# State of the art of the Automated Similarity Judgment Program

*Søren Wichmann*
*(MPI-EVA & Leiden University)*
*& The ASJP Consortium*

**The Swadesh Centenary Conference, MPI-EVA, Jan. 17-18, 2009**

# Structure of the presentation

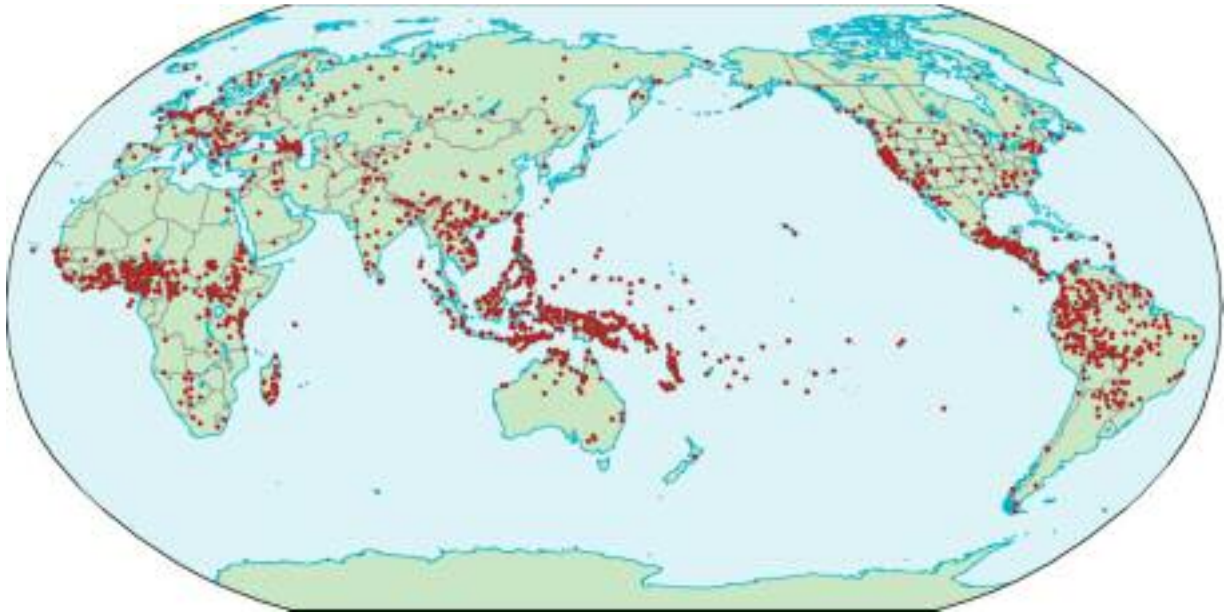
- 1. History of the ASJP project
- 2. Basic methodology
- 3. An assessment of the viability of glottochronology
- 4. Identifying homelands

