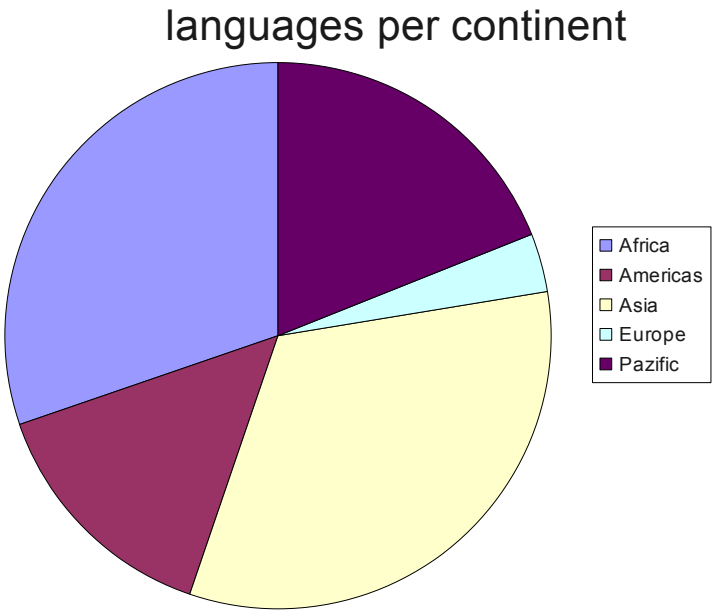
Bioinformatische Methoden in der historischen Linguistik

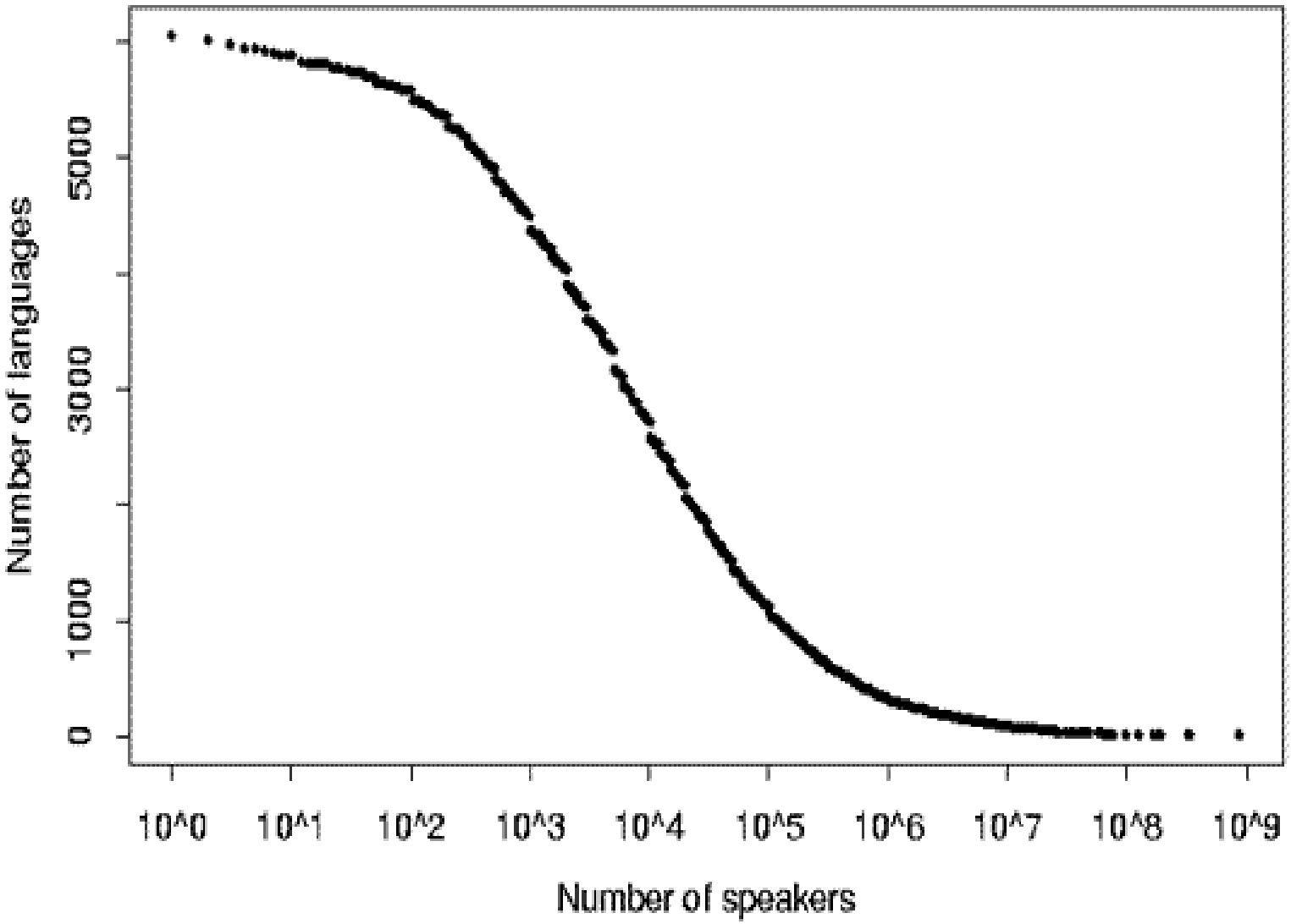
Sprachen und Sprachfamilien:
Etablierte Erkenntnis

Gerhard Jäger
Forum Scientiarum
18. Januar 2013

Introduction

- How many languages are spoken today?
- Ethnologue (2005): 6 912: table 1
- Number of speakers varies substantially



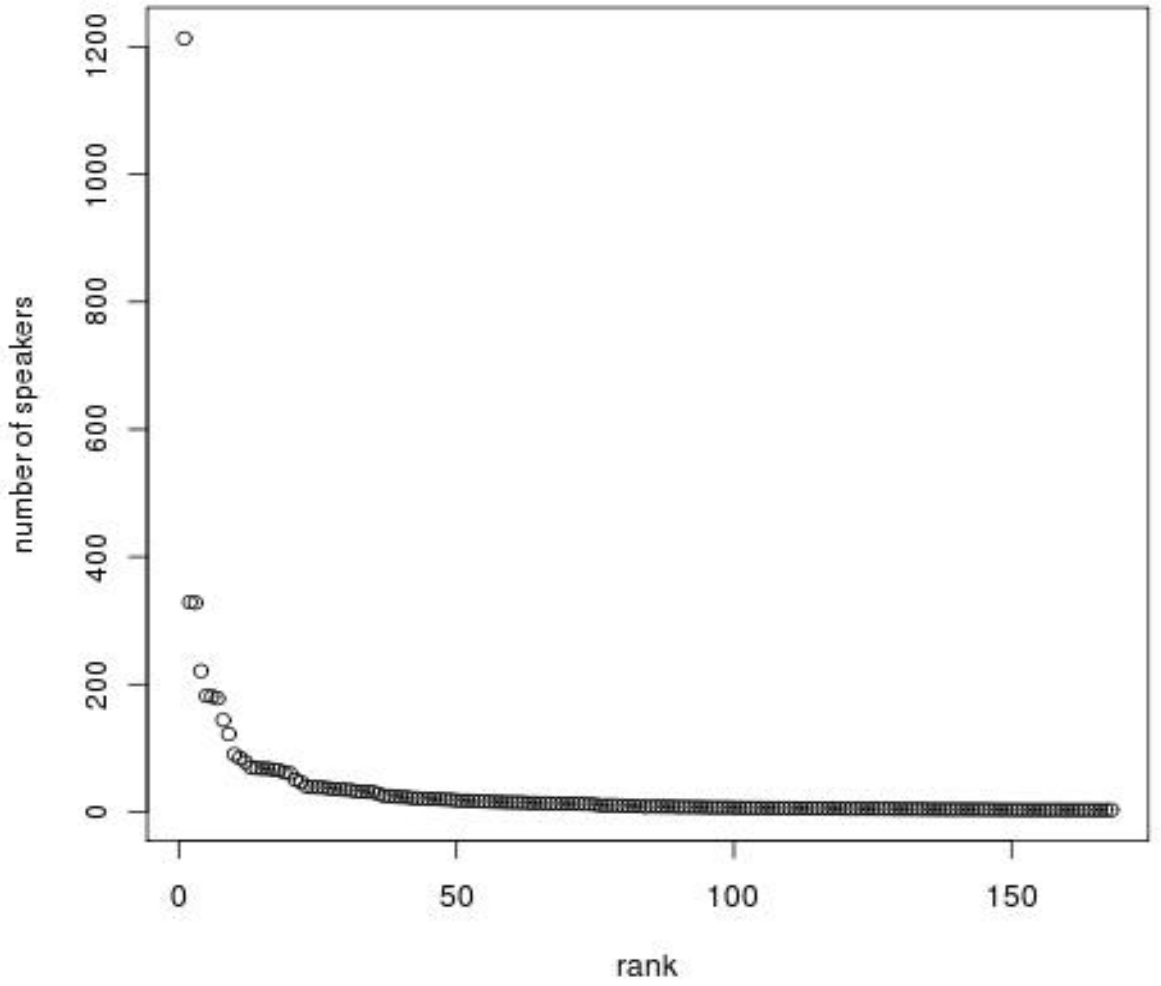


(data from 1999 edition of Ethnologue)

