

der optimale Verbalstammbaum

Ich hatte das LIV² als Wordfile vorliegen und habe daraus eine mysql-Datenbank generiert. Daran lassen sich allerhand Zählungen und Statistiken anstellen. Ich habe mir überlegt, ob sich aus den Daten des LIV nicht eine Darstellung des Verwandtschaftsgrades der einzelnen idg. Sprachfamilien ziehen liesse. Hier ist erstmal ein Versuch zur Berechnung eines Stammbaumes, ausschliesslich nach Kriterien der verbalen Primärstammbildung also.

Um etwaiger Willkür im Ansetzen eines ererbten Stammes im LIV auszuweichen, werden von Anfang an nur Stämme berücksichtigt, die aufgrund von Belegen aus mehr als einer Sprachfamilie rekonstruiert sind. Das sind im LIV² insgesamt 1248. Ihre Verteilung sieht so aus:

	ind	ir	balt	slav	anat	gr	it	germ	kelt	toch	arm	alb
ind	661	343	78	96	61	210	110	125	73	72	51	32
ir	343	421	61	62	32	118	67	78	46	44	35	27
balt	78	61	276	125	14	71	58	94	37	23	26	18
slav	96	62	125	269	23	73	50	71	32	26	24	21
anat	61	32	14	23	124	50	31	17	17	17	9	14
gr	210	118	71	73	50	473	132	96	65	62	62	44
it	110	67	58	50	31	132	281	75	49	31	36	33
germ	125	78	94	71	17	96	75	337	52	36	19	20
kelt	73	46	37	32	17	65	49	52	160	18	22	15
toch	72	44	23	26	17	62	31	36	18	144	14	12
arm	51	35	26	24	9	62	36	19	22	14	110	14
alb	32	27	18	21	14	44	33	20	15	12	14	82

Ein Eintrag in dieser Tabelle ist allerdings kein direktes Mass für den Verwandtschaftsgrad der jeweiligen Sprachen: Die Zahl ist umso grösser, je grösser die totale Anzahl der Belege in den beteiligten Sprachen ist. Man kann die Tabelle normieren, indem man jedes nichtdiagonale Feld mit dem Quotienten der beteiligten Diagonalelemente multipliziert und dann die einzelnen Spalten so skaliert, dass alle Diagonalelemente 1 werden:

	ind	ir	balt	slav	anat	gr	it	germ	kelt	toch	arm	alb
ind	100	51	11	14	9	31	16	18	11	10	7	4
ir	81	100	14	14	7	28	15	18	10	10	8	6
balt	28	22	100	45	5	25	21	34	13	8	9	6
slav	35	23	46	100	8	27	18	26	11	9	8	7
anat	49	25	11	18	100	40	25	13	13	13	7	11
gr	44	24	15	15	10	100	27	20	13	13	13	9
it	39	23	20	17	11	46	100	26	17	11	12	11
germ	37	23	27	21	5	28	22	100	15	10	5	5
kelt	45	28	23	20	10	40	30	32	100	11	13	9
toch	50	30	15	18	11	43	21	25	12	100	9	8
arm	46	31	23	21	8	56	32	17	20	12	100	12
alb	39	32	21	25	17	53	40	24	18	14	17	100

Zu beachten ist hier: (a) Die Tabelle ist nun nicht mehr symmetrisch; (b) Wir gehen implizite von einer sehr hohen Anzahl von Stammbildungen in der Ursprache aus (weil wir damit rechnen, dass mit einer Verdoppelung von Belegen in der einen Sprache die Anzahl der Übereinstimmungen mit einer anderen Sprache sich auch verdoppelt) und (c) Wir verlieren möglicherweise Information, indem wir die absolute Anzahl der Belege nicht berücksichtigen (eine kleinere Zahl von Belegen könnte auf eine grössere Entfernung zur Ursprache deuten. Solche Überlegungen können wir aber nicht anstellen ohne Abschätzungen zur Überlieferungsqualität der einzelnen Sprachen – weil diese enorm unterschiedlich sind – müssten also Information von ausserhalb des LIV heranziehen)

Die Spalten der Tabelle, die wir so gewonnen haben, können als Vektoren in einem 12-dimensionalen Raum aufgefasst werden, die ein 12-dimensionales ‘Rhomboid’ aufspannen. Das Volumen desselben wäre ein Mass für die Stärke der Korrelation zwischen den versammelten Sprachfamilien. Genauer gesagt normiere ich die Spaltenvektoren zuerst auf einheitliche Länge (um Sprachen mit vielen Übereinstimmungen kein Übergewicht zu gewähren), und berechne die zwölfte Wurzel aus der Determinante der resultierenden Matrix (also die Seitenlänge eines 12-dimensionalen ‘Kubus’ desselben Volumens, um den unwesentlichen Umstand, dass wir von gerade 12 Familien ausgehen, nicht in das Resultat einfließen zu lassen). Das Resultat auf einer Skala von 0 (totale Übereinstimmung) bis 1 (keine Übereinstimmung) lautet

$$\mathbf{LIV}^2 = 67.6\% .$$

Wie kommen wir aber nun zu einem Stammbaum? Eine Idee wäre, die Zweige des Stammbaums als Unterräume des 12-dimensionalen Gesamt-Raums zu betrachten und die Verzweigungen immer nach dem Kriterium der jeweils kleinsten Unter-Determinanten vor sich gehen zu lassen. Konkret würde man also alle 8191 Unter-Determinanten (bzw. die jeweils n -ten Wurzeln, um Determinanten verschiedener Dimension miteinander vergleichen zu können) berechnen, und sich dann für die beiden mit dem kleinsten Produkt entscheiden, das wären im ersten Durchgang (im obersten Knoten) $100\% \times 65.21\%$ (Anatolisch vs. der Rest).

