Genetical language classification

Part 2: statistical methods

Time depth of historical reconstruction

- Three possible positions:
  - General limit for time depth that can be reached by the methods of historical linguistics
    traditional position
  - Limits exist, but depend on available information and methods
  - No strict limit (apart from the limit set by biological evolution of human language faculty)
    Greenberg, Ruhlen

Arguments for the pessimistic position

- Reliable proof for genetic relatedness of two languages:
  - Reconstruction of common ancestor language
  - Reconstruction of the diachronic processes from the common ancestor to the languages under discussion
- Requires identification of *cognates*
- Language change obliterates similarities
- Also, cognates may be due to borrowing
- No genetic relatedness can be proved beyond 10,000 years

Arguments for the pessimistic position

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.

Arguments for the pessimistic position

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.

Word lists

- Methods of classical comparative historical linguistics have probably reached their limit
- Alternative method: usage of **word lists**
- Identification of phonetic similarities in the vocabulary of different languages => measure of relatedness of languages
- No attempt is made to reconstruct the common ancestor language

Word lists

- Compilation of *concept lists* – univeral basic vocabulary that is supposedly present in all languges
- Translation of this list into all languages under investigation
- For every translation pair it is tested whether there is relatedness/similarity
- Relatedness corresponds to percentage of similar words

Word lists

- Morris Swadesh (1909-1967)
  - American linguist
  - Studies a.o. The genetic classificaiton of native American languages
  - Pioneer of **lexicostatistics** and **glottochronology**
  - Compiled the so-called **Swadesh-Liste** of 207 concepts that occur in all languages/cultures:

http://www.christianlehmann.eu/fundus/Swadesh_list.html

Lexicostatistics

- Relatedness or chance?
  (Quelle: Maiwald & Willeke)

German	*Latin*	*(engl.)*
Herz	**c**ord-	heart
Horn	**c**ornu	horn
Hund	**c**anis	dog
hundert	**c**entum	hundred

Lexicostatistics

- Relatedness or chance?

English	*Hawaian*
sew	**h**umu
smell	**h**oni
snow	**h**au kea
stab	**h**ou
star	**h**oku
swell	**h**o'opehu

Glottochronology

- Basic assumptions:
  - Every language constantly renews its vocabulary
  - This process takes place continuously
  - Vocabulary of a language has a „half-life“
  - This half-life is approximately constant across languages
  - Comparison of word lists allows reconstruction when languages have split

Glottochronology

- Calibration on the basis of comparison English-Spanish
- Estimate: after 1,000 year, 81% of the 200-Swadesh list are preserved
- Formula:

$$t = \frac{\ln c}{\ln r}$$

- *t: time depth (time since proto-language)*
- *c: percentage of common basic vocabulary (0 < c < 1)*
- *r: glottochronological constant. (81%)*

Glottochronology

- Advantages:
  - Applicable to any language pair
  - Little analytical effort
  - Supplies information about distant genetic relations
  - Supplies information about time depth, not just over relatedness

