algorithm**
  - similar to computation of Levenshtein distance
  - edit operations are *weighted*: algorithm finds optimal alignment, that minimizes total weight
  - $a \sim E$, $d \sim t$ should have lower weight than $t \sim o$
- How to determine these weights?
  - bioinformatics: *log-odds*
  - logarithm of the probability of a replacement, divided by probability of chance co-occurrence of molecula pair in question

Weighted Alignment

- estimation of correspondence probabilities of two sounds in cognates:
    - large sample of pairwise related languages
    - replacement operation under Levenshtein alignment of translation pairs are counted
    - substantive part of word pairs considered are true cognates: replacement operations thus reflect genuine language change processes
    - replacement of sounds between non-cognates is randomly distributed and boils down to an additive constant in the logarithmic term

Weighted Alignment

- log odds:
  - $d \sim t$: 0.69
  - $a \sim E$: 0.07
  - $h \sim m$: -0.61
  - $t \sim o$: -0.80



Weighted Alignment

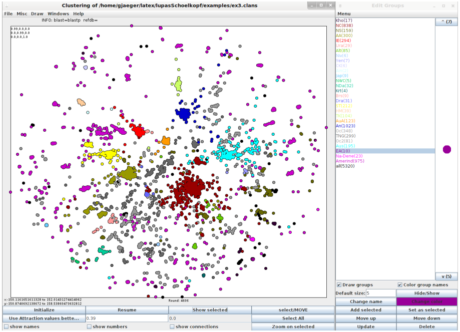
- total value of optimal alignment is interpreted as similarity between strings
- similarity between languages is computed via p-values:

$$\mathrm{nwpv}(x, y, L_1, L_2) \doteq \frac{\#\{(x', y') | \mathrm{nw}(x', y') \geq \mathrm{nw}(x, y)\}}{\#L_1 \times \#L_2}$$

$$\mathrm{nwpv}(L_1, L_2) \doteq \frac{\sum_{x \in L_1, y \in L_2} - \log(\mathrm{pv}(x, y))}{\#\{(x, y) : \|x\| = \|y\|\}}$$

Weighted Alignment

- similarities of English to
  - Dutch 0.60 / 3.38
  - German: 0.68 / 3.16
  - Proto-Indoeuropean: 0.86 / 2.33
  - Latin: 0.88 / 1.85
  - Spanish: 0.93 / 1.59
  - Russian: 0.93 / 1.52
  - Hungarian: 0.95 / 1.30
  - Turkish: 1.03 / 0.83



Clustering of /home/gjaeger/latex/lupasSchoelkopf/examples/ex3.clans

INFO: blast=blastp refdb=

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

x:-150.1161651611328 to 152.01431274414062
y:-150.8740692138672 to 158.53860473632812

Round: 4606

Use Attraction values bette... 0.39 0.0

□ show names □ show numbers □ show connections

☑ Draw groups ☑ Color group names

Default size: 5

Comparison

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

Comparison: cognacy

- **Dyen-Kruskal database:** cognacy judgment for 200-item Swadesh lists from 95 Indo-European languages
- 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)

Comparison: language classification

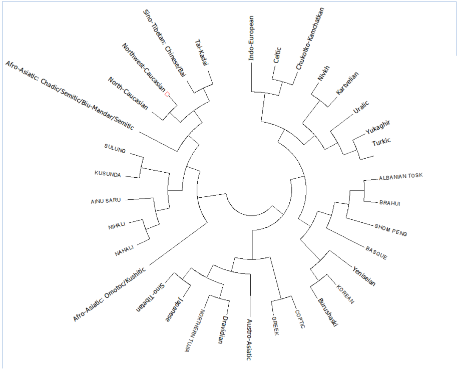
- **Ethnologue:** provides taxonomic classification of virtually all living languages
- **Robinson-Foulds metric:**
  - compares two trees over the same set of leafs
  - returns number of partitions that one of the two trees makes and the other doesn't
- **Neighbor Joining Algorithm:** bottom up cluster algorithm to extract an unrooted tree from a distance matrix

