Lexikostatistik 2.0

Gerhard Jäger

Seminar für Sprachwissenschaft, Tübingen

IDS-Jahrestagung

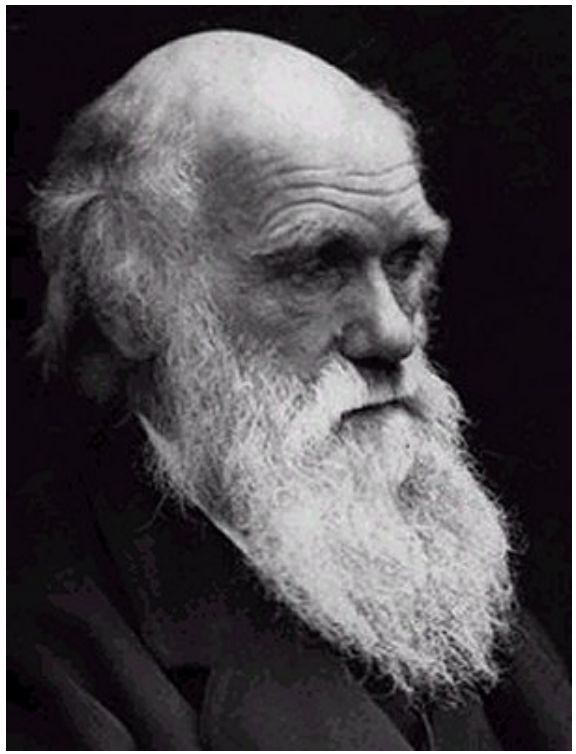
12. März 2013

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)



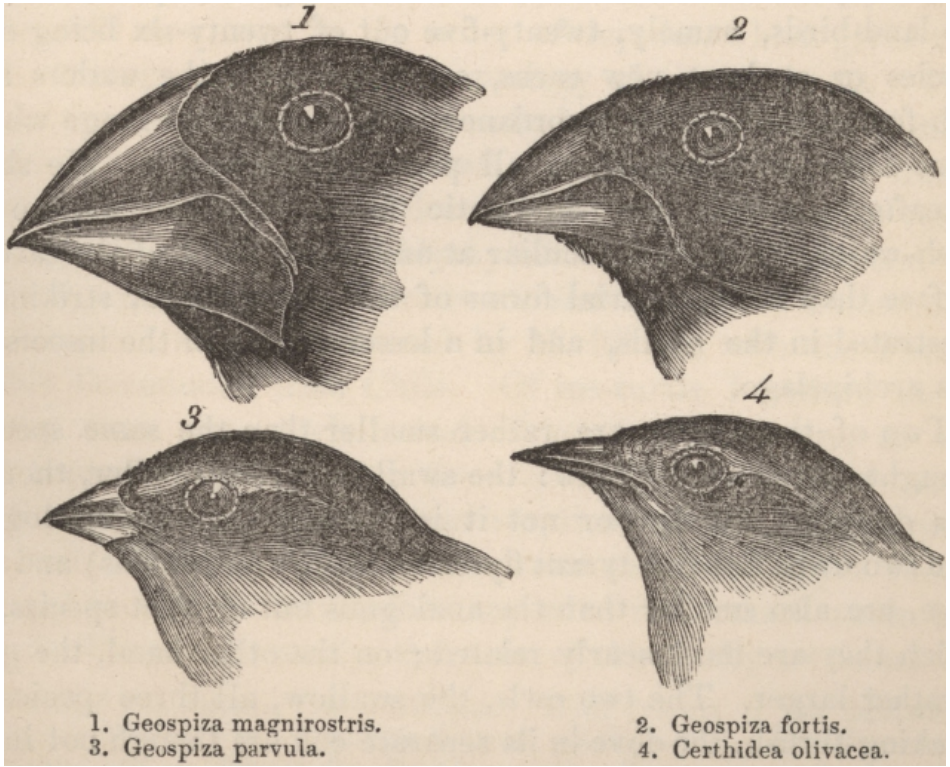
Sprachwandel und Evolution

Vater Unser im Himmel, geheiligt werde Dein Name

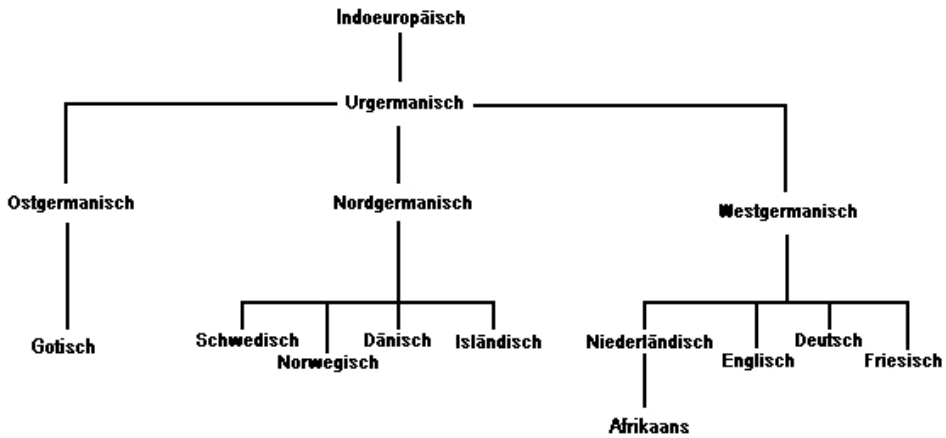
Onze Vader in de Hemel, laat Uw Naam geheiligd worden

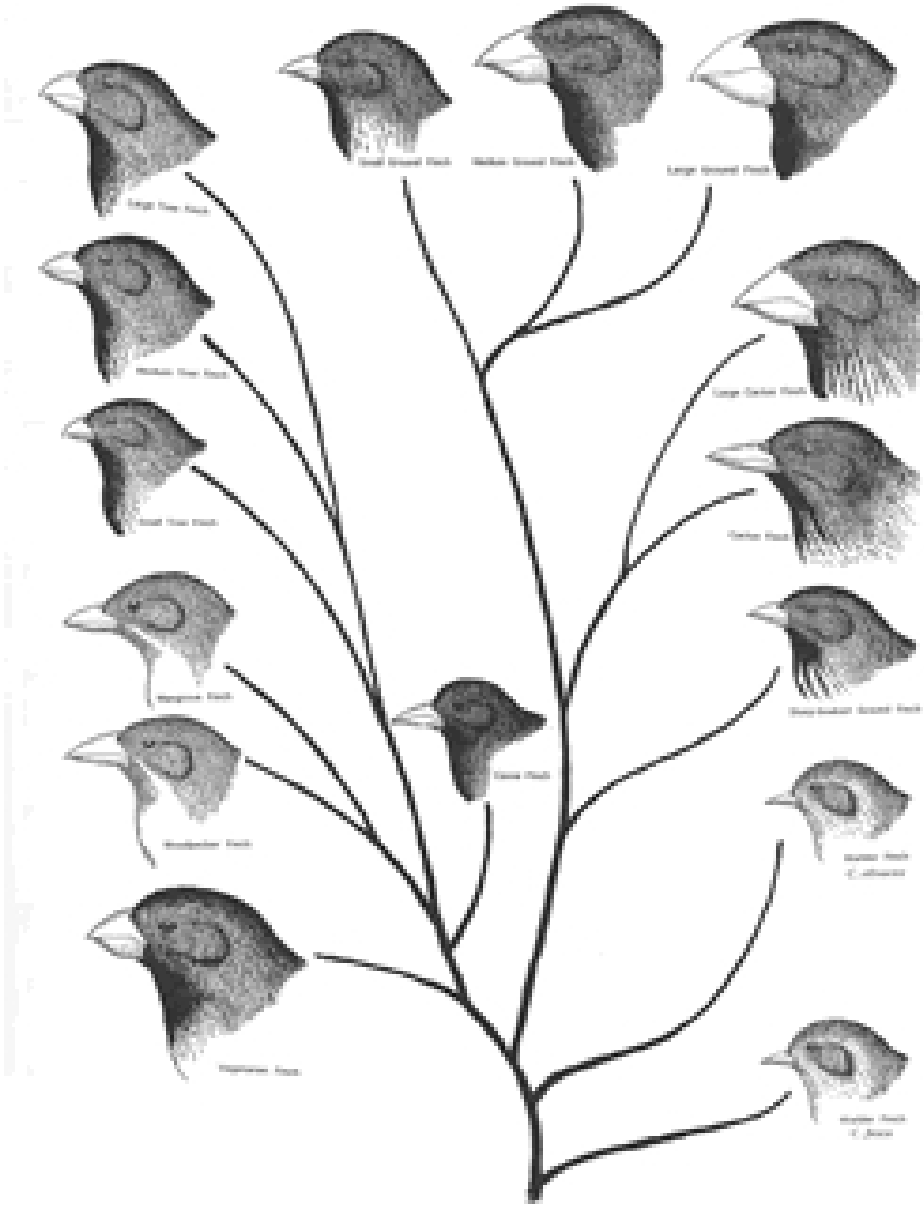
Our Father in heaven, hallowed be your name

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



Sprachwandel und Evolution



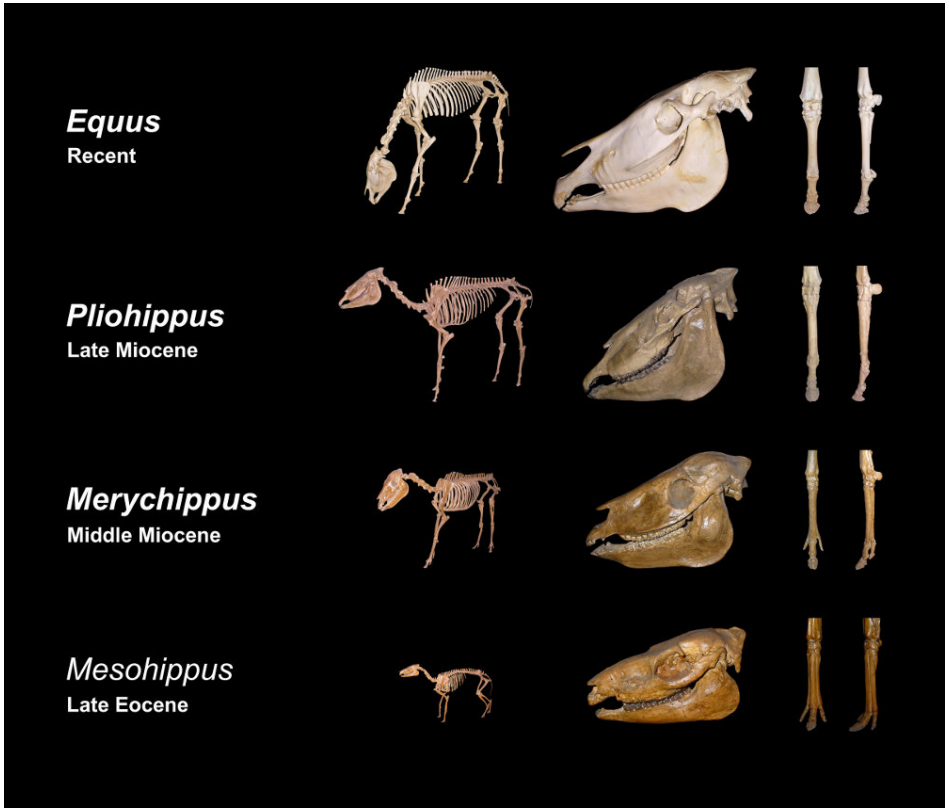


Sprachwandel und Evolution

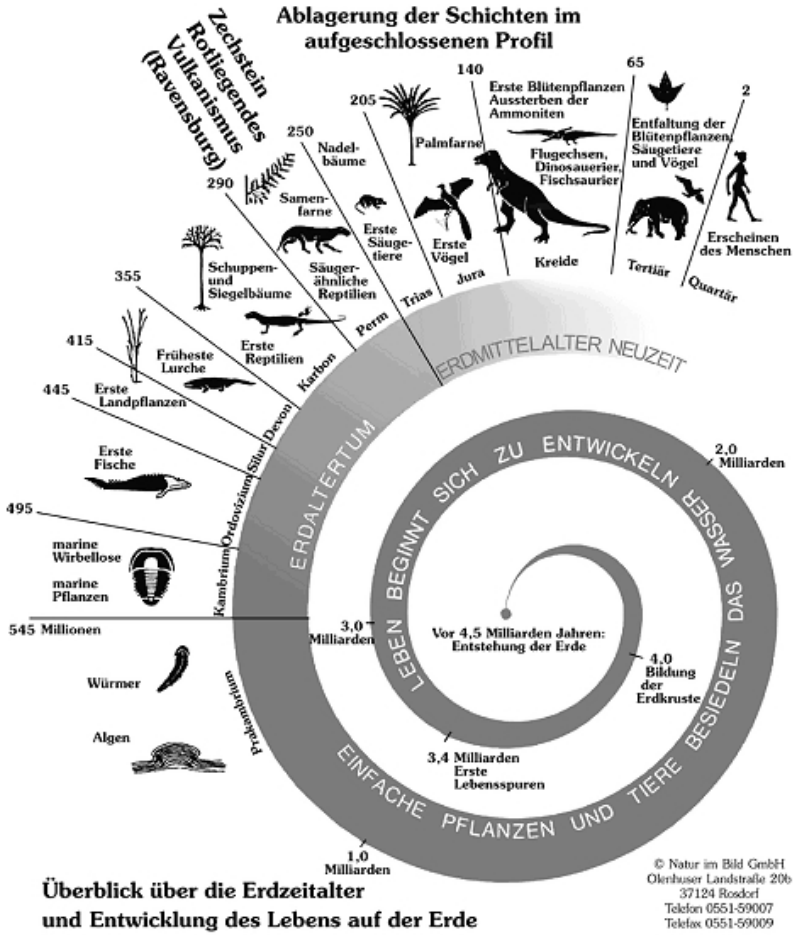
Mittelhochdeutsch:
Got vater unser, dâ du bist in dem himelrîche gewaltic alles des dir ist, geheiliget sô werde dîn nam

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

Gotisch:
Atta unsar þu in himinam, weihnai namo þein



Sprachwandel und Evolution



Höherentwicklung im Sprachwandel

Pidgin- und Kreolsprachen

- eine Indianerin zu einem weißen Verehrer in Pidgin-English:

  You silly. You weak. You baby-hand. No catch horse. No kill buffalo. No good but for sit still—read book.