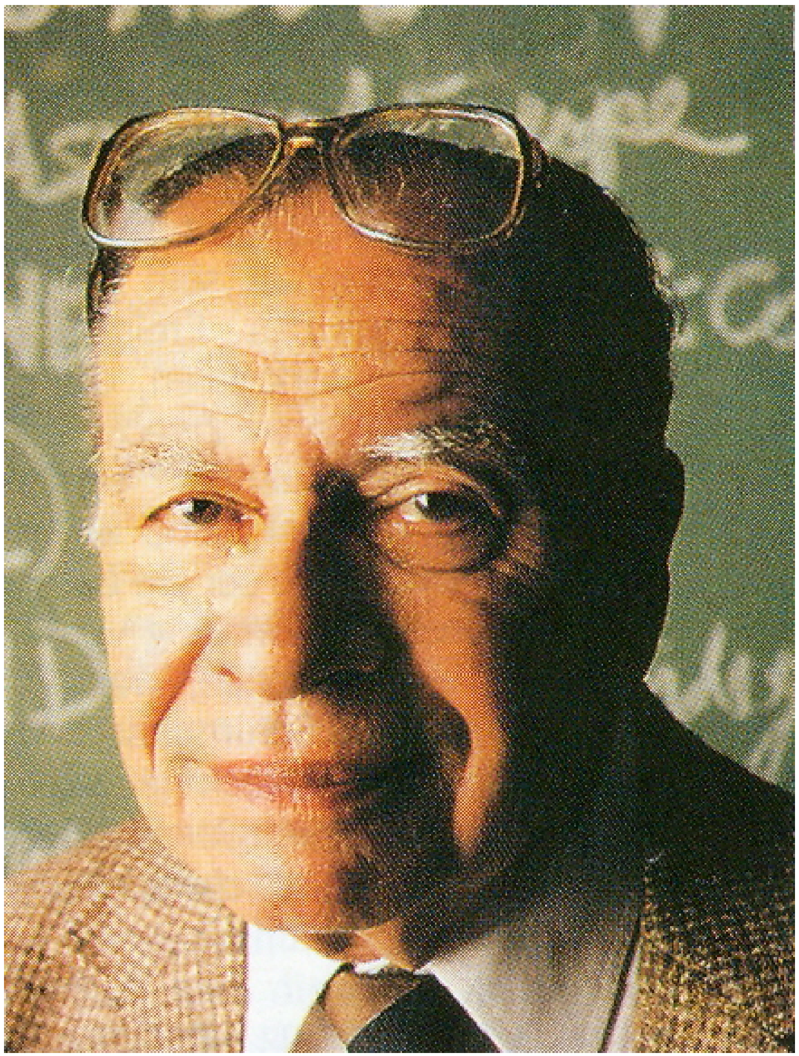
Glottochronology

- Disadvantes
  - Linguistic relationship is mirrored also (or even mainly) in grammar rather than in vocabulary
  - Swadesh list is not universal (some languages lack words for *if, five, smell* etc.)
  - Identification of cognates is questionable if the history of the languages are not known
  - Half-time is not constant, but depends on several factors like intensity of language contact, taboo, literary tradition, national pride, ...

Summary

- „classical“ lexicostatistics & glottochronology
  - Interesting approach
  - Based on questionable background assumptions though
  - Insufficient mathematical foundation to correctly assess the role of chance => statistics

Joseph Greenberg (1915 - 2001)



- One of the most important linguists of the 20th century
- Founder of linguistic typology
- Pathbreaking investigations to linguistic universals
- Classification of African and American languages (controversial)

Mass lexical comparison

- Greenberg:
  - Statistical reliability of lexicostatistics can be improved if the number of data points are increased
  - Rather than comparing two languages, evaluation of the basic vocabulary of entier **language groups**

Mass lexical comparison

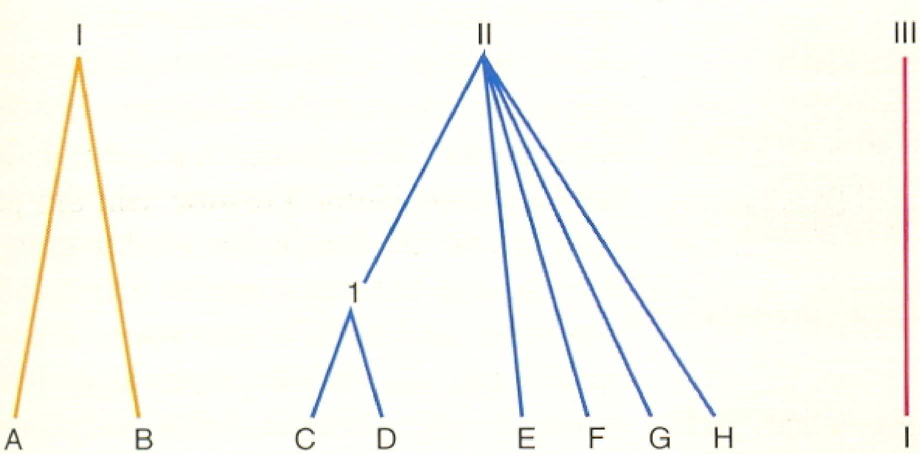
Eine Klassifikationsübung nach der vergleichenden Methode à la Merritt Ruhlen:

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	ɪubb	rijl-	fār	yaḥmil-
B	ʃn-	šaloš	-ni	-ka	mi	lo	ʔem	aβ	šen	leβ	regel	ʕak̠bɔr	nośeh
C	duvā́	tráyas	mā́m	tuvám	kás	ná	mātár	pitár-	dant-	hṛd-	pád	muṣ-	bhárati
D	duva	θrāyō	mąm	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

Mass lexical comparison

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	ɩubb	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	mąm	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



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.

Mass lexical comparison

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	nøs	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

