Bled, February 2006

On the (dis-)similarities of similar things!

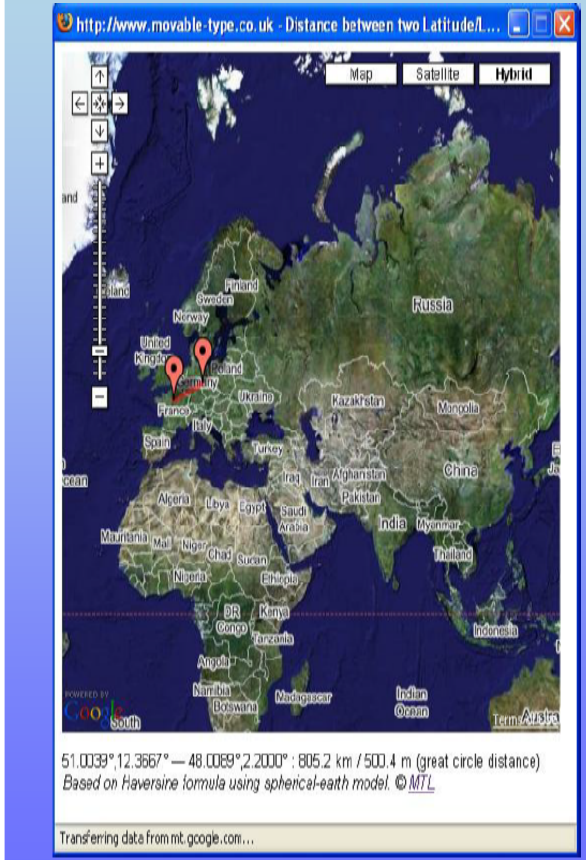
Mihai Albu

MPI-EVA & IZBI

Distance Matrices

	Paris	Leipzig	Shanghai
Paris	0	805	9328
Leipzig	805	0	8543
Shanghai	9328	8543	0

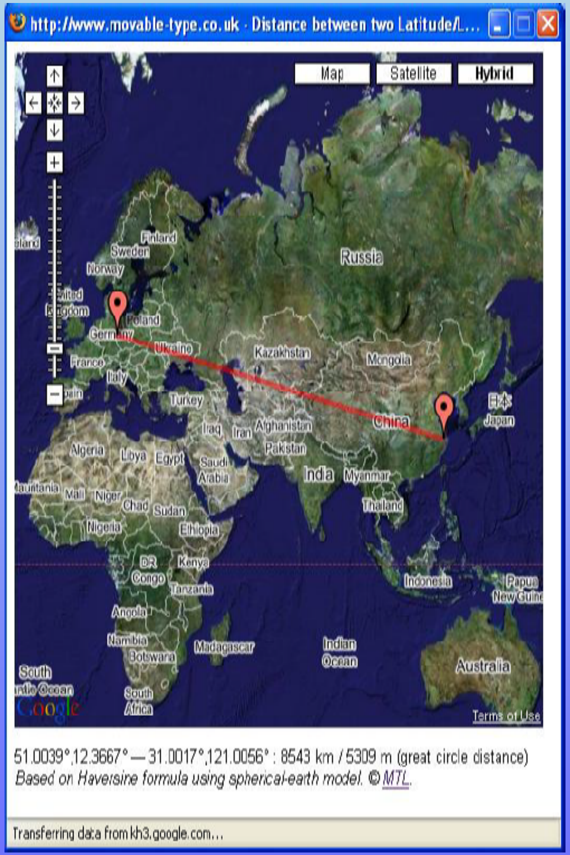
The truth…



http://www.movable-type.co.uk - Distance between two Latitude/L…

51.0039°,12.3667° — 48.0089°,2.2000° : 805.2 km / 500.4 m (great circle distance)
Based on Haversine formula using spherical-earth model. © MTL