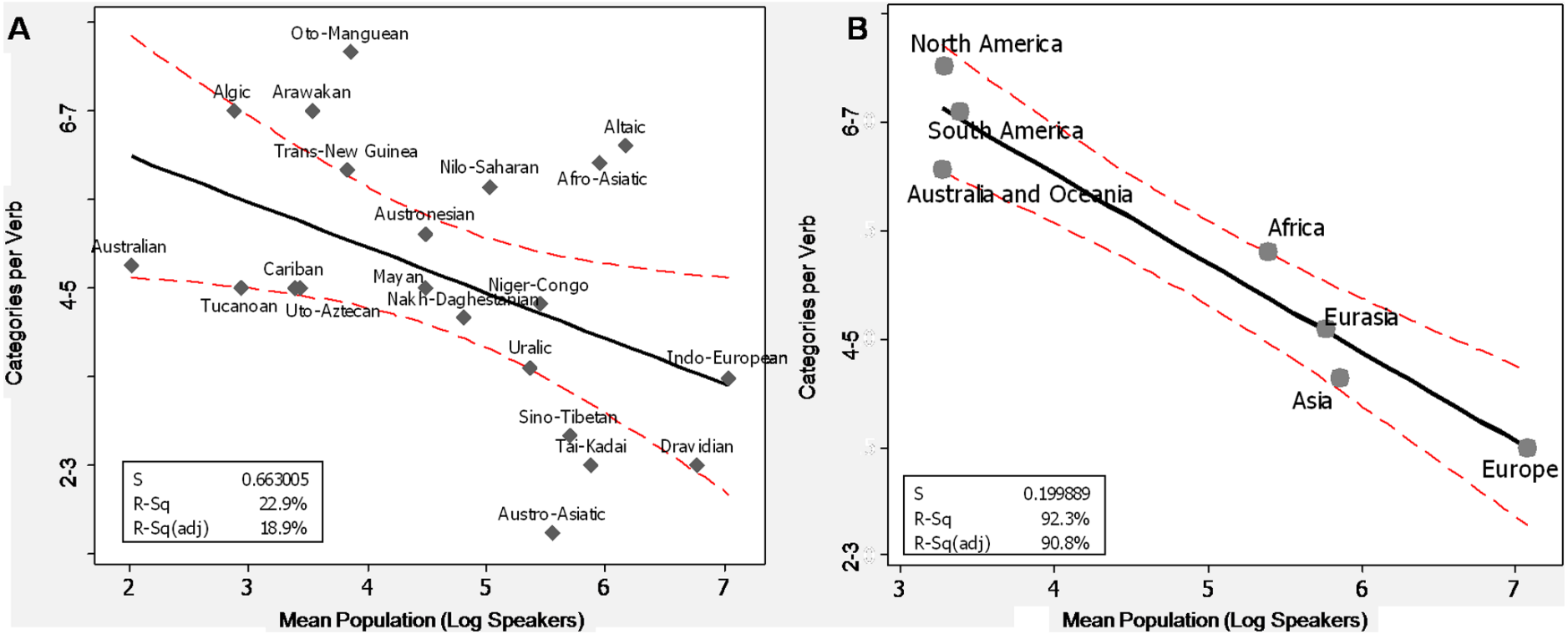
- Satz aus dem Sranan, einer Englisch-basierten Kreolsprache aus Surinam:

  A hondiman datai ben bai wan oso gi en mati.
  'Der Jäger, der ein Haus gekauft hat, gab es seinem Freund.'

(aus John McWhorter, 2003, The Power of Babel)

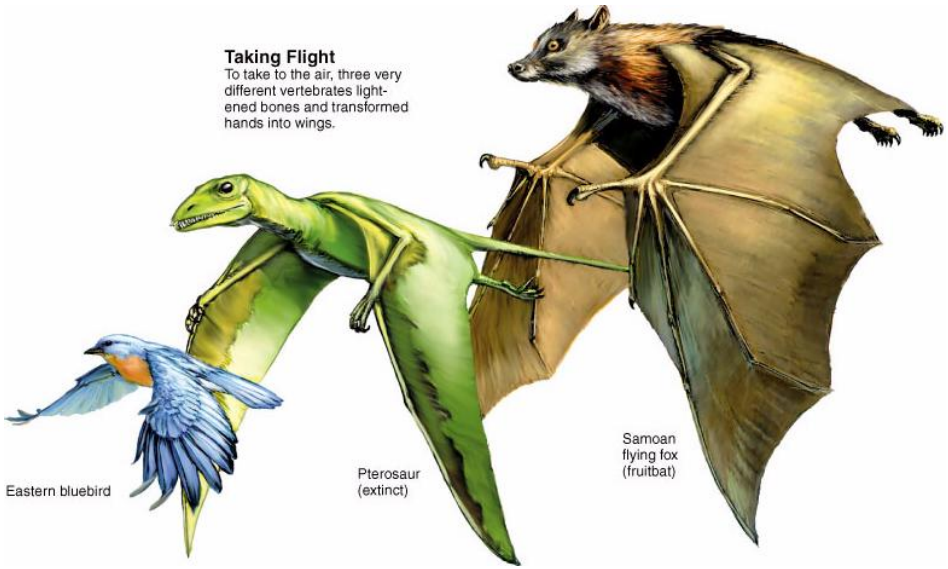
Höherentwicklung im Sprachwandel

Anpassung der Grammatik an soziale Gegebenheiten



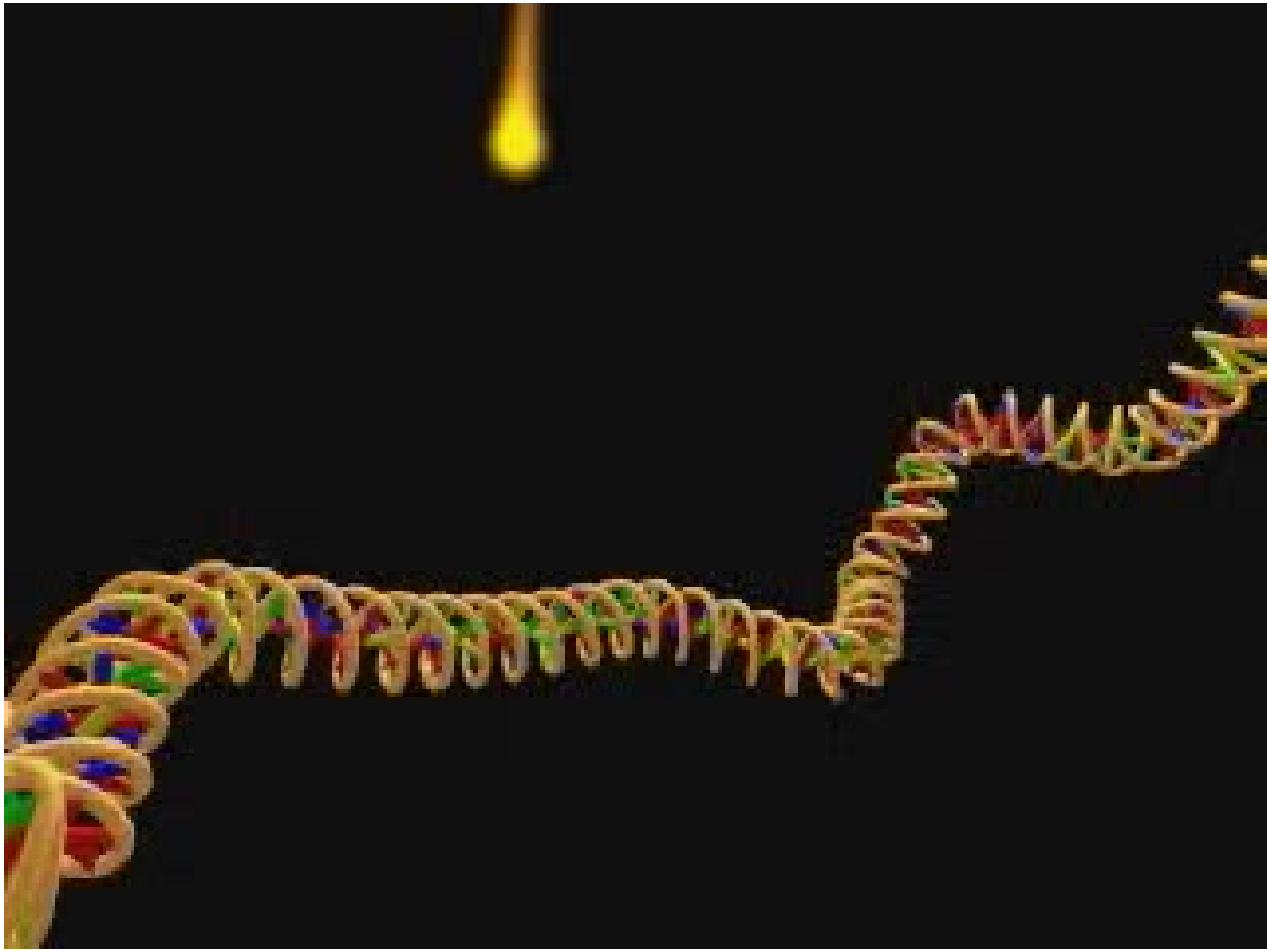
(aus G. Lyupan & R. Dale, 2010, PLoS ONE 5(1))

Konvergente Evolution



- Altenglisch *docga* > Englisch *dog* ('Hund')
- Proto-Paman **gudaga* > Mbabaram *dog* ('Hund')

Evolution via Mutation in der Biologie



Lautgesetze

Erste bzw. Germanische Lautverschiebung (Indoeuropäisch → Germanisch)	Phase	Zweite bzw. Hochdeutsche Lautverschiebung (Germanisch→Althochdeutsch)	Beispiele (Neuhochdeutsch)	Jahrhundert	Dialektgebiete
G: /*b/→/*p/	1	**/*p/→/ff/→/f/**	niederdeutsch: sla**p**en, englisch: slee**p** → schla**f**en; niederdeutsch und englisch: Schi**pp**, shi**p** → Schi**ff** niederdeutsch: scher**p**, englisch: shar**p** → scha**rf**	4/5	oberdeutsch und mitteldeutsch
	2	**/*p/→/pf/**	niederdeutsch: **P**eper, englisch: **p**epper → **Pf**effer; niederdeutsch: **P**lauch, englisch: **p**lough → **Pf**lug; niederdeutsch: scher**p**, englisch: shar**p**, althochdeutsch: scar**ph**, mittelhochdeutsch: schar**pf**	6/7	oberdeutsch
G: /*d/→/*t/	1	**/*t/→/ss/→/s/**	niederdeutsch: da**t**, wa**t**, e**t**en; englisch: tha**t**, wha**t**, ea**t** → da**s**, wa**s**, e**ss**en	4/5	ober- und mitteldeutsch[1]
	2	**/*t/→/ts/**	niederdeutsch: **T**iet, englisch: **t**ide (Gezeiten), schwedisch: **t**id → **Z**eit; niederdeutsch: ver-**t**ellen, englisch: **t**ell → er-**z**ählen; **T**immermann → **Z**immermann	5/6	ober- und mitteldeutsch
G: /*g/→/*k/	1	**/*k/→/xx/→/x/**	niederdeutsch: i**k**, altenglisch: i**c** → i**ch**; niederdeutsch und englisch: ma**k**en, ma**k**e → ma**ch**en; niederdeutsch: au**k** → au**ch**	4/5	ober- und mitteldeutsch[2]
	2	**/*k/→/kx/**	**K**ind → bairisch: **Kch**ind	7/8	südbairisch, hoch- und höchstalemannisch
G: /*b^{h}/→/*b/ V: /*p/→/*b/	3	**/*b/→/p/**	**B**erg, ***b**ist* → *bairisch: **p**erg, **p**ist*	8/9	teilweise bairisch und alemannisch
G: /*d/→/*đ/→/*d/ V: /*t/→/*đ/→/*d/	3	**/*d/→/t/**	niederdeutsch: **D**ag oder **D**ach, englisch: **d**ay → **T**ag; niederfränkisch: *va**d**er* → Va**t**er	8/9	oberdeutsch
G: /*g^{h}/→/*g/ V: /*k/→/*g/	3	**/*g/→/k/**	**G**ott → bairisch: **K**ott	8/9	teilweise bairisch und alemannisch
G: /*t/→/þ/ [ð]	4	**/þ/→/d/** **/ð/→/d/**	englisch: **th**orn, **th**istle, **th**rough, bro**th**er → **D**orn, **D**istel, **d**urch, Bru**d**er	9/10	gesamtes deutsches Dialektkontinuum

Lautgesetze

- Lautgesetze sind spezifisch für eine bestimmte Sprachwandel-Periode
- gelten nahezu universell für alle Instanzen des betroffenen Lautes in der betroffenen Sprache
- im Idealfall gibt es schriftliche Zeugnisse der älteren und der jüngeren Sprachstufe (z.B. Latein/romanischen Sprachen, Althochdeutsch/Mittelhochdeutsch)
- meistens müssen Lautgesetze durch systematischen Vergleich verwandter Sprachen identifiziert werden
- erlaubt partielle Rekonstruktion der gemeinsamen Ursprungssprache

The Indo-European language family

- William Jones 1786:
  „The Sanskrit Language, whatever be its antiquity, is of wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either; yet bearing to both of them a stronger affinity both in the roots of verbs and the forms of grammar, than could possibly have been produced by accident; so strong indeed that no philologer could examine them at all without believing them to have sprung from some common source, which perhaps no longer exists: there is similar reason, so not quite so forcible, for supposing that both the Gothic and the Celtic, though blended with a different idiom, had the same origin with the Sanskrit; and the old Persian might be added to the same family, if this were the place for discussing any question concerning the antiquities of Persia.“

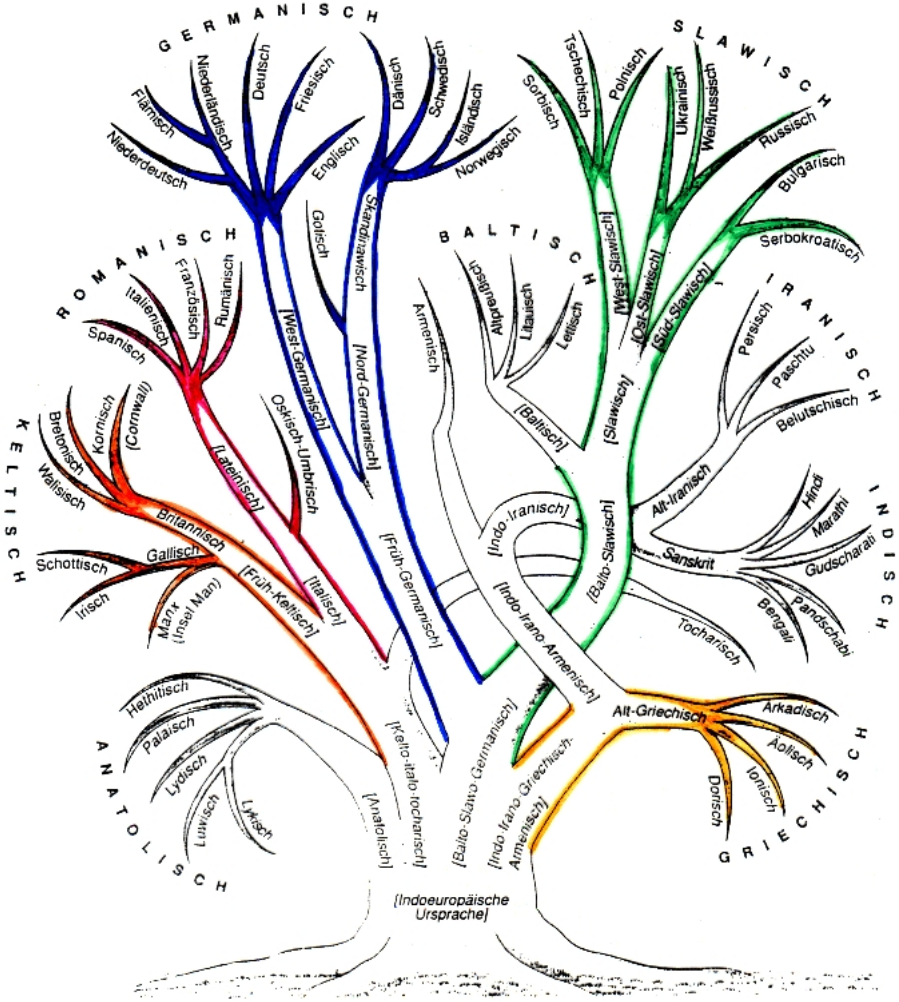
Sprachrekonstruktion durch die komparative Methode

- erste erfolgreiche Anwendung auf *Indo-europäisch* im 19. Jhd.

