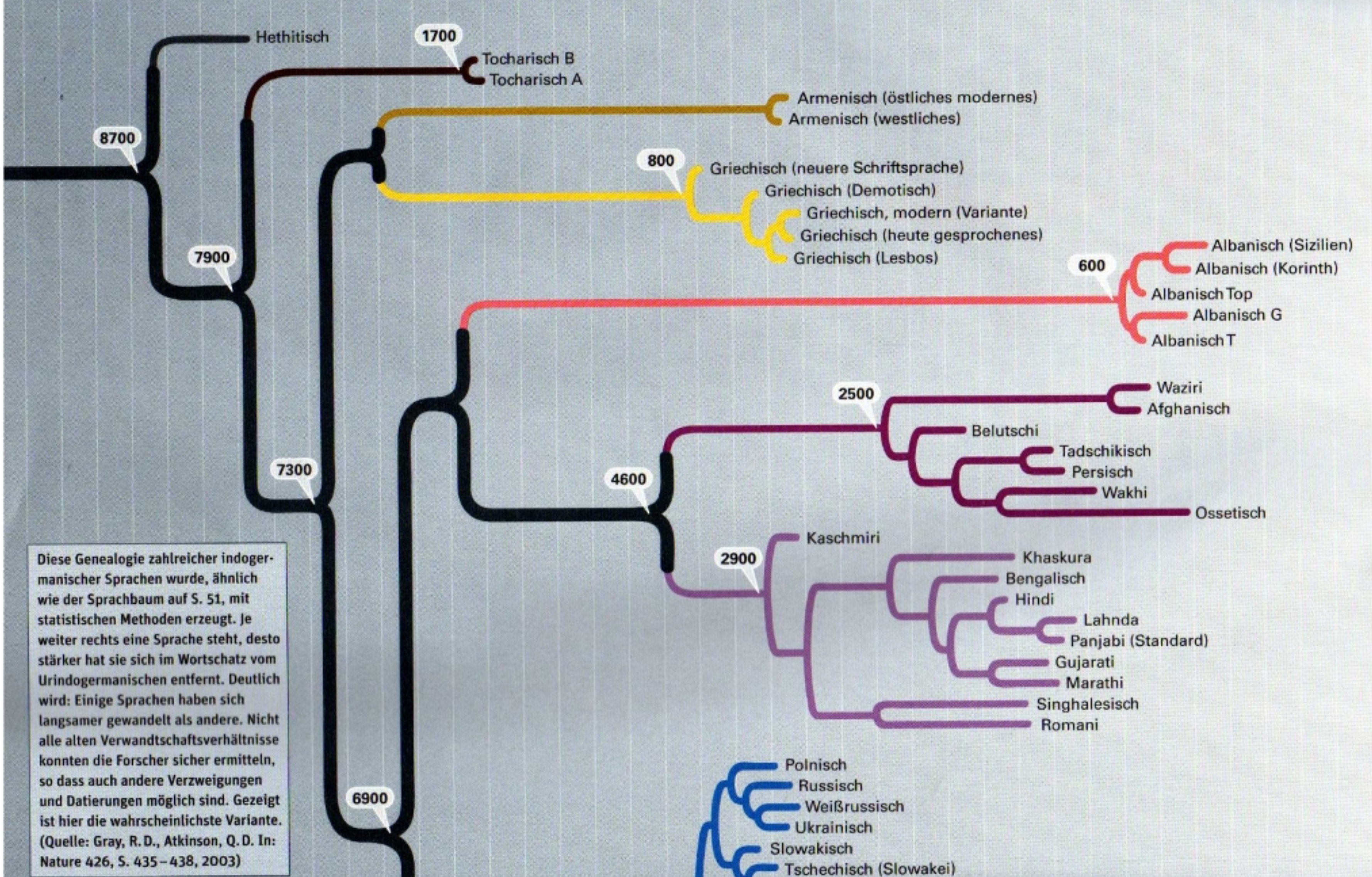