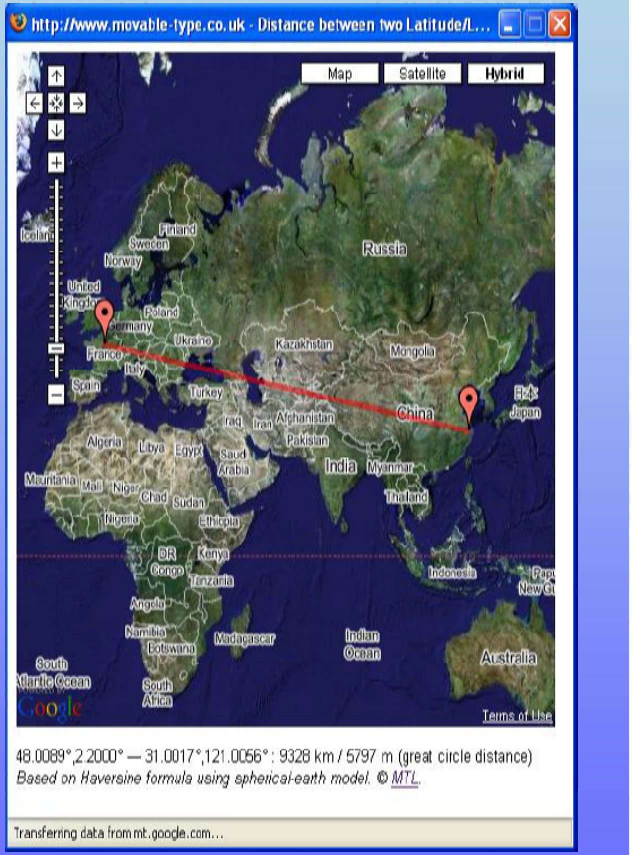
Transferring data from mt.google.com…



http://www.movable-type.co.uk - Distance between two Latitude/L…

51.0039°,12.3667° — 31.0017°,121.0056° : 8543 km / 5309 m (great circle distance)
Based on Haversine formula using spherical-earth model. © MTL.

Transferring data from kh3.google.com…



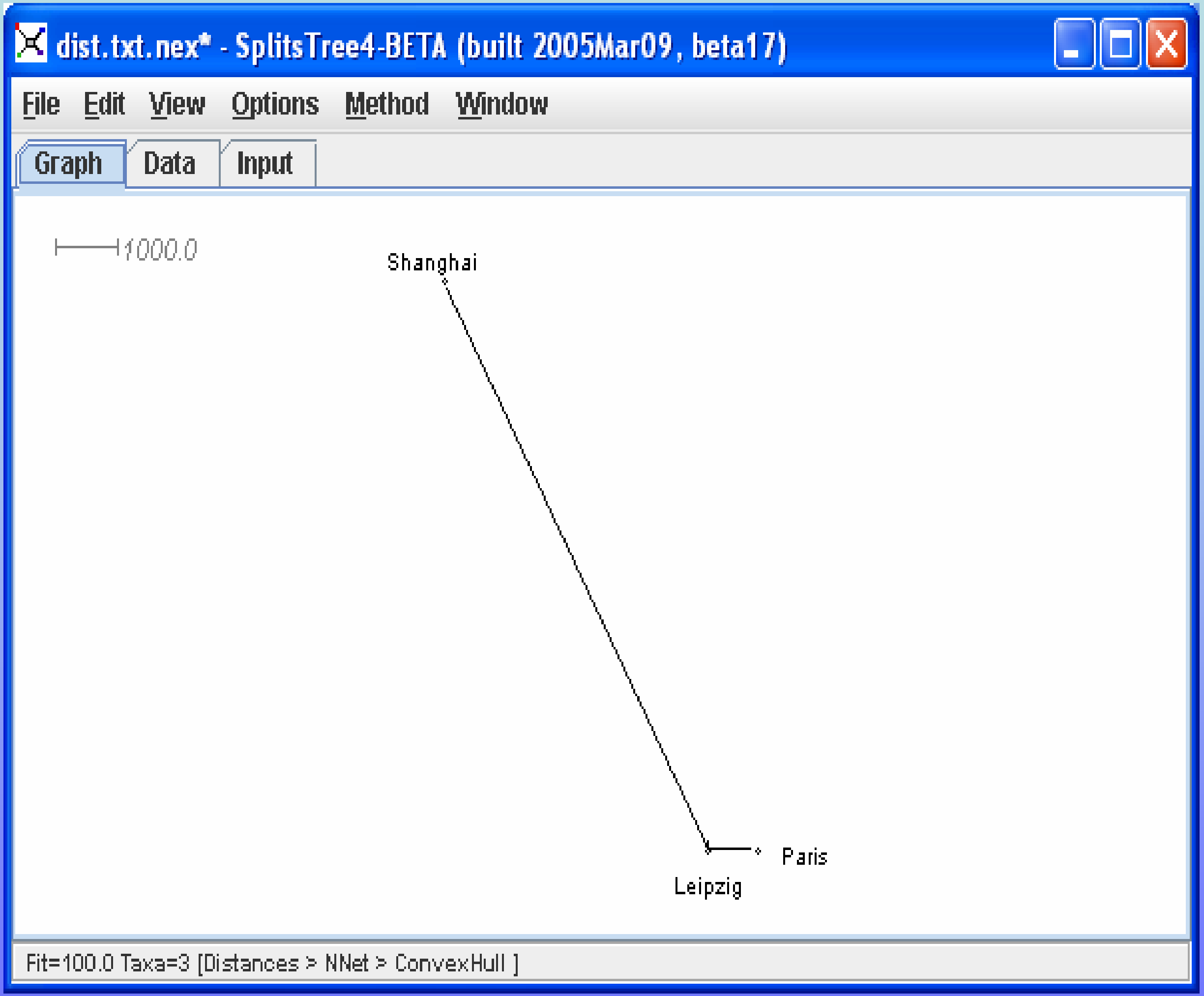
http://www.movable-type.co.uk - Distance between two Latitude/L…

48.0089°,2.2000° — 31.0017°,121.0056° : 9328 km / 5797 m (great circle distance)
Based on Haversine formula using spherical-earth model. © MTL.

