Computational Historical Linguistics

Gerhard Jäger

Current Trends in Linguistics

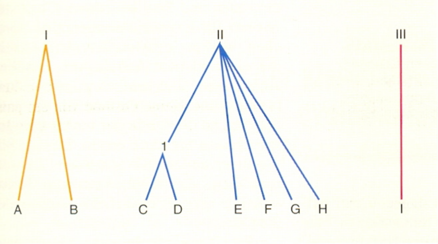
November 3, 2016

Similarity between languages

Eine Klassifikationsübung nach der vergleichenden Methode à la Merritt Ruhlen:

Sprache	zwei	drei	ich	du	wer?	nicht	Mutter	Vater	Zahn	Herz	Fuß	Maus	er trägt
A	ʔiθn-	θalāθ-	-ni	-ka	man	lā	ʔumm-	abū	sinn	ıubb	rijl-	fār	yaḥmil-
B	ʃn-	šaloš	-ni	-ka	mi	lo	ʔem	aβ	šen	leβ	regel	ʕak̲bɔr	nośeh
C	duvā́	tráyas	mā́m	tuvám	kás	ná	mātár	pitár-	dant-	hṛd-	pád	muṣ-	bhárati
D	duva	θrāyō	mạm	tuvəm	čiš	naē-	mātar-	pitar-	dantan-	zərəd	paiðya		baraiti
E	duo	treîs	eme	sú	tís	ou(k)	māter	pater	odṓn	kardiā	pod-	mûs	phérei
F	duo	trēs	mē	tū	kwis	ne-	māter	pater	dent-	kord-	ped-	mūs	fert
G	twai	θreis	mik	θu	hwas	ni	aiθei	faðar	tunθus	haírtō	fōt		baíriθ
H	dó	trí	-m	tú	kía	ní-	máθir	aθir	dēt	kride	traig	lux	berid
I	iki	üč	ben-i	sen	kim	deyil	anne	baba	diš	kalp	ayak	sıčan	taşıyor

Similarity between languages



Klassifizieren Sie die angegebenen neun Sprachen (von A bis I) in Familien und Unterfamilien und vergleichen Sie den Wortschatz für die 13 Wörter, die hier in phonetischer Umschrift geboten werden. Lösung: Sprache A und B (Arabisch und Hebräisch) gehören zur Familie der semitischen Sprachen. Die sechs Sprachen C bis H (Sanskrit, Awestisch, Altgriechisch, Latein, Gotisch und Altirisch) sind indogermanische Sprachen. I (Türkisch) läßt sich keiner Familie zuordnen. Mit einer längeren Wortliste kann man nach demselben Verfahren die Familien wieder in Überfamilien einteilen usw. Der Stammbaum, den man so erhält, würde dann beweisen, daß alle Sprachen von einer Muttersprache abstammen.

Similarity between languages

Multilateraler Sprachenvergleich

Schlichtes Vergleichen einiger Allerweltswörter erhellt bereits die Verwandtschaftsverhältnisse unter den Sprachfamilien Indoeuropäisch (mit den Zweigen Germanisch, Romanisch und Slawisch) sowie Uralisch-Jukagirisch und Baskisch.

Sprachfamilie	Sprache	eins	zwei	drei	Kopf	Auge	Nase	Mund
Germanisch	Schwedisch	en	tvo	tre	hyvud	øga	næsa	mun
	Niederländisch	ēn	tvē	drī	hōft	ōx	nø̄s	mont
	Englisch	wən	tū	θrī	hɛd	ai	nouz	mauθ
	Deutsch	ains	tsvai	drai	kopf	augə	nāzə	munt
Romanisch	Französisch	œ̃/yn	dø	trwa	tɛt	œj	ne	buš
	Italienisch	uno	due	tre	tɛsta	ok̄jo	naso	bok̄a
	Spanisch	uno	dos	tres	kabesa	oxo	naso	boka
	Rumänisch	un	doi	trei	kap	oki	nas	gurə
Slawisch	Polnisch	jeden	dva	tři	gwova	oko	nos	usta
	Russisch	adin	dva	tri	galava	oko	nos	rot
	Bulgarisch	edin	dva	tri	glava	oko	nos	usta
Uralisch-Jukagirisch	Finnisch	yksi	kaksi	kolme	pǣ	silmæ	nenæ	sū
	Estnisch	yks	kaks	kolm	pea	silm	nina	sū
Baskisch	Baskisch	bat	bi	hiryr	byry	begi	sydyr	aho

Sound laws

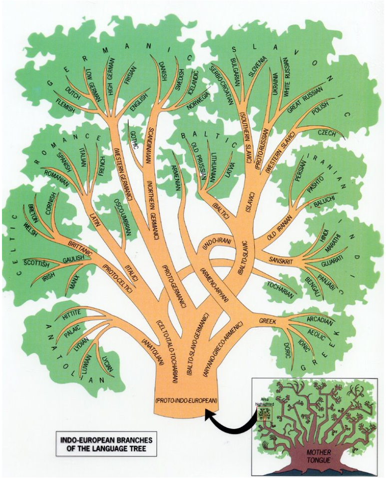
Erste bzw. Germanische Lautverschiebung (Indoeuropäisch → Germanisch)	Phase	Zweite bzw. Hochdeutsche Lautverschiebung (Germanisch→Althochdeutsch)	Beispiele (Neuhochdeutsch)	Jahrhundert	Dialektgebiete
G: /*b/→/*p/	1	**/*p/→/ff/→/f/**	niederdeutsch: sla**p**en, englisch: slee**p** → schla**f**en; niederdeutsch und englisch: Schi**pp**, shi**p** → Schi**ff** niederdeutsch: scher**p**, englisch: shar**p** → schar**f**	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

Sound laws

- sound laws are specific for a particular period in language change
- they hold nearly universally for all occurrences of the sound in question in the language in question
- ideally we have written records of both stages (Latin/Romance languages, Old High German, Middle High German)
- in most cases, sound laws must be reconstructed via systematic comparison of related languages
- applying sound laws backwards leads to reconstructed vocabulary of common mother language

Language trees

- comparative method gives rise to pyhlogenetic trees of historic development



Limits of the comparative method

- Similarities between languages may be due to horizontal transfer (loans)
- limited time depth ($\leq$ 10,000 years)

Hock & Joseph (1996):

Let us pursue this issue a little further by taking a closer look at the relationship between Modern Hindi and English – pretending that we do not yet know that they are related, and trying to establish their relationship by vocabulary comparison. This is actually more difficult than it appears. It is all too easy to be influenced by one's knowledge of the historical relationship between the two languages and therefore to notice the genuine cognates, or even to underestimate the effects of linguistic change on the recognizability of genuine cognates.