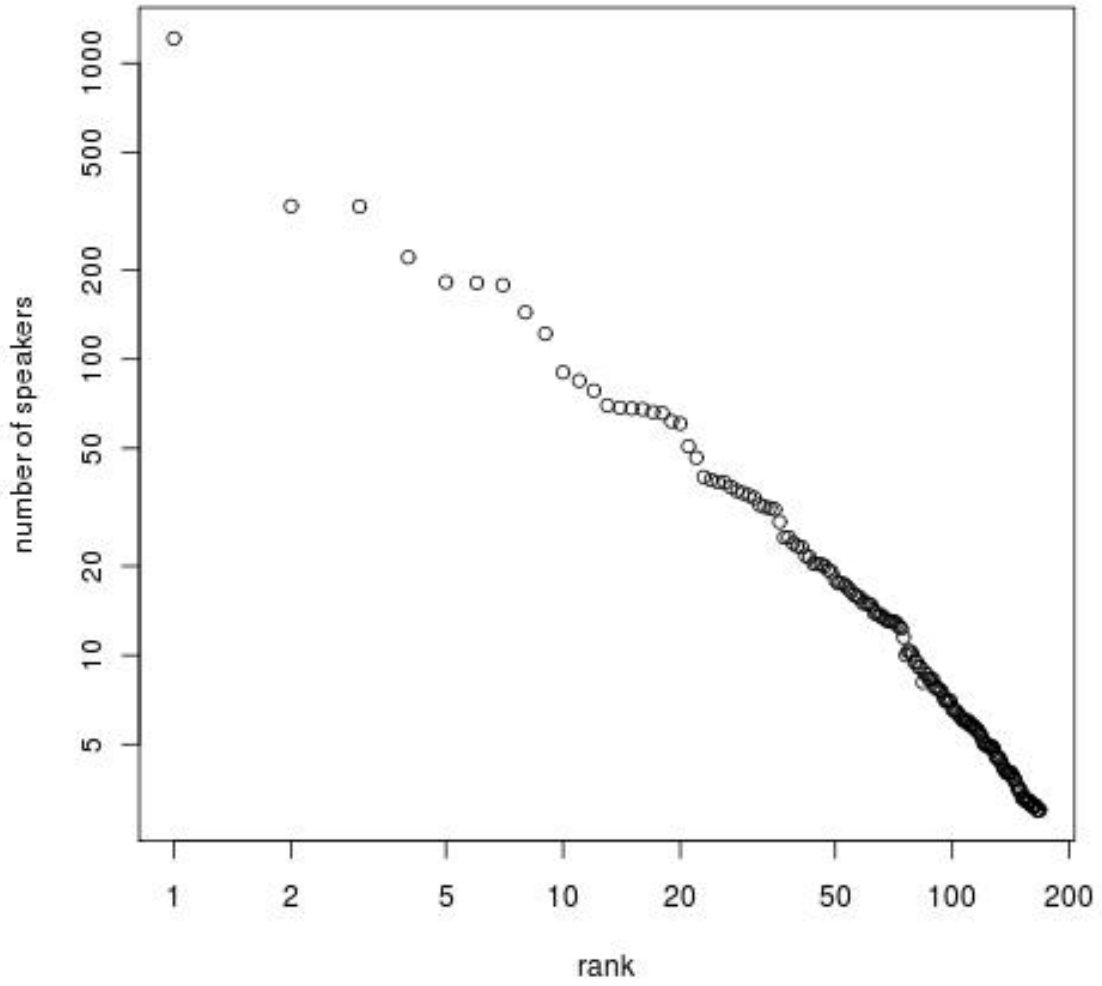
language	number of native speakers (Mill.)
Mandarin	873
Spanish	322
English	309
Hindi	181
Portuguese	177
Bengali	171
Russian	145
Japanese	122
German	95
Wu (China)	77
Javanese	76
Telugu	70
Marathi	68
Vietnamese	67
Korean	67
Tamil	66
French	65
Urdu	61
Yue (Kantonese)	55
Turkish	51

More recent data source

Quantitative distribution



Quantitative distribution



- Zipfian distribution
- Number of speakers is inversely proportional to rank of a language
- Frequent distribution in linguistics/social sciences

What counts as „speaker“?

- 1996 edition of Ethnologue: 266 million speaker of Spanish
- 1999 edition: 322 million
- Does not correspond to population growth
- Data sources are sometimes unreliable

What counts as a language?

- Arabic does not belong to „top twenty“
  - Arabic (including all variants): 202 mill. speaker (would amount to 4th rank)
  - Ethnologue treats different variants of Arabic as different languages
  - Justification: variants are mutually unintelligible. Algerian and Egyptian Arabic are as different as Spanish and Portuguese.

What counts as a language?

- Hindi and Urdu are the same language
  - History/politics: differernt writing systems, different strata of loan words
  - Regular speakers understand each other fairly well
  - If counted as one language, Hindi/Urdu would be on 4th place.

What counts as a language?

- Depending on how you count, Turkish might have higher number of speakers
  - 51 millionen speakers (46 million in Turkey)
  - However, more than 80 million people speak a language that is mutually intelligible with Turkish
  - Counting them in would bring Turkish to 10th rank

What counts as a language?

- **Serbo-Croatian**
  - Before Balkan wars of the nineties:
    - Serbo-Croation counted as one language
    - Two writing systems – Latin alphabet in Croatia, kyrillic alphabet in Serbia
    - Continuum of dialectal variants
  - Now:
    - Three languages – Serbian, Croation, Bosnian

What counts as a language?