# 1. History of the ASJP project

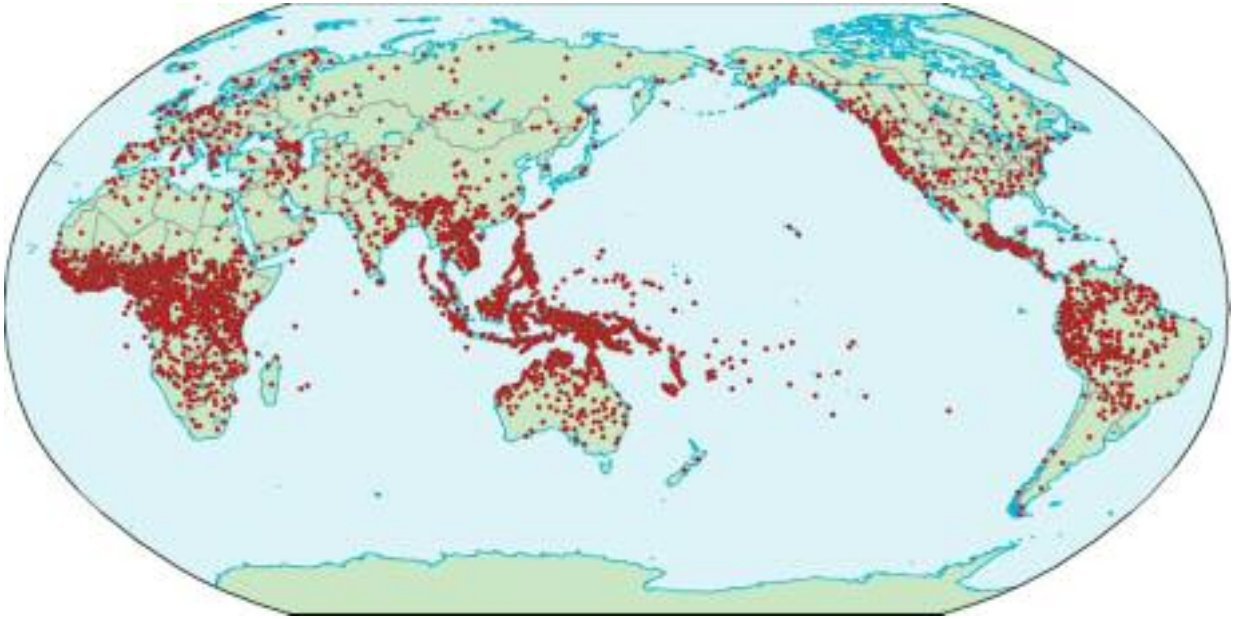
- Jan. 2007:
  - Cecil Brown (US linguistic anthropologist) comes up with idea of comparing languages automatically and communicates this to
  - Eric Holman (US statistician) and me. Brown and Holman work on rules to identify cognates implemented in an „automated similarity judgement program“ (ASJP).
- May 2007:
  - Cecil Brown is in Leipzig and explains to me what the two of them have come up with and I begin to take more active part, adding ideas.
- Aug. 2007:
  - Viveka Velupillai (Giessen-based linguist) joins in.
  - A first paper is written up (largely by Brown and Holman) showing that the classifications of a number of families based on a 245 language sample conform pretty well with expert classification.

- Sept. 2007:
  - Andre Müller (linguist, Leipzig) joins.
  - Pamela Brown (wife of Cecil Brown) joins.
  - Dik Bakker (linguist, Amsterdam & Lancaster) joins, and begins to do automatic data-mining, an implementation in Pascal, and to look at ways to identify loanwords.
- Oct. 2007:
  - Hagen Jung (computer scientist, MPI, makes a preliminary online implementation).
  - I take over the „administration“ of the project.
  - A second paper is finished about stabilities of lexical items, defining a shorter Swadesh list, etc.
- Nov. 2007:
  - Robert Mailhammer (linguist, BRD) joins.
- Dec. 2007:
  - Anthony Grant (linguist, GB) joins.
  - Dmitry Egorov (linguist, Kazan) joins.
  - Levenshtein distances are implemented instead of old „matching rules“ identifying cognates.

- Jan. 2008:
  - Kofi Yakpo (linguist) joins.
- Febr. 2008
  - The two papers are accepted for publication without revision (in respectively *Sprachtypologie und Universalienforschung* and *Folia Linguistica*).
- April 2008:
  - Oleg Belyaev (linguist, Moscow) joins.
- 2008:
  - Papers presented at conferences in Tartu, Helsinki, Cayenne, Forli, and Amsterdam.
  - Work on the structure of phylogenetic trees, glottochronology, onomatopeitic phenomena, homelands.
- Jan. 2009:
  - Paper accepted for *Linguistic Typology*
  - The database expanded to hold around 2500 languages. Another 1000 or so in the pipeline.



2432 fully processed languages in the ASJP database (~1000 are in the pipeline)



6000+ Languages in the world

# 2. Basic Methodology

# The database

- Encoding: a simplifying transcription
- Contents: 40-item lists

# Transcriptions

- 7 vowel symbols
- Nasalization indicated but not length, tone, stress
- Some rare distinctions merged
- „Composite“ sounds indicated by a modifier
- Vx sequences where x = velar-to-glottal fricative, glottal stop or palatal approximant reduced to V

Example of transcription: Havasupai (Yuman)

| | | |
|---|---|---|
| 30. Blood | hʷáte | hw~ate |
| 31. Bone | ʧija:k | Ciyak |
| 51. Breast | XXX | XXX |
| 66. Come | mijúwa | miyuwa |
| 61. Die | pí:ka | pika |
| 21. Dog | ʔaháte | ahate |
| 54. Drink | θí:ka | 8ika |
| 39. Ear | smárk | smark |
| 40. Eye | júʔ | yu7 |
| 82. Fire | ʔaʔóʔ | a7o7 |
| 19. Fish | ʔitʃí:ʔ | iCi7 |
| 95. Full | timʔórika | tim7orika |
| 48. Hand | sále | sale |
| 58. Hear | ʔé:vka | evka |
| 34. horn | ʔkwáʔa | kw~a7a |