	Griechisch	Vedisch	Awestisch	Latein	Walisisch	Gotisch	Armenisch	Tocharisch A	A.K.Slawisch	Litauisch	Indogermanisch (rekonstruiert)
1	heîs (< *hens < *sems)	éka	aēuua	ūnus (Altlatein: oinos)	un	ains	mi	sas	inŭ	vienas	*oyno-, oyko-, sem-
2	dúō	dvā́	duua	duō	dau	twai	erkow	wu	dŭva	dù	*duwóh₁
3	treîs	tri	θrāiiō	trēs	tri	þreis	erek`	tre	trije	trỹs	*tréyes
4	téttares	catvā́ras	caθuuārō	quattuor	pedwar	fidwor	čork`	śtwar	četyre	keturì	*kʷetwóres
5	pénte	páñca	panca	quīnque	pump	fimf	hing	päñ	pętĭ	penkì	*pénkʷe
6	héks	ṣát	xšuuaš	sex	chwech	saihs	več	ṣäk	šestĭ	šešì	*swéks
7	heptá	saptá	hapta	septem	saith	sibun	ewt`n	ṣpät	sedmĭ	septynì	*septḿ̥
8	oktṓ	aṣṭā́	ašta	octō	wyth	ahtau	owt`	okät	osmĭ	aštuonì	*oḱtō
9	ennéa	náva	nauua	novem	naw	niun	inn	ñu	devętĭ	devynì	*néwn
10	déka	dáśa	daśa	decem	deg	taihun	tasn	śäk	desętĭ	dẽšimt	*déḱm̥
20	wikati (dorisch)	vimśati	vīsaiti	vigintī	ugeint (Mittelwalisisch)		k`san	wiki			*wīḱmtī
100	hekatón	śatám	satəm	centum	cant	hund		känt	sũto	šimtas	*ḱm̥tóm

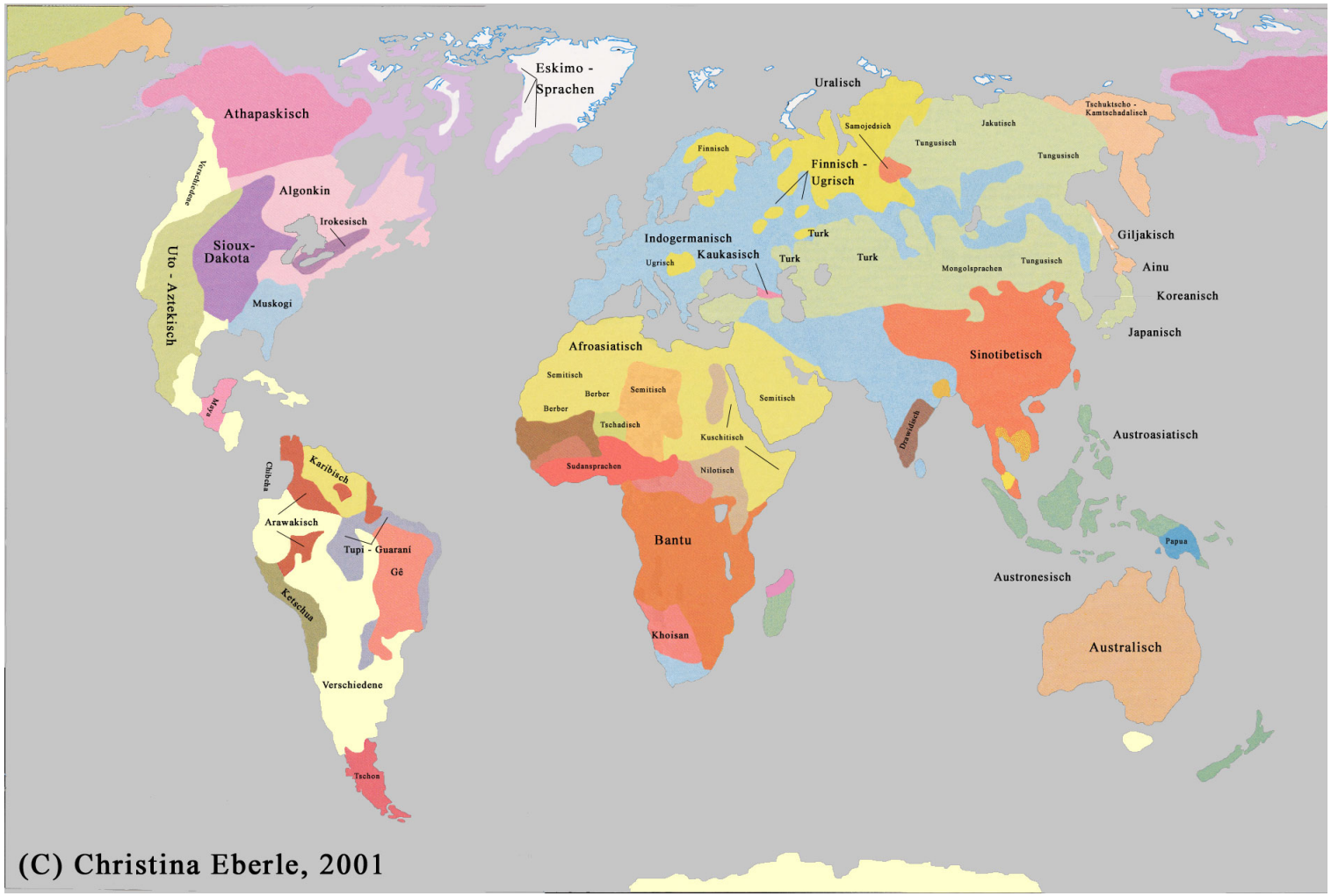
Language trees

- komparative Methode ergibt Abstammungsbaum einer Sprachfamilie



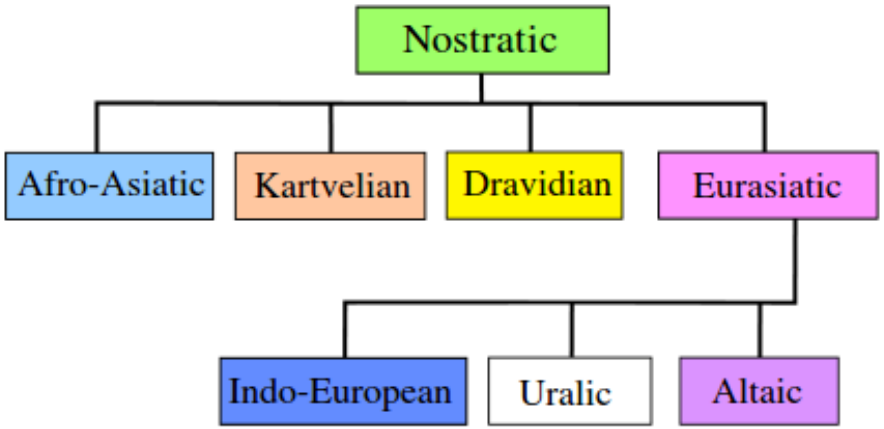
Grenzen der komparativen Methode

- Zeittiefe beschränkt auf 2 000 bis 8 000 Jahre



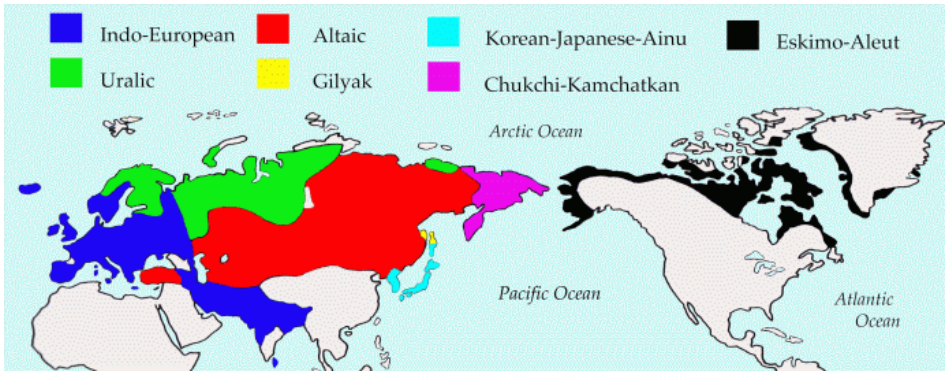
Tiefe Sprachverwandtschaften

- Vielzahl von Vorschlägen für Meta-Familien
  - Nostratisch:
    - erstmals von Pedersen (1903) vorgeschlagen
    - ursprünglicher Vorschlage: Indo-europäisch, Finno-ugrisch, Samoyedisch, Turk-Sprachen, Mongolisch, Manchu, Yukaghir, Eskimo, Semitisch und Hamitisch
    - weiterentwickelt durch „Moskauer Schule“ in den 1960ern
    - Versuch der Rekonstruktion von Wortschatz



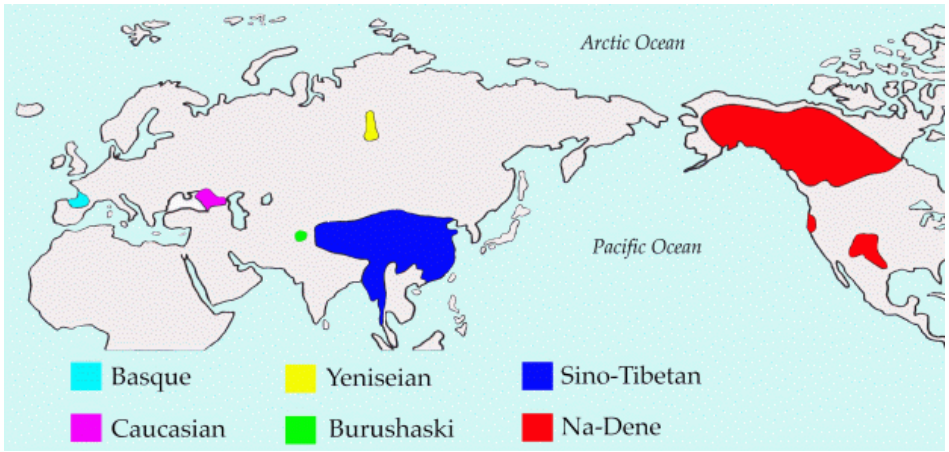
Tiefe Sprachverwandtschaften

- Vielzahl von Vorschlägen für Meta-Familien
  - Eurasiatisch
    - vorgeschlagen von Greenberg (2000)
    - umfasst Indo-europäisch, Uralisch-Yukaghirisch, Altaisch, Tschuktscho-Kamtschadalisch, Eskimo-Aleutisch, Koreanisch-Japanisch-Ainu, Gilyak, Etruskisch
    - diverse Argumente, v.a. Morphologie und Phonologie



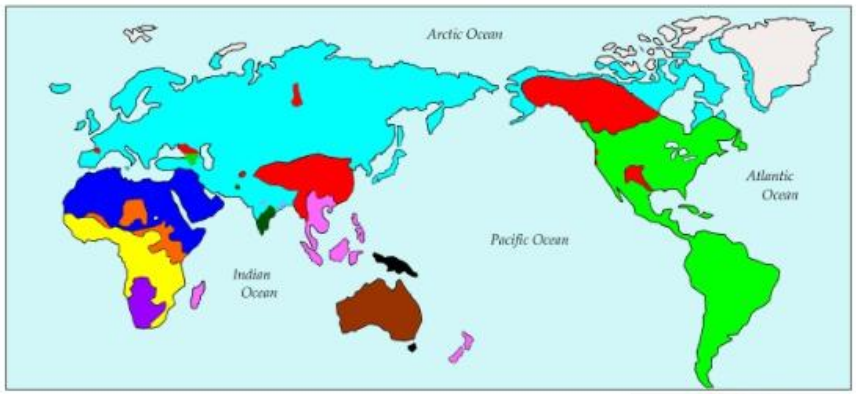
Tiefe Sprachverwandtschaften

- Vielzahl von Vorschlägen für Meta-Familien
  - Dene-Kaukasisch
    - umfasst Ne-Dene, Kaukasisch, Sino-Tibetisch, Jenniseiisch, Burushaski, manchmal auch Baskisch



Tiefe Sprachverwandtschaften

- Vielzahl von Vorschlägen für Meta-Familien
  - Amerindisch
    - vorgeschlagen von Greenberg (1987)
    - umfasst alle Indianersprachen außer Na-Dene



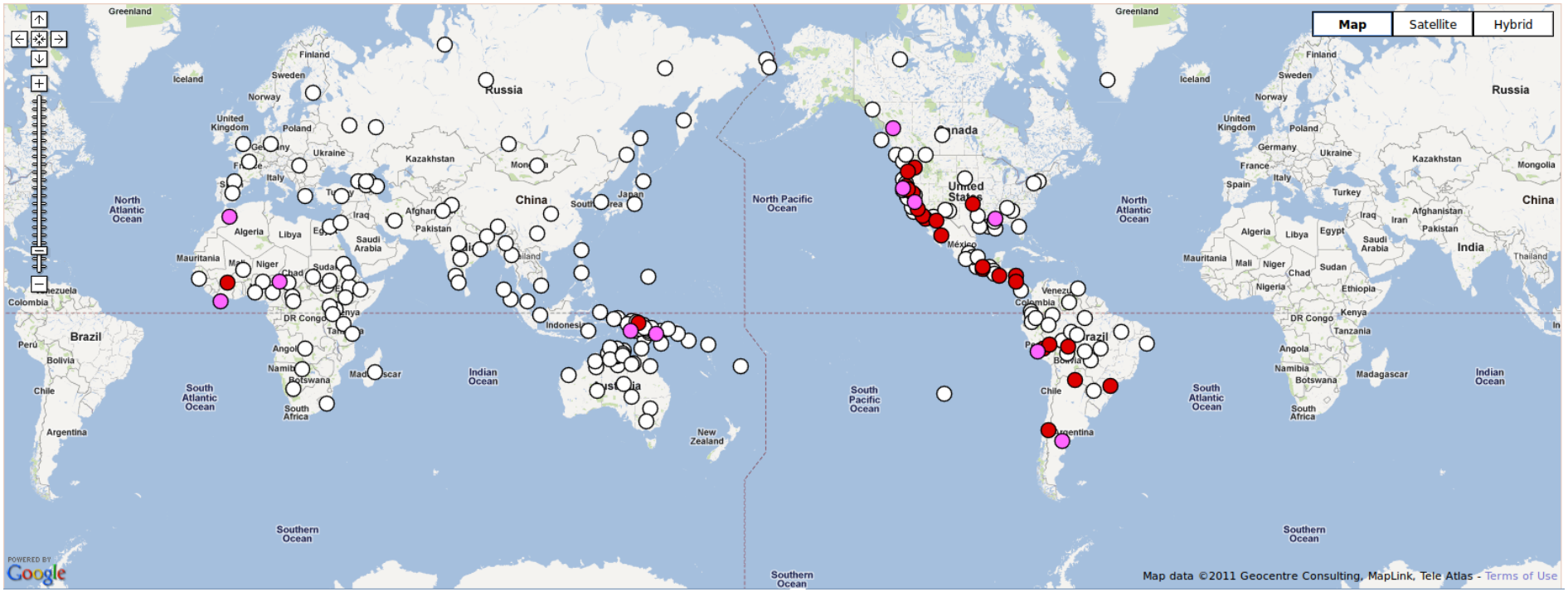
Language Families of the World (after Greenberg)