JOHNNY JOHNSON NACH ANGABEN VON MERRITT RUHLEN

Africa

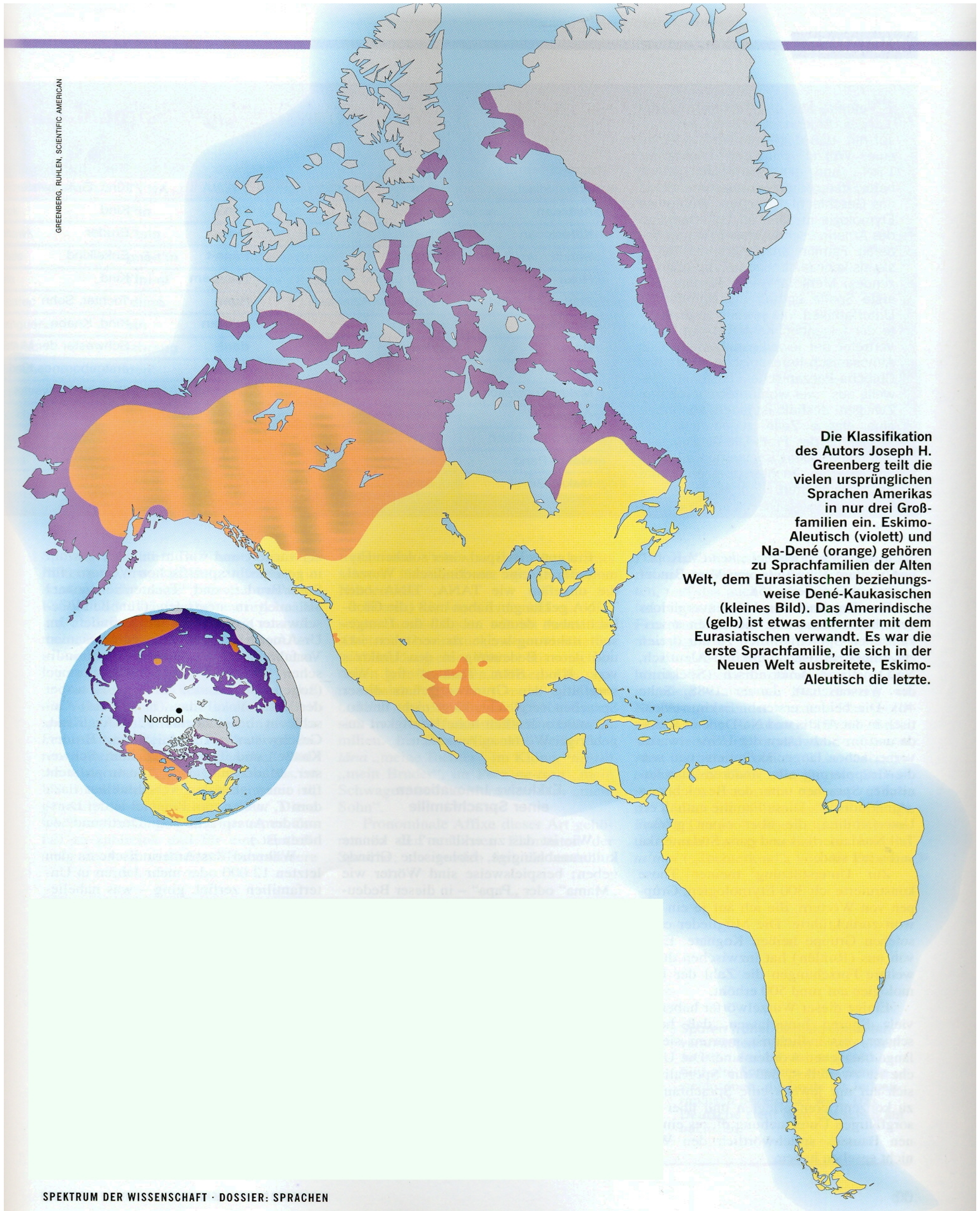
- Greenberg 1963 „The languages of Africa“
  - Only four language families in Africa
    - Afroasiatic (replaces traditional. „Hamito-Semitic“)
    - Niger-Kongo
    - Nilosaharian
    - Khoisan
  - Not uncontroversial, but largely accepted

Pazific/Oceania

Location of Indo-Pacific languages: Kusunda (K), Andaman Islands (A), Halmahera (H), Timor-Alor-Pantar (T), New Guinea (NG), New Britain (NB), New Ireland (NI), Solomon Islands (SI), Santa Cruz Islands (SC), Rossel Island (R), Tasmania (TS)