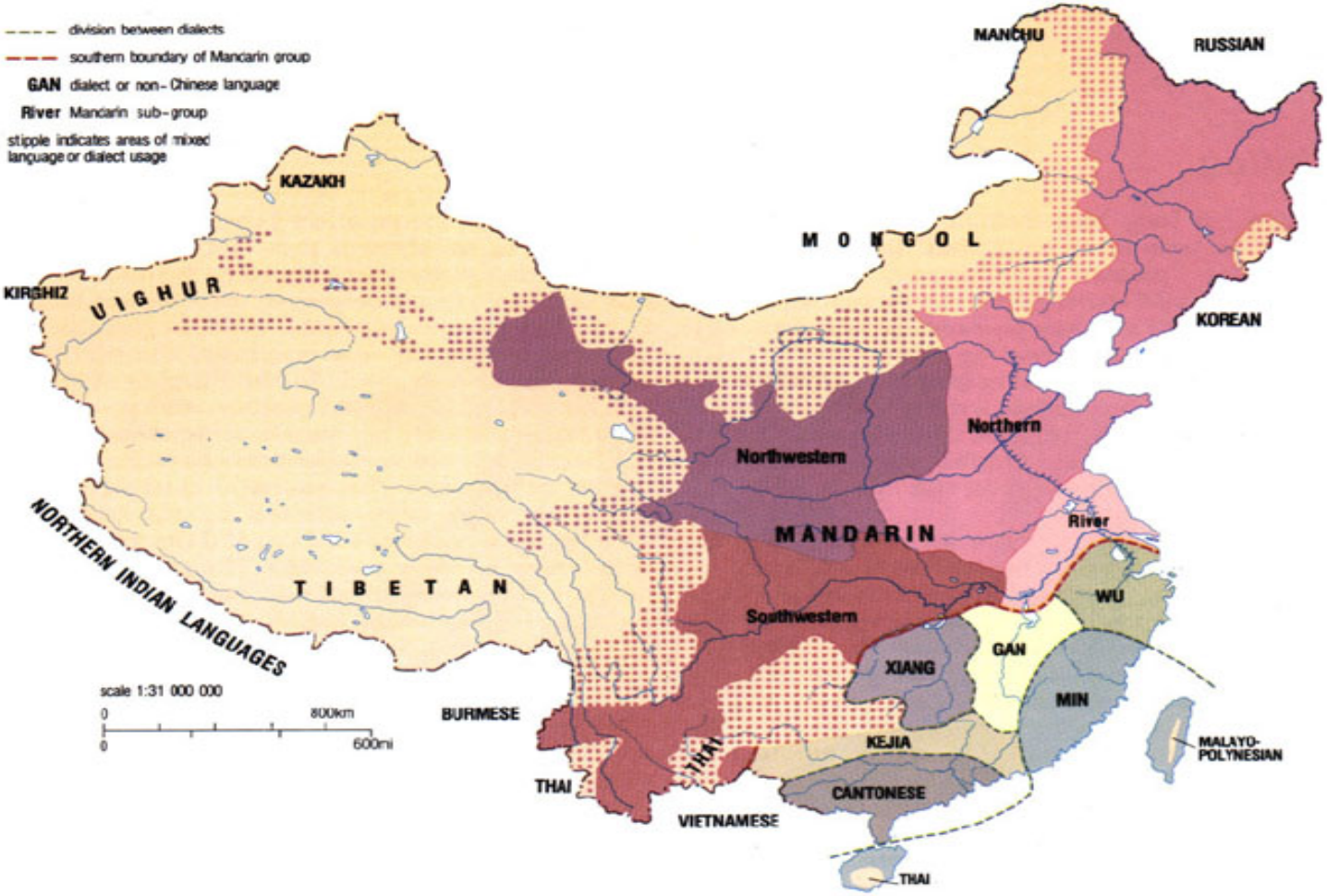
- **Skandinavian**
  - Norwegian and Swedish – and, up to a point, also Danish, are mutually intellibible
  - Count as different languages though, because they are associated with different countries

What counts as a language?

- **Chinese**
  - Is frequently considered a single language
  - Consists of at least seven different languages (with considerable internal dialectal variation)
  - Chinese is considered as a unit for cultural and political reasons, like the common writing system

What counts as a language?

- **Chinese**



What counts as a language

- Dialect continua
  - Portugese, Spanish, French and Italian are counted as different languages
  - Nonetheless, local dialects changes only gradually if you travel from town to town from Portugal to Italy.
  - The same holds for German and Dutch.

What counts as a language

What counts as a language

- Cynically speaking: *A language is a dialect with an army and a navy.*
- Distinction between language and dialect cannot be done by purely linguistic criteria
- In the end, it is a political and cultural decision of a linguistic community about its identity
- Criteria from Ethnologue

Language families

- Languages: no clearly separated unites, rather a hierarchy/tree structure.
  - Categories can be split into ever smaller units, until the level of the single speaker
  - Assumption of a meta-unit is justified if there is evidence for a common origin

Language families

- German belongs to the family of Indo-European
- Sometimes also called (obsolete now) „Indo-Germanic“
- It is the language family that was discovered first and is best studied

The Indo-European language family

- Ancient times: little interest in comparative linguistic research
- Middle ages:
  - Written documents from many European languages
  - Wide-spread assumption that all languages originate from Hebrew
  - No real concept of language change
- Real starting point of comparative linguistics was the discovery of Sanskrit

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.“

The Indo-European language family

- Cœurdoux 1767

Sanskrit	*devah*	„god“	Latin	*deus*	Greek	*theós*
	padam	„foot“		*pes, ped-is*		poús, podo-ós
	maha	„large“				*mégas*
	viduva	„widow“		*viduva*		

- Also grammatical similarity between Greek and Sanskrit
- Partially incorrect according to modern insights (for instance, the Greek cognate to lat. *deus* is *Zeus*, not *theos*

The Indo-European language family

Sanskrit		Latin
as-mi	I am	*s-um*
as-i	you(sg.) are	*es*
as-ti	he is	*es-t*
s-mas	we are	*s-umus*
s-tha	you(pl) are	*es-tis*
s-anti	they are	*s-unt*

The Indo-European language family

- Sanskrit *as-* and lat. *es-* both mean „to be“
- Both have allomorph *s-*
- Inflectional paradigm comprises both variants
- Sanskrit has additional suffix *-i*; otherwise the suffixes are virtually identical

➔ *Sufficient evidence to establish genetic relatedness*

The Indo-European language family

- Reconstructed paradigm of the Indo-European proto language

(V)s-(V)m(i)
Vs-(i)
Vs-t(i)
s-(V)mVs
(V)s-t(h)V
s-Vnt(i)

The Indo-European language family

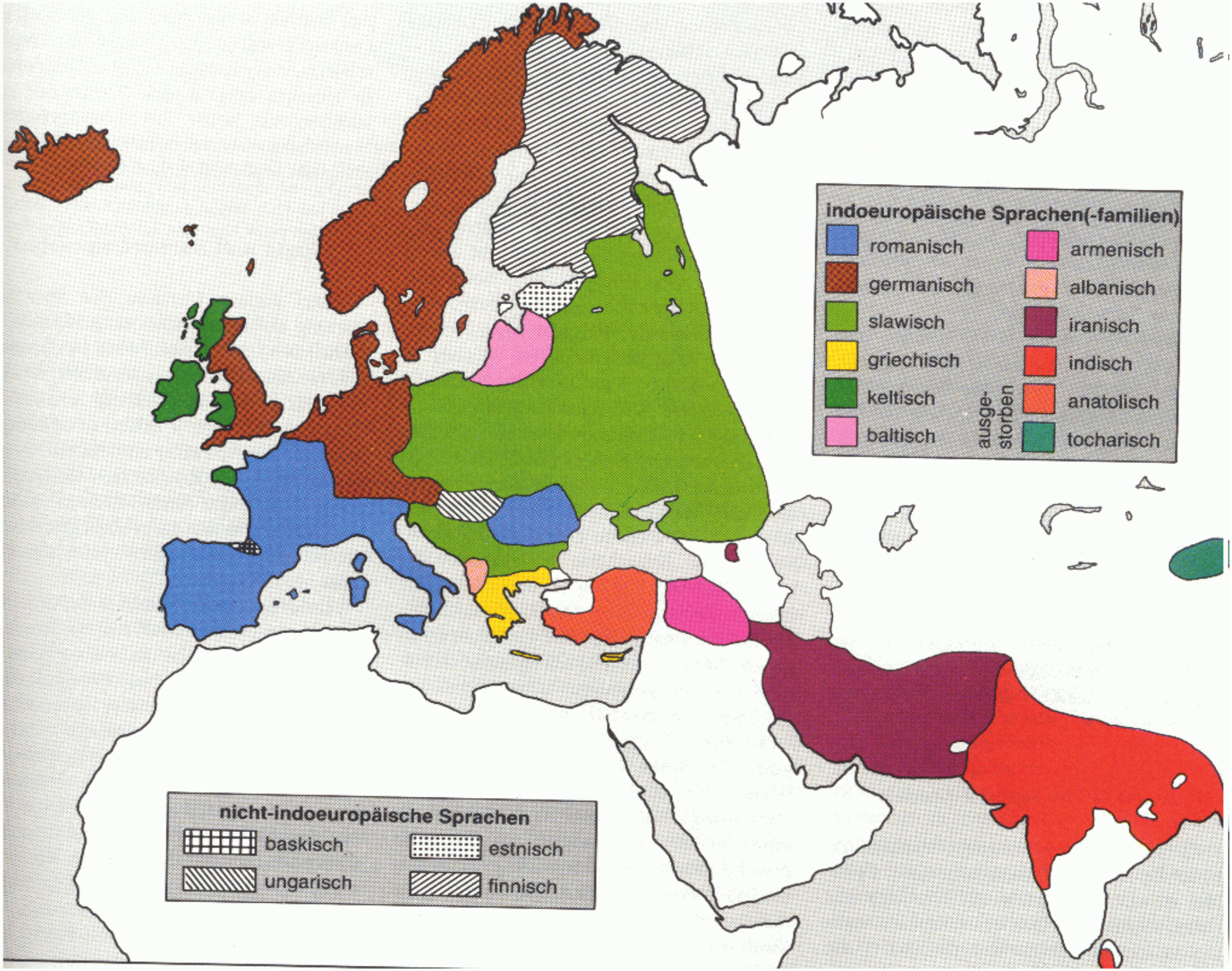
- Middle of 19th century: discovery of **sound laws**
- Phonological change is not arbitrary, but applies essentially to all words of a language
- For instance Grimm's Law (applies to all Germanic languages),
  High German consonant shift (applies to all High-German dialects)