Limits of the comparative method

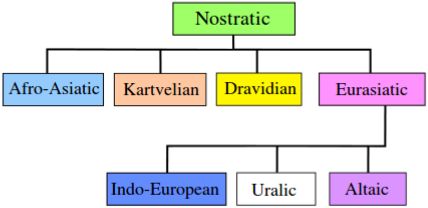
- Similarities between languages may be due to horizontal transfer (loans)
- limited time depth ($\leq$ 10,000 years)

Hock & Joseph (1996):

Clearly, one correspondence is not enough; nor are twenty. And just as clearly, a thousand correspondences with systematic recurrences of phonetic similarities and differences would be fairly persuasive. Are 500 enough, then? And if not, are 501 sufficient? Nobody can give a satisfactory answer to these questions. And this is no doubt the reason that linguists may disagree over whether a particular proposed genetic relationship is sufficiently supported or not.

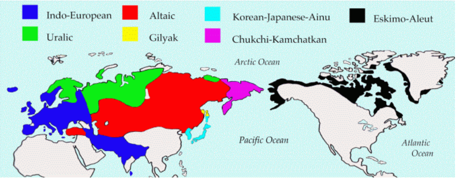
Deep genetic relationships

- Plethora of proposals beyond well-established families:
  - Nostratic:
    - proposed by Pedersen (1903)
    - original proposal: Indo-European, Finno-Ugric, Samoyed, Turkish, Mongolian, Manchu, Yukaghir, Eskimo, Semitic, and Hamitic
    - revived by "Moscow school" in 1960
    - traditional comparative method, including reconstruction of proto forms



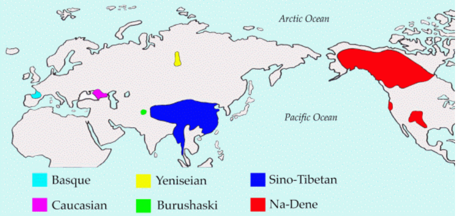
Deep genetic relationships

- Plethora of proposals beyond well-established families:
  - Eurasiatic
    - proposed by Greenberg (2000)
    - comprises Indo-European, UralicYukaghir, Altaic, Chukotko-Kamchatkan, EskimoAleut, Korean-Japanese-Ainu, Gilyak, Etruscan
    - multitude of arguments, mostly from morphology and phonology



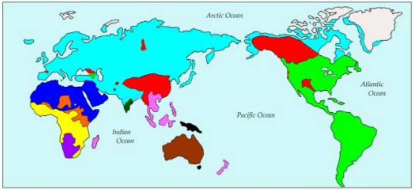
Deep genetic relationships

- Plethora of proposals beyond well-established families:
  - Dene-Caucasian
    - based on work by Sapir, Starostin, Swadesh and others
    - comprises Ne-Dene, Caucasian, Sino-Tibetan, Yeniseian, Burushaski, perhaps Basque and other languages
    - also multitude of arguments, mostly from morphology and phonology



Deep genetic relationships

- Plethora of proposals beyond well-established families:
  - Amerind
    - proposed by Greenberg (1987)
    - comprises all American languages except Na-Dene and Eskimo-Aleut
    - arguments based on mass lexical comparison



Language Families of the World (after Greenberg)

Deep genetic relationships

- Merritt Ruhlen, a student of Greenberg, even claims to have reconstructed a few words of "Proto-World" (for instance the word *aqua* for water, which miraculously didn't change from the dawn of time till Cicero)
- such deep connection are mostly based on suggestive salient features of the languages involved, like pronoun forms
- Nostratic pronouns
- Amerind pronouns
- generally, these approaches neither quantify the probability of chance resemblances nor do they take negative evidence into account

Computational methods

- **this project:**
  - starting from raw word lists (phonetic strings)
  - automatically assess string similarity
  - automatically control for chance resemblances
  - quantify (dis)similarity between word lists
  - evaluate results by
    - comparison to expert language classification
    - correlation with phenotypical distances between populations

The Automated Similarity Judgment Program

- Project at MPI EVA in Leipzig around Søren Wichmann
- covers more than 6,000 languages and dialects
- 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*

Automated Similarity Judgment Project

concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei

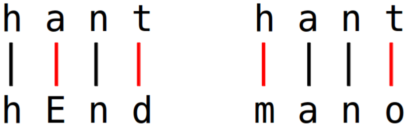
concept	Latin	English
nose	nasus	nos
tooth	dens	tu8
tongue	liNgw∼E	t3N
knee	genu	ni
hand	manus	hEnd
breast	pektus, mama	brest
liver	yekur	liv3r
drink	bibere	drink
see	widere	si
hear	audire	hir
die	mori	dEi
come	wenire	k3m
sun	sol	s3n
star	stela	star
water	akw∼a	wat3r
stone	lapis	ston
fire	iNnis	fEir

Determining distances between word lists

- two steps:
  - compute similarity/distance between individual word forms
  - aggregate word distances to doculect distances

Word distances

- based on string *alignment*
- baseline: Levenshtein alignment ⇒ count matches and mis-matches



- too crude as it totally ignores sound correspondences

Capturing sound correspondences

- weighted alignment using **P**ointwise **M**utual **I**nformation (PMI, a.k.a. *log-odds*):

$$s(a,b) = \log \frac{p(a,b)}{q(a)q(b)}$$

  - $p(a,b)$: probability of sound a being etymologically related to sound b in a pair of cognates
  - $q(a)$: relative frequency of sound a
- **Needleman-Wunsch algorithm:** given a matrix of pairwise PMI scores between individual symbols and two strings, it returns the alignment that maximizes the aggregate PMI score
- but first we need to estimate $p(a,b)$ and $q(a), q(b)$ for all soundclasses a and b
- $q(a)$: relative frequency of occurence of segment a in all words in ASJP
- $p(a,b)$: that's a bit more complicated...

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5				
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5				
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5				
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13			
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13			
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13			
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53		
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1				
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53			
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65		
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7				
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03			
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	−	m	E	n	S
−	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05		
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3				
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47			
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75		
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9				

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97			

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15		

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	8.84

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	↖ 8.84

- memorizing in each step which of the three cells to the left and above gave rise to the current entry lets us recover the corresponing optimal alignment

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2 ↑	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1 ↖	8.84

- memorizing in each step which of the three cells to the left and above gave rise to the current entry lets us recover the corresponing optimal alignment

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	8.84

- memorizing in each step which of the three cells to the left and above gave rise to the current entry lets us recover the corresponing optimal alignment

Computing the weighted alignment score

- Dynamic Programming

	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	8.84

- memorizing in each step which of the three cells to the left and above gave rise to the current entry lets us recover the corresponing optimal alignment

Computing the weighted alignment score

- Dynamic Programming

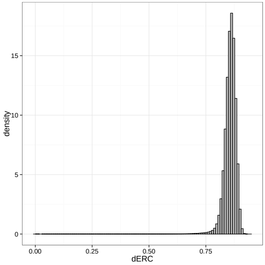
	–	m	E	n	S
–	0	−2.5	−4.1	−5.7	−7.3
m	−2.5	4.13	1.53	0.03	−1.47
e	−4.1	1.53	5.65	3.05	1.55
n	−5.7	0.03	3.05	9.2	6.6
E	−7.3	−1.47	4.75	6.6	7.62
s	−8.9	−2.97	2.15	5.1	8.84