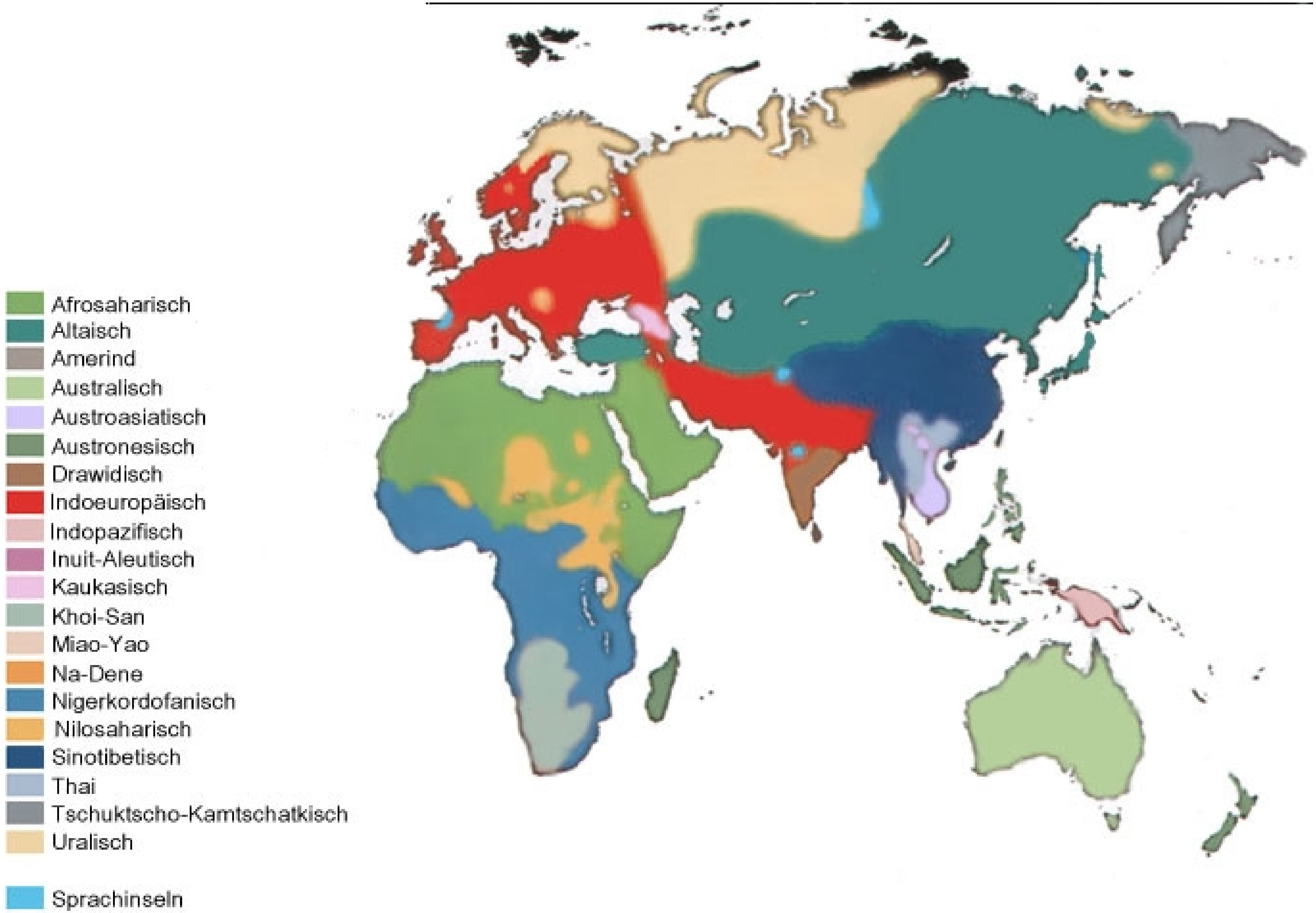
- 1971: „The Indo-Pacific Hypothesis“
  - Indo-pacific macro family
  - Comprises Papua languages, andamanian and tasmanian languages
- Nowadays generally rejected

America



- 1987: „Language in the Americas“
- Three macro-families:
  - Eskimo-Aleut
  - Na-Dené
  - Amerind
- Latter class is heavily contested
- Standard wisdom: Greenberg's Amerind consists of about 200 families

Eurasia


Afrosaharisch
Altaisch
Amerind
Australisch
Austroasiatisch
Austronesisch
Drawidisch
Indoeuropäisch
Indopazifisch
Inuit-Aleutisch
Kaukasisch
Khoi-San
Miao-Yao
Na-Dene
Nigerkordofanisch
Nilosaharisch
Sinotibetisch
Thai
Tschuktscho-Kamtschatkisch
Uralisch
Sprachinseln

Eurasien

- 2000/2002: „Indo-European and Its Closest Relatives“
  - Macro-family „Eurasiatic“
  - sub-families:
    - Indoeuropean
    - uralic languages
    - Altaic (Turkic, Mongolian, Tungusian, Korean, Japanese)
    - Eskimo-Aleut
    - Various isolated languages (like Etruskian)
  - Also highly controversial

Ruhlen

- Merritt Ruhlen
  - Student of Greenberg
  - Even more radical application of Greenberg's method of genetic classification
  - Hypothesis that partial reconstruction of „Proto-World“ is possible – the original language of humankind
  - Generally rejected by experts

Criticism of Greenberg/Ruhlen

- Fundamental objections against the usage of word lists
  - Not culture-independent
  - Assumptions of „universal“ concepts is questionable
- Specific objections against Greenberg's method
  - Very loose interpretation of „semantic correspondence“
  - No serious statistical evaluation
  - Boë et al. 2003: Ruhlen's reconstruction of Proto-World is based on statistically non-significant data (similarities could be due to chance)
  - Greenberg repeatedly reached valuable results, but so far nobody else managed to apply his method successfully => success perhaps more due to his extremely good intuition than to the validity of his method Intuition als aufgrund einer validen Methode