Tiefe Sprachverwandtschaften

- Merritt Ruhlen, ein Schüler von Greenberg, behauptet sogar, „Proto-World“ z.T. rekonstruieren zu können, z.B. das Wort *akwa* für Wasser (das sich faszinierenderweise von Adam und Eva über Cicero bis zu Umberto Eco im Indoeuropäisch/Italisch/Lateinisch/Italienischen Zweig nicht verändert hat)
- derartige Vorschläge basieren häufig auf geographischen Häufungen einzelner Merkmale, wie z.B. Pronominalformen

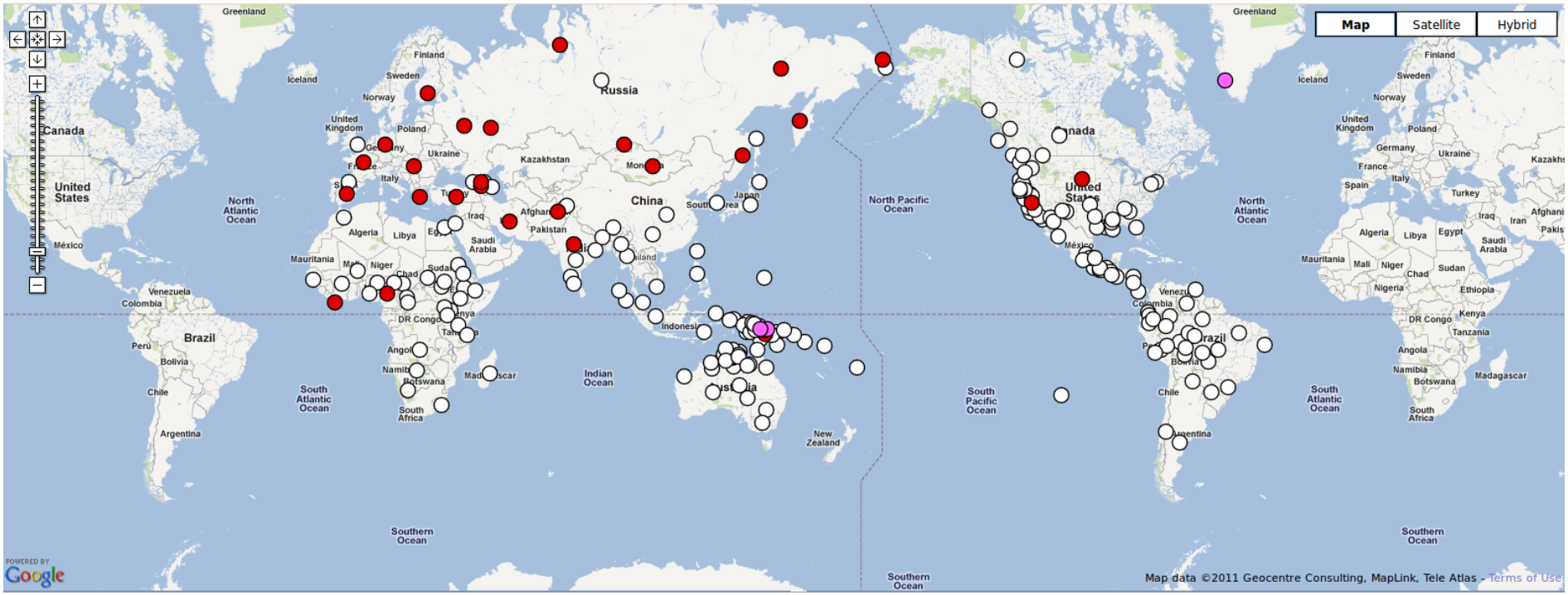
Tiefe Sprachverwandtschaften

- N/M-Pronomina



Tiefe Sprachverwandtschaften

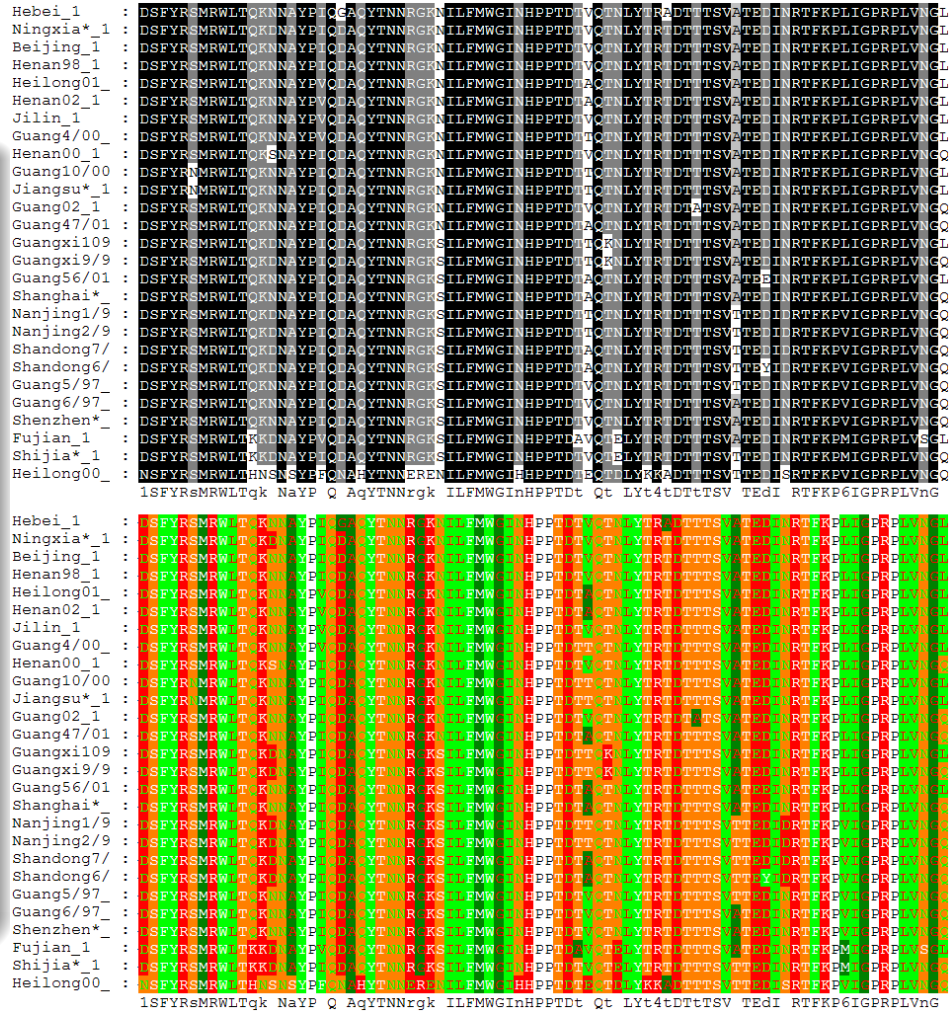
- M/T-Pronomina



Phylogenetische Rekonstruktion in der Bioinformatik

Sequenzalinierung

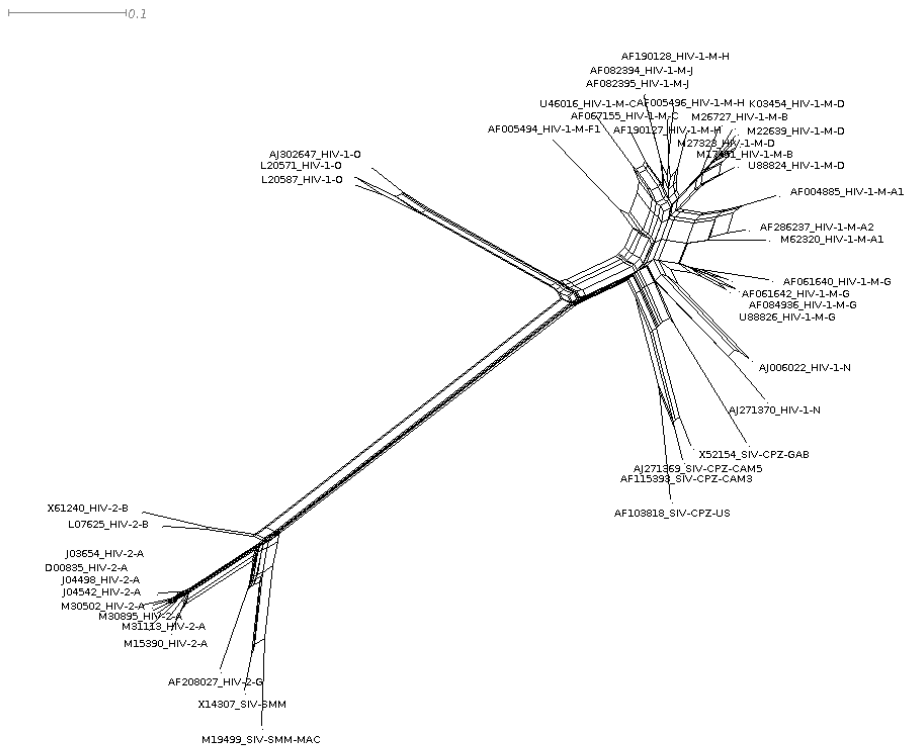
- Algorithmus findet optimale Alinierung zwischen Sequenzen
- Anzahl der Mutationen wird somit abgeschätzt
- ergibt Abschätzung des evolutionären Abstands zwischen den entsprechenden Organismen



Phylogenetische Rekonstruktion in der Bioinformatik

Phylogenetische Bäume

- statistische Verfahren zur Rekonstruktion des wahrscheinlichsten Stammbaums
- häufig konfligierende Information wegen:
  - konvergenter Evolution
  - Rück-Mutation
  - lateraler Gen-Transfer
- Darstellung alternativer Rekonstruktionen in Netzwerk-Strukturen



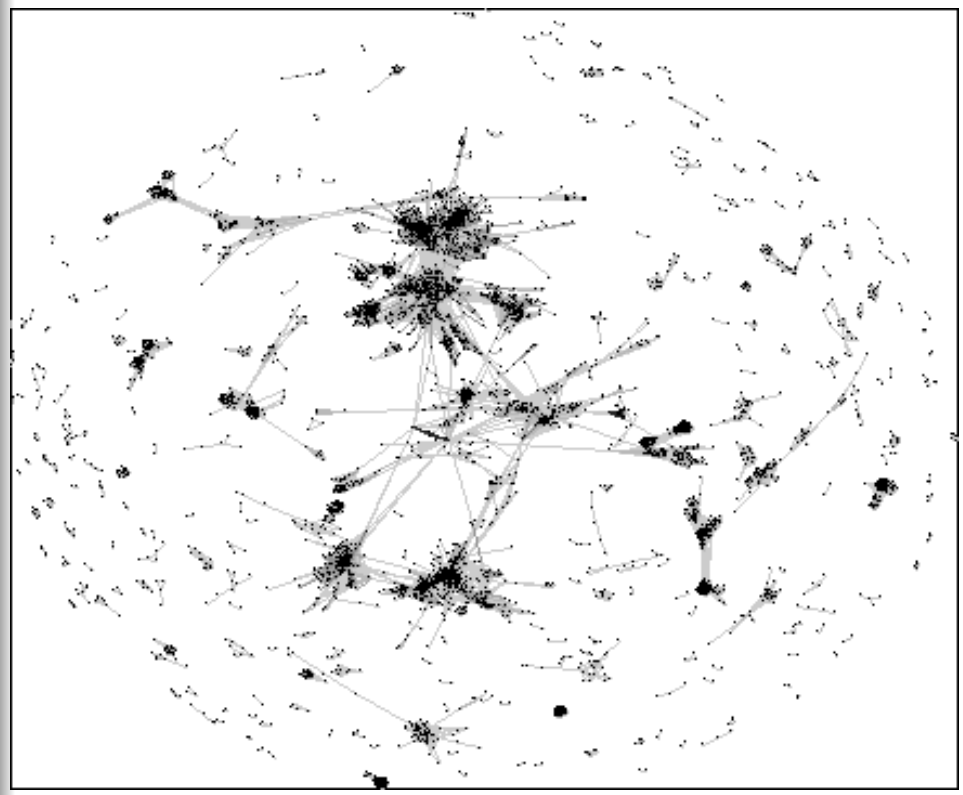
SplitsTree Software, Huson & Bryant, MatNat-Fakultät

Phylogenetische Rekonstruktion in der Bioinformatik

Alternative: Cluster-Karten

- Organisation aller Datenpunkte (=Molekularsequenzen) in 2- oder 3-dimensionalen Raum
- größere Ähnlichkeit entspricht (simulierter) physikalischer Anziehungskraft und umgekehrt
- Algorithmus findet Energie-Minimum

$\Rightarrow$ verwandte Sequenzen bilden Cluster



Software: Frickey & Lupas, MPI für Entwicklungsbiologie

Die Daten des Automated Similarity Judgment Project

- Projekt am MPI EVA in Leipzig um Sören Wichmann
- erfasst inzwischen über 5 000 Sprachen
- für jede Sprache Grundwortschatz von 40 Wörtern in (vereinfachter) phonetischer Umschrift
- frei elektronisch verfügbar

verwendete Konzepte: *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*

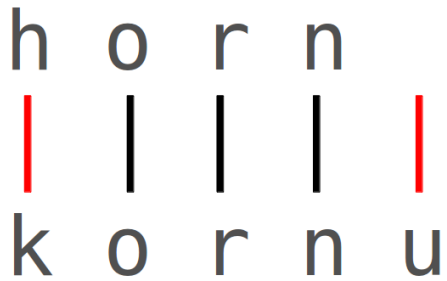
Automated Similarity Judgment Project

Konzept	Deutsch	Englisch
I	iX	Ei
you	du	yu
we	vir	wi
one	ains	8is
two	cvai	8Et
person	mEnS	pers3n
fish	fiS	fiS
dog	hunt	dag
louse	laus	laus
tree	baum	tri
leaf	blat	lif
skin	haut	skin
blood	blut	bl3d
bone	knoX3n	bon
horn	horn	horn
ear	XXX	ir
eye	aug3	Ei

Konzept	Deutsch	Englisch
nose	naz3	nos
tooth	ch~an	tu8
tongue	ch~uN3	t3N
knee	kni	ni
hand	hant	hEnd
breast	brust	brest
liver	leb3r	liv3r
drink	triNk3n	drink
see	ze3n	si
hear	her3n	hir
die	Sterb3n	dEi
come	kh~om3n	k3m
sun	zon3	s3n
star	StErn	star
water	vas3r	wat3r
stone	Stain	ston
fire	foia	fEir