Transferring data from mt.google.com…

http://www.movable-type.co.uk/scripts/LatLong.html

The hopes


dist.txt.nex* - SplitsTree4-BETA (built 2005Mar09, beta17)
File Edit View Options Method Window
Graph Data Input
1000.0
Shanghai
Paris
Leipzig

Fit=100.0 Taxa=3 [Distances > NNet > ConvexHull]

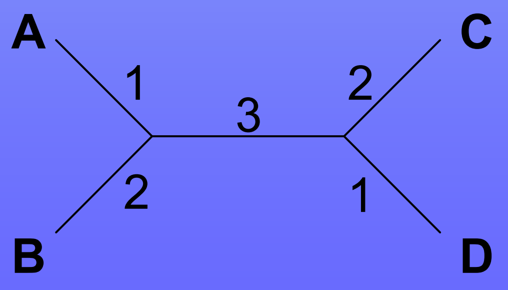
- Phylogenetic trees
- Multidimensional Scaling (MDS)

Question

- How do you build a distance matrix?
  - Use km measure and walk (drive a car?) from each town to each town ☺ / use latitude & longitude parameters together with the Haversine formula.
  - Good, but….
    - How to build a distance matrix for species/languages/any object of interests?

Problems….☹

- **Most of the data have contradictions because:**
  - Species/languages/ etc are not towns
  - Missing data
  - ‘bad’ distance measurement
  - Wrong data/default mistakes.
  - Some data need rescaling
- **Why is this so important?**
  - Distance matrices are one of the most used input files for phylogenetic algorithms.



	A	B	C	D
A	0	3	6	5
B	3	0	7	6
C	6	7	0	3
D	5	6	3	0

	A	B	C	D
A	0	3	1	5
B	3	0	7	6
C	1	7	0	3
D	5	6	3	0

Answer

- Usually distance matrices are built by comparing multiple characteristics for the objects analyzed.
- These characteristics may be:
  - Physical aspects of species
  - DNA /RNA/ Protein alignments
  - Typological features of languages
- At least two important aspects must be considered:
  - Dissimilarities
  - Similarities

Based on this…

	Feat3	Feat9	Feat13	Feat19	Feat63
German	1	2	1	3	?
English	1	2	1	4	1
Romanian	3	?	1	6	1

Feat 3	Description
1	Low (x<2.0)
2	Moderately Low (2.0<x<2.75)
3	Average (2.75<x<4.5)
4	Moderately High (4.75<x<6.5)
5	High (x>6.5)

Feat3: Consonant-Vowel Ratio
Feat9: The Velar Nasal
Feat13: Tone
Feat16: Weight Factors in Weight Sensitive Stress Systems
Feat63: Noun Phrase Conjunction

Andoke: **10** consonants and **9** vowels
Abkhaz: **58** consonants and **2** vowels

Haspelmarth, Gill, Dryer, Comrie, The World Atlas of Language Structures, 2005

Let’s try it….

- Basic Hamming distance (D1)
- Treat ‘?’ as different.
- Ignore ‘?’ but count the available data (D2)
- Replace ‘?’ with the most probable value
- Refine the similarities (D5)
- Refine the dissimilarities
  - NormD (D3)
  - NormDform (D6)
- Both refinements

First approaches

- Treat ‘?’ as different.
  - D(G,E) = 0+0+0+1+1=2
  - D(G,R) = 1+1+0+1+1=4
  - D(E,R) = 1+1+0+1+0=3
- Ignore ‘?’.
  - D(G,E) = 0+0+0+1+0=1
  - D(G,R) = 1+0+0+1+0=2
  - D(E,R) = 1+0+0+1+0=2
- Ignore ‘?’ but count the available data.
  - D(G,E) = 1/4 = 0.25
  - D(G,R) = 2/3 = 0.66
  - D(E,R) = 2/4 =0.5