Sound laws and the reconstruction of language families

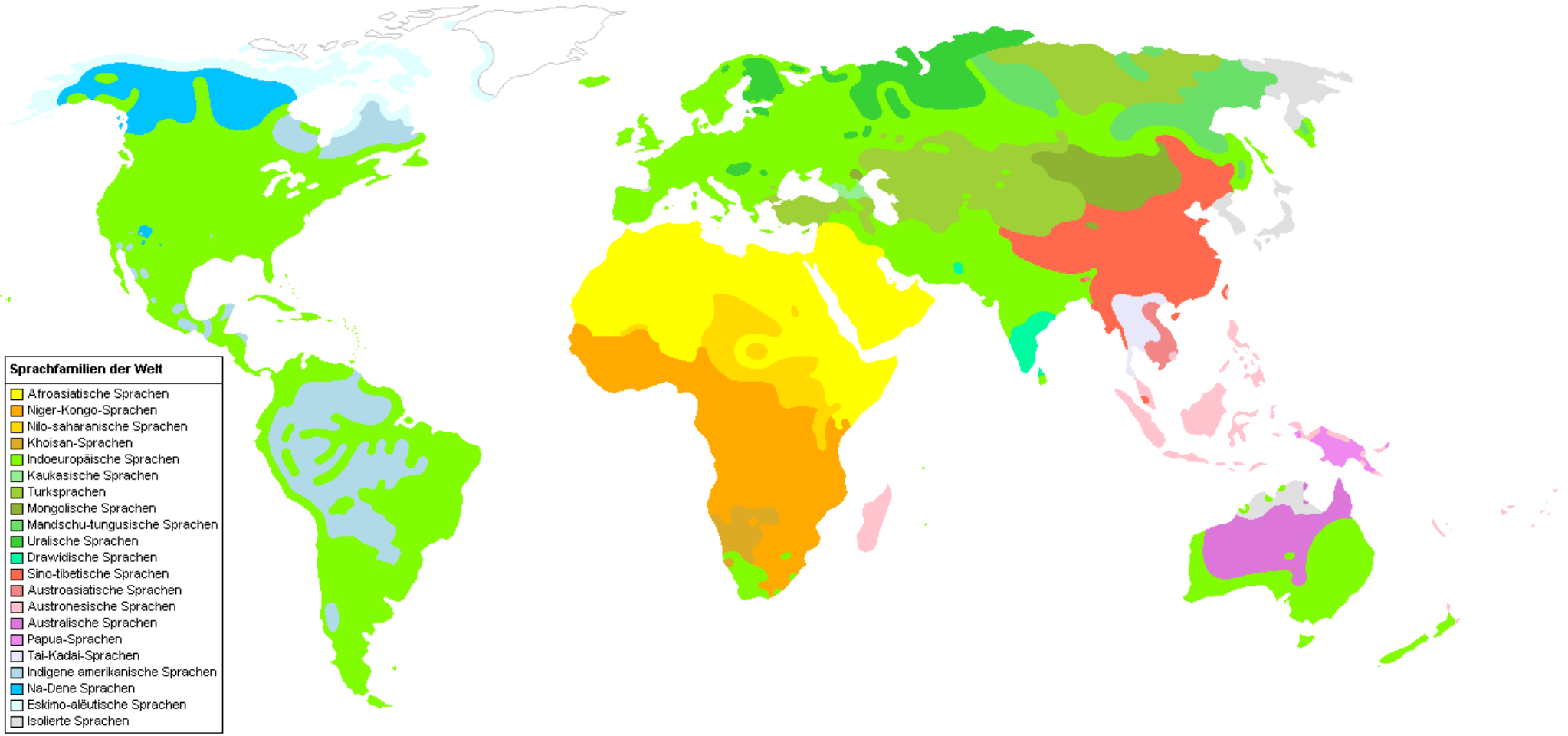
- Applicable to other languages as well (example from Austronesian)
- Reconstruction is usually possible at most until 8,000 years into the past

The Indo-European language family

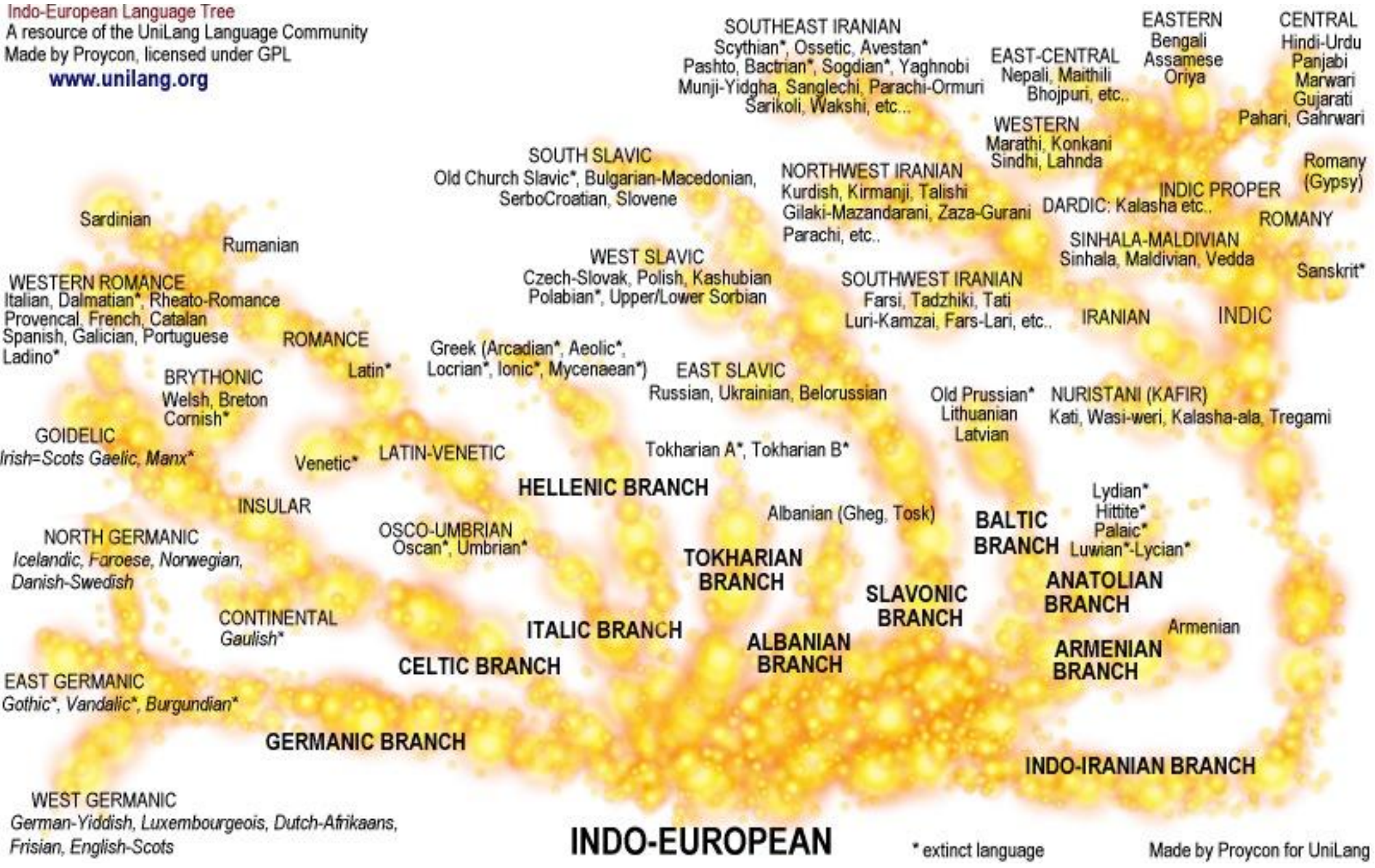
- Modern Indo-European languages are
  - All European languages except Hungarian, Finnish, Estonian, and Basque
  - Many West Asian and South Asian languages


indoeuropäische Sprachen(-familien)
romanisch
germanisch
slawisch
griechisch
keltisch
baltisch
armenisch
albanisch
iranisch
indisch
anatolisch
tocharisch
ausge-storben
nicht-indoeuropäische Sprachen
baskisch
ungarisch
estnisch
finnisch