Another transcription example: Abaza (Northwest Caucasian)

| | | |
|---|---|---|
| 18 person | ʕʷɨɬʃʼʲʷʕʷɨs | Xw~3Cw"y$Xw~3s |
| 19 fish | pslaɬʃʷa | pslaCw~a |
| 21 dog | la | la |
| 22 louse | ʦʼa | c"a |
| 23 tree | ʦʼla | c"la |
| 25 leaf | bɣʲɨ | bxy~3 |
| 28 skin | ɬʃʷazʲ | Cw~azy~ |
| 30 blood | ʃʲa | Sy~a |
| 31 bone | bʕʷɨ | bXw~3 |
| 34 horn | ɬʃʼʷɨʕʷa | Cw"~3Xw~a |
| 39 ear | lɨmha | l3mha |
| 40 eye | la | La |
| 41 nose | pɨnʦʼa | p3nc"a |
| 43 tooth | pɨʦ | p3c |
| 44 tongue | bzɨ | bz3 |
| 47 knee | ʃʲamqa | Sy~amqa |

# Towards a shorter Swadesh list

Procedure:

- Measure stabilities of items on the Swadesh list
- Find the shortest list among the most stable items that gives adequate results

# Measure stabilites

- count proportions of matches for pairs of words with similar meanings among languages within genera
- add corrections for chance agreement
- weighted means

# Check whether it actually makes sense to assume that items have inherent stabilites by

- seeing whether the rankings obtained correlate across different areas (in this case New World vs. Old World is convenient)

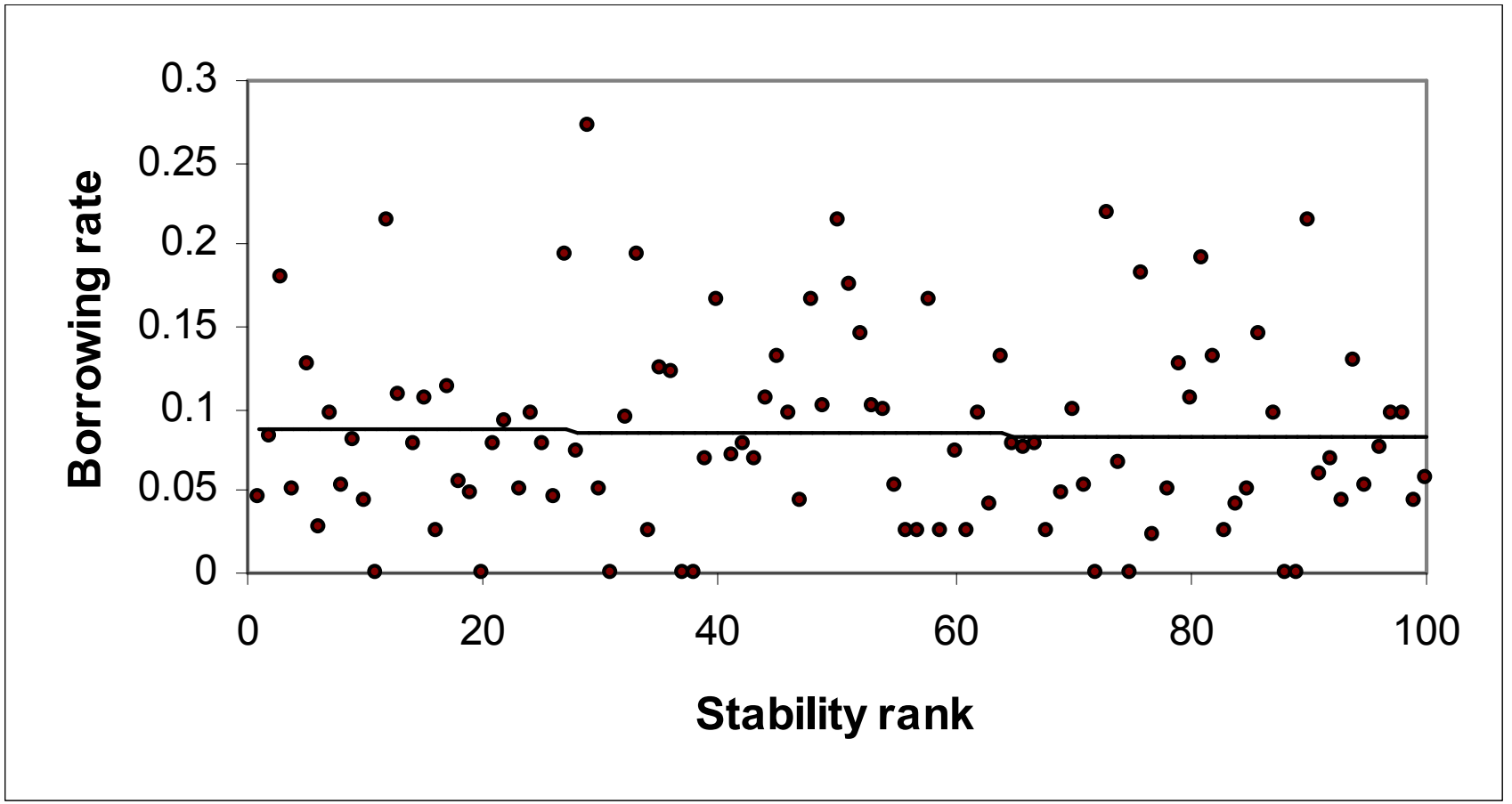
**Items on the Swadesh 100-item list in order of descending stability**