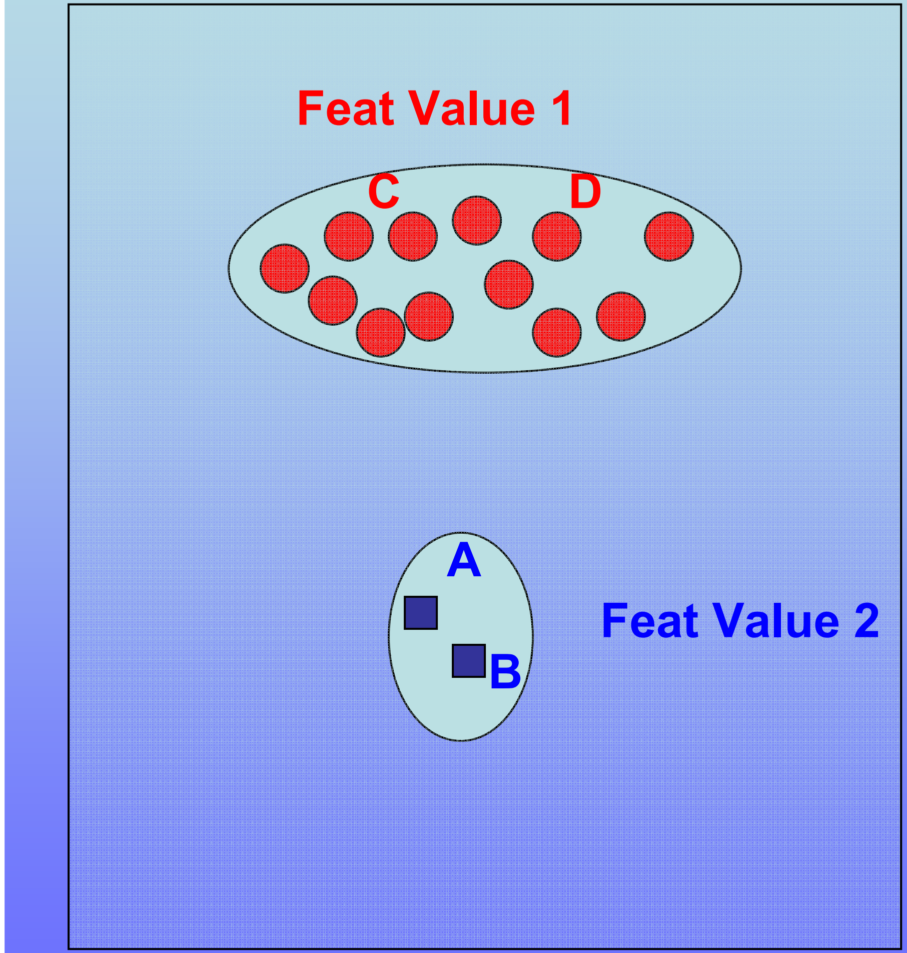
	Feat3	Feat9	Feat13	Feat19	Feat63
German	1	2	1	3	?
English	1	2	1	4	1
Romanian	3	?	1	6	1

One at a time

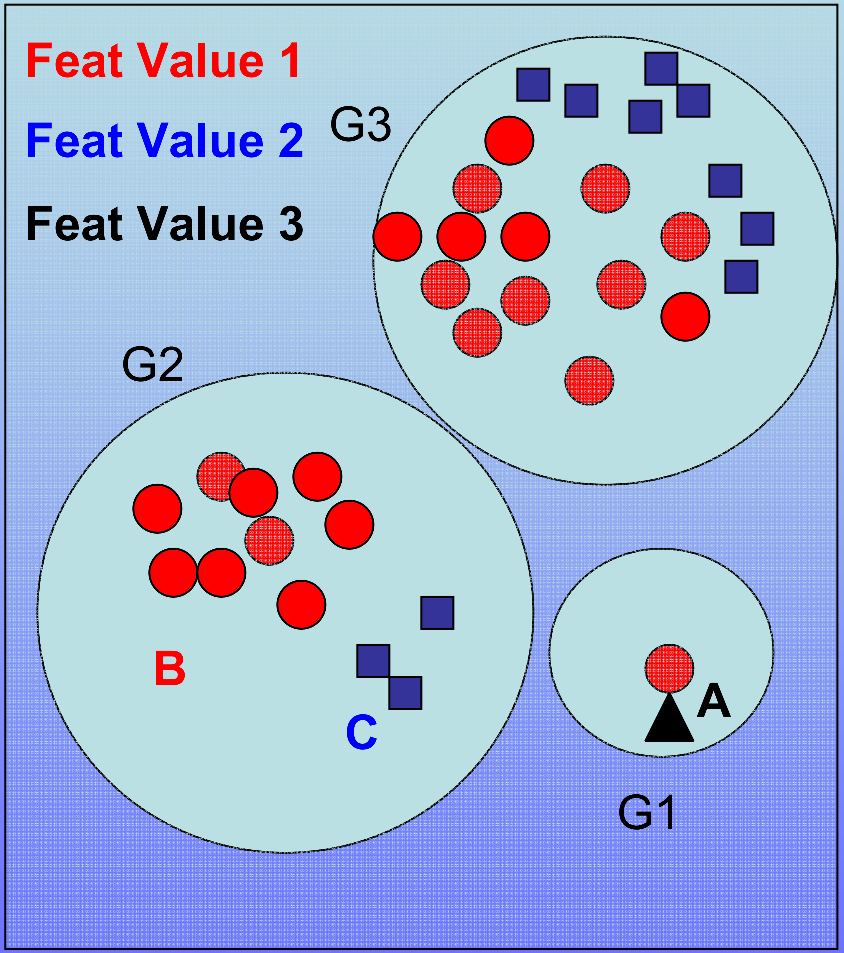
Replace ‘?’ with the most probable value