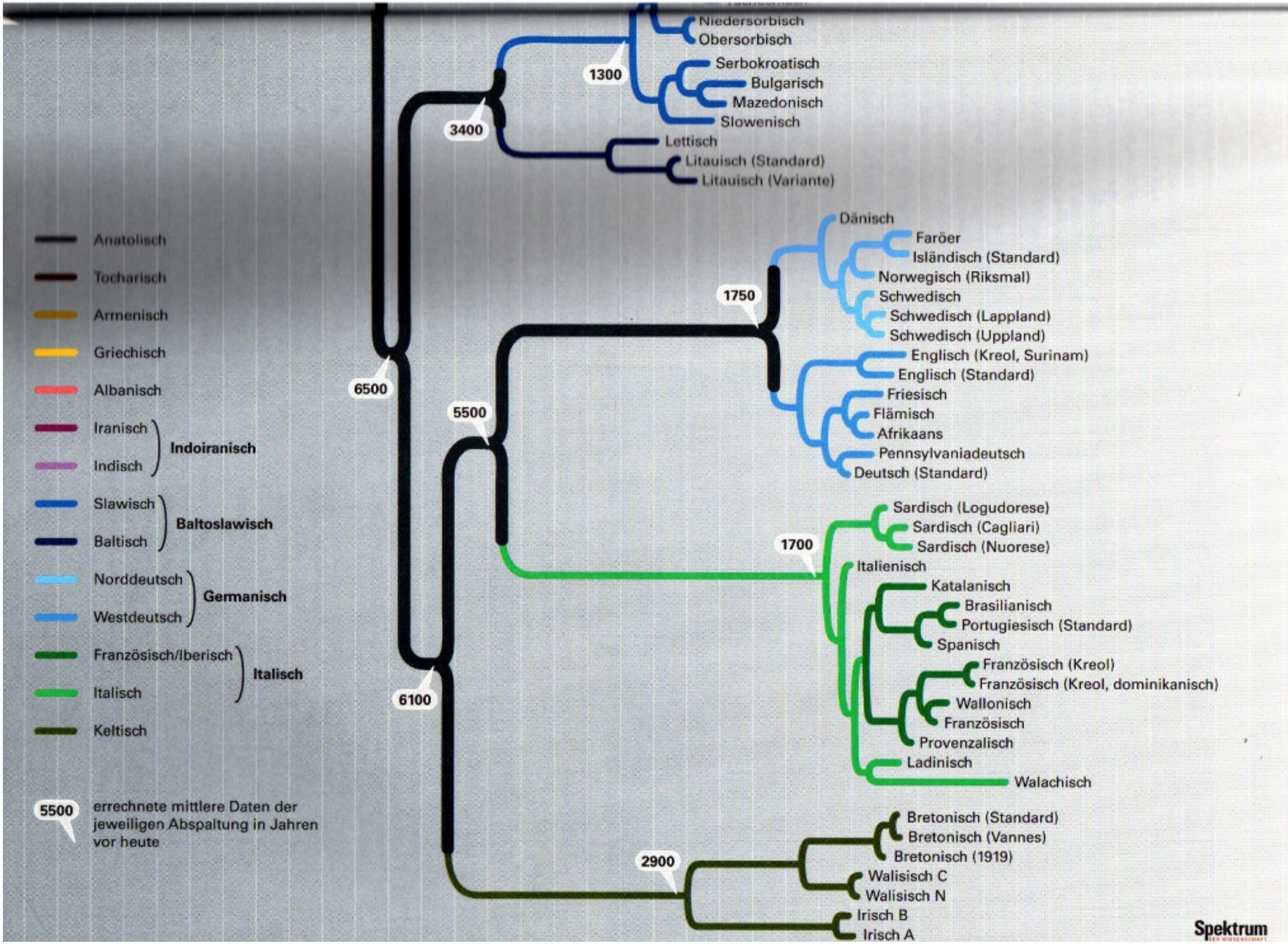
Hierarchical clustering - Dendrogram