The genetic family tree of humankind

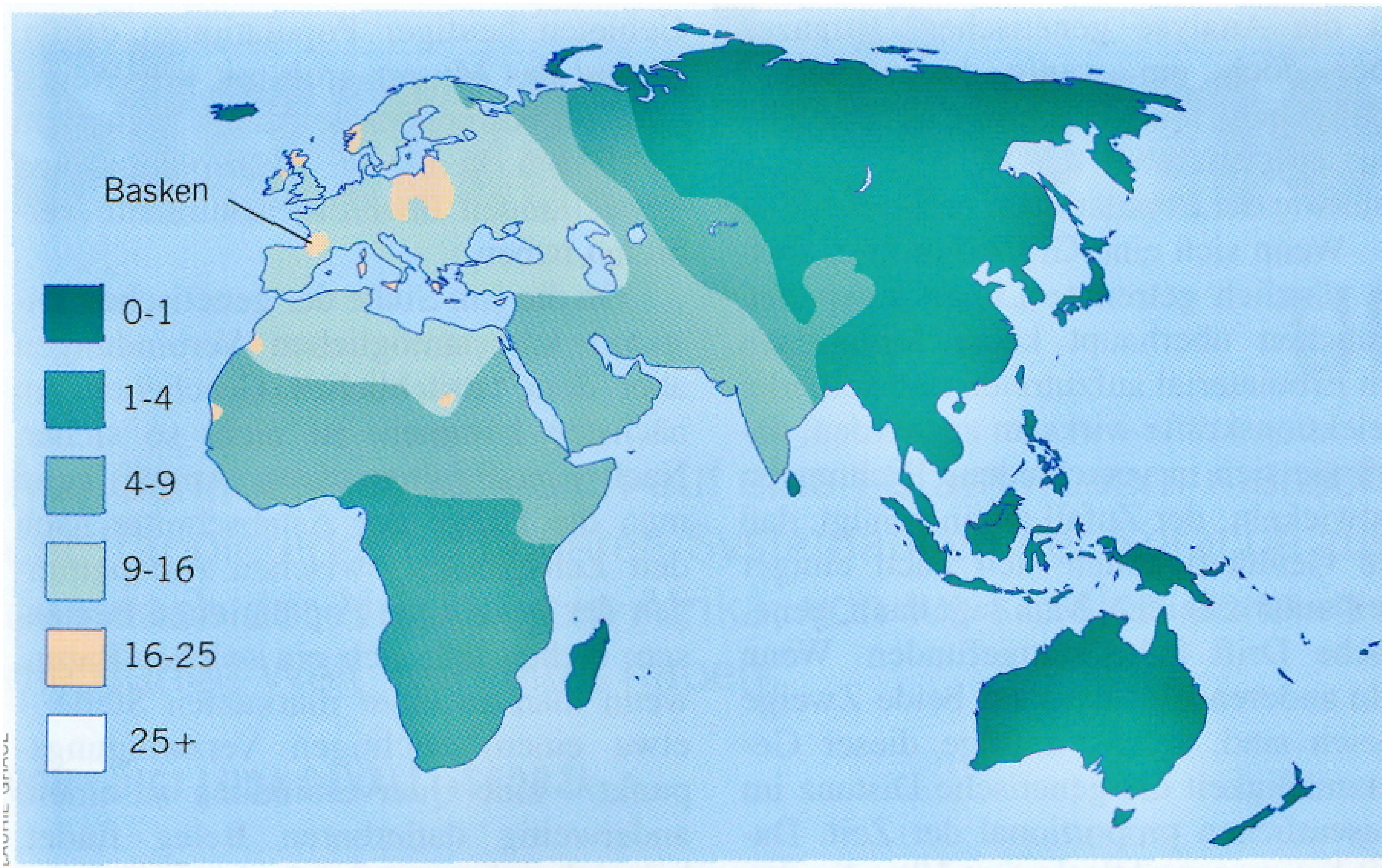
- Luigi Luca Cavalli-Sforza (1922*)
  - Evolutionary biologist (colleague of Greenberg at Stanford)
  - Human Genome Diversity Project
  - Attempt to develop a biological family tree of modern humankind with the help of genetic analysis