- Look at the world wide distribution of the feature.
- Get the most probable values.
- Replace ‘?‘ with this value.
- ‘Good’ for good distribution
  - 98% = value 1
  - 2% = value2
  - => replace ‘?’ with value1
- ‘Bad ‘for equal distribution
  - 51% = value 1
  - 49% = value2
  - => replace ‘?’ with ……

Refinements



d(A,B) < d(C,D)



d(B,C) < d(A,B)

Refine the similarities

- For each feature
  - F_A = frequency of value A
  - F_T = frequency of available code point for all values of this feature
- The similarity is defined now as:
  - NormS = F_A/F_T

- Sample: map 51
  - value 1 = 431 cases out of 934 available. => NormS = 0.4603
  - value 2 = 35 cases out of 934 available. => NormS = 0.036
  - => two languages that share the value 2 for feature 51 are more similar than two languages that share the value 1!!!!

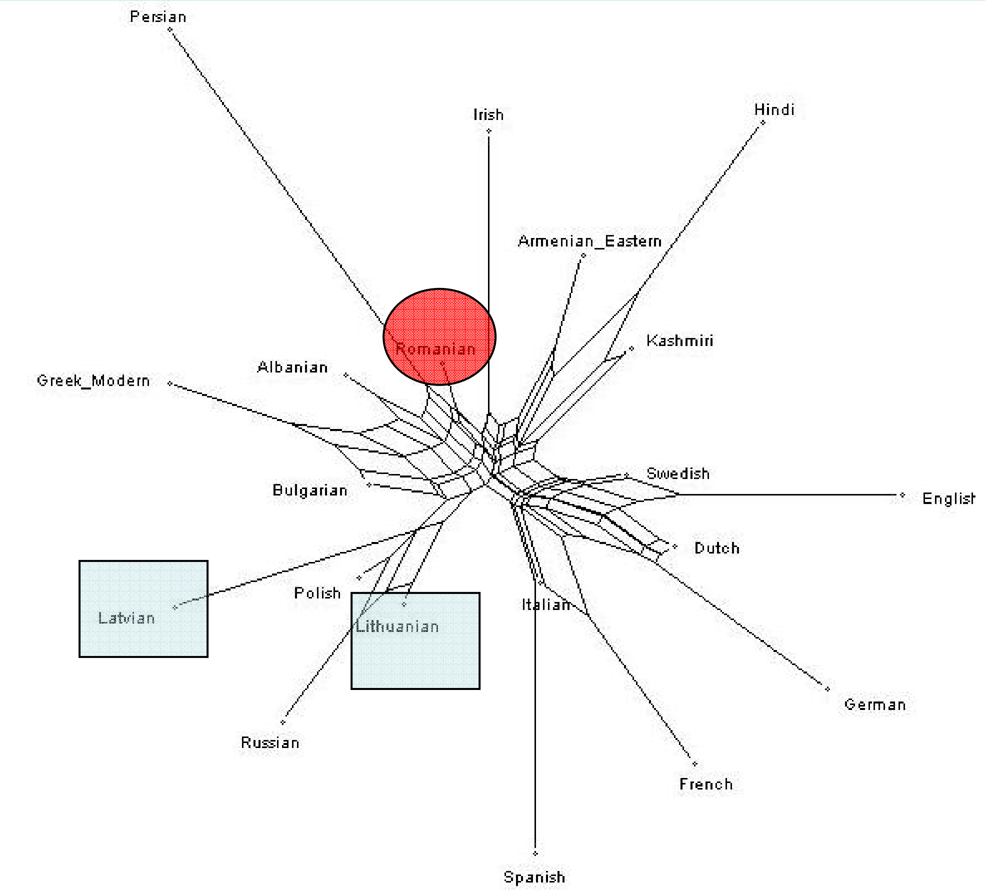
Refine the dissimilarities

- N = the number of genera for which we have information on multiple languages.
- G_A = the number of genera in N that contain a language with value A
- G_B = the number of genera in N that contain a language with value B
- G_{AB} = the number of genera in N that contain both A and B
- Expected coincidence $E = G_A * G_B / N$
- Standard deviation $S = Sqrt(E * (N-E)/N)$
- Difference value $D = (G_{AB}-E)/S$
- NormD = 1 – ((D – Dmin)/(Dmax-Dmin))