Einfache Sequenz-Alinierung

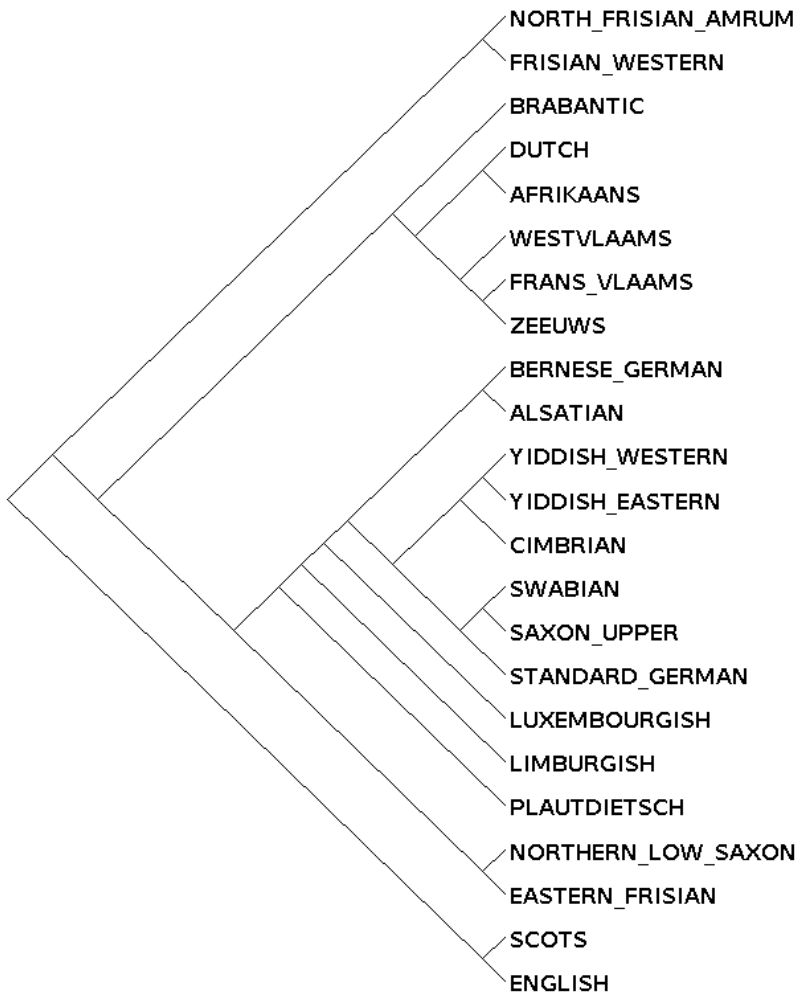
- z.B. Deutsch $\leftrightarrow$ Latein



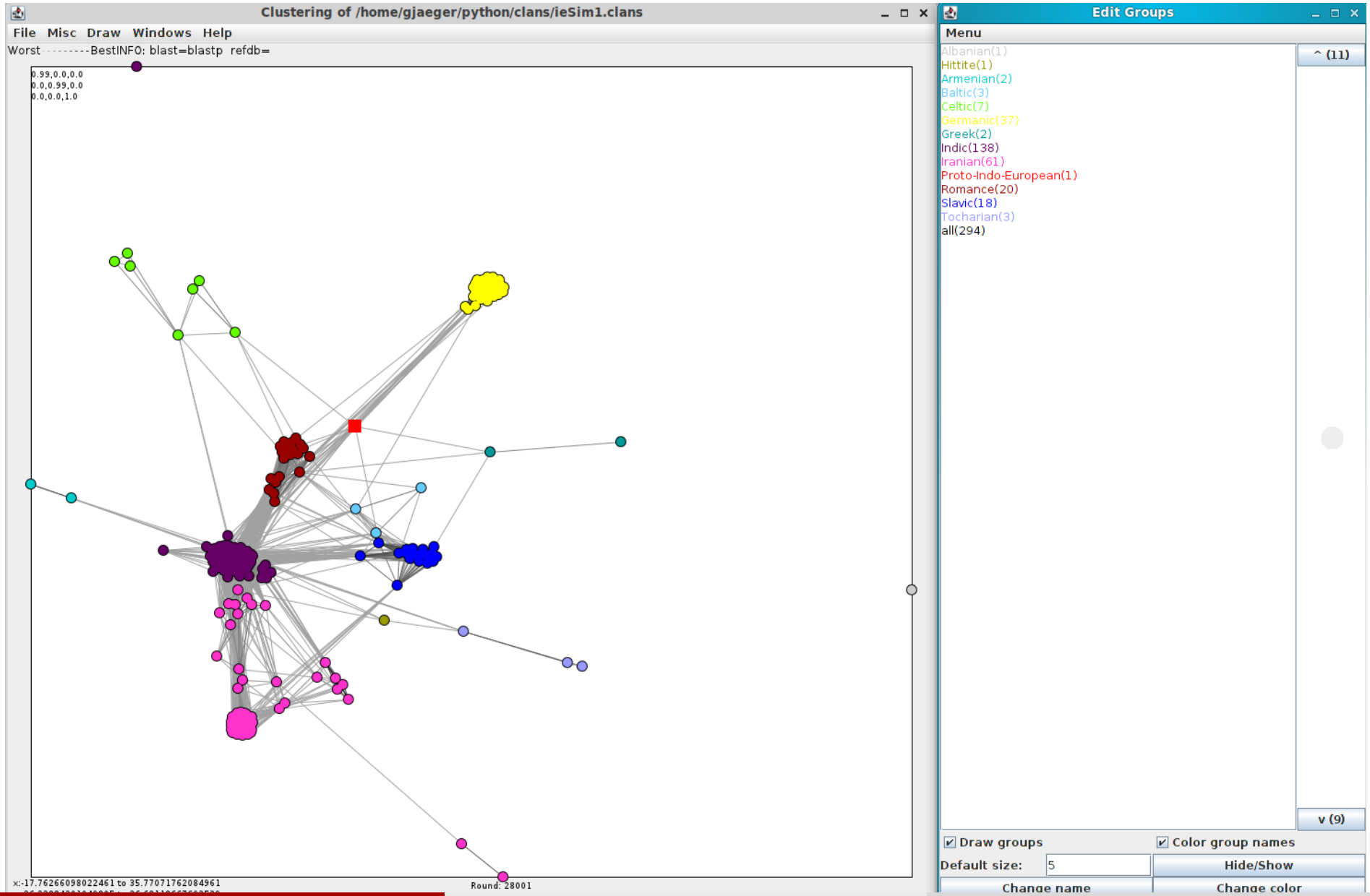
- Anzahl der Unterschiede: 2
- durchschnittliche Abweichung = $^2/_5$
- Ähnlichkeit zwischen zwei Sprachen wird durch durchschnittliche Abweichung der jeweiligen Übersetzungspaare geschätzt

Einfache Sequenz-Alinierung

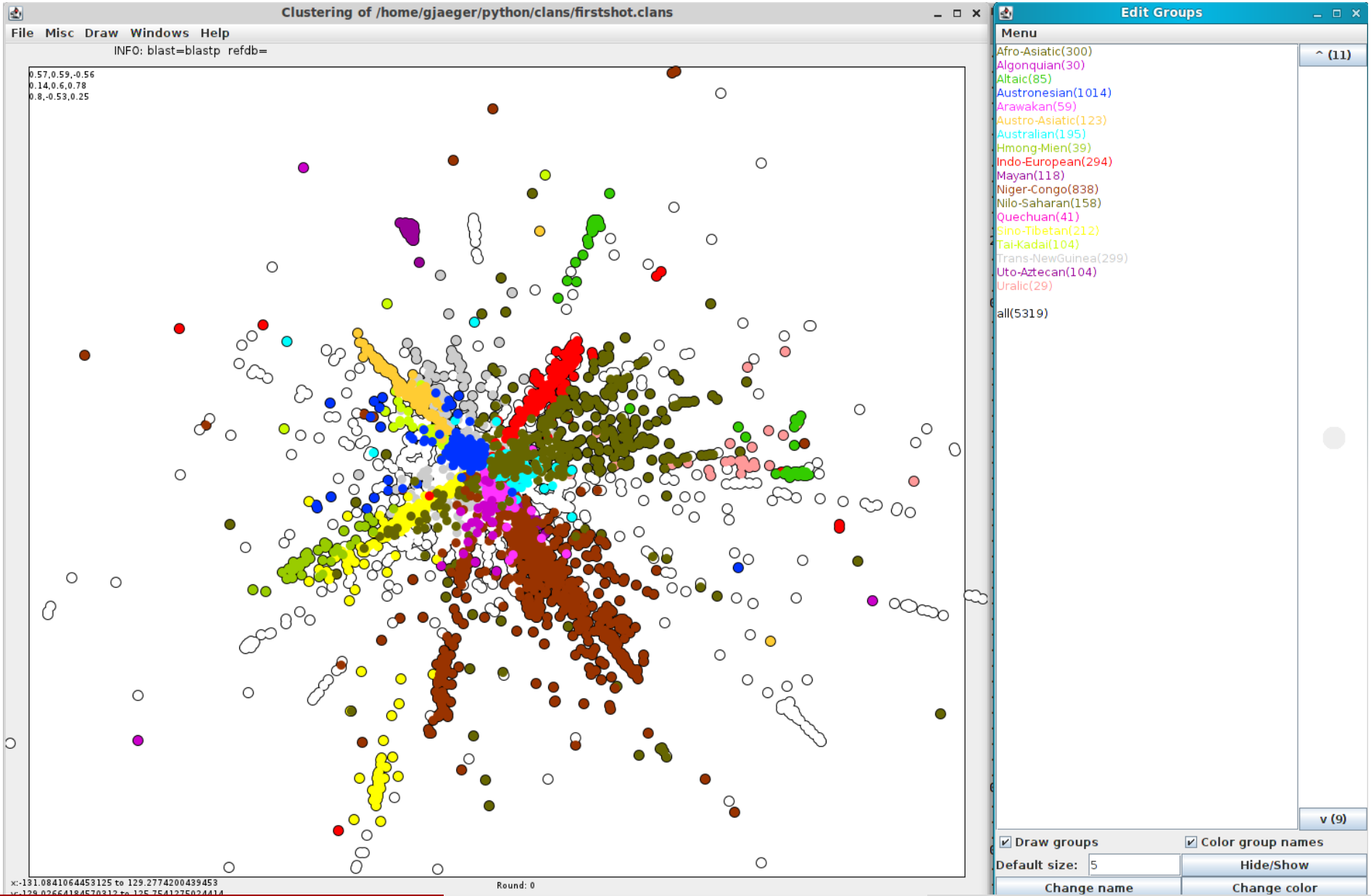
- für die westgermanischen Sprachen/Dialekte ergibt sich die Phylogenie



Einfache Sequenz-Alinierung

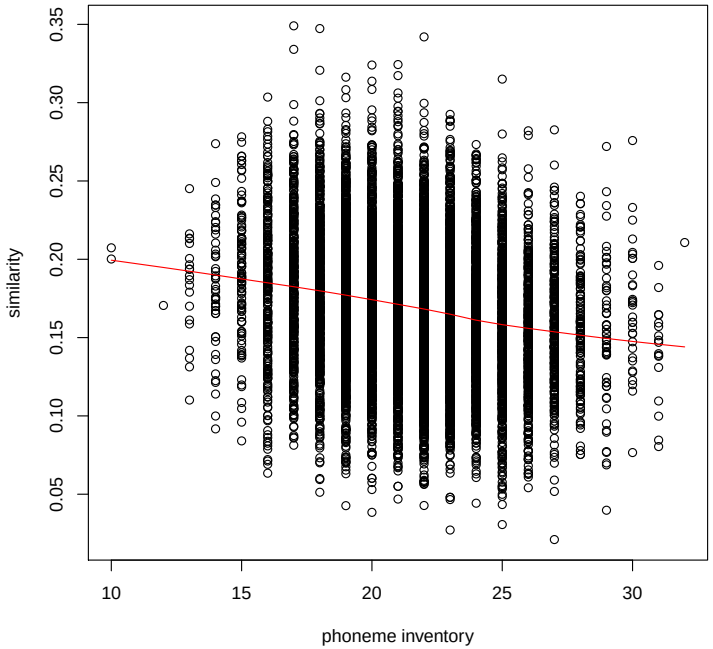


Einfache Sequenz-Alinierung



Einfache Sequenz-Alinierung

- Störeffekt: bei Sprachen mit kleineren Lautinventaren ergeben sich mehr Zufallsähnlichkeiten also bei Sprachen mit vielen verschiedenen Lauten
- Daher erscheinen Sprachen mit wenigen Lauten einander ähnlicher, als sie es tatsächlich sind.



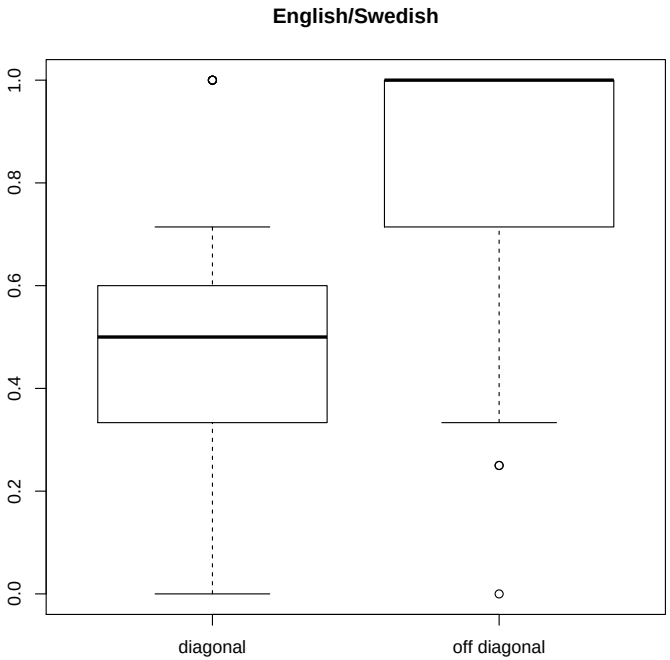
Kalibrierte Alinierung

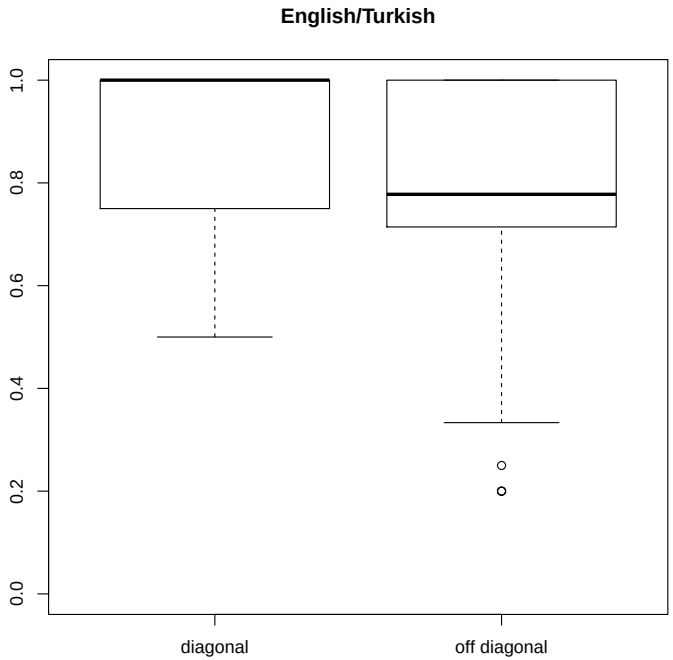
Englisch / Swedisch

	Ei	yu	wi	w3n	tu	fiS	...
yog	1	$^2/_3$	1	1	1	1	
du	1	$^1/_2$	1	1	$^1/_2$	1	
vi	$^1/_2$	1	$^1/_2$	1	1	$^2/_3$	
et	1	1	1	1	1	1	
tvo	1	1	1	1	$^2/_3$	1	
fisk	$^3/_4$	1	$^3/_4$	1	1	$^1/_2$	
⋮							

- je näher zwei Sprachen verwandt sind, umso mehr unterscheiden sich die Werte auf der Diagonale von den restlichen Werten