Cavalli-Sforza

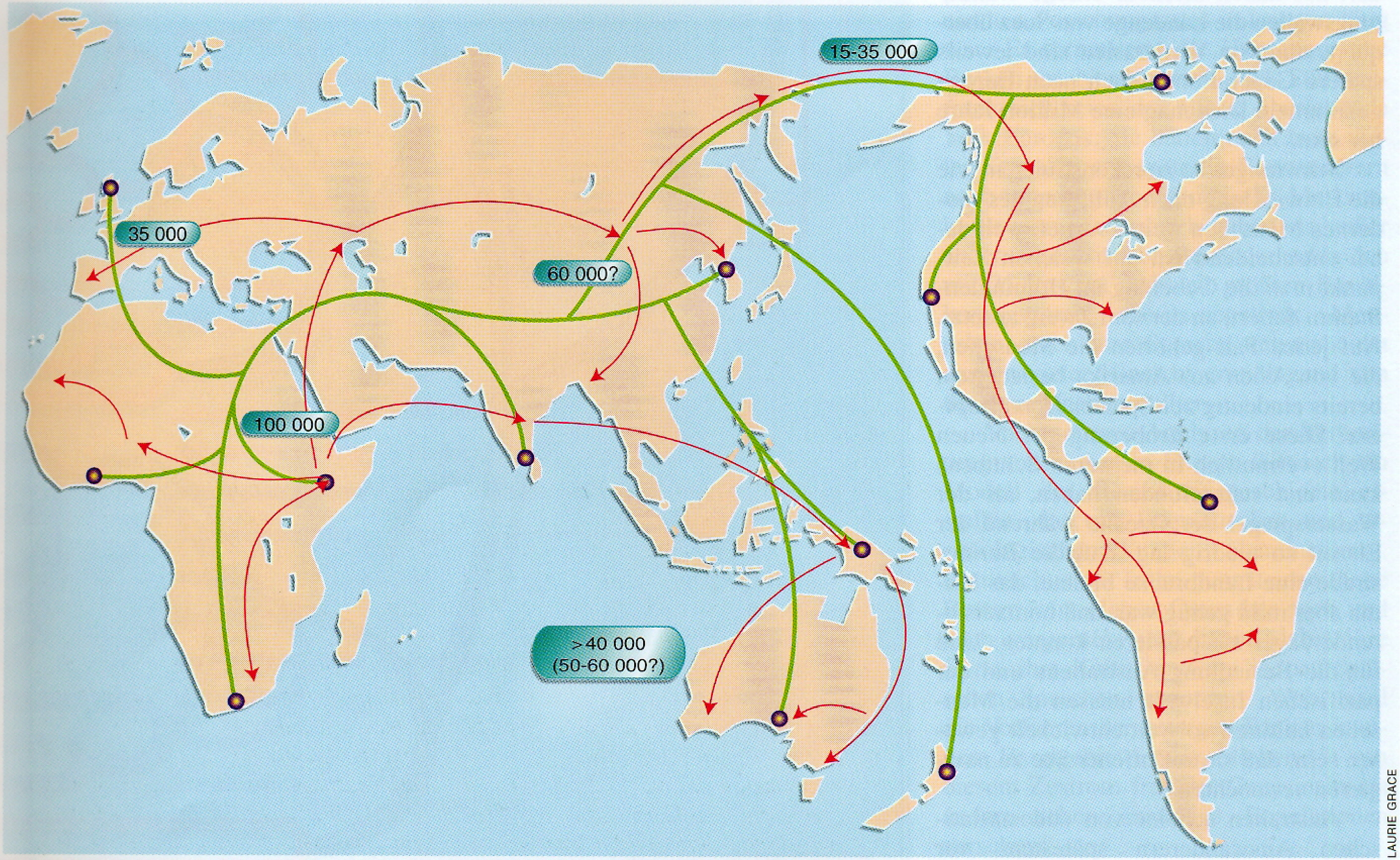
- Basic idea:
  - Evolution is based on **natural selection**: alleles that increase fitness (= ability to survive and replicate) will spread in the population
  - Variation develops via genetic **mutations**
  - Many mutations do not have any effect on phenotype
  - Therefore no selective pressure
  - Whether or not such a neutral mutation will spread is due to chance => so-called **genetic drift**

Cavalli-Sforza

- example:
  - Rhesus factor of the blood
  - Has no impact on fitness
  - heritable
  - Connected populations have a specific percentage of Rh-
  - If a population is separated, these values drift apart
  - Difference in percentage of Rh- is thus a crude measure of the relatedness of populations
  - Cavalli-Sforza used several hunders of such neutral genetic markers



Anteile der Menschen mit dem Blutfaktor rhesus-negativ in der Alten Welt und in Australien. Bei der baskischen Bevölkerung kommt er am häufigsten vor; nach Osten und Süden zu wird er immer seltener. Die Basken scheinen eine sehr alte Bevölkerung zu sein, die erst spät Kontakt zu Einwanderern aus dem asiatischen Raum bekam und deshalb viele ihrer ursprünglichen Merkmale bewahrt hat, so auch ihre Sprache. Die Zahlen geben den Prozentsatz rhesus-negativer Menschen an.



Rekonstruktion der Ausbreitung des Menschen in vorgeschichtlicher Zeit. Ein erster genetischer Stammbaum (grün) wurde derart auf eine Weltkarte projiziert, daß die Endpunkte der Zweige in den heutigen Regionen der einzelnen Populationen liegen. Das Ergebnis paßt recht gut zu einer Rekonstruktion nach archäologischen und fossilen Funden (die Zahlen bezeichnen das Auftauchen des anatomisch modernen Menschen in Jahren vor der Gegenwart). Neuere genetische Untersuchungen (rot) lassen vermuten, daß der *Homo sapiens* auf zwei Routen nach Asien gelangte; die Details der Wege beruhen aber auf Spekulation.