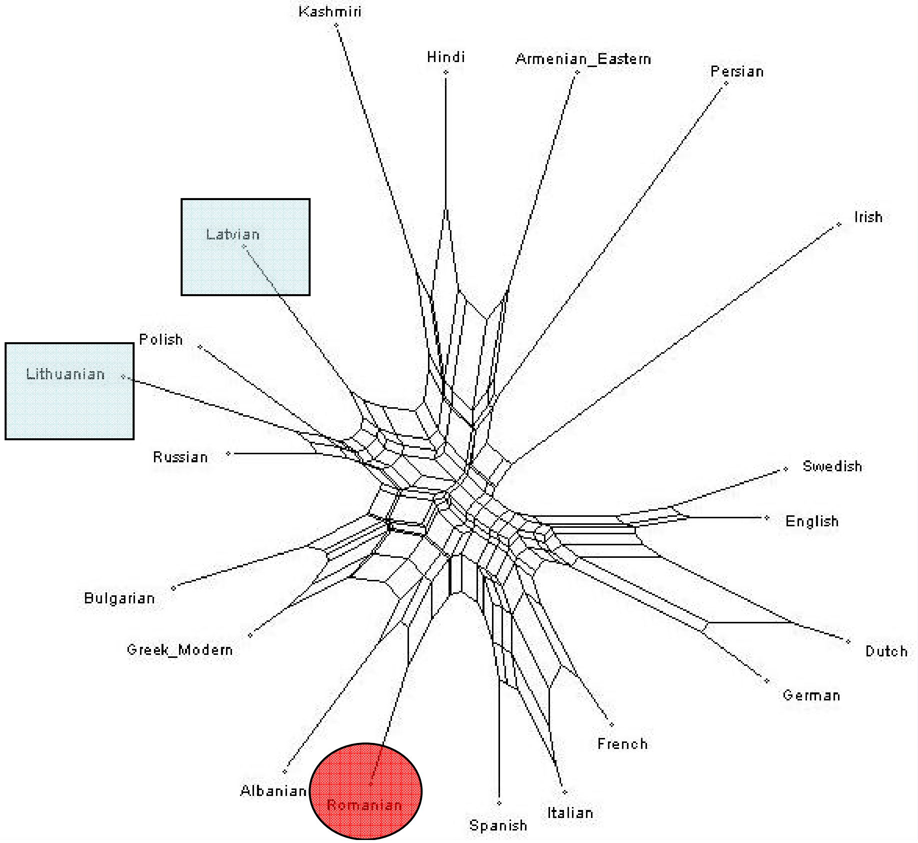
Dissimilarities -formula

- N = the number of genera for which we have information on multiple languages.
- G_A = the number of genera in N that contain a language with value A
- G_B = the number of genera in N that contain a language with value B
- G_{AB} = the number of genera in N that contain both A and B
- $E = G_A * G_B / N$
- $NewNormD=(E - G_{AB} * Log(E) * Log(G_{AB}!)) / Log(N)$

Different methods / different results



Counting the available data



Basic Hamming distance

Improvement measurement

- 6 distance measurements x 5 families x 2 based comparison matrices (P,G) = 30 x 2
- Build geographical and phylogenetic distance matrices
- Calculate the Pearson correlation coefficient for each distance measurement versus each of E/G matrices
- The coefficient is still appropriate, as we don’t need the significance of the measurements, just how much better they became.

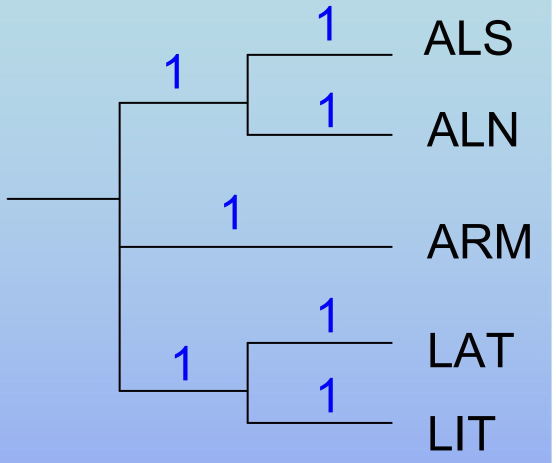
Geographical distances

- Haversine formula
  - Δlat = lat2 - lat1
  - Δlong = long2 – long1
  - $a = \sin^2(\Delta lat/2) + \cos(lat1) * \cos(lat2) * \sin^2(\Delta long/2)$
  - $C = 2\ \mathrm{atan2}(\sqrt{a}, \sqrt{(1-a)})$
  - D = R * C, R = 6.371km
- Very appropriate also for small distances.