INTRODUCTION TO INTELLIGENT SYSTEMS 17/18

STAMMBAUM DER INDOGERMANISCHEN SPRACHEN



Diese Genealogie zahlreicher indogermanischer Sprachen wurde, ähnlich wie der Sprachbaum auf S. 51, mit statistischen Methoden erzeugt. Je weiter rechts eine Sprache steht, desto stärker hat sie sich im Wortschatz vom Urindogermanischen entfernt. Deutlich wird: Einige Sprachen haben sich langsamer gewandelt als andere. Nicht alle alten Verwandtschaftsverhältnisse konnten die Forscher sicher ermitteln, so dass auch andere Verzweigungen und Datierungen möglich sind. Gezeigt ist hier die wahrscheinlichste Variante. (Quelle: Gray, R.D., Atkinson, Q.D. In: Nature 426, S. 435–438, 2003)



Niedersorbisch
Obersorbisch
1300
Serbokroatisch
Bulgarisch
Mazedonisch
Slowenisch
3400
Lettisch
Litauisch (Standard)
Litauisch (Variante)
Anatolisch
Tocharisch
Armenisch
Griechisch
Albanisch
Iranisch
Indisch
Indoiranisch
Slawisch
Baltisch
Baltoslawisch
Norddeutsch
Westdeutsch
Germanisch
Französisch/Iberisch
Italisch
Italisch
Keltisch
5500
errechnete mittlere Daten der jeweiligen Abspaltung in Jahren vor heute
6500
5500
6100
1750
Dänisch
Faröer
Isländisch (Standard)
Norwegisch (Riksmal)
Schwedisch
Schwedisch (Lappland)
Schwedisch (Uppland)
Englisch (Kreol, Surinam)
Englisch (Standard)
Friesisch
Flämisch
Afrikaans
Pennsylvaniadeutsch
Deutsch (Standard)
1700
Sardisch (Logudorese)
Sardisch (Cagliari)
Sardisch (Nuorese)
Italienisch
Katalanisch
Brasilianisch
Portugiesisch (Standard)
Spanisch
Französisch (Kreol)
Französisch (Kreol, dominikanisch)
Wallonisch
Französisch
Provenzalisch
Ladinisch
Walachisch
2900
Bretonisch (Standard)
Bretonisch (Vannes)
Bretonisch (1919)
Walisisch C
Walisisch N
Irisch B
Irisch A
Spektrum

How can we build such diagrams (called dendrograms) in which the objects cluster in groups of different size at different levels?

Dissimilarity between words

Let

S_1 = {a, e, i, o, u, y} (vowels)

S_2 = {b, c, d, f, g, h, j, k, l, m, n, p, q, r, s, t, v, w, x, z} (consonants)

We define the dissimilarity d(x,x') between two characters x and x' as follows:

d(x,x') = 0 if x == x'

= 1 if x != x' and both in S_1 (vowels)

= 2 if x != x' and both in S_2 (consonants)

= 5 if x != x' and in different sets (one is vowel, the other consonant)

= 7 if x != empty and x' == empty OR

x == empty and x' != empty

Dissimilarity between words

Next we define the dissimilarity of two words as the sum of the dissimilarities of counterpart characters

- Example 1:
  - dissimilarity(‘boy’, ‘bay’) = 0 + 1 + 0 = 1
- Example 2:
  - dissimilarity(‘boss’, ‘bayes’) = 0 + 1 + 5 + 5 + 7 = 18

General requirements on (any) dissimilarity

- non-negativity $d(\mathrm{x}, \mathrm{x}') \geq 0$
- reflexivity $d(\mathrm{x}, \mathrm{x}') = 0 \quad \text{if and only if} \quad \mathrm{x} = \mathrm{x}'$
- symmetry $d(\mathrm{x}, \mathrm{x}') = d(\mathrm{x}', \mathrm{x})$
- triangle inequality: $d(x, x') + d(x', x'') \geq d(x, x'')$

Dissimilarity matrix

	Baby	***Day***	***Disc***	***Human***	***Mucus***	***Music***	***Mainly***	***People***
Baby		12	10	11	11	11	22	23
Day			11	18	18	18	18	19
Disc				15	15	13	20	20
Human					7	7	20	20
Mucus						5	18	20
Music							18	20
Mainly								7
People								

Agglomerative clustering

Starting from individual objects, produce a sequence of clusters of increasing size.