| | | | | |
|---|---|---|---|---|
| louse | skin | new | root | cold |
| two | night | dog | claw | flesh |
| ear | leaf | sun | bite | neck |
| die | rain | fly | ash | say |
| water | blood | heart | red | burn |
| liver | horn | give | egg | tail |
| eye | kill | grease | eat | sand |
| hand | person | feather | who | that |
| I | knee | moon | hair | sit |
| hear | nose | yellow | dry | all |
| tree | full | white | smoke | many |
| fish | star | bird | not | know |
| name | come | head | seed | walk |
| stone | mountain | earth | woman | cloud |
| breasts | one | foot | this | belly |
| path | fire | black | round | big |
| tongue | we | mouth | long | swim |
| tooth | drink | green | stand | hot |
| you | bark | what | good | lie |
| bone | see | sleep | man | small |

# Stability and borrowability

| Meaning | Stability | Attestations | Borrowings ["probable" or "clear"] | Proportion (%) |
|---|---|---|---|---|
| louse-H | 42.8 | 43 | 2 | 4.7 |
| louse-B | | 36 | 3 | 8.3 |
| two | 39.4 | 39 | 7 | 17.9 |
| ear | 37.2 | 40 | 2 | 5.0 |
| die | 36.3 | 47 | 6 | 12.8 |
| water | 36.2 | 37 | 1 | 2.7 |
| liver | 35.4 | 41 | 4 | 9.8 |
| eye | 35.1 | 38 | 2 | 5.3 |
| hand | 34.9 | 37 | 3 | 8.1 |
| I | 34.1 | 46 | 2 | 4.3 |
| hear | 33.8 | 39 | 0 | 0 |
| tree | 33.6 | 42 | 9 | 21.4 |
| fish | 33.4 | 37 | 4 | 10.8 |
| name | 32.4 | 38 | 3 | 7.9 |
| stone | 32.1 | 47 | 5 | 10.6 |
| breasts | 30.2 | 41 | 1 | 2.4 |