Kalibrierte Alinierung





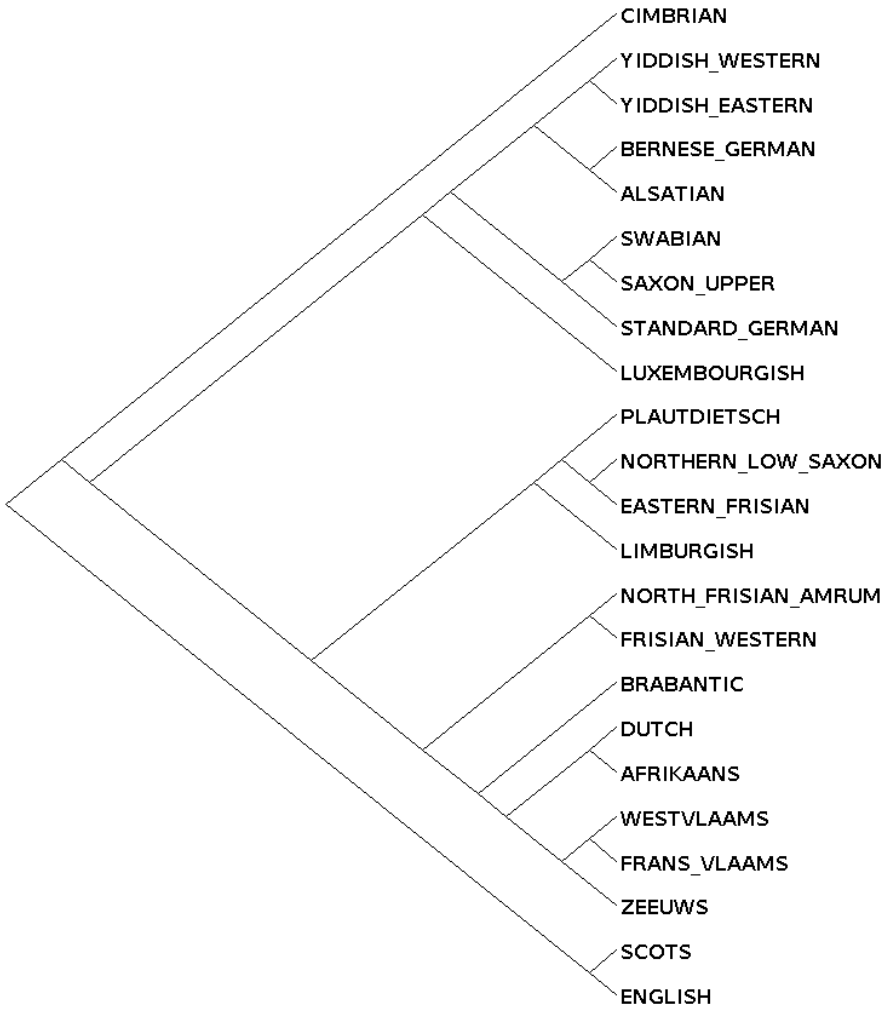
Kalibrierte Alinierung

- Ähnlichkeit des Standard-Deutschen zu:

Schwäbisch	26,13
Zimbrisch	20,28
Niederländisch	23,75
Englisch	17,45
Ur-Indoeuropäisch	10,26
Latein	9,23
Spanisch	8,95
Hindi	8,70
Russisch	8,36
Türkisch	6,33
Ungarisch	6,84

Kalibrierte Alinierung

- für die westgermanischen Sprachen/Dialekte ergibt sich die Phylogenie

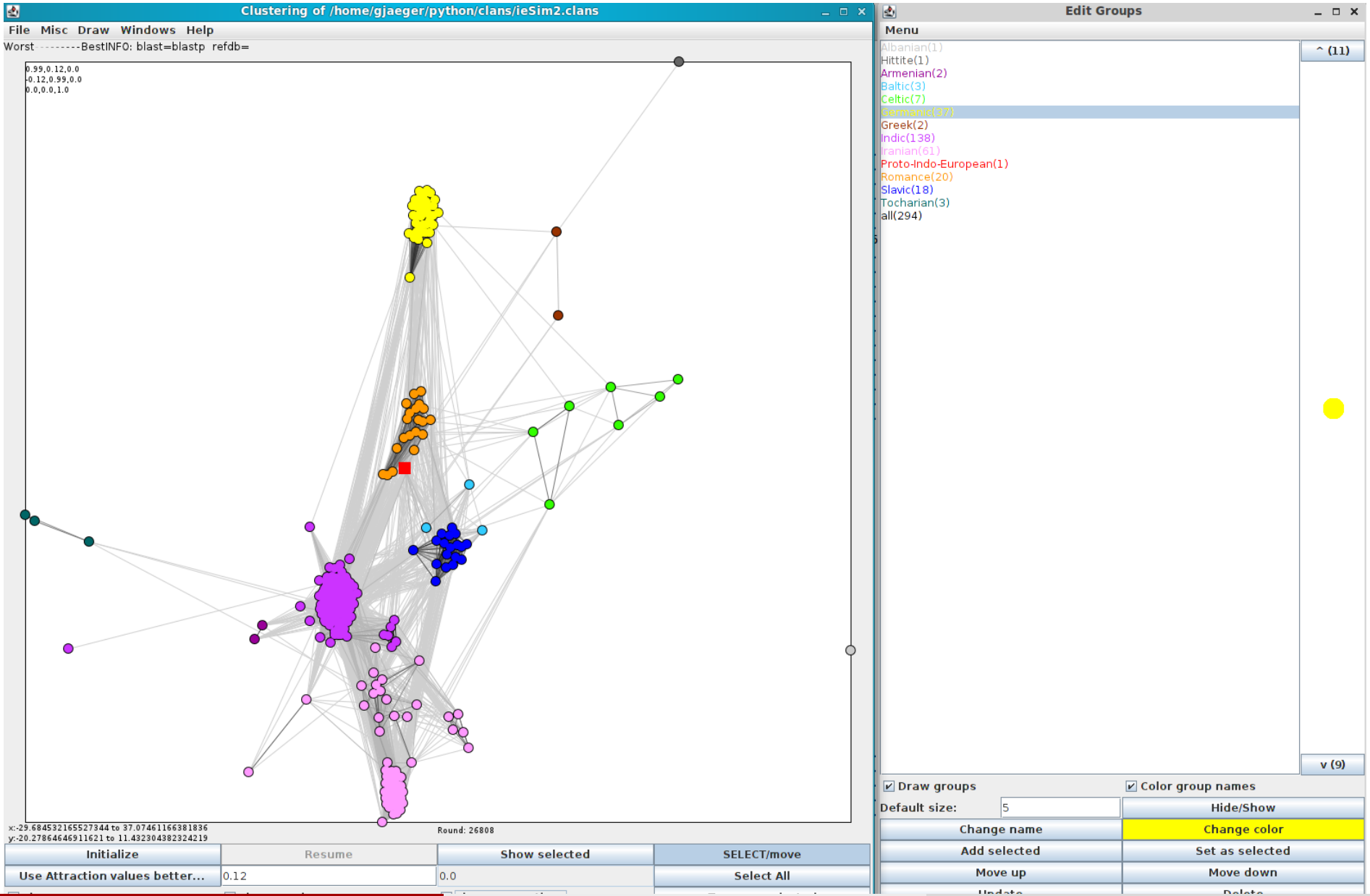


Problemfall: Zimbrisch

Konzept	Deutsch	Zimbrisch
I	iX	ix
you	du	du
we	vir	bar
one	ains	XXX
two	cvai	sben
person	mEnS	menEs
fish	fiS	XXX
dog	hunt	hunt
louse	laus	laus
tree	baum	pom
leaf	blat	placa
skin	haut	XXX
blood	blut	plut
bone	knoX3n	poan
horn	horn	horn
ear	XXX	oar
eye	aug3	ogh~E

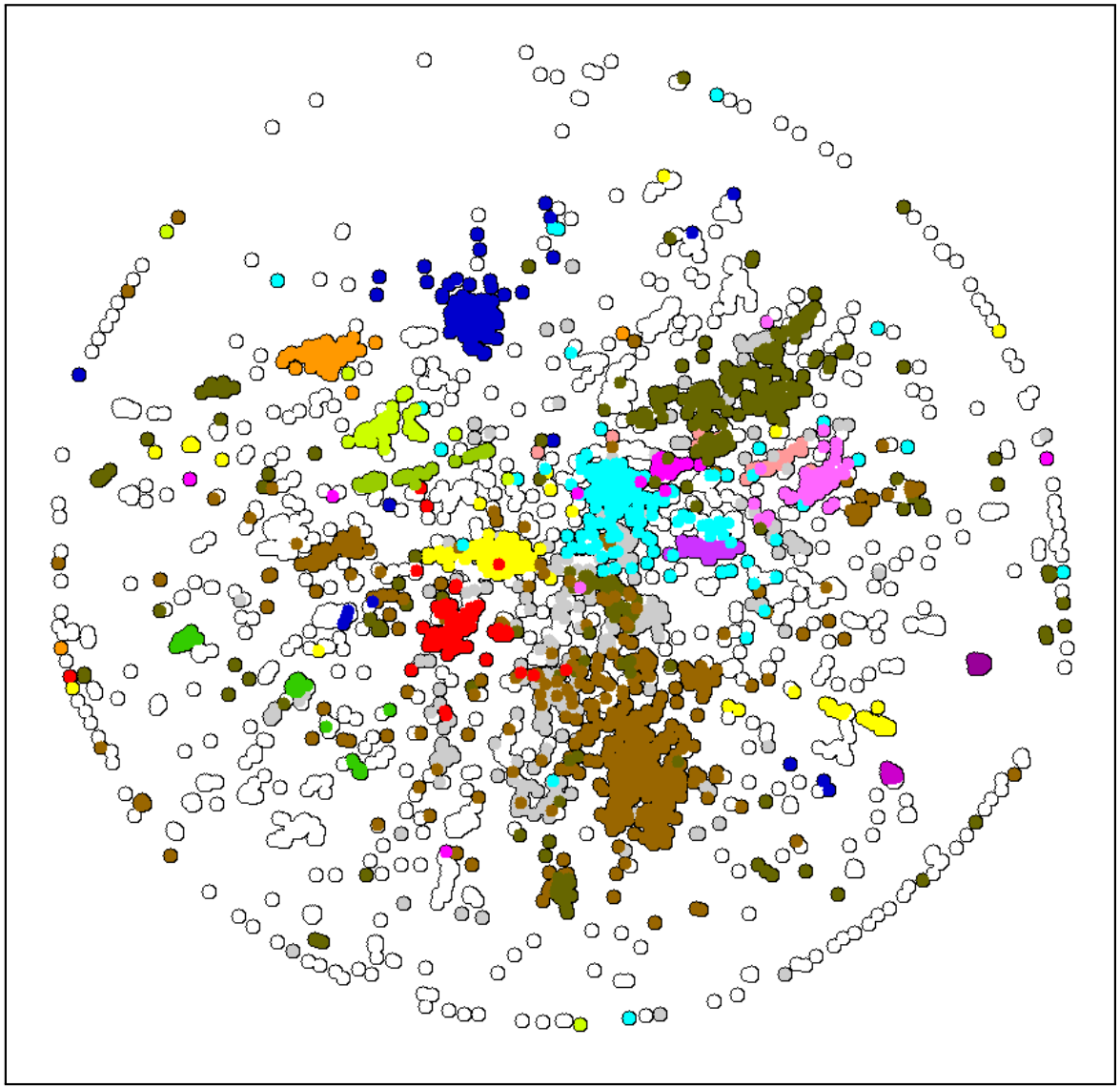
Konzept	Deutsch	Zimbrisch
nose	naz3	naza
tooth	ch~an	XXX
tongue	ch~uN3	suNa
knee	kni	XXX
hand	hant	hant
breast	brust	prust
liver	leb3r	lEbara
drink	triNk3n	trinkh~
see	ze3n	zeg
hear	her3n	hor
die	Sterb3n	sterb
come	kh~om3n	kh~Em
sun	zon3	zuna
star	StErn	stErna
water	vas3r	basar
stone	Stain	stoan
fire	foia	boar

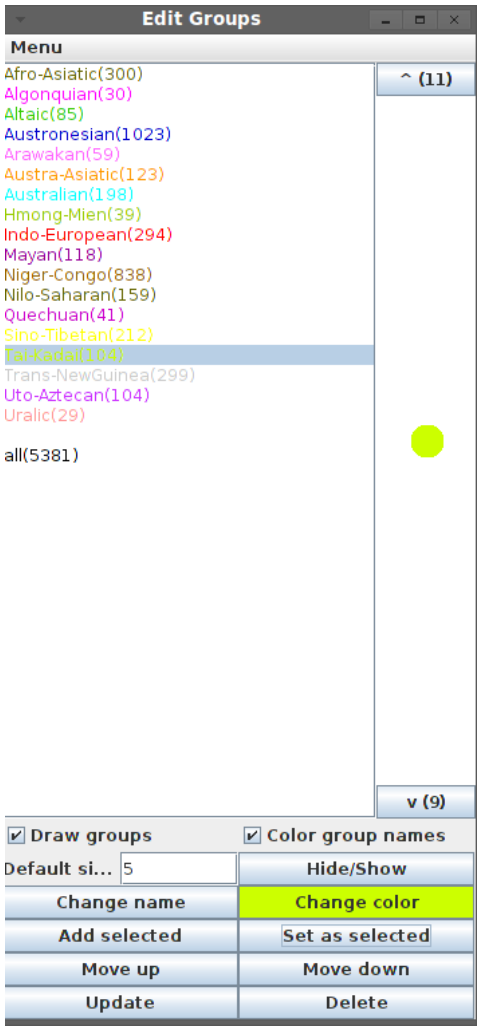
Kalibrierte Alinierung



Kalibrierte Alinierung

Alle 5 000 Sprachen:



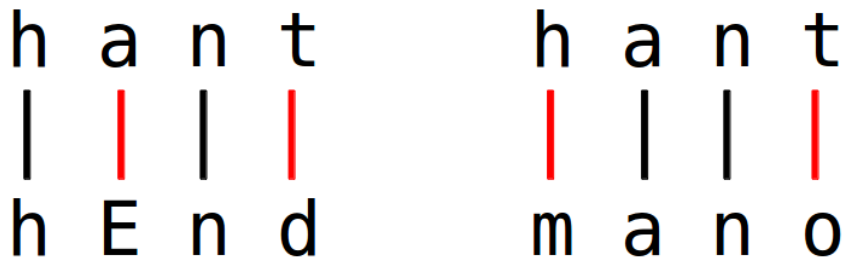


Kalibrierte Alinierung

- etablierte Sprachfamilien bilden stabile Cluster
- keine darüber hinausgehenden sichtbaren Muster

Kalibrierte Alinierung

- Methode ist relativ grobkörnig



- Ähnlichkeit ist in beiden Fällen 50%
- Korrespondenz $a \sim E$, $t \sim d$ sind nach linguistischen Kriterien viel natürlicher als $h \sim m$ or $t \sim o$
- Deutsch/Englisch und Deutsch/Spanisch erscheinen hier äquidistant, obwohl die Ähnlichkeit zwischen Deutsch und Englisch intuitiv viel größer ist