We define the dissimilarity between two clusters as the smallest pair-wise dissimilarity of objects from these clusters, one object form each cluster (single linkage)

$$d_{\min}(D_i, D_j) = \min_{\mathrm{x} \in D_i, \mathrm{x}' \in D_j} \| \mathrm{x} - \mathrm{x}' \|$$

Dissimilarity matrix

	Baby	***Day***	***Disc***	***Human***	***(Mucus, Music)$_5$***	***Mainly***	***People***
Baby		12	10	11	11	22	23
Day			11	18	18	18	19
Disc				15	13	20	20
Human					7	20	20
(Mucus, Music)$_5$						18	20
Mainly							7
People							

Dissimilarity matrix

	Baby	Day	Disc	$((Mucus, Music)_5 \ Human)_7$	$(Mainly, People)_7$
Baby		12	10	11	22
Day			11	18	18
Disc				13	20
$((Mucus, Music)_5 \ Human)_7$					18
$(Mainly, People)_7$					

Dissimilarity matrix

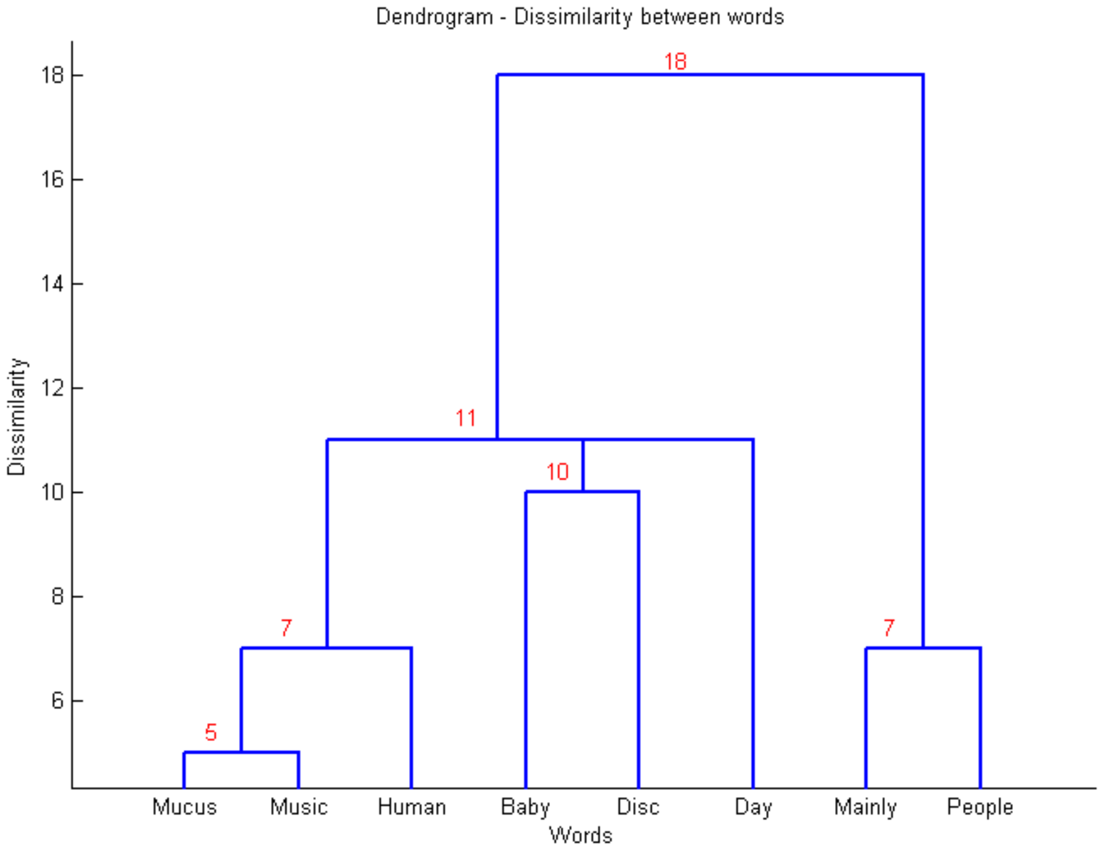
	(Baby,Disc)$_{10}$	***Day***	***((Mucus,Music)$_{5}$ Human)$_{7}$***	***(Mainly, People)$_{7}$***
(Baby,Disc)$_{10}$		11	11	20
Day			18	18
((Mucus,Music)$_{5}$ Human)$_{7}$				18
(Mainly,People)$_{7}$				