- memorizing in each step which of the three cells to the left and above gave rise to the current entry lets us recover the corresponing optimal alignment

```
m E n - S
m e n E s
```

Capturing sound correspondences

- **First step:** automatically compile a list of language pairs that are (fairly) certain to be related
- start with a measure for language dissimilarity based on Levenshtein alignment



- all language pairs with dissimilarity ≤ 0.7 (ca. 1% of all pairs) qualify as *probably related*

Capturing sound correspondences

- doculects *probably related* (in this sense) to English:

AFRIKAANS, ALSATIAN, BERNESE_GERMAN, BRABANTIC, CIMBRIAN, DANISH, DUTCH, EASTERN_FRISIAN, FAROESE, FRANS_VLAAMS, FRISIAN_WESTERN, GJESTAL_NORWEGIAN, ICELANDIC, JAMTLANDIC, LIMBURGISH, LUXEMBOURGISH, NORTH_FRISIAN_AMRUM, NORTHERN_LOW_SAXON, NORWEGIAN_BOKMAAL, NORWEGIAN_NYNORSK_TOTEN, NORWEGIAN_RIKSMAL, PLAUTDIETSCH, SANDNES_NORWEGIAN, SAXON_UPPER, SCOTS, STANDARD_GERMAN, STELLINGWERFS, SWABIAN, SWEDISH, WESTVLAAMS, YIDDISH_EASTERN, YIDDISH_WESTERN, ZEEUWS

- these are all and only the Germanic languages
- 99.9% of all probably related pairs belong to the same family, and 60% to the same genus

Capturing sound correspondences

- **Second step:**
  - let L_1 and L_2 be *probably related*
  - every pair of words w_1/w_2 from L_1/L_2 sharing the same meaning are considered *potentially cognate*
  - all potential cognate pairs are (Levenshtein-)aligned
  - relative frequency of a being aligned with b is used as estimate of $s(a, b)$
  - all potential cognate pairs are Needleman-Wunsch aligned using PMI scores obtained in the previous step
  - all potential cognate pairs with an aggregate PMI score ≥ 5.0 are considered *probable cognates*
  - $s(a, b)$ is re-estimated using only probable cognate pairs
  - this is repeated ten times

Capturing sound correspondences

- only probabe cognate between English and Latin: `pers3n/persona`
- probable cognates English/German:

fiS	fiS
laus	laus
bl3d	blut
horn	horn
brest	brust
liv3r	leb3r
star	StErn
wat3r	vas3r
ful	fol

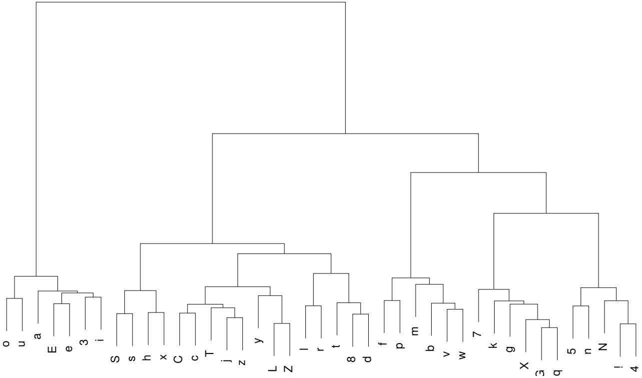
Capturing sound correspondences

- procedures results in pairwise PMI scores for each pair from the 41 ASJP sound classes
- positive PMI-score between a and b: evidence for etymological relatedness
- negative PMI-score between a and b: evidence against etymological relatedness

	a	**e**	**i**	**o**	**u**	**p**	**b**	**d**	**t**	**8**	**s**	**h**
a	**1.88**	−1.35	−2.35	−1.66	−2.54	−8.49	−8.82	−7.07	−7.03	−4.64	−8.78	−8.40
e	−1.35	**2.40**	−0.48	−1.52	−2.88	−7.47	−7.80	−7.66	−6.01	−5.01	−7.76	−7.38
i	−2.35	−0.48	**2.37**	−2.81	−1.32	−6.75	−8.46	−8.33	−8.98	−3.48	−7.04	−6.66
o	−1.66	−1.52	−2.81	**2.48**	−0.27	−7.08	−8.10	−7.96	−8.61	−5.31	−8.06	−7.68
u	−2.54	−2.88	−1.32	−0.27	**2.76**	−6.62	−8.05	−7.91	−8.56	−5.26	−8.01	−7.63
p	−8.49	−7.47	−6.75	−7.08	−6.62	**3.69**	**0.36**	−6.59	−4.30	−3.94	−2.70	−0.49
b	−8.82	−7.80	−8.46	−8.10	−8.05	**0.36**	**3.62**	−4.84	−5.09	−3.58	−5.63	−3.24
d	−7.07	−7.66	−8.33	−7.96	−7.91	−6.59	−4.84	**3.41**	−0.10	**2.52**	−2.29	−2.81
t	−7.03	−6.01	−8.98	−8.61	−8.56	−4.30	−5.09	−0.10	**3.15**	**2.11**	−1.67	−1.76
8	−4.64	−5.01	−3.48	−5.31	−5.26	−3.94	−3.58	**2.52**	**2.11**	**5.49**	**1.92**	−0.85
s	−8.78	−7.76	−7.04	−8.06	−8.01	−2.70	−5.63	−2.29	−1.67	**1.92**	**3.50**	**0.26**
h	−8.40	−7.38	−6.66	−7.68	−7.63	−0.49	−3.24	−2.81	−1.76	−0.85	**0.26**	**3.50**

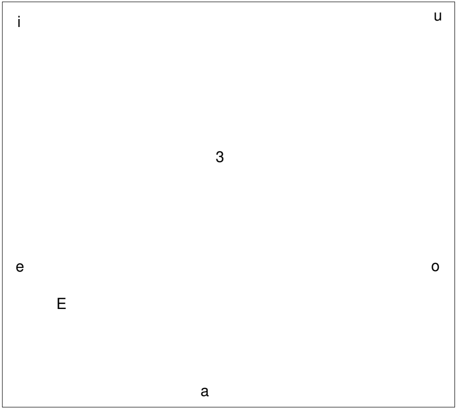
Capturing sound correspondences

- hierarchical clustering of sound classes according to PMI scores:

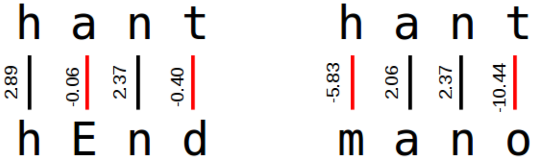


Capturing sound correspondences

- multidimensional scaling of vowel classes according to PMI scores:



Weighted alignment



$\Sigma = 4.80$ $\Sigma = -11.85$

Weighted alignment

- alignments German/Latin:

```
iX-
ego

du
tu

vir--
--nos

ain-s
-unus

cvai
d-uo

--mEnS
homo--

fiS---
piskis

hun-t
kanis

--la-u--s
pedikulus

--baum
arb-or

b-lat
folu-

haut--
k-utis

--blut
saNgis

knoX3n
--os--

-or--
auris

a-ug3-
okulus

naz3-
nasus

can-
dens

cuN-3
liNgE

k-ni
genu

han-t
manus

b--rust
pektus-

leb3r
yekur

triNk3n-
b-i-bere

--ze-3n
widere-

--her3n
audire-

Sterb3n
-mor-i-

kom3n---
w--enire

zon3
sol-

StErn-
ste-la

vas3r
-aka-

Sta-in
-lapis

foi--a-
--iNnis

p--at
viya-

bErk
mons

naxt
noks

f---ol
plenus

no-i-
nowus

nam3-
nomen
```

Weighted alignment

- alignments German/Cimbrian:

```
iX
ix

du
dE

vir
bar

cvai-
sb-en

mEn-S
menEs

hunt
hunt

laus
laus

baum
p-om

blat
-lop

blut
plut

knoX3n
-po-an

horn
horn

o-r
oar

aug3
-ogE

--n--az3
kanipa--

cuN3-----
--gaprext

hant
hant

brus---t
p-uzamEn

leb3r-
lEbara

triNk3n
trink--

ze3n
ze-g

her3n
hor--

Sterb3n
sterb--

kom3n
kEm--

zon3
zuna

StE-rn
stEarn

vas3r
basar

St-ain
stoa-n

foia-
bo-ar

vek---
bEgale

bErk
perg

naxt
naxt

--fol--
gabasEt

noi
noy

nam3
namo
```

Aggregating word similarites

- Needleman-Wunsch alignment returns a *similarity score* for each word pair
- not too reliable to identify cognates:
  - often low scores for genuine cognate pairs ('false negatives'):
    - lat. *genu*/eng. *knee*: -3.39
    - lat. *unus*/eng. *one*: -5.00
  - occasionally high scores for non-cognates ('chance similarities'/'false positives'):
    - grm. *Blatt* ('leaf')/Tilquiapan *bldag* ('leaf'): 0.22
    - lat. *oculus* ('eye)/Lachixio *ikulu* ('eye'): 6.72
- approach pursued here:
  - for each language pair, estimate amount of chance similarities
  - quantify to what degree the observed similarities exceed expected chance similarities

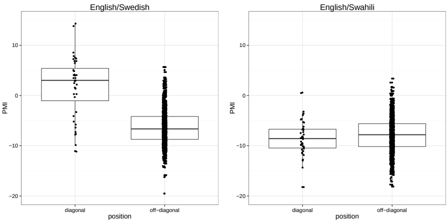
Aggregating word distances

English / Swedish

	Ei	**yu**	**wi**	**w3n**	**tu**	**fiS**	…
yog	**−7.77**	0.75	−7.68	−7.90	−8.57	−10.50	
du	−7.62	**0.33**	−5.71	−7.41	2.66	−8.57	
vi	−2.72	−2.83	**4.04**	−1.34	−6.45	0.70	
et	−5.47	−7.87	−5.47	**−6.43**	−1.83	−4.70	
tvo	−7.91	−4.27	−3.64	−4.57	**0.39**	−6.98	
fisk	−7.45	−11.2	−3.07	−9.97	−8.66	**7.58**	
⋮							

- values along diagonal give similarity between candidates for cognacy (possibility of meaning change is disregarded)
- values off diagonal provide sample of similarity distribution between non-cognates

Aggregating word distances



- distance between two word lists is a measure for how much the distribution along the diagonal differs from the distribution off the diagonal

Aggregating word distances

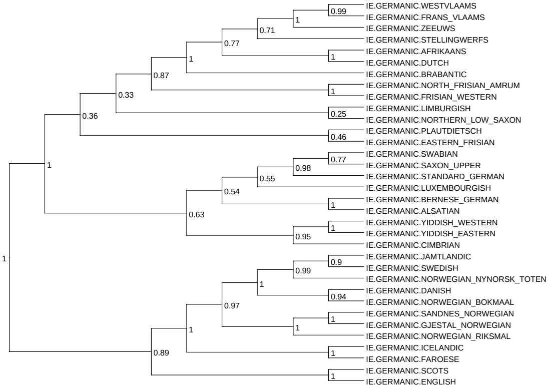
- some examples

A	B	$d(A, B)$
English	Scots	0.2139
Danish	Swedish	0.2773
English	Swedish	0.3981
English	Frisian	0.4215
English	Dutch	0.4040
Hindi	Farsi	0.6231
English	French	0.7720
English	Hindi	0.7735
Amharic	Vietnamese	0.8566
Swahili	Warlpiri	0.8573
Navajo	Dyirbal	0.8436
Japanese	Haida	0.8504
English	Swahili	0.8901

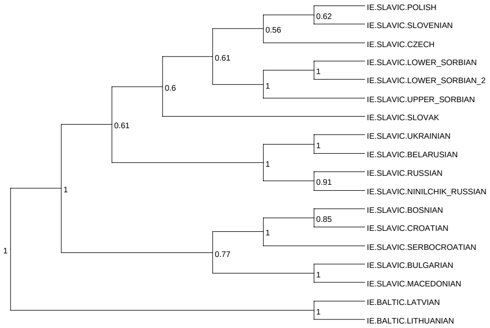
Phylogenetic inference

- pairwise distances for all (extant) languages present in ASJP are computed
- resulting distance matrix is fed into distance-based phylogenetic algorithm (*Neighbor Joining + Ordinary Least Square Nearest Neighbor Interchange Optimization*)
- outcome recognizes language families and their internal structure remarkably well