R.W.Sinnott, "Virtues of the Haversine", Sky and Telescope, vo.68, no.2, 1984, p.159

Phylogenetic distances

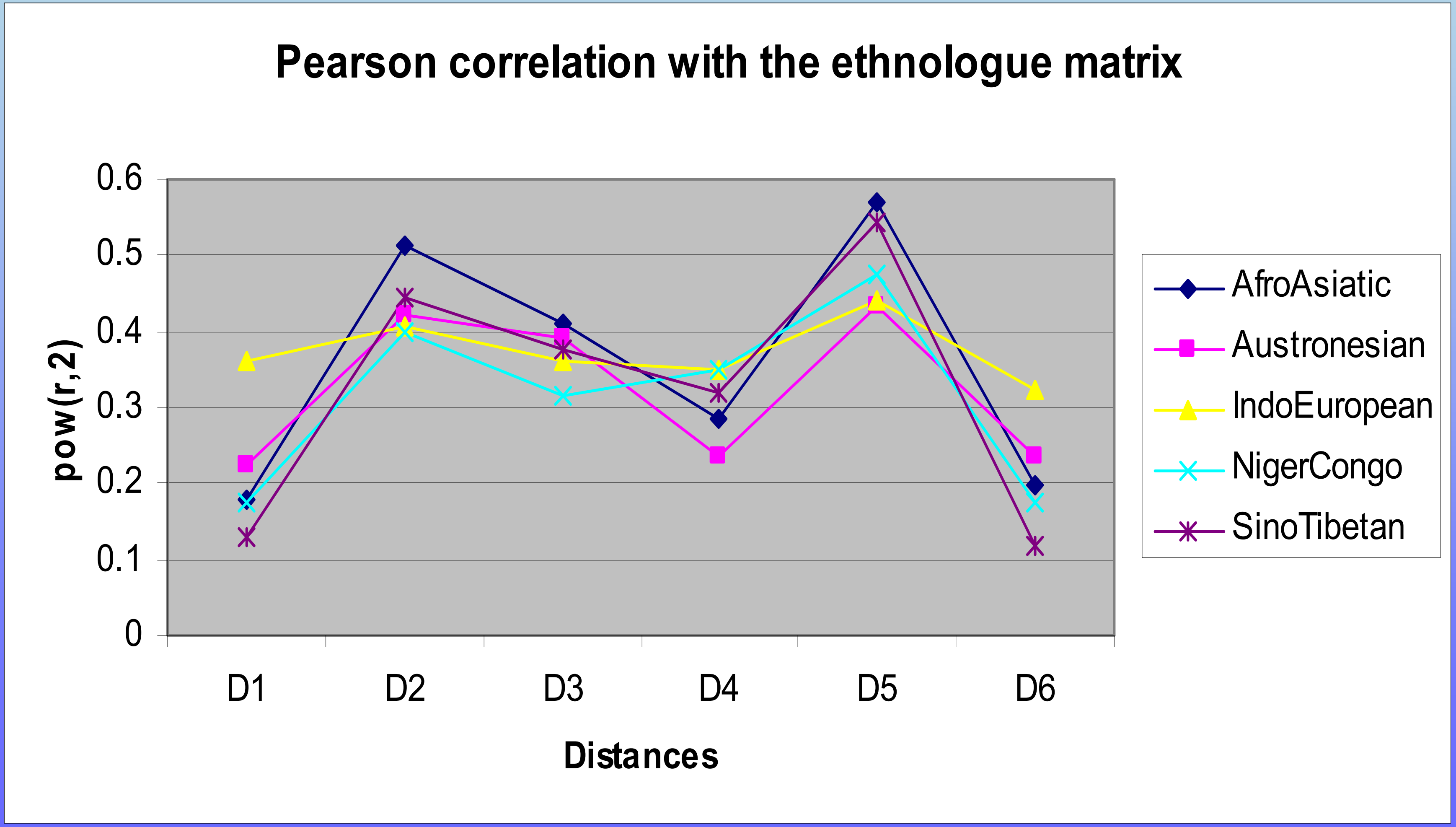
Indo-European (443)
- Albanian (4)
- - Gheg (1)
- - - ALBANIAN, GHEG [ALS] Yugoslavia
- - Tosk (3)
- - - ALBANIAN, TOSK [ALN] Albania
- Armenian (2)
- - ARMENIAN [ARM] Armenia
- Baltic (3)
- - Eastern (2)
- - - LATVIAN [LAT] Latvia
- - - LITHUANIAN [LIT] Lithuania
- - Western (1)
- Celtic (7)
- - Insular (7)
- - - Brythonic (3)
- - - - BRETON [BRT] France
- - - - CORNISH [CRN] United Kingdom
- - - - WELSH [WLS] United Kingdom
- - - Goidelic (4)
- - - - GAELIC, IRISH [GLI] Ireland
- - - - GAELIC, SCOTS [GLS] United Kingdom
- - - - MANX [MJD] United Kingdom



D(ALS,ALN)=2
D(ALS,ARM)=3
D(ARM,LAT)=3
….