Distribution of IE languages



Family tree of the IE languages



Branches of the IE family

- **Indo-Iranian**
  - **Indo-Aryan:** Sanskrit, Hindi, Urdu, Bengali, Marathi, Sinhala, …
  - **Iranian:** Avestan, ancient Persian (cuneiform documents), Farsi, Pashto, Kurdish, Balochi, ...
  - **Nuristani:** Kati, Prasuni, Ashkunu, Waigali, Gambiri, … (small languages, mostly spoken in Pakistan/Afghanistan)

Branches of the IE family

- **Tocharian** (extinct):
  - Was spoken in second half of the first millenium in present day China
  - About 5,000 written documents survive

Branches of the IE family

- **Armenian:**
  - Old Armenian, Eastern Armenian, Western Armenian

Branches of the IE family

- **Anatolian languages** (extinct):
  - Hittite, Lydian, Palaic, Luwian, Lycian, Carian, Pisidian, Sidetic
- **Phrygian** (extinct)
- **Thracian** (extinct)
- **Macedonian** (extinct; was spoken during antiquity, unrelated to modern Macedonian, which is a Slavic language)

Branches of the IE family

- **Balto-Slavic**:
  - **Slavic**:
    - **East Slavic:** Russian, Belarussian, Ukrainian, Ruthenian
    - **West Slavic:** Sorbian (Upper Sorbian, Lower Sorbian), Polabian (extinct), Polish, Pomeranian (Kashubian, Slovincian (extinct)), Czech, Slovak
    - **South Slavic:** Burgenland Croatian, Bosnian, Croatian, Molise Croatian, Macedonian, Montenegrin, Serbian, Slovenian

Branches of the IE family

- **Balto-Slavic**:
  - **Baltic:**
    - **Eastern Baltic:** Lithuanian, Latvian, Curonian, Selonian (extinct), Semigallian (extinct)
    - **Western Baltic (extinct):** Old Prussian, Sudovian, Galindian, Skalvian

Branches of the IE family

- **Hellenic**
- **Albanian**
- **Illyrisch** (extinct)
- **Venetic** (extinct)
- **Lusitanian** (extinct)

Branches of the IE family

- **Celtic**:
  - **Continental Celtic (extinct):** Gaulish, Galatian, Lepontian, Celtiberian
  - **Insular Celtic:**
    - British languages: Cumbric (extinct), Welsh, Cornish (extinct), Breton
    - Goidelic languages: Irish, Scottish Gaelic, Manx

Branches of the IE family

- **Germanic:**
  - **East Germanic** (extinct): Burgundian, Vandalic, Gothic
  - **North Germanic**: Norwegian, Faroese, Jamtlandic, Norn (extinct), Swedish, Danish, Gutnish
  - **West Germanic:** English, Scots, Frisian, Dutch, Low German, German, Swiss German, Yiddish, ...

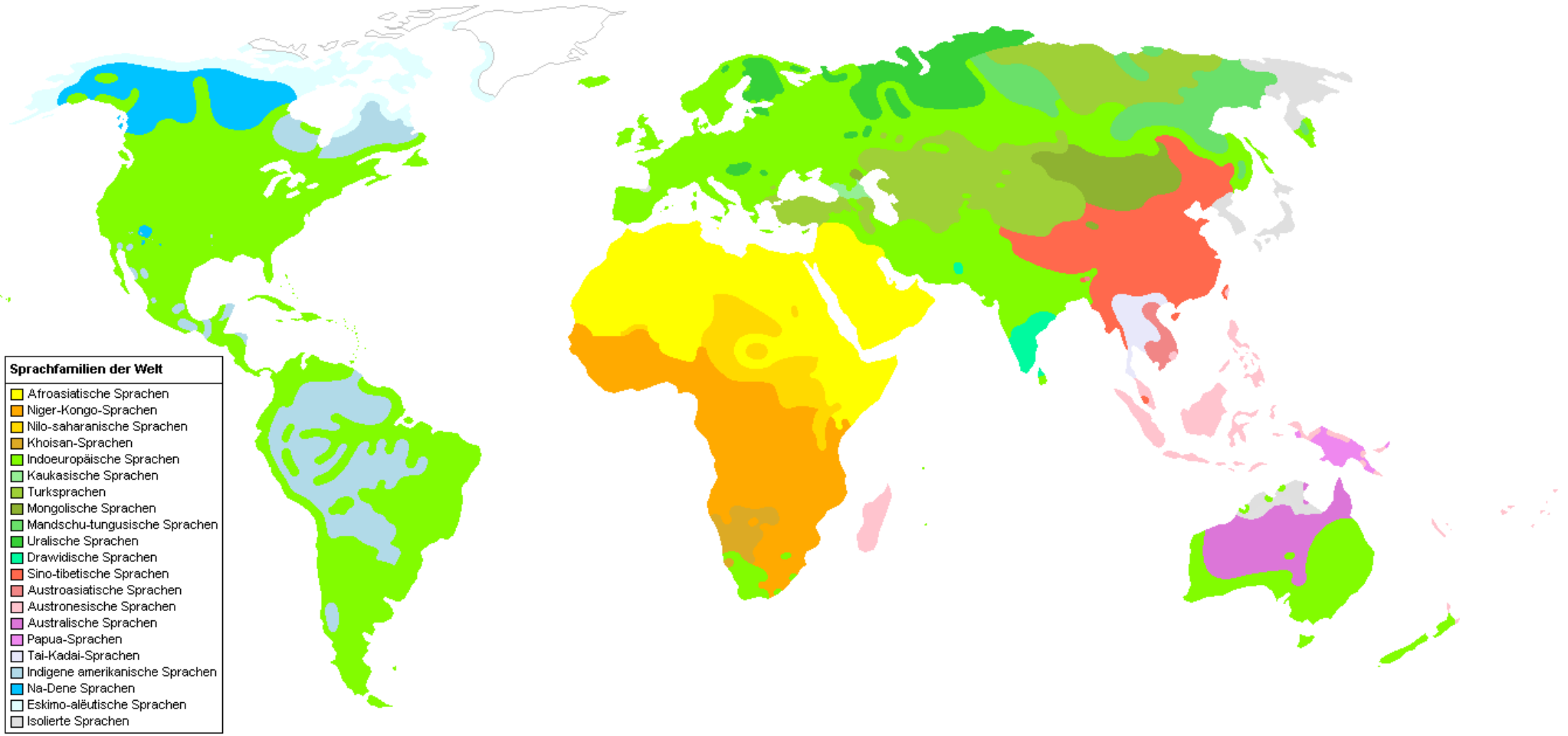
Branches of the IE family

- **Italic:**
  - **Latino-Faliscan:** Latin (extinct), Faliscan (extinct), Spanish, Portuguese, French, Italian, Romanian, Moldovan, Catalan, Galician, Occitan, Sardinian, Ladin, Romansh
  - **Osco-Umbrian** (extinct)