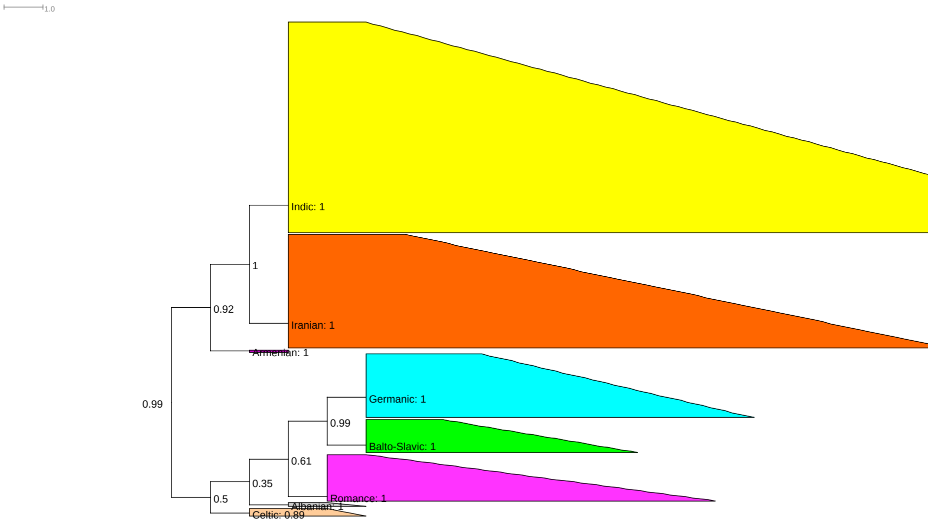
Phylogenetic inference



Phylogenetic inference

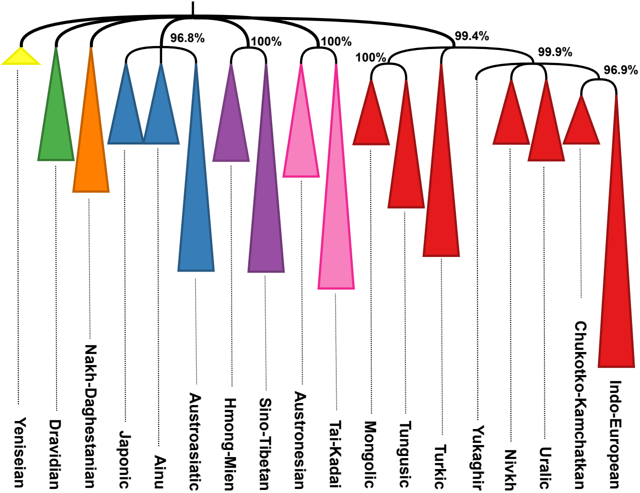


Phylogenetic inference



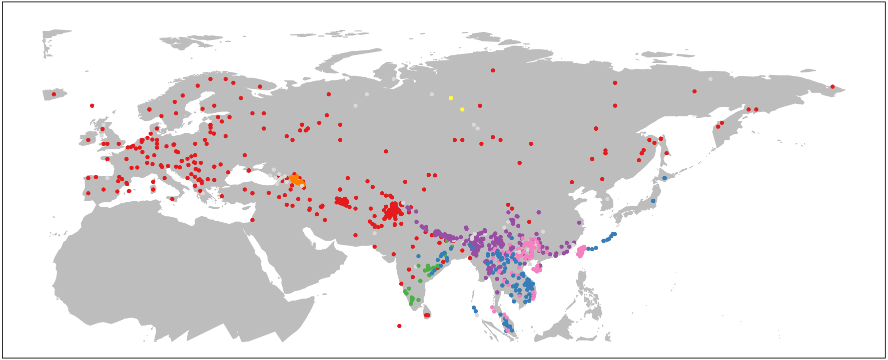
Phylogenetic inference

Languages of Eurasia

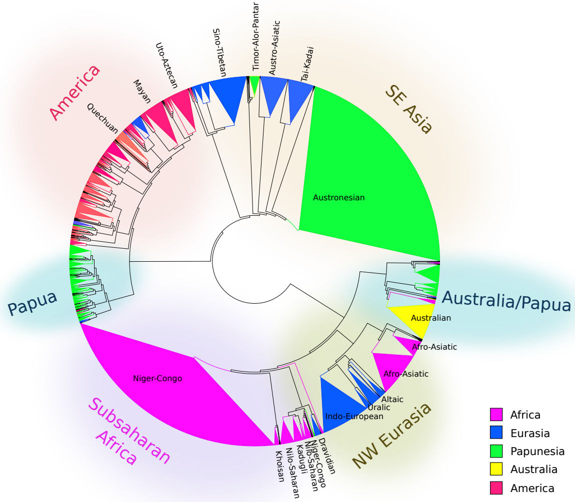


Phylogenetic inference

Languages of Eurasia



Phylogenetic inference

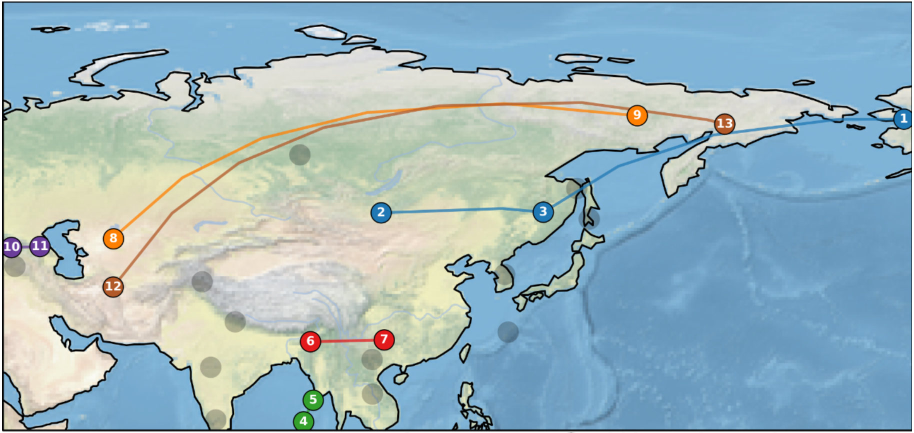


Distant relationships

(joint work with Cecil Brown, Eric Holman, Johann-Mattis List and Søren Wichmann)

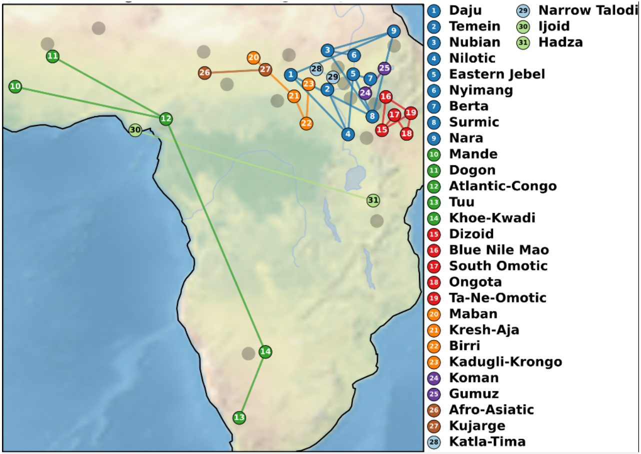
- compute aggregate distances between language families
- find threshold with *false discovery rate* of 5%: all families pairs with a distance below this threshold are genuinely related (due to common descent or contact) with a confidence or 95%

Distant relationships



1 Eskimo-Aleut	4 Jarawa-Onge	7 Hmong-Mien	10 Abkhaz-Adyge	13 Chukotko-Kamchatkan
2 Mongolic	5 Great Andamanese	8 Turkic	11 Nakh-Daghestanian	
3 Tungusic	6 Sino-Tibetan	9 Yukaghir	12 Indo-European	