Cavalli-Sforza

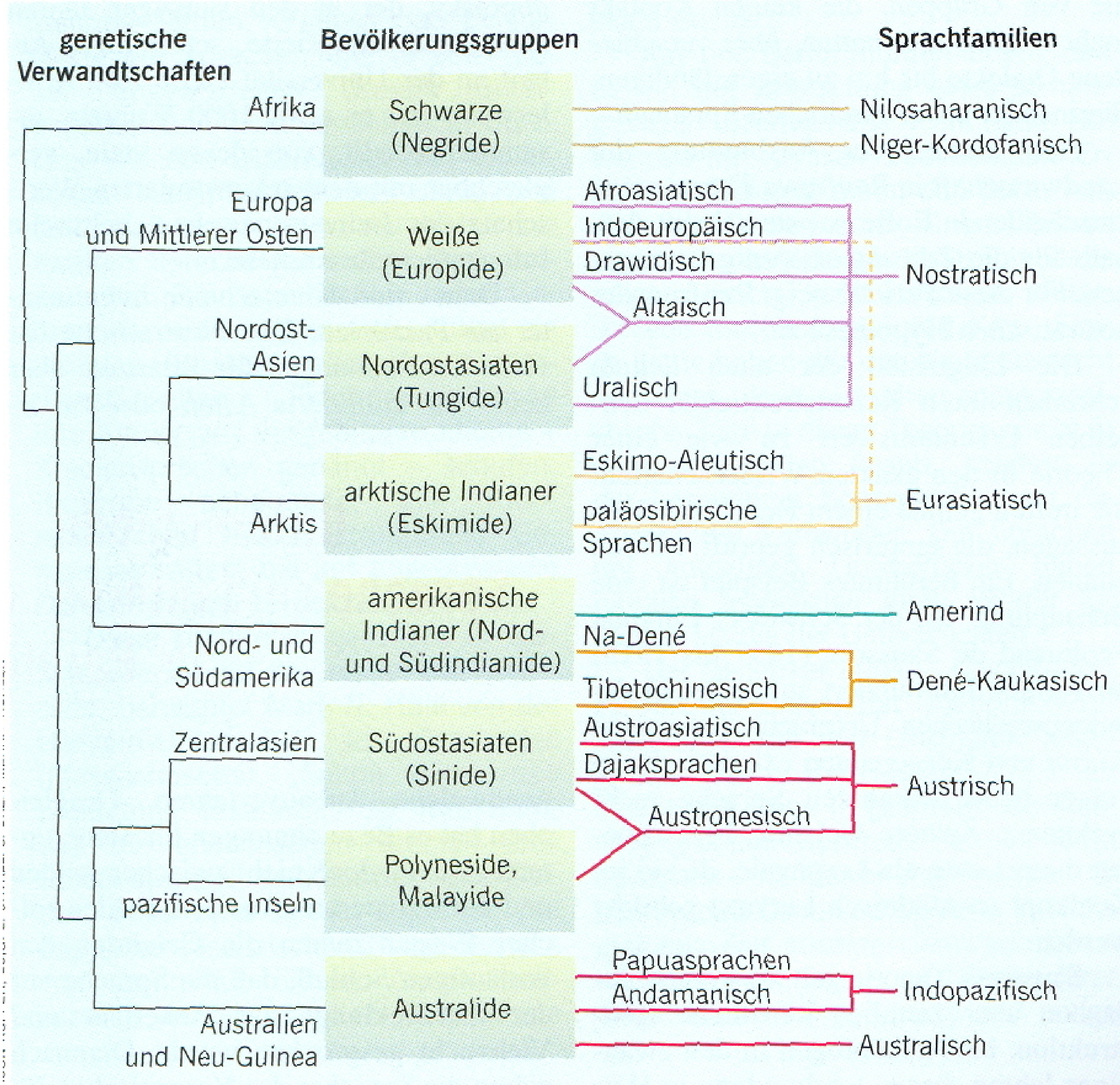
- Project was partially criticized because it allegedly revived the biological concept of human races
- Criticism is not valid, quite the contrary:
  - Results show that there are no human races in the biological sense (of compete reproductive isolation)
  - Split of human sub-populations occured a – according to evolutionary standards – very short time ago: about 60,000 years
  - Genetic variation within a population is sometimes larger than between populations
  - Visible features like skin color, hair consistency etc. are determined by small number of genes