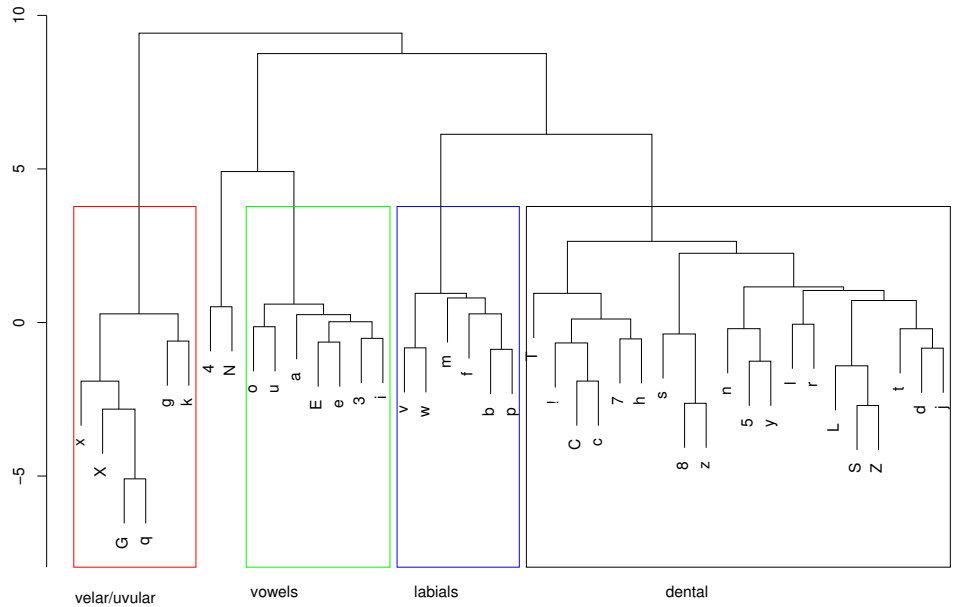
Gewichtete Alinierung

Needleman-Wunsch-Algorithmus

- Analogie zur Bioinformatik: Mutationen zwischen verschiedenen Aminosäuren-Paaren sind unterschiedlich wahrscheinlich
- Algorithmus sucht *wahrscheinlichste* Übereinstimmungen zwischen Sequenzen
- $a \sim E$, $d \sim t$ sind im Sprachwandel wahrscheinlicher als $t \sim o$

Gewichtete Alinierung

- automatisch bestimmte Gewichte:
  - $d \sim t$: $0,38$
  - $a \sim E$: $-0,23$
  - $h \sim m$: $-3,65$
  - $t \sim o$: $-8,89$
  - *hant* $\sim$ *hEnd*: 5.77
  - *hant* $\sim$ *mano*: -7.63



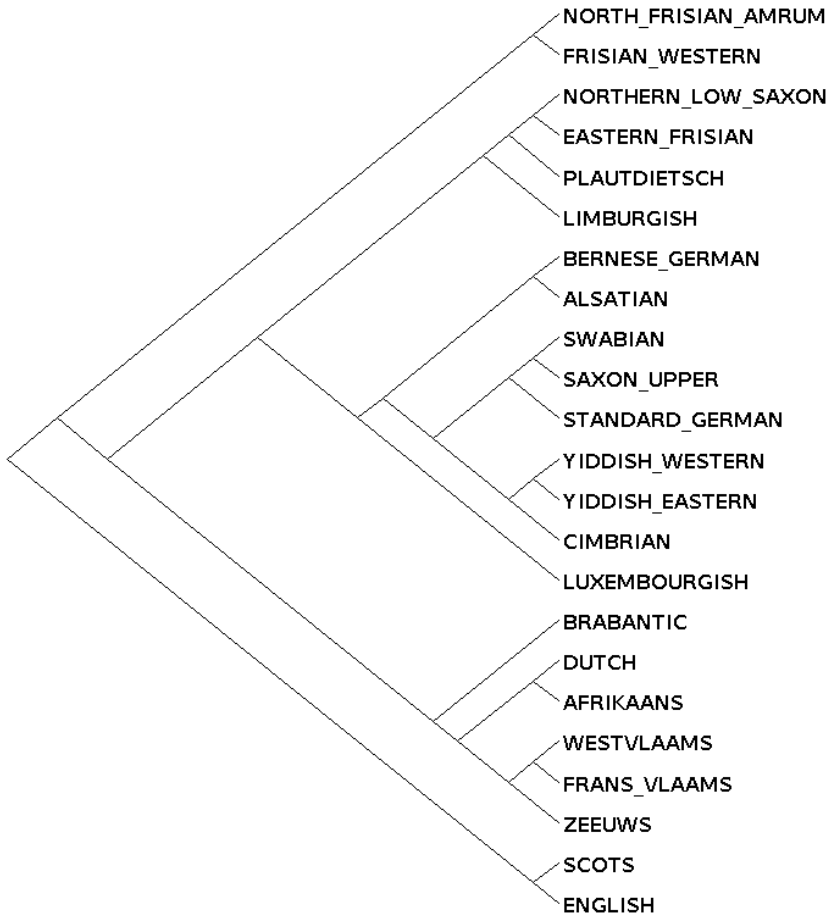
Gewichtete Alinierung

- Ähnlichkeit des Standard-Deutschen zu:

	ungewichtet	gewichtet
Schwäbisch	26,13	35,44
Zimbrisch	20,28	31,86
Niederländisch	23,75	29,76
Englisch	17.45	22,14
Ur-Indoeuropäisch	10,26	15,86
Latein	9,23	12,54
Spanisch	8,95	9,48
Hindi	8,70	12,35
Russisch	8,36	11,89
Türkisch	6,33	5,76
Ungarisch	6,84	7,57