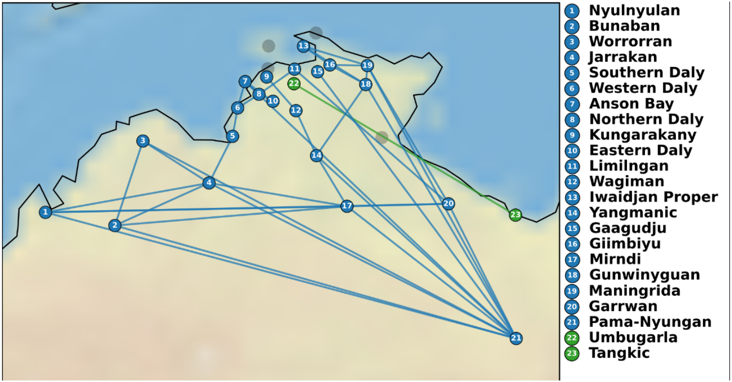
Distant relationships

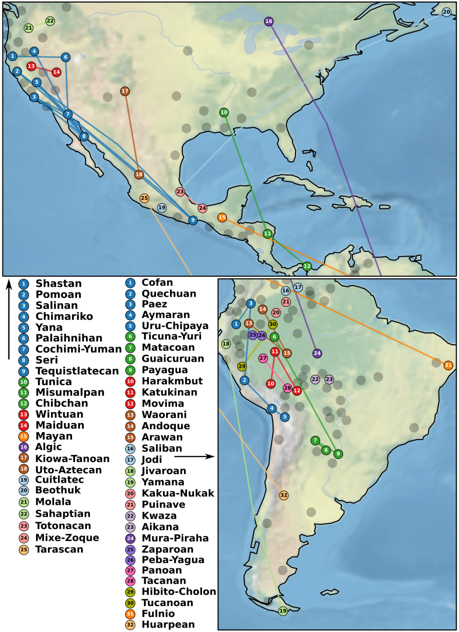


Distant relationships

1. Nimboran
2. Kosare
3. Elseng
4. Border
5. Suki-Gogodala
6. Nuclear Torricelli
7. Tirio
8. Waia
9. Kiwaian
10. Taiap
11. Nuclear Trans-New Guinea
12. Lepki-Murkim
13. Namla-Tofanma
14. Kimki
15. Biksi
16. Pauwasi
17. Koiarian
18. East Strickland
19. Dibiyaso
20. Bosavi
21. Fasu
22. East Kutubu
23. Turama-Kikori
24. Austronesian
25. Bilua
26. Touo
27. Lavukaleve
28. Savosavo
29. Walio
30. Sepik
31. Ndu
32. Morehead-Wasur
33. Pahoturi
34. Eastern Trans-Fly
35. Alor-Pantar
36. East Timor-Bunaq
37. West Bomberai
38. Marindic
39. Awin-Pa
40. Kamula
41. Bogaya
42. Duna
43. Amto-Musan
44. Left May
45. Greater Kwerba
46. Kapauri
47. Maybrat
48. Anem
49. Mpur
50. Yawa
51. Kolopom
52. Bulaka River
53. Kaure-Narau
54. Yale

Distant relationships





1 Shastan
2 Pomoan
3 Salinan
4 Chimariko
5 Yana
6 Palaihnihan
7 Cochimi-Yuman
8 Seri
9 Tequistlatecan
10 Tunica
11 Misumalpan
12 Chibchan
13 Wintuan
14 Maiduan
15 Mayan
16 Algic
17 Kiowa-Tanoan
18 Uto-Aztecan
19 Cuitlatec
20 Beothuk
21 Molala
22 Sahaptian
23 Totonacan
24 Mixe-Zoque
25 Tarascan
1 Cofan
2 Quechuan
3 Paez
4 Aymaran
5 Uru-Chipaya
6 Ticuna-Yuri
7 Matacoan
8 Guaicuruan
9 Payagua
10 Harakmbut
11 Katukinan
12 Movima
13 Waorani
14 Andoque
15 Arawan
16 Saliban
17 Jodi
18 Jivaroan
19 Yamana
20 Kakua-Nukak
21 Puinave
22 Kwaza
23 Aikana
24 Mura-Piraha
25 Zaparoan
26 Peba-Yagua
27 Panoan
28 Tacanan
29 Hibito-Cholon
30 Tucanoan
31 Fulnio
32 Huarpean

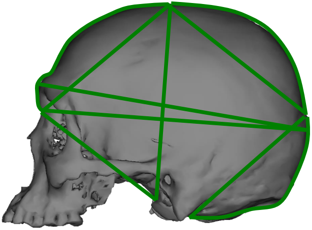
Words and bones

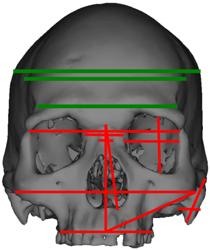
(joint work with Katerina Harvati and Hugo Reyes-Centeno)

- Since Cavalli-Sforza's work: lot of interest in correlations between genetic and linguistic features of human populations
- our work: correlations between phenotypical (cranial) and linguistic (vocabulary-based) features
- motivation:
  - different parts of the cranium respond to different selective pressures
  - ASJP provides data for computing linguistic distances on an unprecedented scale; this study provides (additional) evidence for the reliability of ASJP-based distances across language family boundaries
  - part of the general endeavor to disentangle human bio-historical co-evolution

Cranial Phenotype Data

- Whole Cranium: 30 variables
- Face: 15 variables
- Neurocranium: 15 variables



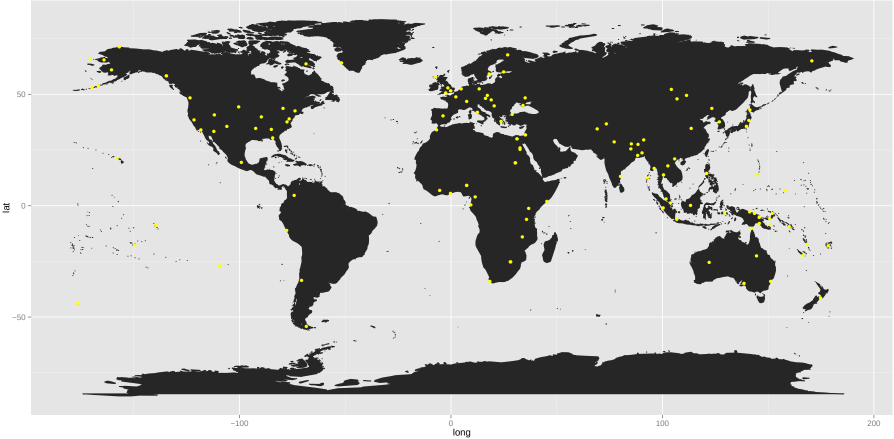


Does language track population history?