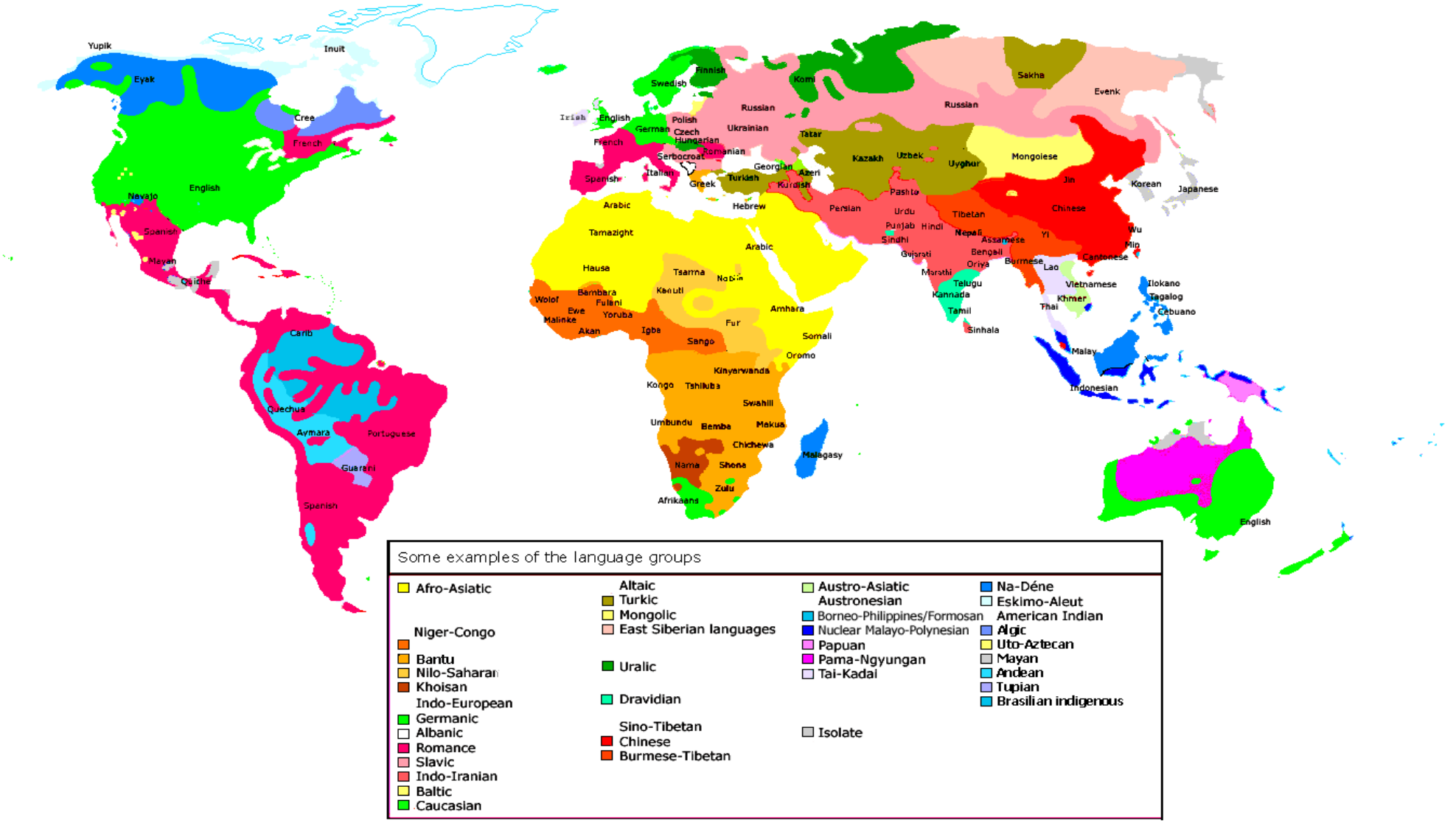
Language families

- Language family: group of genetically (i.e. historically) related languages
- Descent from a common proto language
- Descent has to be established via generally accepted methods
- Classification is (unavoidably) variable and sometimes subjective
- Ethnologue counts more then 100 language families

Language families



Language families



Some examples of the language groups

- Afro-Asiatic
- Niger-Congo
  - Bantu
  - Nilo-Saharan
  - Khoisan
- Indo-European
  - Germanic
  - Albanic
  - Romance
  - Slavic
  - Indo-Iranian
  - Baltic
  - Caucasian
- Altaic
  - Turkic
  - Mongolic
  - East Siberian languages
- Uralic
- Dravidian
- Sino-Tibetan
  - Chinese
  - Burmese-Tibetan
- Austro-Asiatic
- Austronesian
  - Borneo-Philippines/Formosan
  - Nuclear Malayo-Polynesian
  - Papuan
  - Pama-Ngyungan
  - Tai-Kadai
- Isolate
- Na-Déne
- Eskimo-Aleut
- American Indian
  - Algic
  - Uto-Aztecan
  - Mayan
  - Andean
  - Tupian
  - Brasilian indigenous

Language families

- **Afro-Asiatic**
  - Also called „Hamito-Semitic“ (obsolete)
  - subgroups:
    - Semitic (Arabic, Hebrew, Amharic, ...)
    - Berber (Tuareg, ...)
    - Egyptian (extinct)
    - Cushitic (Somali, Oromo, ...)
    - Chadic (Hausa, ...)

Sprachfamilien

- **Nilo-Saharan**
  - Comprises about 200 African languages
  - Nubian, Fur, ...

Sprachfamilien

- **Niger-Congo languages**
  - Most important subgroup: Bantu languages
  - Swahili, Rwanda, Zulu, Yoruba

Sprachfamilien

- **Khoisan languages**
  - Languages of the bushmen in Southern Africa
  - Use click sounds (which are typologically uncommon)

Language families

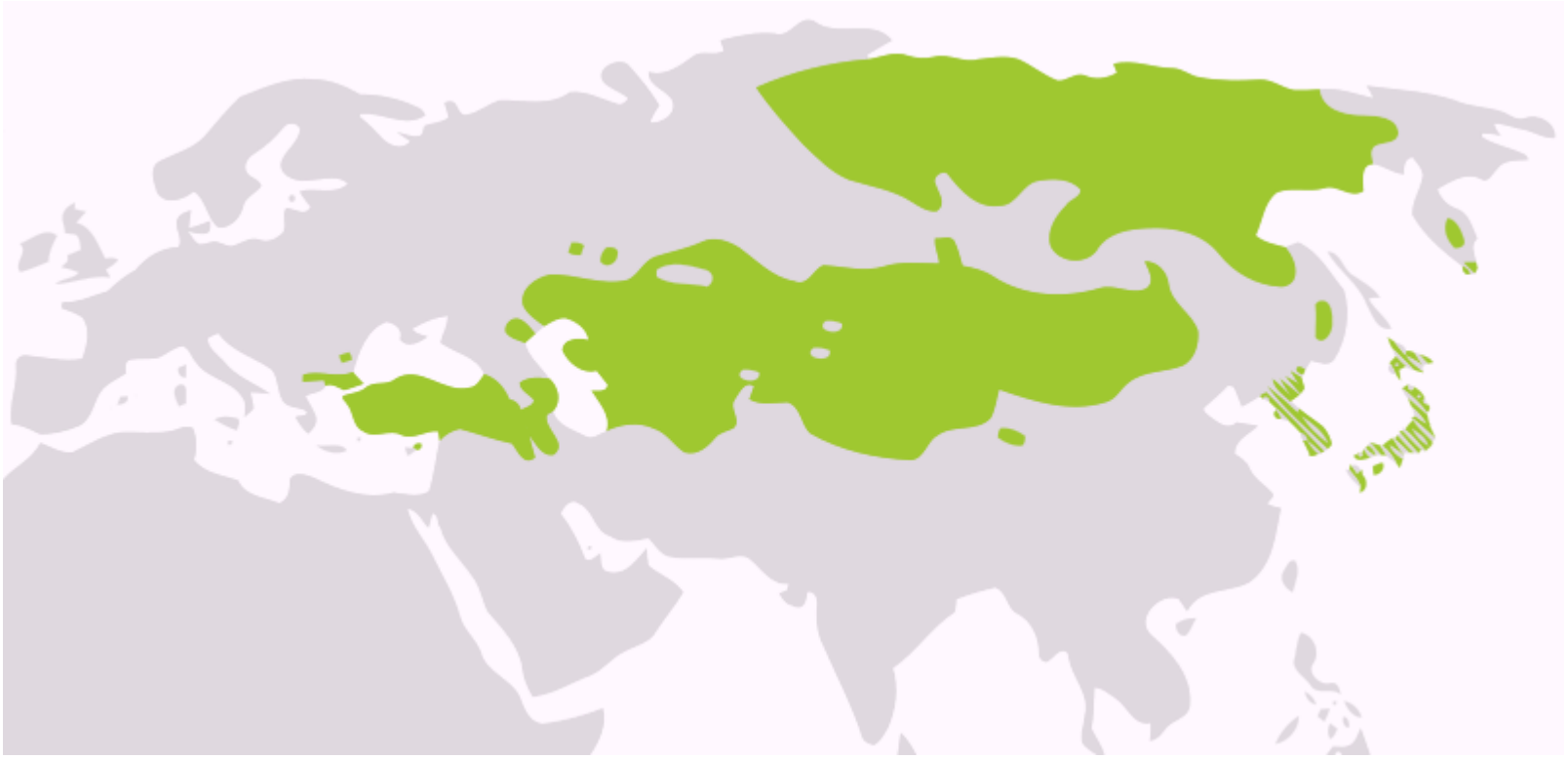
- **Uralic**
  - subgroups
    - Finno-ugric: Hungarian, Estonian, Sami, Karelian
    - Samoyedic (< 30,000 speaker in Nothern Eurasia)