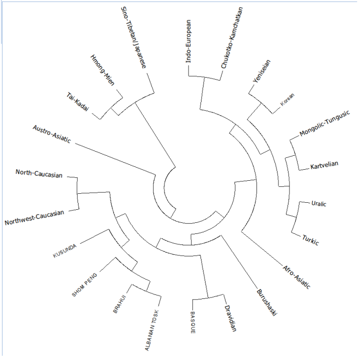
Comparison: language classification

- **Experiment:**
  - compute NJ-tree for all languages in ASJP based on LDND and NWPV distances
  - extract sub-tree of Ethnologue tree for the languages in ASJP
  - compute Robinson-Foulds metric between Ethnologue tree and each of the two NJ trees
- **Outcome:**
  - LDND: **5,522** (4,550 false positives, 972 false negatives)
  - NWPV: **5,476** (4,527 false positives, 949 false negatives)

Qualitative comparison

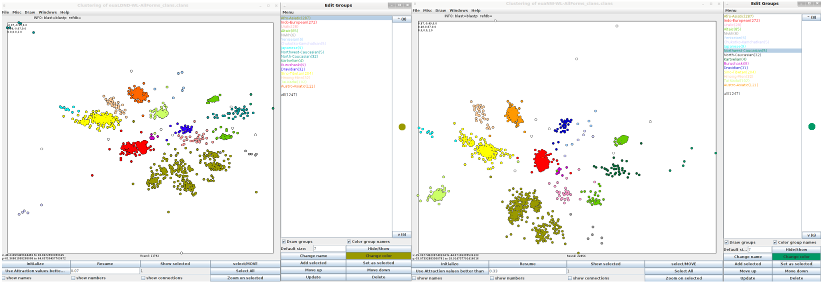
- NJ trees for the languages of Eurasia (left: LDND; right: NWPV)





Qualitative comparison

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



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

☐ show names ☐ show numbers ☐ show connections

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

Clustering of euaNW-WL-AllForms_clans.clans

INFO: blast=blastp refdb=

0.87,-0.49,0.0
0.49,0.87,0.0
0.0,0.0,1.0

x:-25.067745208740234 to 44.07106399536133
y:-33.07392883300781 to 18.914737701416016

Round: 22856

Use Attraction values better than 0.33 1

☐ show names ☐ show numbers ☐ show connections

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

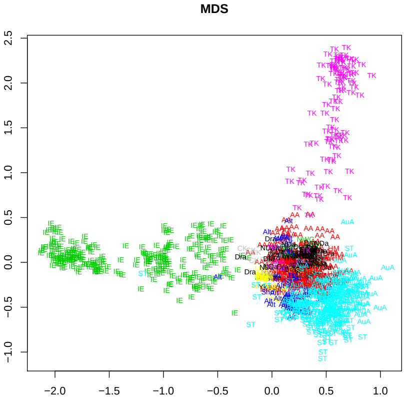
Default si... 7

CLANS and dimensionality reduction

- CLANS performs a kind of (non-deterministic) dimensionality reduction
- How does this relate to more established methods?

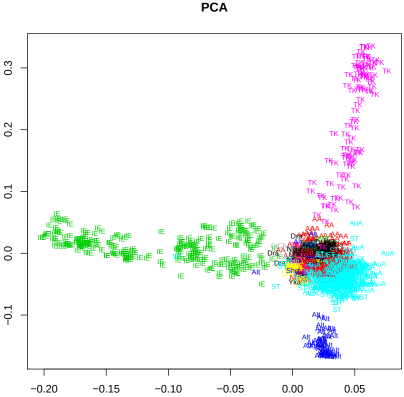
CLANS vs. Multi-Dimensional Scaling

- MDS applied to NWPV-matrix of the Eurasian languages



CLANS vs. Principal Component Analysis

- PCA applied to NWPV-matrix of the Eurasian languages



CLANS and dimensionality reduction

- language families massively vary in size
- MDS and PCA only provide information about the largest families
- CLANS is sensitive to local patterns

Conclusion

- weighted alignment improves results of lexico-statistical language classification
- more powerful methods from bioinformatics (such as progressive multiple alignment) are likely to lead to further improvement
- CLANS is a useful exploratory tool

Frickey, T. and A. N. Lupas (2004). Clans: a java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**(18):3702–3704.