# No correlation between borrowability and stability



# Potential explanations

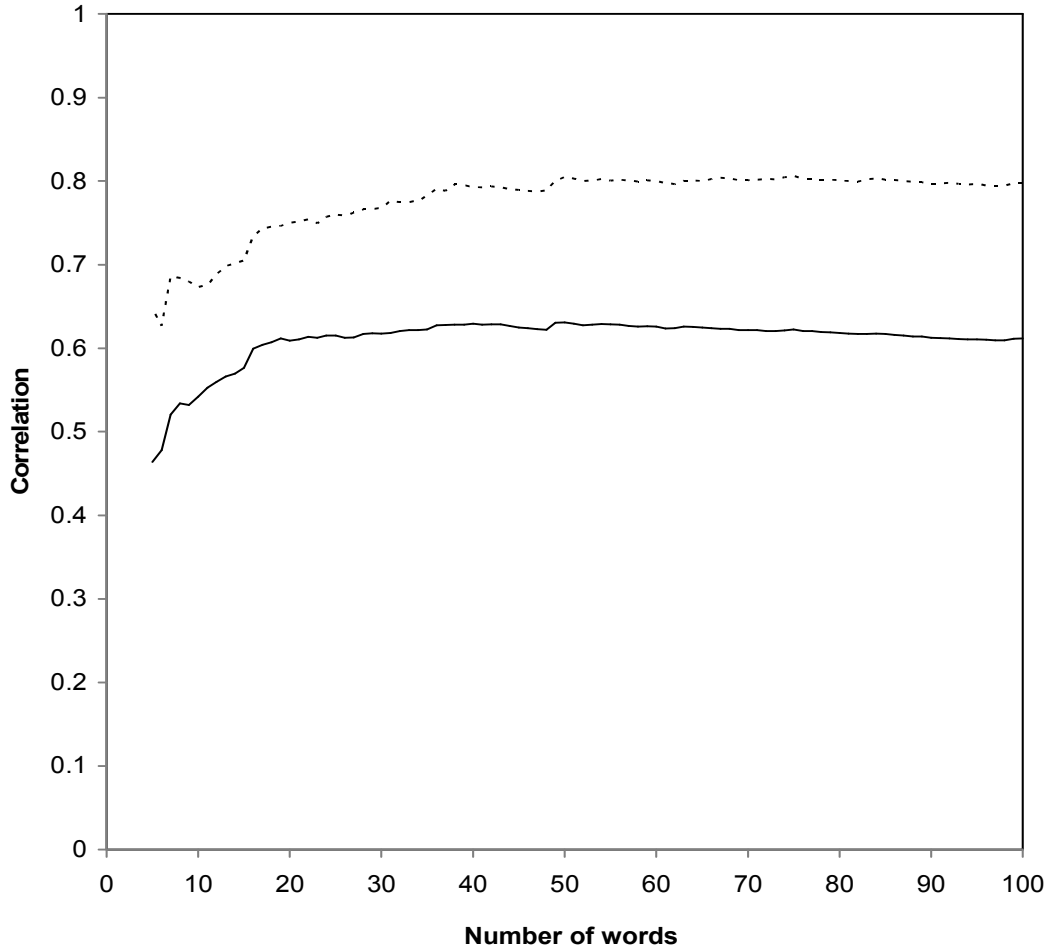
- Borrowability may be more variable for given lexical items across areas than stability and not be an inherent property of lexical items (similar to typological features).

- Borrowability is not a significant contributor to stability, at least as the segment constituted by the Swadesh 100-item list is concerned.

- There are still far too little data on borrowability to be conclusive (the sample for studying stability was constituted by 245 languages, whereas we had only 36 language at our disposal for the study of borrowability).

# Selecting a shorter list

Correlation between distances in the automated approach and other classifications as a function of list lengths




Ethnologue
(Goodman-Kruskal gamma )

WALS/Dryer
(Pearson product-moment correlation)

# Automating the similarity measure

Levenshtein distances: the minimum number of steps—substitutions, insertions or deletions—that it takes to get from one word to another

Germ. Zunge → Eng. tongue

tsuŋə
tuŋə (substitution)
tɔŋə (substitution)
tɔŋ (deletion)

Or tongue → Zunge

tɔŋ
tɔŋə (insertion)
tuŋə (substitution)
tsuŋə (substitution)

= 3 steps, so LD = 3

# Weighting Levenshtein distances

Serva & Petroni (2008): divide by the lengths of the strings compared. Takes into account that LD‘s grow with word length

ASJP:

1. divide LD by the length of the longest string compared to get LDN (takes into account typical word lengths of the languages compared);
2. then divide LDN by the average of LDN‘s among words in Swadesh lists with different meanings to get LDND (takes into account accidental similarity due to similarities in phonological inventories)

# Results for classification

Two methods of evaluation:

- Looking at statistical correlations with WALS or Ethnologue classification
- Comparing tree with „expert trees“/expert knowledge