- **Hypothesis 1:** Language reflects genetic population history if there is a significant relationship with neurocranial morphology and geography
- **Hypothesis 2:** Language reflects other factors if there is a significant relationship with facial morphology

Mapping bones to languages

- cranial data from 135 populations



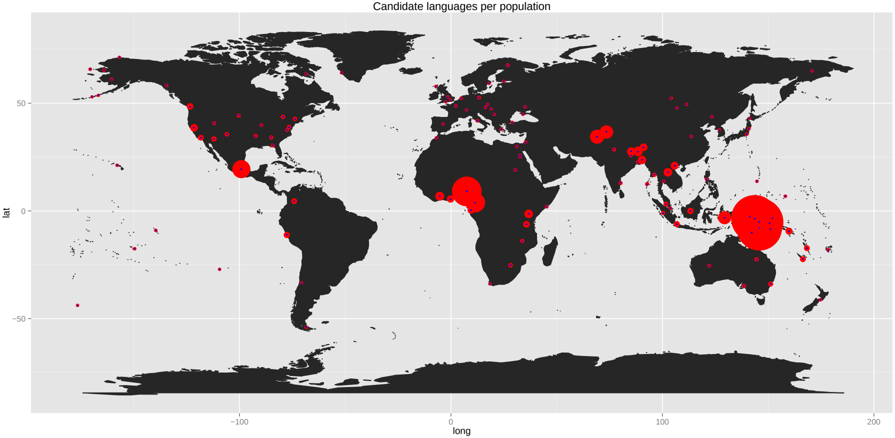
Assigning languages to populations

- in some cases, assignment is straightforward:
  - WestAleut → Aleut
  - South West Alaska → Central Yupik
  - Serbia → Serbo-Croatian
  - Gyzeh → Late Egyptian
- sometimes, several candidate languages from the same language family or genus
  - North East Asia → Inupiaq, 3 dialects of Yupik (all Eskimo languages)
  - Germany → Standard German + 6 German dialects
  - Recent Italy → Corsican, Friulian, Italian, Sardinian

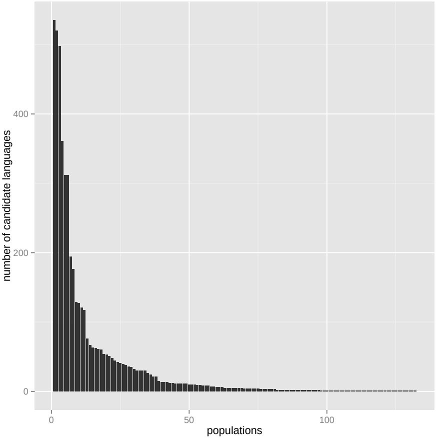
Assigning languages to populations

- in many cases, assignment is pure guesswork (based on geography)
- PNG, Australia, sub-Saharan Africa, America, India
- criteria:
  - geographic location (according to ASJP) $\leq$ 300 km from coordinates of cranial data
  - for islands (New Caledonia, Hebrides, Torres Strait, ...): Ethnologue information
  - if cranial data contain ethnic information, these override geography
    - Han North is mapped to Mandarin, even though several Turkic languages are closer
    - only Khoisan languages are considered for South Africa
- number of candidate languages assigned to single populations range from 1 to 535 (for Madang/PNG)
- average: 37 languages per population

Assigning languages to populations

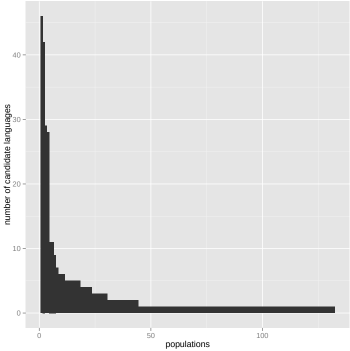


Assigning languages to populations

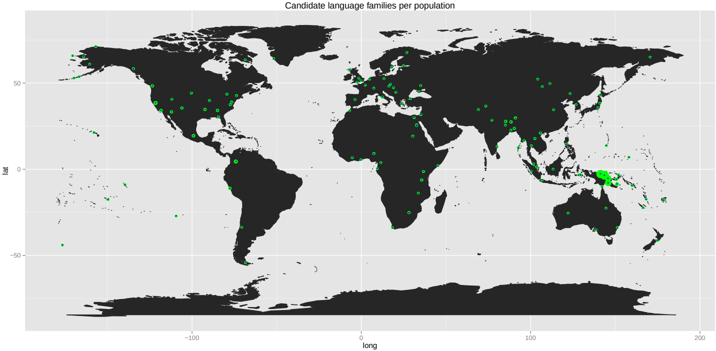


Assigning languages to populations

- in most cases, candidate languages belong to the same language families
- maximum number of candidate families: 46 (for East Sepik, PNG)
- mean number of candidate families per population: 3 (median: 1)

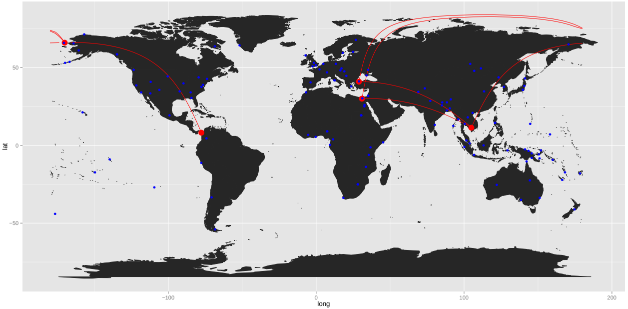


Assigning languages to populations



- in the sequel, the linguistic distance between two populations is computed as the average distance between the corresponding candidate languages

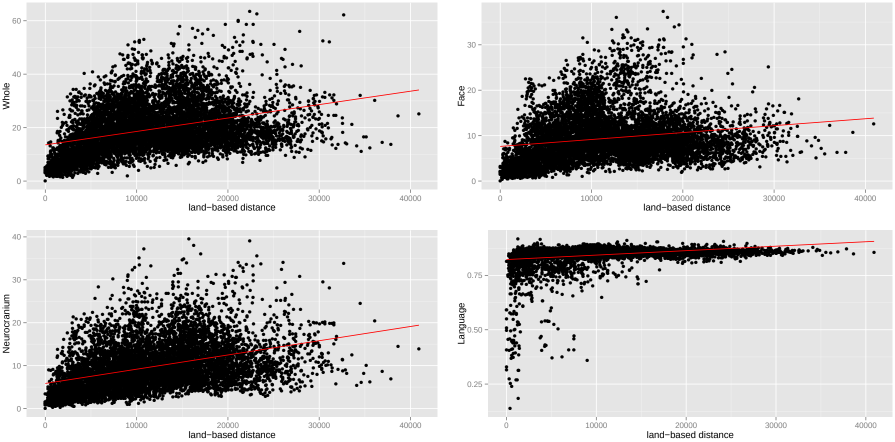
Land-based distances



- following Atkinson 2011:
  - Africa/Asia: *Cairo*
  - Asia/Europ: *Istanbul*
  - Asia/Oceania: *Phnom Phen*
  - Asia/North America: *Bering Strait*
  - North America/South America: *Panama*

Correlations

- correlations between land-based geographic distances phenotypical/linguistic distances