Language families

- **Altaic**
  - subgroups
    - Turkic: Turkish, Turkmen, Kyrgyz, Kazakh
    - Mongolic
    - Tungusic (Northern China, East Siberia)
    - Korean
    - Japanese
  - Partially controversial, especially the inclusion of Korean and Japanese

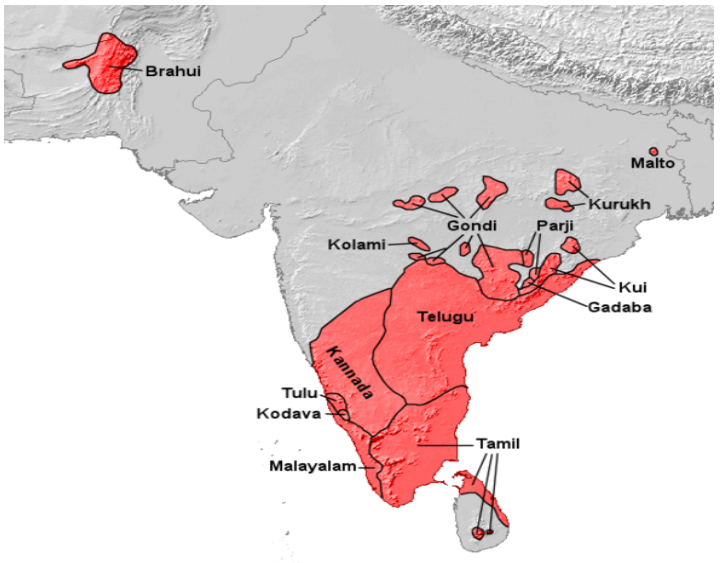
Language families

- **Altaic**



Language families

- **Dravidian**
  - Telugu, Tamil, Kannada, ...
  - Spoken mainly in Southern India and Sri Lanka



Language families

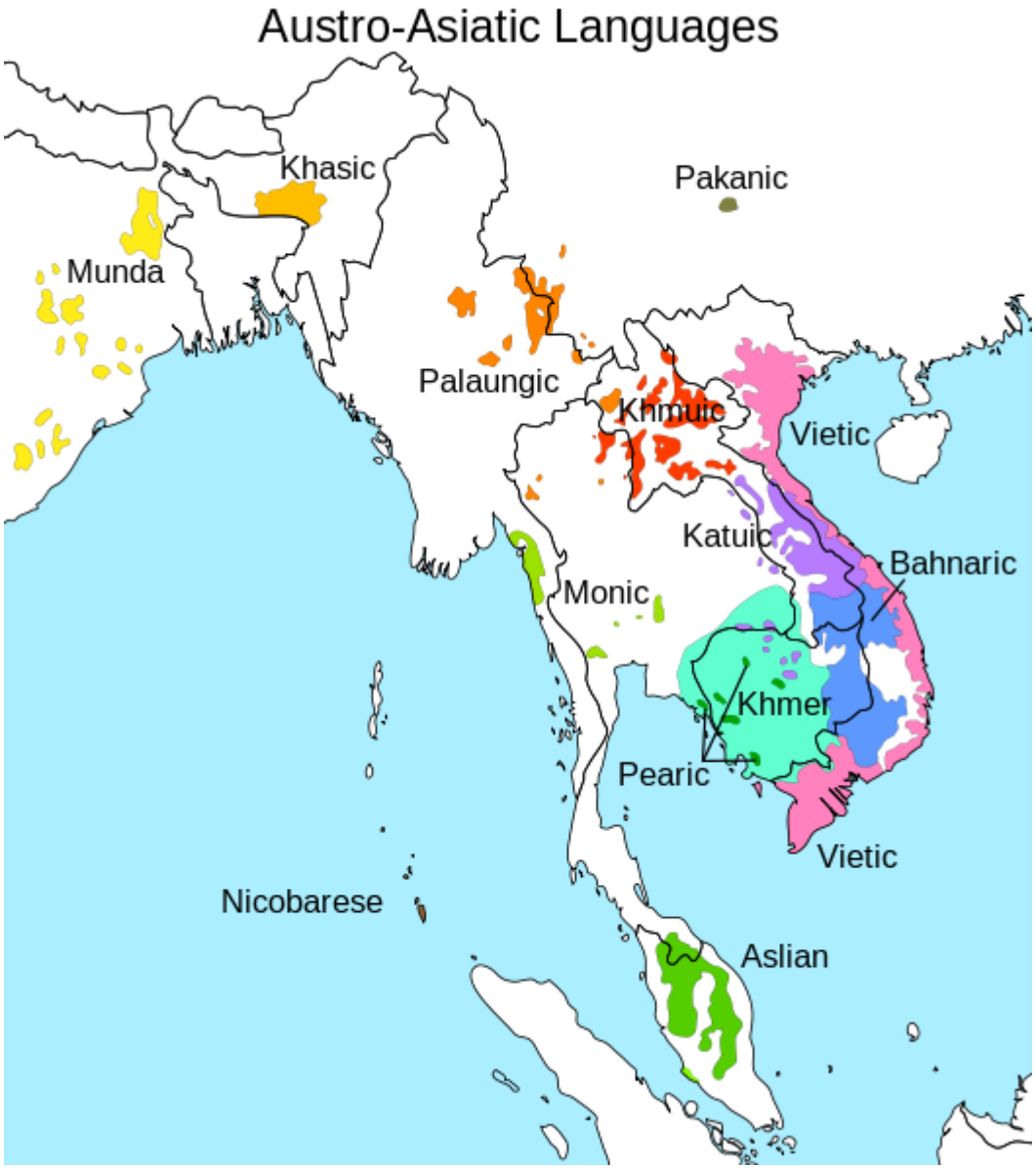
- **Sino-Tibetan**
  - subgroups
    - Sinitic (chinese languages)
    - Tibeto-Burman (spoken in Myanmar, Northern Thailand, Nepal, Bhutan, parts of China, India and Pakistan): Tibetan, Brahmaputran, ...