# Performance of classification: a correlation with *Ethnologue*

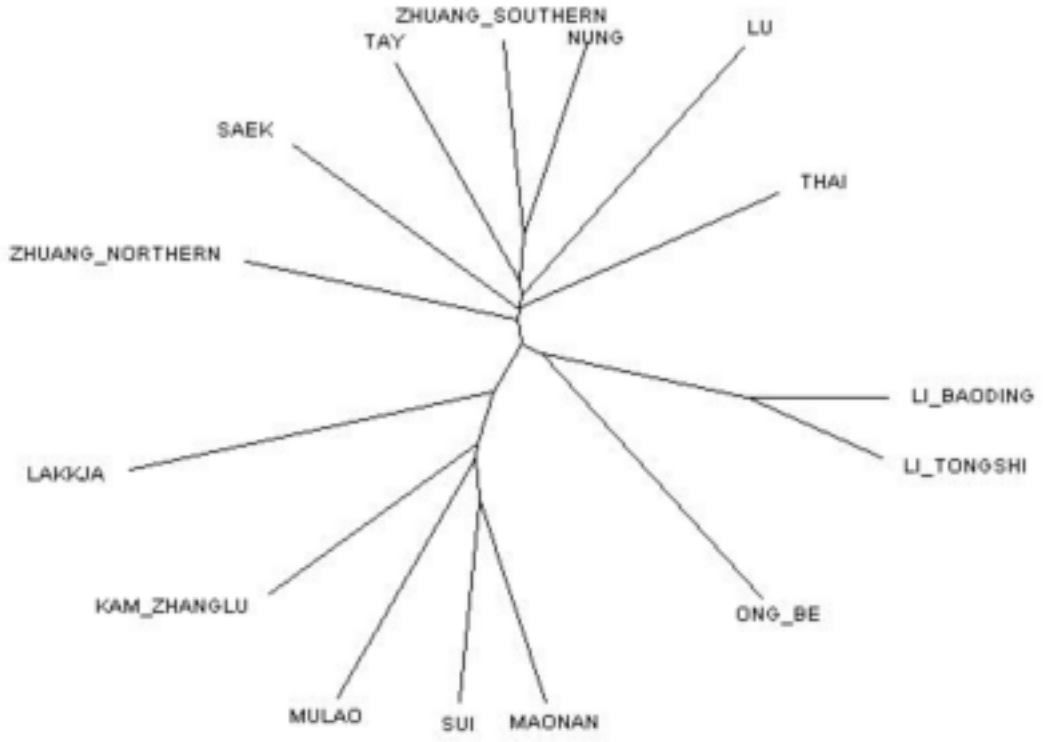
| | | | |
|---|---|---|---|
| MIXE-ZOQUE | 0.9803 | URALIC | 0.7021 |
| OTO-MANGUEAN | 0.9793 | TAI-KADAI | 0.6955 |
| INDO-EUROPEAN | 0.9332 | AUSTRO-ASIATIC | 0.6475 |
| ALTAIC | 0.8552 | HOKAN | 0.6223 |
| NAKH-DAGHESTANIAN | 0.8515 | KADUGLI | 0.5725 |
| MACRO-GE | 0.8447 | ALGIC | 0.5477 |
| MAYAN | 0.8276 | KHOISAN | 0.5069 |
| PENUTIAN | 0.8062 | TRANS-NEW GUINEA | 0.5047 |
| TUPIAN | 0.7867 | NIGER-CONGO | 0.4404 |
| TUCANOAN | 0.7565 | ARAWAKAN | 0.393 |
| NILO-SAHARAN | 0.7475 | AUSTRALIAN | 0.3866 |
| UTO-AZTECAN | 0.7356 | CARIBAN | 0.3169 |
| CHIBCHAN | 0.7333 | PANOAN | 0.2733 |
| SINO-TIBETAN | 0.7318 | AUSTRONESIAN | 0.2553 |
| AFRO-ASIATIC | 0.7246 | | |

- Disadvantages of automated method:
  - blind to anything but lexical evidence
  - not always accurate
  - has a swallower limit of application than the comparative method

- Advantages:
  - extremely quick
  - consistent and objective
  - provides information on the amount of changes, and therefore a time perspective