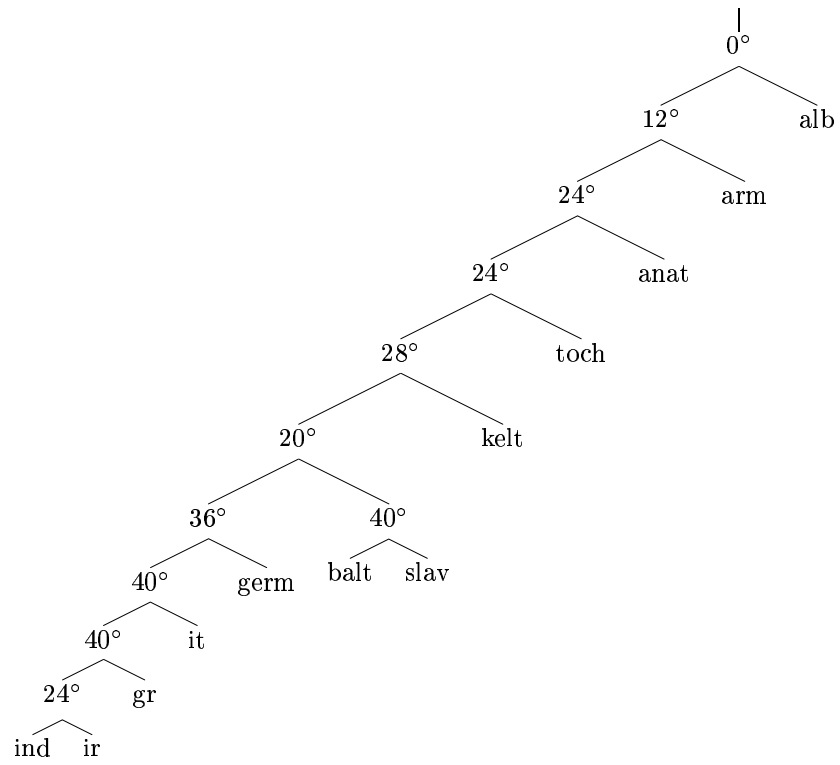
Soweit so gut, aber es stellt sich heraus, dass Paare mit grosser Korrelation (also v.a. Indo-Iranisch) dazu neigen, Clusters zu bilden, d.h. sie saugen andere Familien an sich, die an sich nicht allzu nahe verwandt wären, nur um mehr Dimensionen zu gewinnen, damit sie besser von ihrem kleinen Produkt profitieren können.¹ Der Stammbaum würde also ungehörigerweise davon beeinflusst, wie detailliert wir Dialekte unterscheiden (d.h. ob wir von Indo-Iranisch ausgehen oder aber von Indisch und Iranisch hätte Konsequenzen für die Struktur des ganzen Baumes).

Besser ist es, den Stammbaum von unten aufzubauen: Ich berechne alle 121 zweidimensionalen Unterdeterminanten der Matrix und wähle die kleinste aus (Indo-Iranisch, 39.7%). Dann baue ich die gesamte Matrix neu auf aus der Stammformen-Datenbank, diesmal aber nur noch 11-dimensional, mit einer Kolumne 'Indisch+Iranisch', berechne alle 110 2-dimensionalen Determinanten und so weiter.

Um zu vermeiden, dass wieder Effekte auftreten, die nur mit der Wahl der Familienzahl zusammenhängen, muss ich aber bei jedem Aufbau einer neuen Matrix diejenigen Formen ausschliessen, die in nur einem Zweig des bisherigen Baumes vorkommen (ganz analog wie ich anfangs die Stämme fortwarf, die in nur einer Familie belegt sind). Das hat die zwingende Konsequenz, dass zuoberst an meinem Baum die Zahl Null stehen wird (es werden nur Stämme aus beiden Zweigen berücksichtigt, also totale Korrelation).

Noch ein kosmetischer Punkt, die zweidimensionale Determinante lässt sich als Fläche eines Rhombus interpretieren: Die beiden beteiligten Sprachen stellen zwei Seiten des Rhombus dar, die sich mehr oder weniger zugeneigt sind, eine kleinere Fläche bedeutet also grössere Korrelation. Um die Korrelation intuitiver darzustellen, kann ich aber auch den Winkel zwischen den beiden Sprach-Vektoren angeben: 0° bedeutet totale, 90° gar keine Korrelation. Von der Determinante D auf den Winkel α komme ich also mit dem Zusammenhang $D = 2 \sin \frac{\alpha}{2} \cos \frac{\alpha}{2}$. Dementsprechend sind im folgenden Baum-Diagramm an den Vertizes die 'Öffnungswinkel' zwischen den Ästen angegeben.

¹Ok, diese Erklärung ist etwas kurz. Die Quintessenz ist, die Methode ist schlecht.



Dieses Dokument, bzw. Weiterentwicklungen davon, befindet sich hier:
www.flaez.ch/pdf/baum.pdf