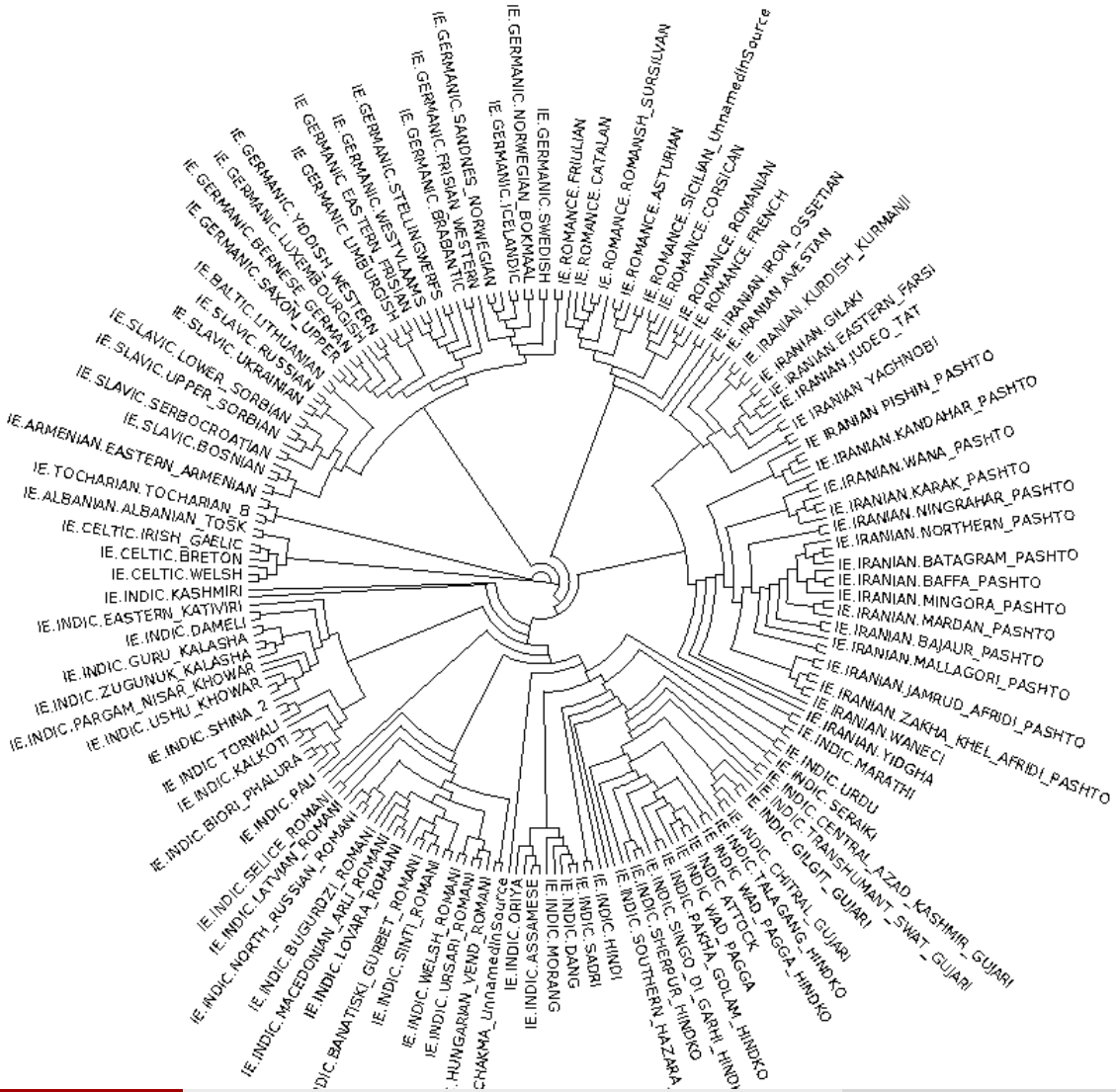
Gewichtete Alinierung

- für die westgermanischen Sprachen/Dialekte ergibt sich die Phylogenie



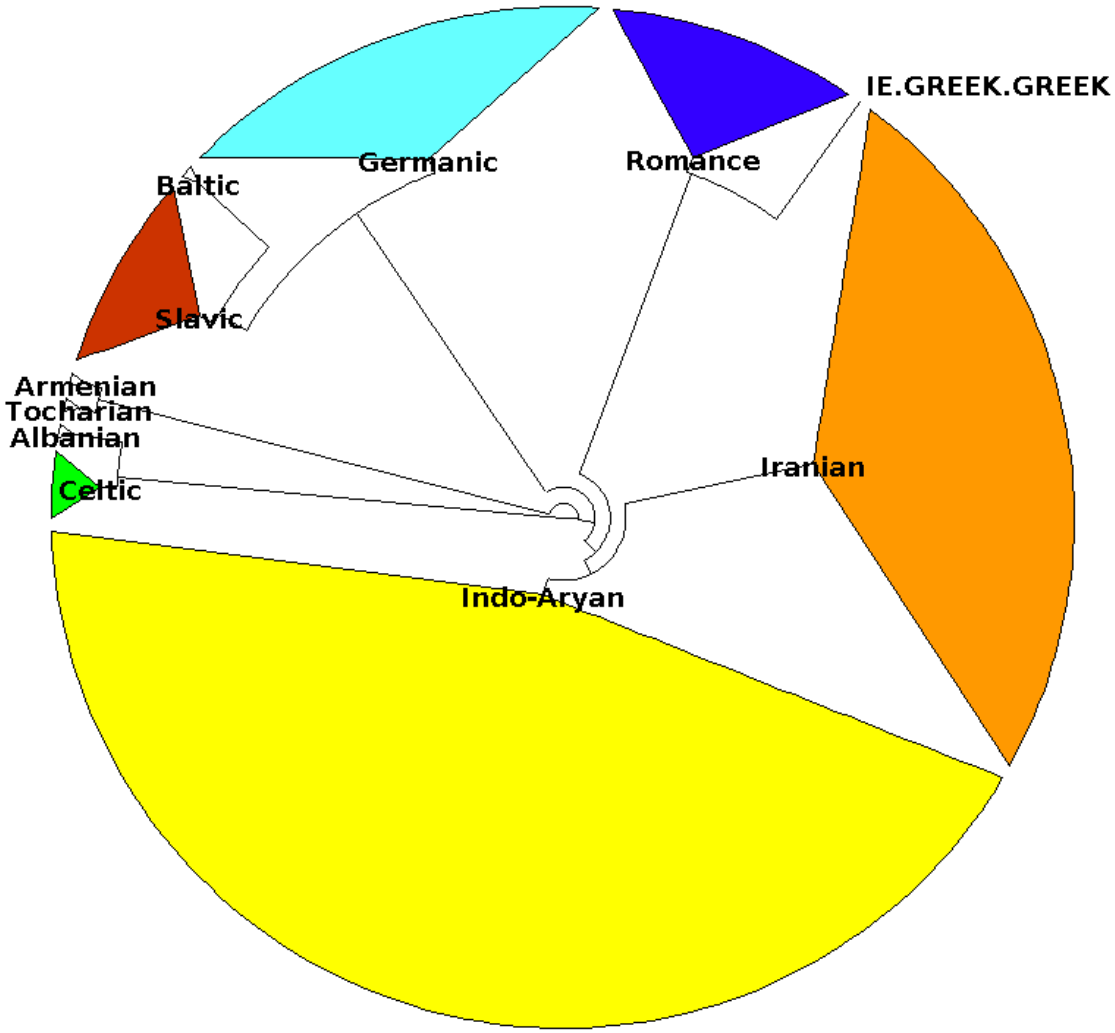
Gewichtete Alinierung

automatische Klassifikation der indo-europäischen Sprachen



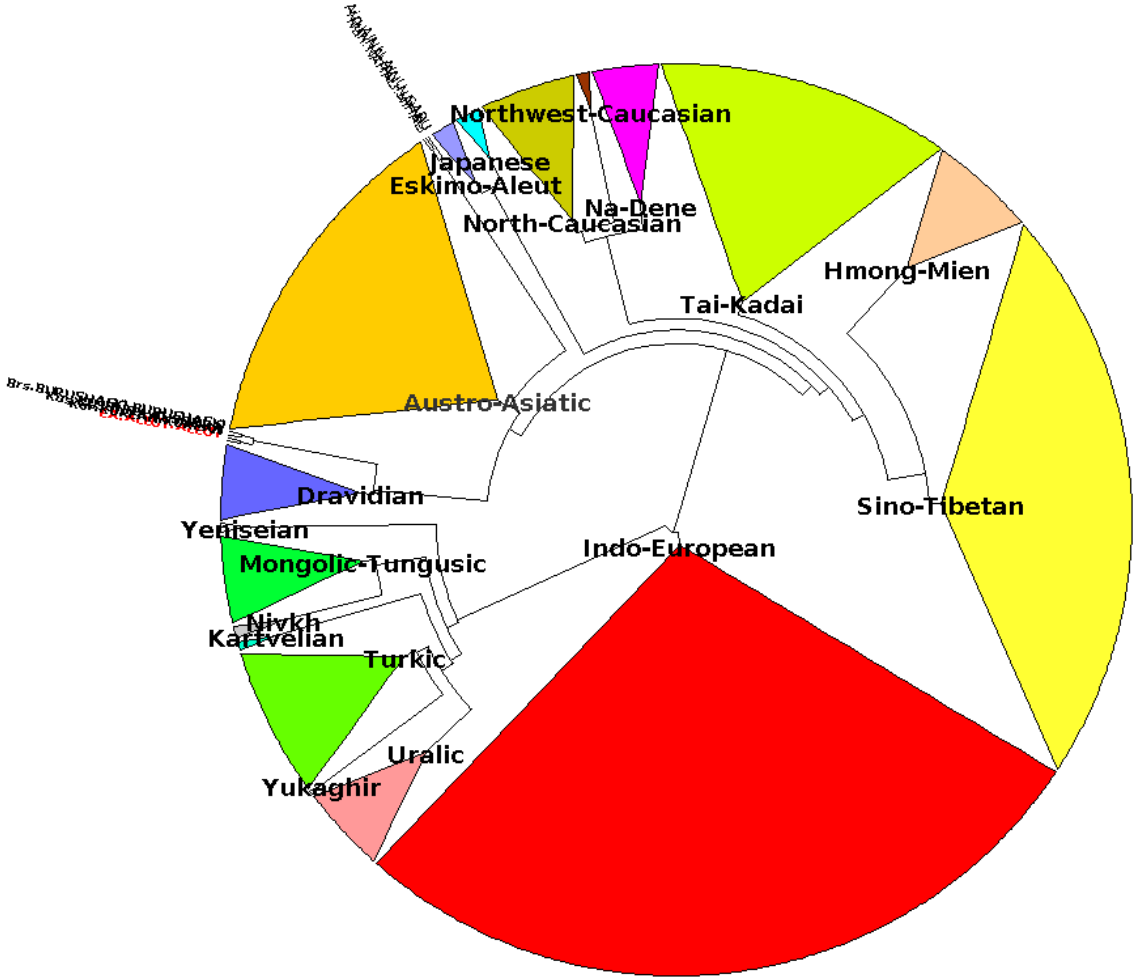
Gewichtete Alinierung

automatische Klassifikation der indo-europäischen Sprachen



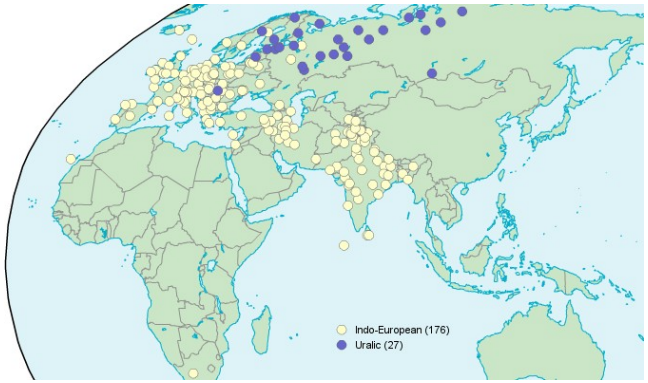
Gewichtete Alinierung

Sprachen Eurasiens (ohne Afro-Asiatisch) + Na-Dene + Eskimo-Aleutisch



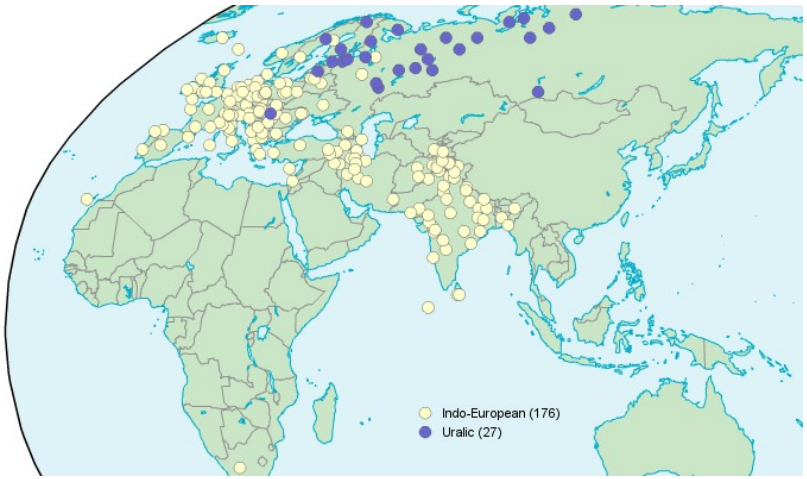
Gewichtete Alinierung

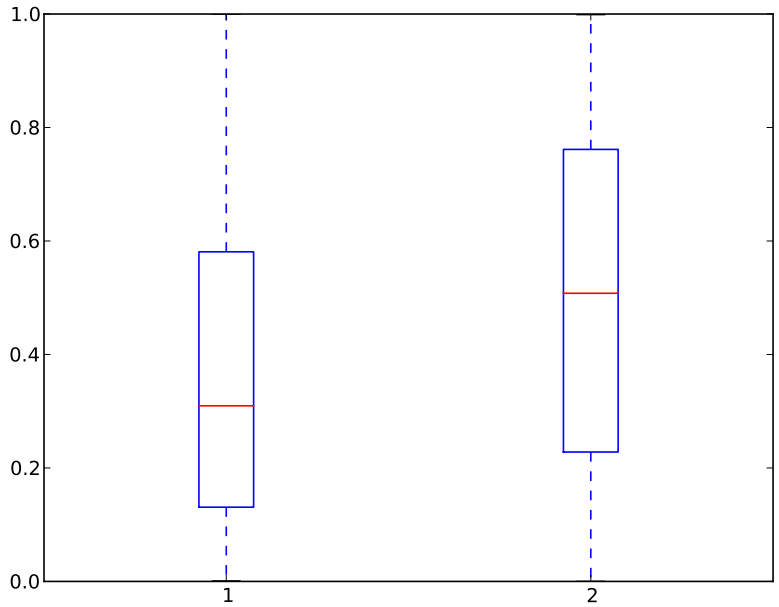
- einige interessante Meta-Verwandtschaften werden sichtbar, v.a.
  - Indo-europäisch/Uralisch
  - Austronesisch/Tai-Kadai



Indo-europäisch/Uralisch

- p-Werte für Vergleich Ähnlichkeiten IE/Ura vs. Zufallspaarungen

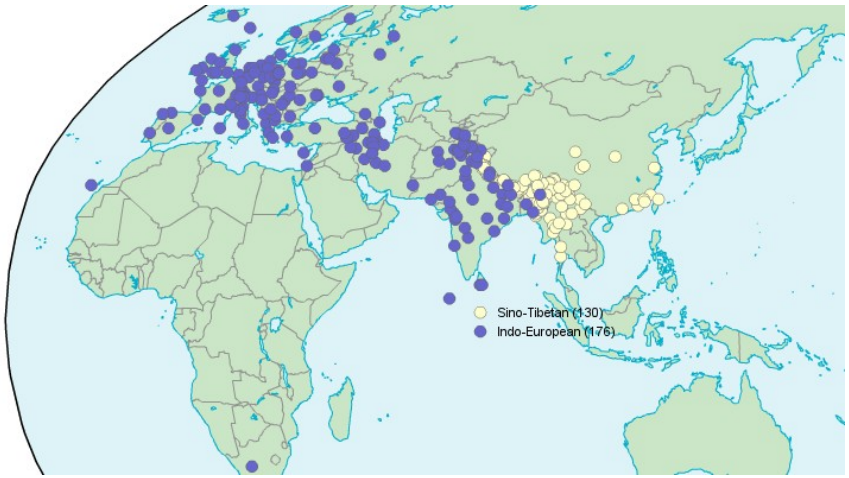


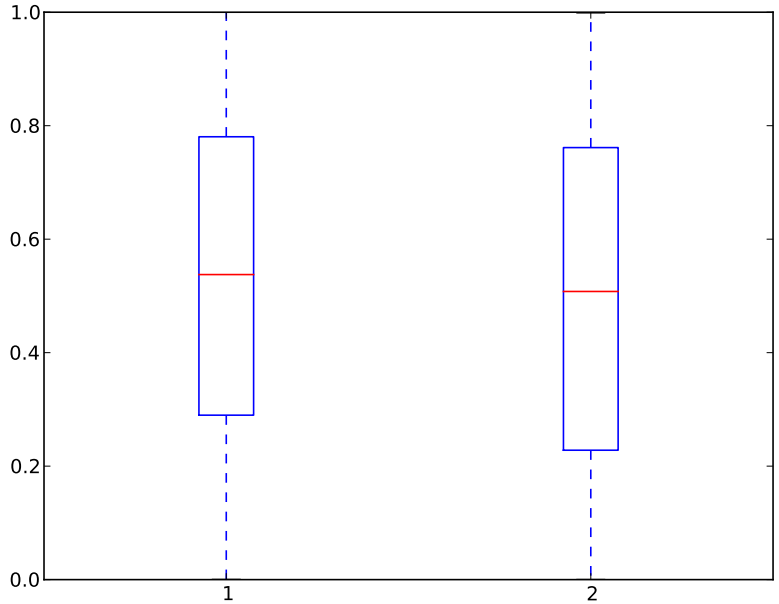


p-Wert: $1,5 \times 10^{-20}$

Indo-europäisch/Uralisch

- p-Werte für Vergleich Ähnlichkeiten IE/Sino-Tibetisch vs. Zufallspaarungen

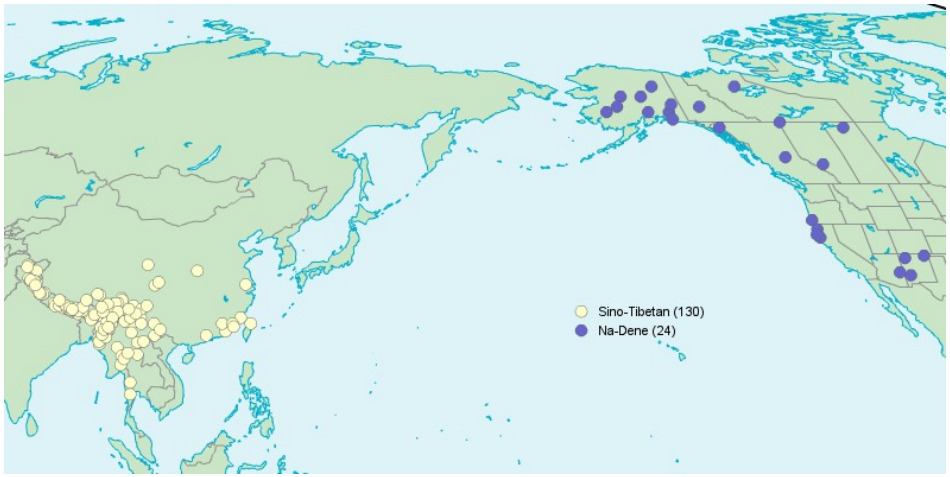


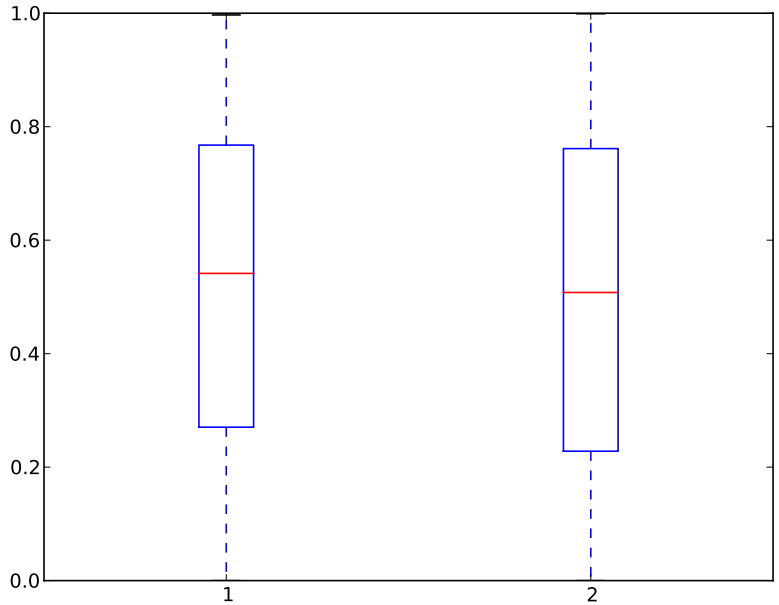


p-Wert: 1

Indo-europäisch/Uralisch

- p-Werte für Vergleich Ähnlichkeiten Sino-Tibetisch/Na-Dene vs. Zufallspaarungen

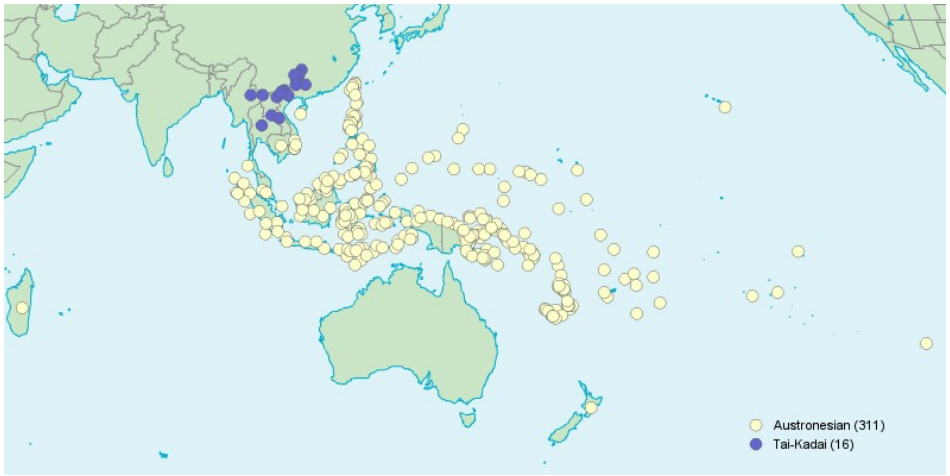


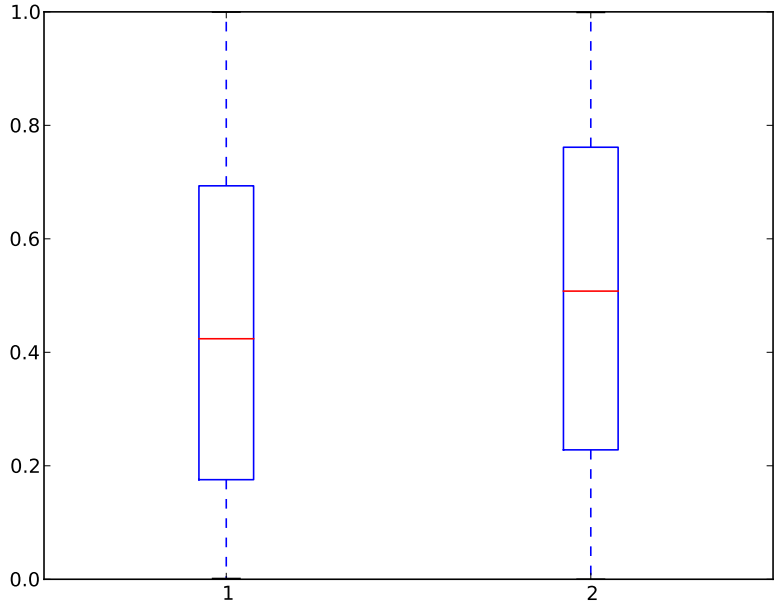


p-Wert: 1

Indo-europäisch/Uralisch

- p-Werte für Vergleich Ähnlichkeiten Tai-Kadai/Austronesisch vs. Zufallspaarungen





p-Wert: 5×10^{-5}

Zusammenfassung

- Erfolgsbilanz:
  - von ca. 1.000 eurasiatischen Sprachen + Na-Dene + Eskimo-Aleutisch wird genau eine Sprache falsch klassifiziert (Aleutisch)
  - Altaisch wird nicht als Einheit erkannt, ist aber auch unter Experten umstritten
  - für alle 5,000 Sprachen ist die Fehlklassifikations-Quote bei 4%

Lexikostatistik funktioniert.

- suggestive Hinweise auf tiefe Verwandtschaften (Nostratisch, Dene-Kaukasisch etc.)
- statistisch wesentlich weniger robust als etablierte Sprachfamilien
- Einfluss von Sprachkontakt?

Fazit

- starke lexikostatistische Evidenz für die Klassifikationseinheiten, die unter Experten unkontrovers sind
- schwache Evidenz für umstrittene Klassifikationen

$\Rightarrow$ Lexikostatistik detektiert die selbe Evidenz, die auch von der komparativen historischen Linguistik genutzt wird