# 3. Assessing the viability of glottochronology (or Levenshtein chronologies)

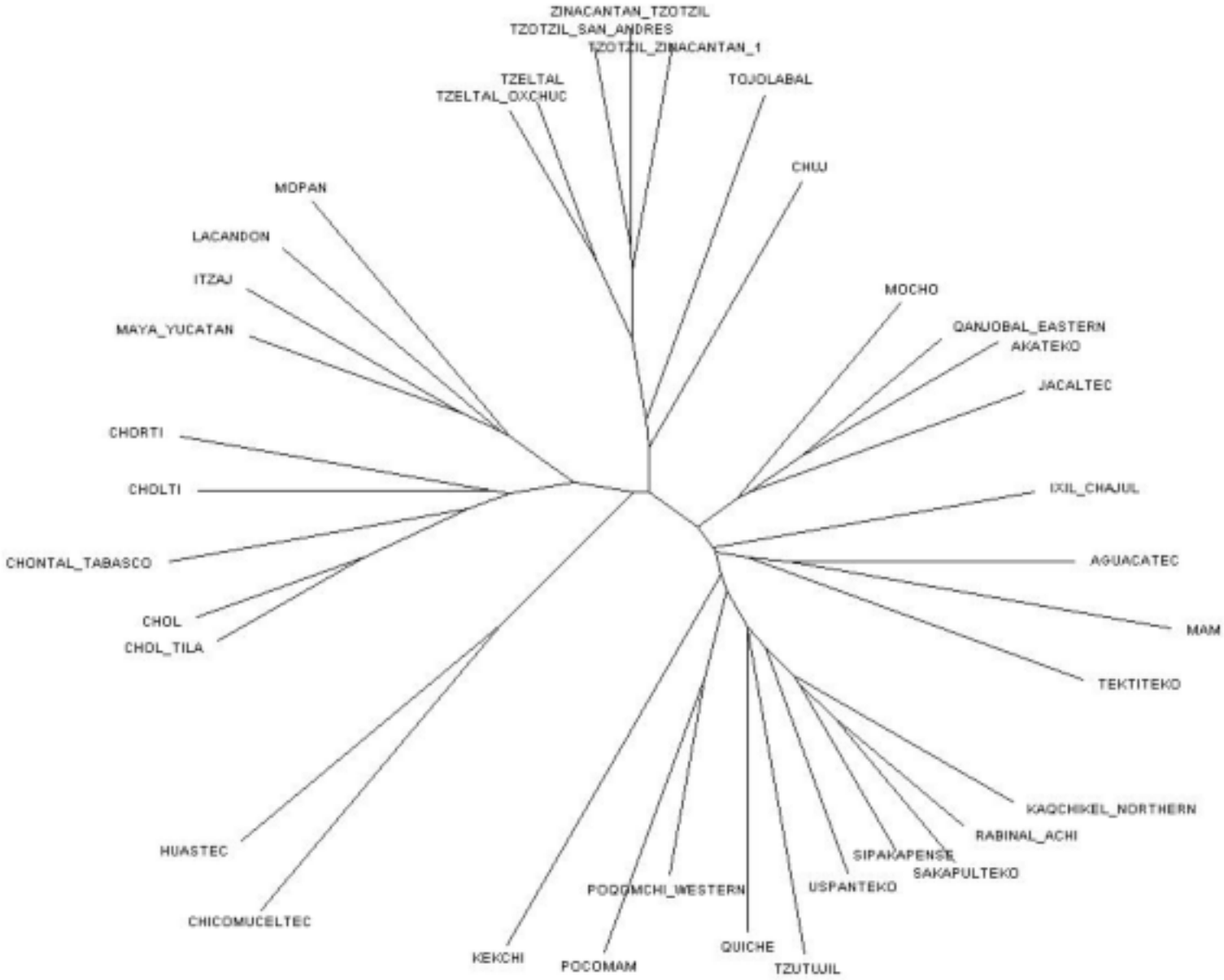
- The assumption of a (fairly) constant rate of change can be checked by looking at branch lengths for lexicostatistical trees. Let‘s see some examples:


ZHUANG_SOUTHERN
TAY
NUNG
LU
SAEK
THAI
ZHUANG_NORTHERN
LI_BAODING
LI_TONGSHI
LAKKJA
KAM_ZHANGLU
ONG_BE
MULAO
SUI
MAONAN

Tai-Kadai

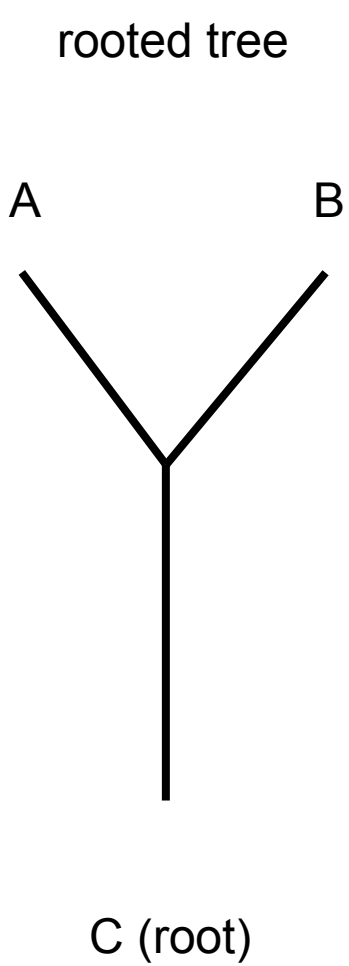

SHOSHONI
COMANCHE
NORTHERN_PAIUTE
UTE
SOUTHERN_PAIUTE
TUBATULABAL
CAHUILLA
HOPI
TOHONO_OODHAM
PIMA_BAJO
TEPECANO
MAYO
YAQUI
HUARUIO
EL_NAYAR_CORA
POCHUTLA_NAHUATL
TETELCINGO_NAHUATL
HIGHLAND_PUEBLA_NAHU
PIPIL