Language families

- **Sino-Tibetan**

Language families

- **Austro-Asiatic**
  - Vietnamese, Khmer, Santali
  - Spoken in South-East Asia and Northern India



Language families

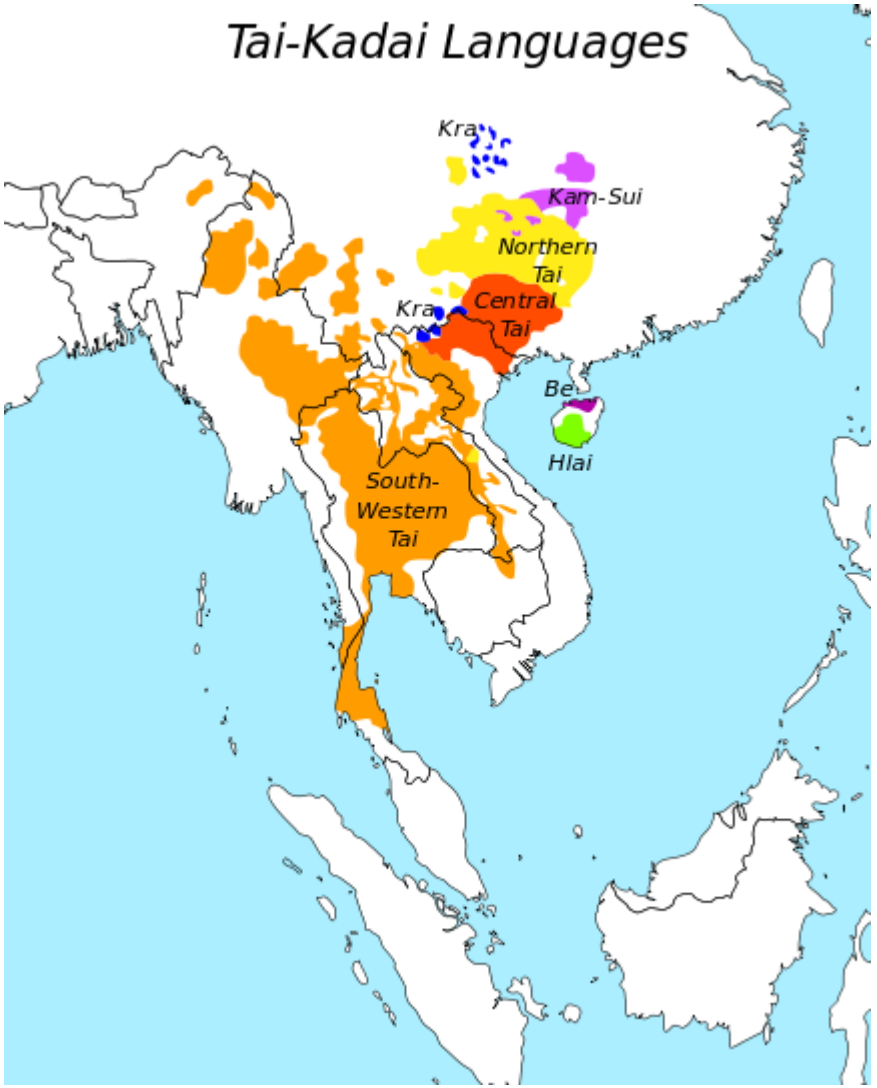
- **Austronesian**
  - Family with the largest geographical expansion (from Madagaskar in the West until Hawaii in the East)
  - Malagasy, Javanese, Bahasa Indonesian, Tagalog, Taiwanese languages, Maori (language of the aborigines of New Zealand), polynesian languages, ...

Language families

- **Austronesian**

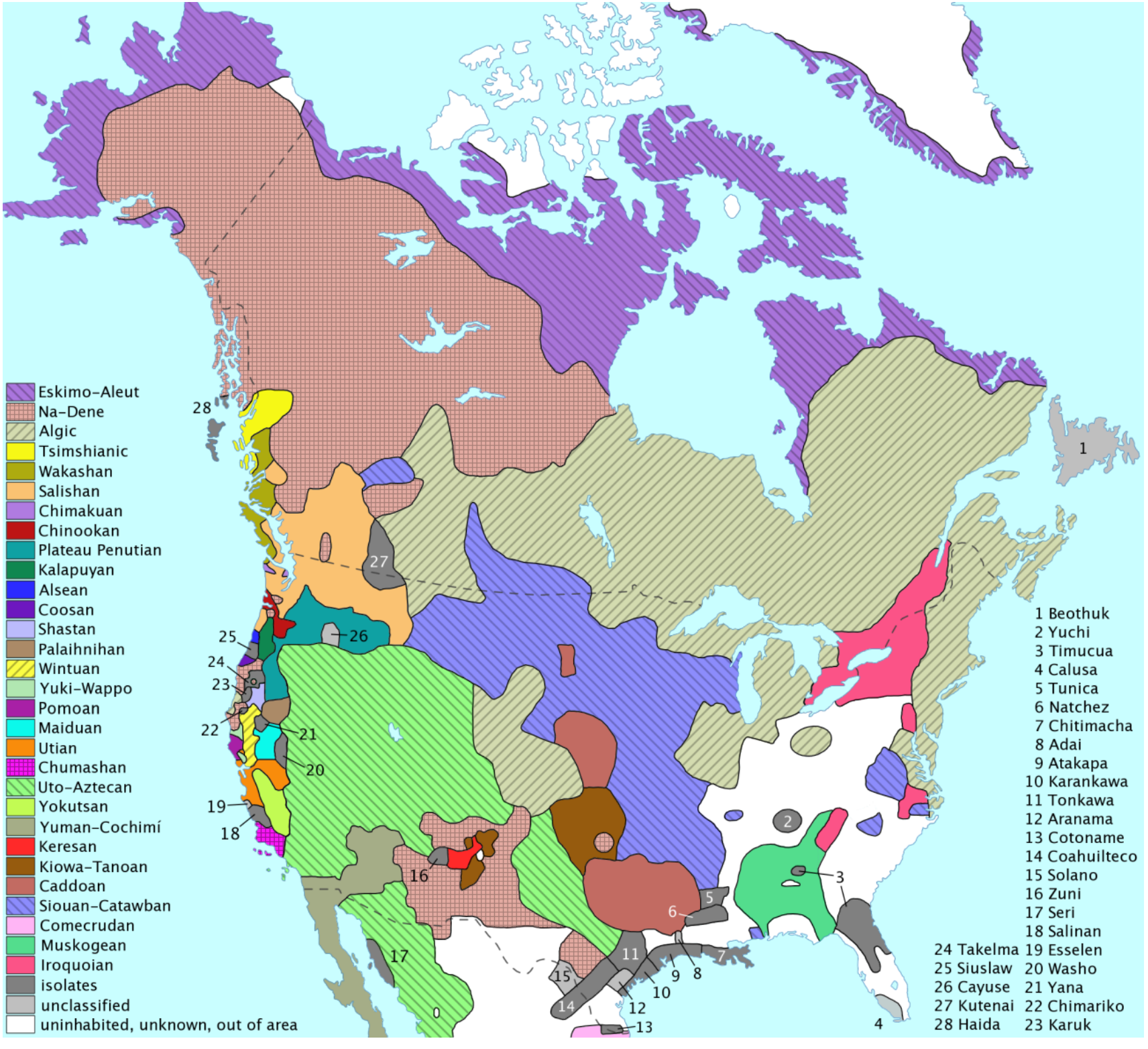
Language families

- **Tai-Kadai languages**
  - Thai, Isan, Lao, ...
  - Speculations, that Austronesian and Tai-Kadai form a single family („Austro-Thai“)



Paleo-American language families

- Classification according to Greenberg:
  - Eskimo-Aleut
  - Na-Dene (Northern and Western North-America)
  - Amerindian (rest of North-America and South-America)
- „Amerindian“ is heavily contested
- Using traditional methods, only many much smaller families can be established


Eskimo-Aleut
Na-Dene
Algic
Tsimshianic
Wakashan
Salishan
Chimakuan
Chinookan
Plateau Penutian
Kalapuyan
Alsean
Coosan
Shastan
Palaihnihan
Wintuan
Yuki-Wappo
Pomoan
Maiduan
Utian
Chumashan
Uto-Aztecan
Yokutsan
Yuman-Cochimí
Keresan
Kiowa-Tanoan
Caddoan
Siouan-Catawban
Comecrudan
Muskogean
Iroquoian
isolates
unclassified
uninhabited, unknown, out of area
1 Beothuk
2 Yuchi
3 Timucua
4 Calusa
5 Tunica
6 Natchez
7 Chitimacha
8 Adai
9 Atakapa
10 Karankawa
11 Tonkawa
12 Aranama
13 Cotoname
14 Coahuilteco
15 Solano
16 Zuni
17 Seri
18 Salinan
19 Esselen
20 Washo
21 Yana
22 Chimariko
23 Karuk
24 Takelma
25 Siuslaw
26 Cayuse
27 Kutenai
28 Haida

Language families

- In many cases, it is impossible to come up with a clear classification
  - 700 languages in Papua-New Guinea, often unrelated to each other
  - Several hundred languages of Australian aborigines; genetic classification is unclear
  - Many „isolated“ language (i.e. no genetic relationship to any other language can be established), for instance Basque