Cavalli-Sforza & Greenberg

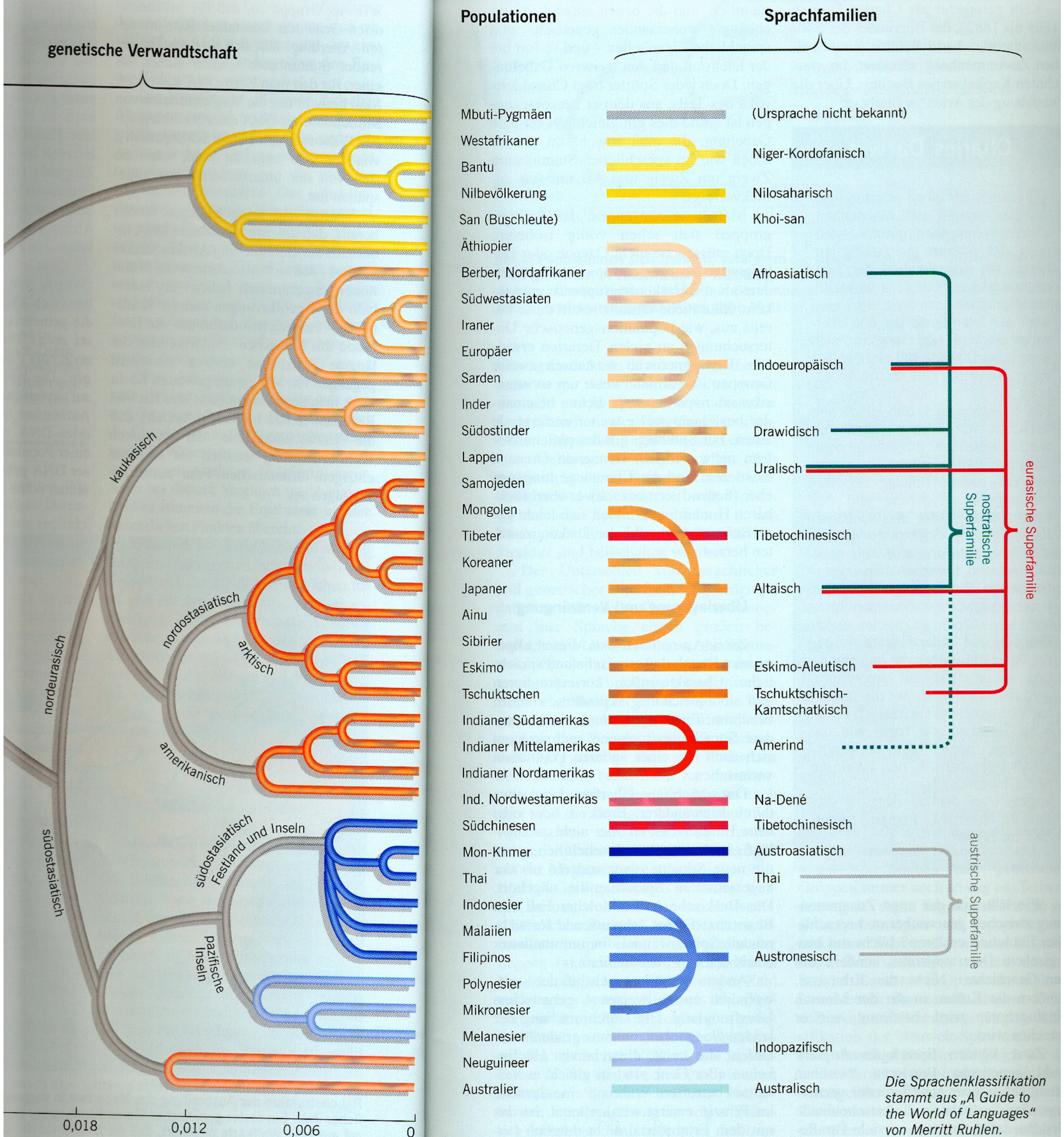
- Surprisingly good correspondence between Cavalli-Sforza's and Greenberg's classifications of human populations
- Additional argument in favor of Greenberg's results

Cavalli-Sforza & Greenberg

- Minor divergences:
  - Hungarians speak an Uralic language, but are genetically indistinguishable from neighboring populations
  - Sami also speak Uralic language, but are genetically closer related to other Skandinavians and mongoloid siberians
  - Ethiopians speak Afro-Asiatic language, but are genetically closer related to Southern African than to Northern African populations


genetische Verwandtschaften
Bevölkerungsgruppen
Sprachfamilien
Afrika
Schwarze (Negride)
Nilosaharanisch
Niger-Kordofanisch
Europa und Mittlerer Osten
Weiße (Europide)
Afroasiatisch
Indoeuropäisch
Drawidisch
Altaisch
Uralisch
Nostratisch
Nordost-Asien
Nordostasiaten (Tungide)
Arktis
arktische Indianer (Eskimide)
Eskimo-Aleutisch
paläosibirische Sprachen
Eurasiatisch
Nord- und Südamerika
amerikanische Indianer (Nord- und Südindianide)
Amerind
Na-Dené
Tibetochinesisch
Dené-Kaukasisch
Zentralasien
Südostasiaten (Sinide)
Austroasiatisch
Dajaksprachen
Austronesisch
Austrisch
pazifische Inseln
Polyneside, Malayide
Australien und Neu-Guinea
Australide
Papuasprachen
Andamanisch
Indopazifisch
Australisch

Der biologische und der linguistische Stammbaum



Die Sprachenklassifikation stammt aus „A Guide to the World of Languages" von Merritt Ruhlen.

…tische Stammbaum (links) und der linguistisch ermittelte Sprachenstammbaum (rechts) heute lebender Völker in der Gegenüberstellung

Using methods from bio-informatics

- problem of reconstructing language trees is similar to problems in evolutionary biology
- bio-informaticians use statistical techniques to induce family relationships between groups of DNA or protein sequences
- these methods are increasingly being applied to linguistic data

Gray & Atkinson (2003) on reconstructing the indo-european tree

- data: 200-word Swadesh list of 95 indo-euroean languages (quality is heavily contested, see this Wikipedia entry)
- estimation both of most likely tree and most likely time depth of branching nodes
- method is still experimental but has high potential in the future
- main problem: how to distinguish true cognates from borrowings

Sources

- http://www.christianlehmann.eu/ling/wandel/Glottochronologie.html
- Die Evolution der Sprache, Spektrum der Wissenschaft -- Dossier 1/2000
- Ruhlen, M., On the Origin of Languages: Studies in Linguistic Taxonomy. Stanford University Press, 1996.
- Ruhlen, M., The Origin of Language: Tracing the Evolution of the Mother Tongue. Wiley, 1996.
- Wade, N., A biological dig for the roots of language, New York Times, March 16, 2004