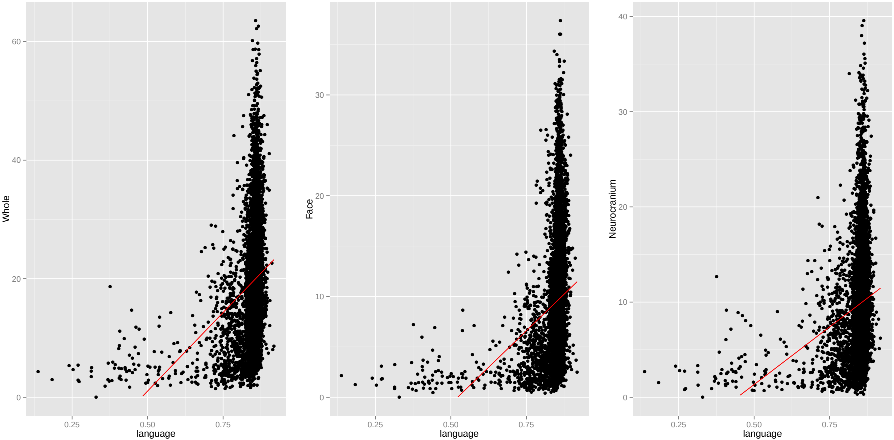
Correlations

- correlations between land-based geographic distances phenotypical/linguistic distances
- determined via Mantel test

	(Spearman) correlation
Whole	$0.399\ (10^{-4})$
Face	$0.250\ (10^{-4})$
Neurocranium	$0.457\ (10^{-4})$
Language	$0.246\ (10^{-4})$

Correlations

- Correlation of linguistic distances to various cranial distances



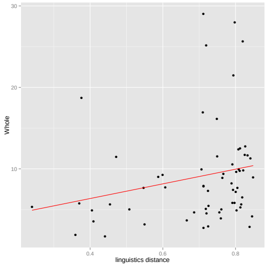
Correlations

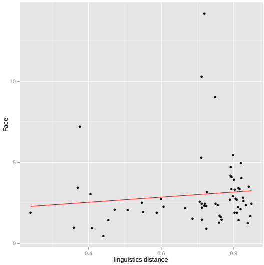
- Correlation of linguistic distances to various cranial distances

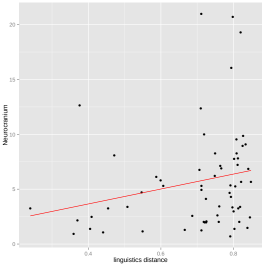
	unconditional	conditioned on geography
Whole	$0.296(10^{-4})$	$0.222(10^{-4})$
Face	$0.321(10^{-4})$	$0.276(10^{-4})$
Neurocranium	$0.246(10^{-4})$	$0.155(10^{-4})$

Correlations within language families

- intra-family correlation of language with
  - Whole: 0.290
  - Face: 0.200
  - Neurocranium: 0.272

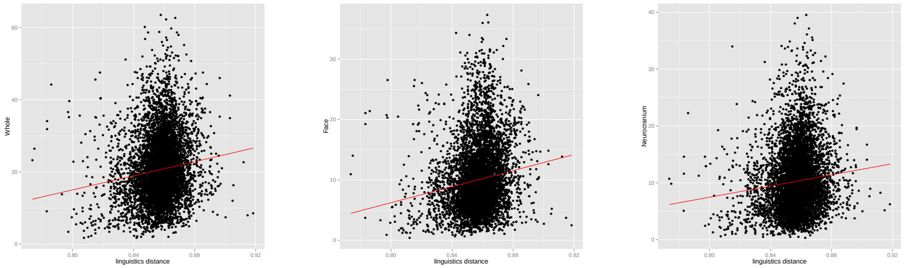






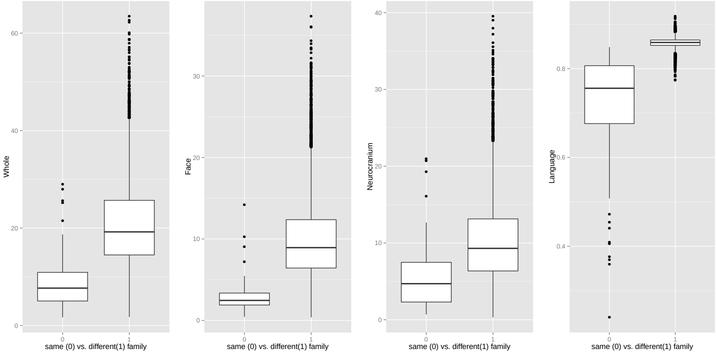
Correlations across language families

- inter-family correlation of language with
  - Whole: 0.139
  - Face: 0.177
  - Neurocranium: 0.120



Separating language families

- correlation of degree on non-overlap of the candidate language families of a population with
  - Whole: 0.365
  - Face: 0.351
  - Neurocranium: 0.299

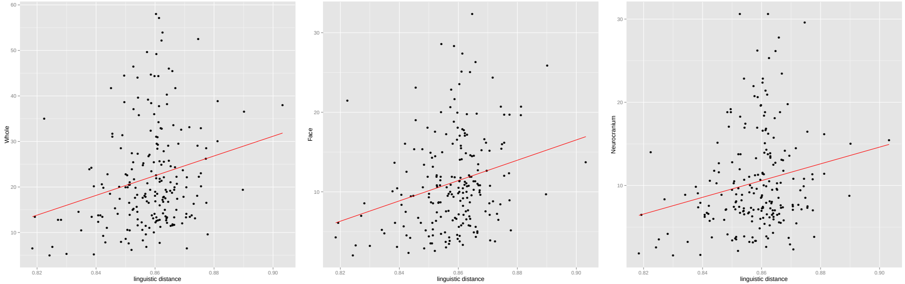


Aggregating language families

- a population "belongs" to a given language family f if all candidate languages for that population belong to f
- the phenetic (Whole, Face, Neurocranium)/geographical distance between the families f_1 and f_2 is defined as the average distance between the populations belonging to f_1/f_2 respectively
- the linguistic distance between f_1 and f_2 is the average distance between all languages assigned to populations that belong to f_1/f_2 respectively

Aggregating language families

- aggregated correlations of language with
  - Whole: 0.198 ($p = 0.013$)
  - Face: 0.256 ($p < 0.001$)
  - Neurocranium: 0.178 ($p = 0.028$)
- partial correlations, conditioned on land-based distance
  - Whole: 0.141 ($p = 0.089$)
  - Face: 0.219 ($p = 0.003$)
  - Neurocranium: 0.116 ($p = 0.155$)



Considerations and hypotheses

- Evolutionary rate of change
  - Genes and neurocranium evolve slowly
  - Language and face evolve faster?
- Depth of population history
  - Genes and neurocranium track deep history
  - Language and face track recent history?
- Modes of transmission
  - Genes and neurocranium are vertically transmitted
  - Language and face are horizontally transmitted?
- Selection on face and language?