Dissimilarity matrix

	(Day, (Baby, Disc)$_{10,}$((Mucus, Music)$_5$ Human)$_7$)$_{11}$	***(Mainly,People)$_7$***
(Day, (Baby, Disc)$_{10,}$((Mucus, Music)$_5$ Human)$_7$)$_{11}$		18
(Mainly,People)$_7$		

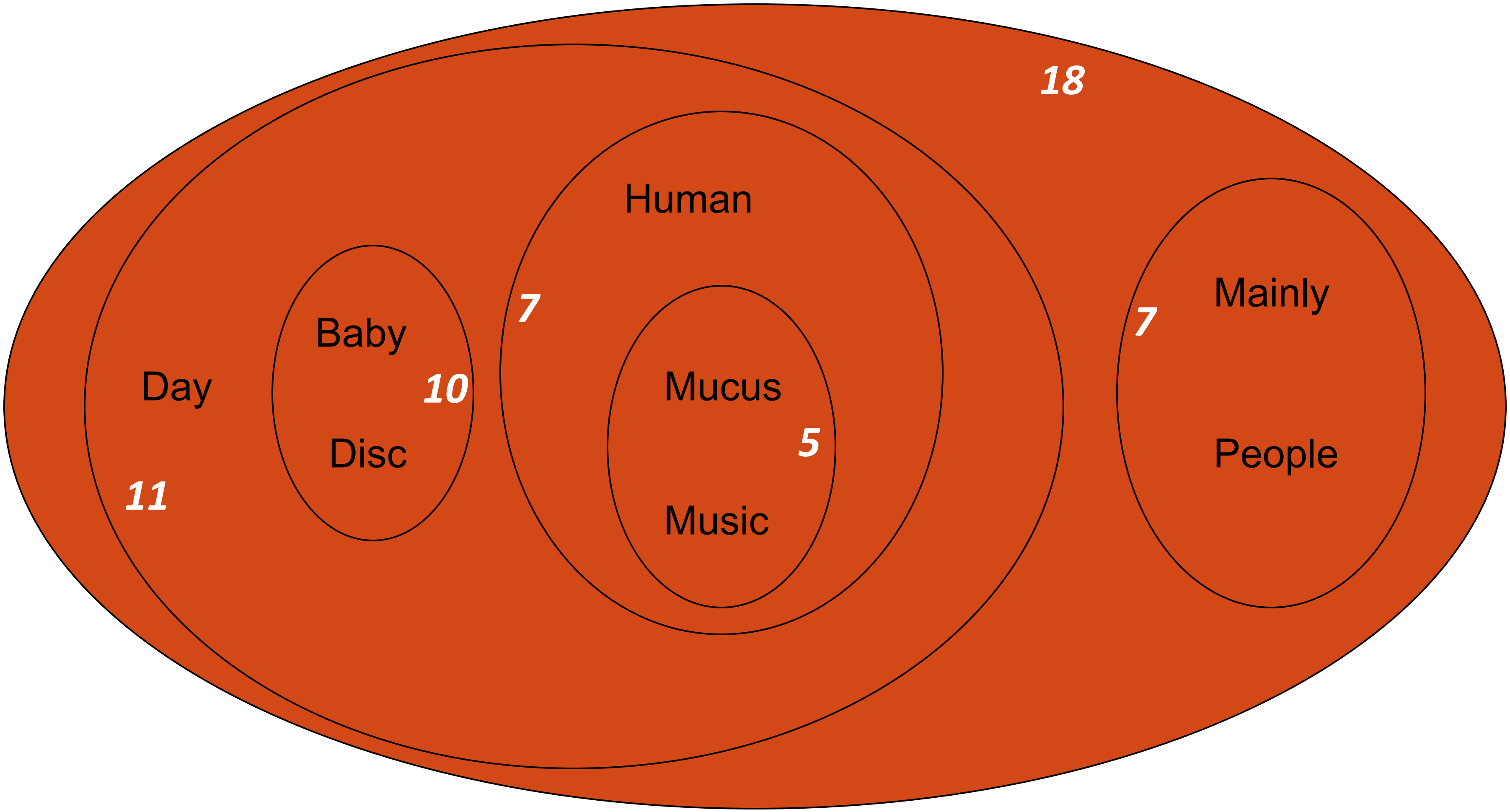
Dendrogram

Representation of the clustering hierarchy

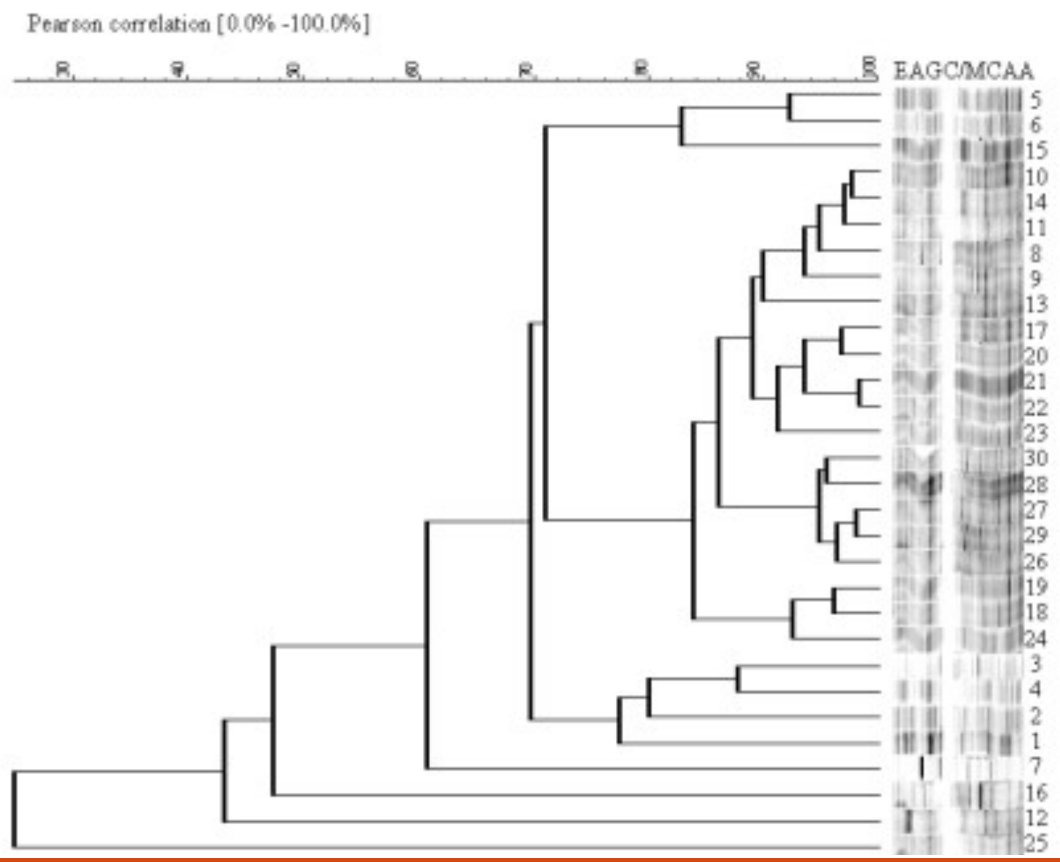


Venn diagram

$((((\text{Baby, Disc})_{10}, ((\text{Mucus, Music})_{5}, \text{Human})_{7})_{11}, \text{Day})_{11}, (\text{Mainly,People})_{7})_{18}$



Black pepper cultivars using AFLP analysis along with digital fingerprint profile (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1948014)



The clustering pattern obtained for the major cultivars of black pepper.