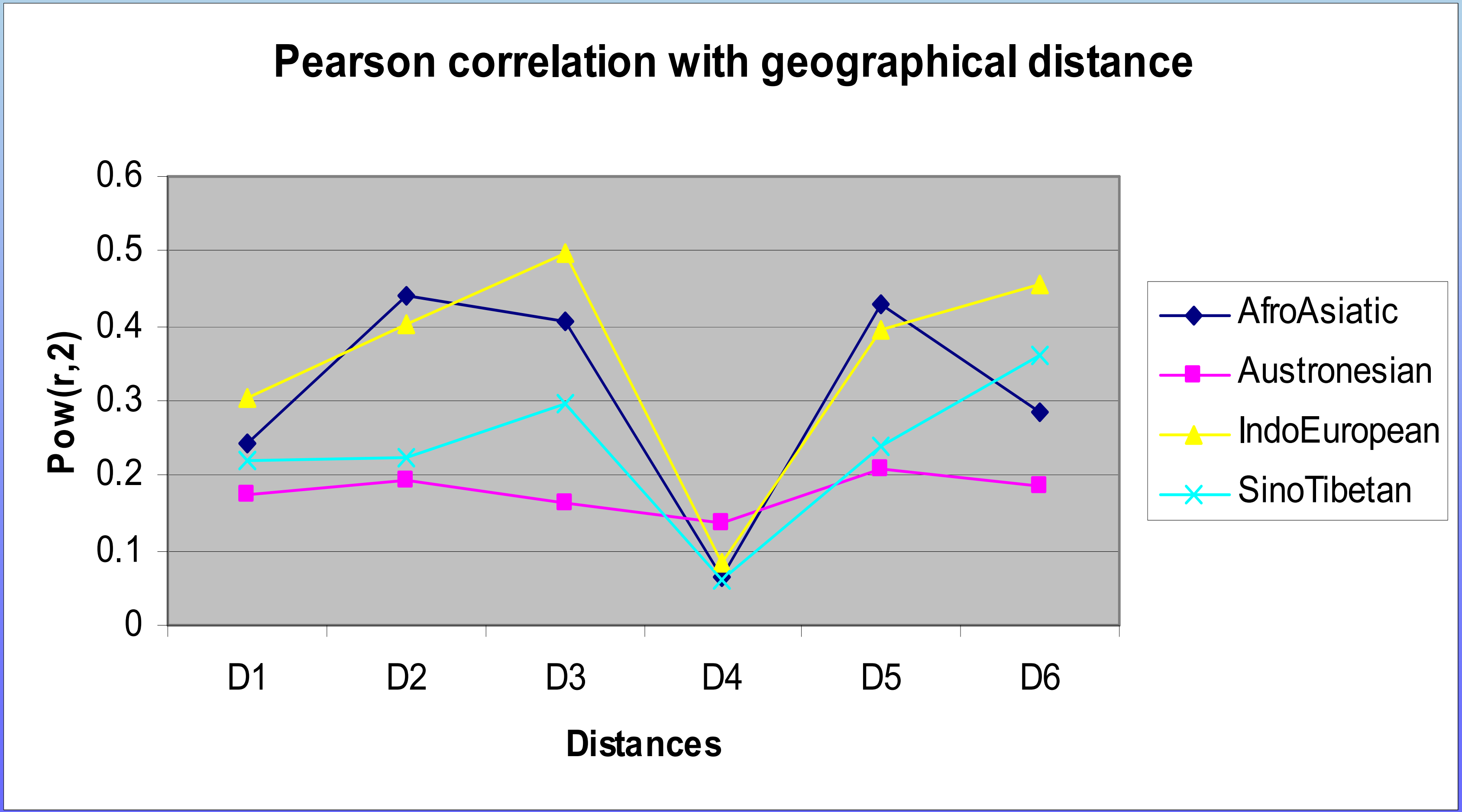
How much better are the methods?

Comparison with the typological distance matrix



How much better are the methods?

Comparison with the geographical distance matrix



Conclusions part 1

Family	Number of data points for the languages analyzed
Sino-Tibetan	1340
Afro-Asiatic	1536
Niger-Congo	1752
Austronesian	1932
Indo-European	1980

- Indo-European (most datapoints) is the most ‘resistant’ to different methods
- Distance measure 2(Hamm depending on the available data) seems to be from the beginning one of the best
- Distance measure 5(measuring the similarities) slightly better
- Distance measure 6 approaches the geographical (it really depends on the data)
- Austronesian has the largest geographical distances.

Conclusions part2

- The methods relate different to the two types of matrices (E&G).
- We still need to understand better why some measurements are getting closer.
- ‘Cleaner’ results might be obtained if different phylogenetic measurement will be applied.
- We need to realize the important of the number of datapoints. Based on this, we should be able to specify the appropriate method depending on each data set.
- Why similarity measure improves the results might be explained because we are analyzing languages located in the same family, so they should have mostly the same feature values.
- Combining both similarities and dissimilarities measurements might produce an even better result.
    - ❖ $D = \mathbf{\Sigma} \mathrm{NormD}/(\mathbf{\Sigma} \mathrm{NormD}+ \mathbf{\Sigma}(1-\mathrm{NormS}))$

Thanks to…

- MPI-EVA
- IZBI
- YOU
- Sebastian for accepting my sleeping habits ;-)

albu@eva.mpg.de