(http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1948014)

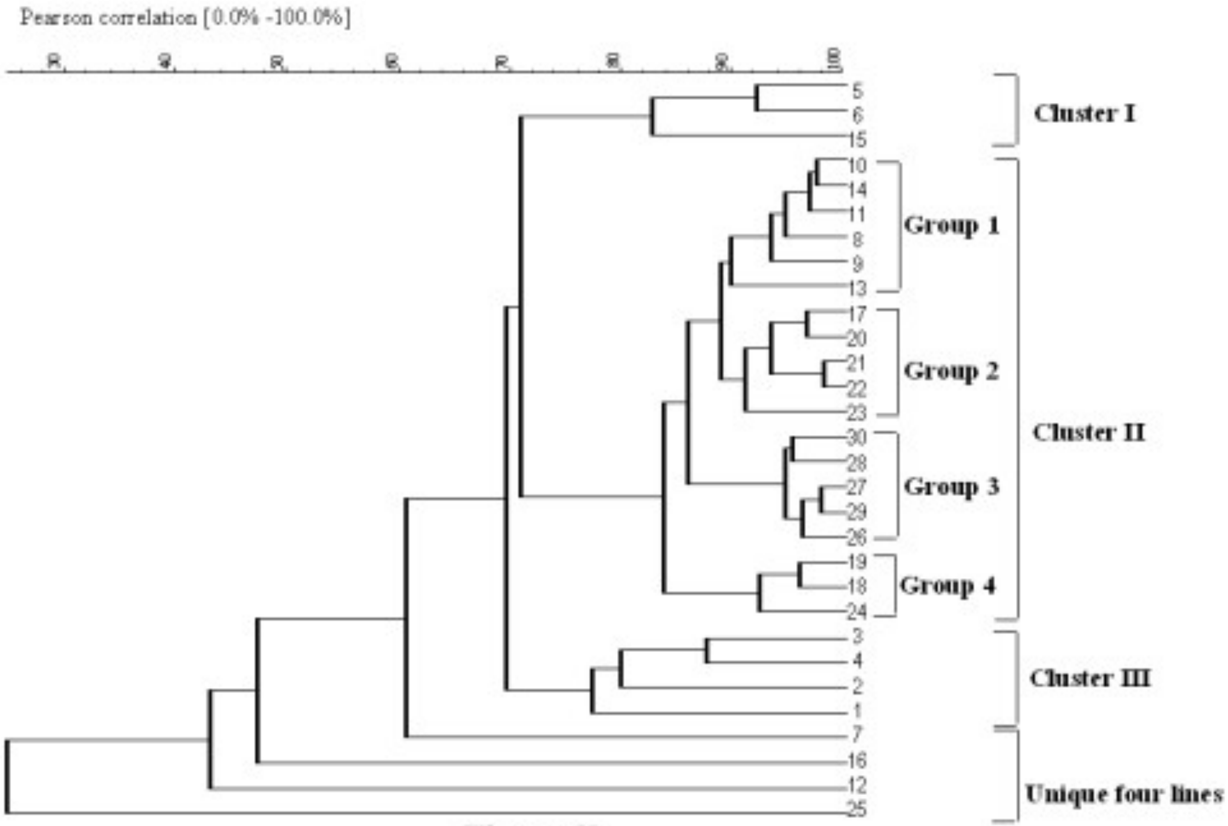


Figure 3b

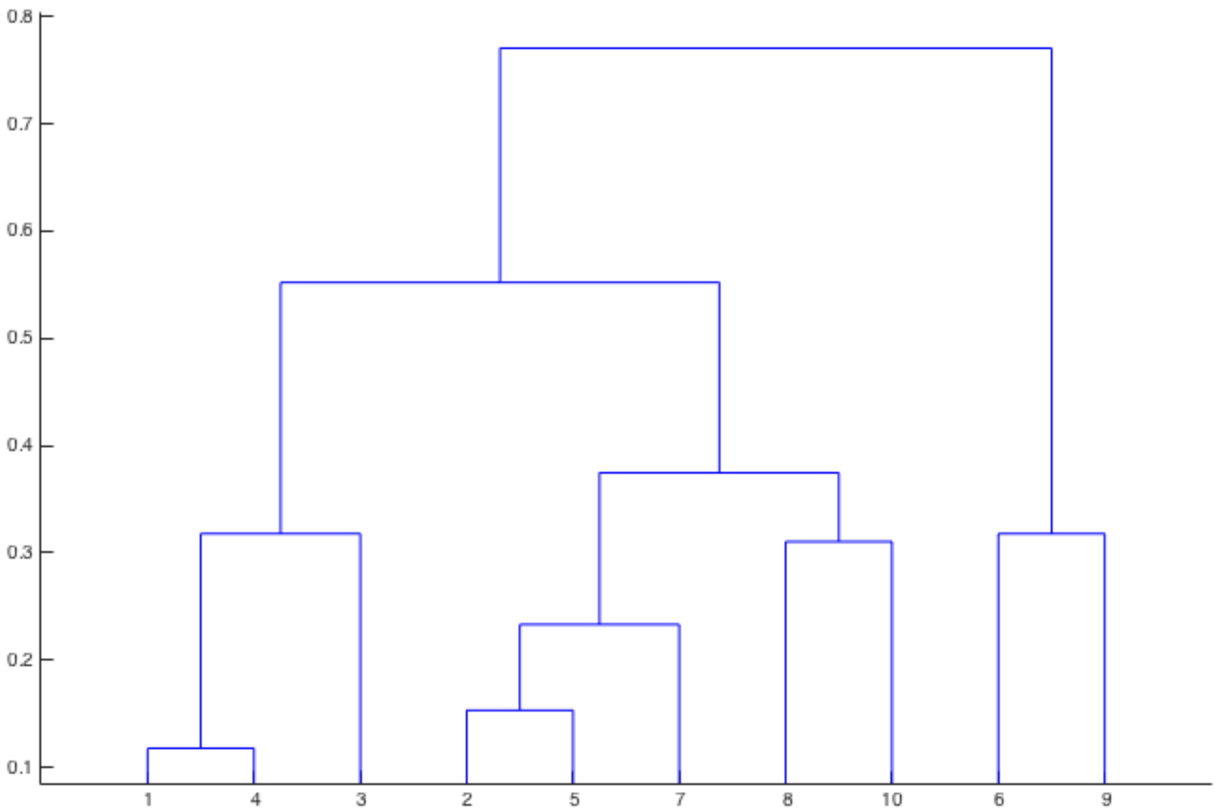
UPGMA dendrograms. A: black pepper cultivars using AFLP analysis along with digital fingerprint profile. B: the clustering pattern obtained for the major cultivars of black pepper.

Dendrogram in Matlab (example)

Let x_i, i = 1 … 10, be n = 10 feature vectors representing 10 objects. We will make a dendrogram for this set.

1. Put x_i, i = 1 … 10, to become the rows of a matrix X.
   (E.g. X = rand(10,2); (10 2-dim feature vectors)
2. Compute the pairwise distances of all observations in matrix X using a given distance metric: *Y = pdist(X, 'cityblock');*
   (Y is a row vector that includes the off-diagonal elements of a pair-wise distance matrix. It has n(n-1)/2 elements.)
3. Using the pair-wise distances, compute a (n-1)*3 matrix Z that represents a hierarchical binary cluster tree:
   Z = linkage(Y, 'average'); % ‘average’ – type of linkage used
4. Compute and plot a dendrogram using
   *[H, T] = dendrogram(Z); (T – n*1; H – vector of line handles)*

Dendrogram in Matlab (example)



X =[0.3477 0.7363;
0.1500 0.3947;
0.5861 0.6834;
0.2621 0.7040;
0.0445 0.4423;
0.7549 0.0196;
0.2428 0.3309;
0.4424 0.4243;
0.6878 0.2703;
0.3592 0.1971]