Uto-Aztecan


ZINACANTAN_TZOTZIL
TZOTZIL_SAN_ANDRES
TZOTZIL_ZINACANTAN_1
TZELTAL
TZELTAL_OXCHUC
TOJOLABAL
CHUJ
MOPAN
LACANDON
ITZAJ
MAYA_YUCATAN
MOCHO
QANJOBAL_EASTERN
AKATEKO
JACALTEC
CHORTI
CHOLTI
IXIL_CHAJUL
CHONTAL_TABASCO
AGUACATEC
CHOL
CHOL_TILA
MAM
TEKTITEKO
KAQCHIKEL_NORTHERN
RABINAL_ACHI
SIPAKAPENSE
SAKAPULTEKO
USPANTEKO
POQOMCHI_WESTERN
HUASTEC
CHICOMUCELTEC
KEKCHI
POCOMAM
QUICHE
TZUTUJIL

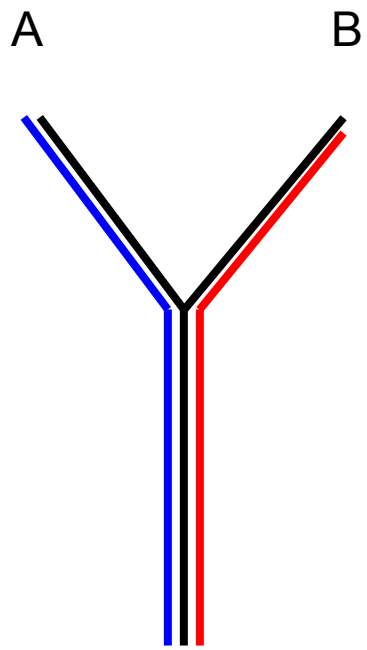
Mayan

# The ultrametric inequality condition



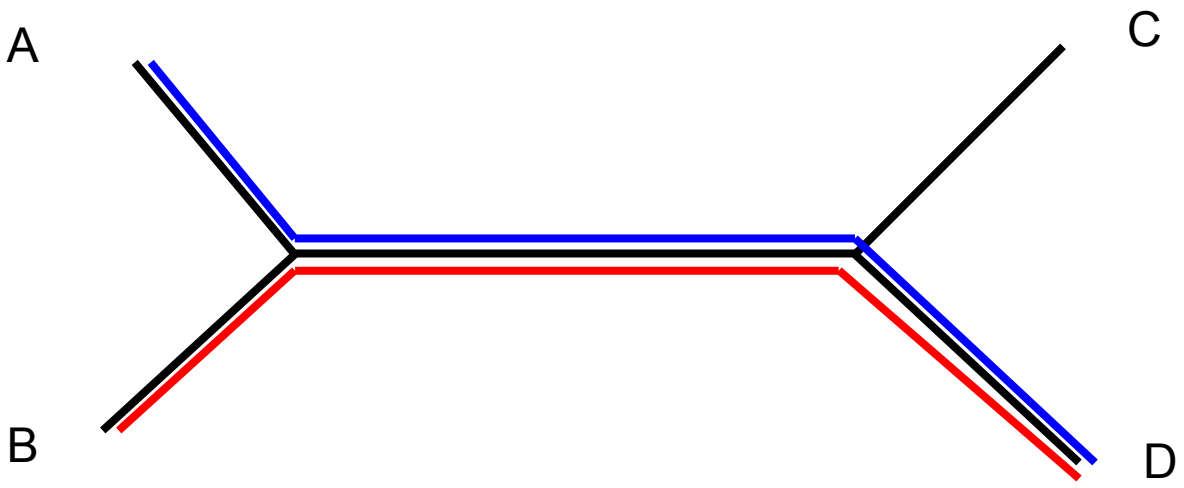
# The ultrametric inequality condition

rooted tree

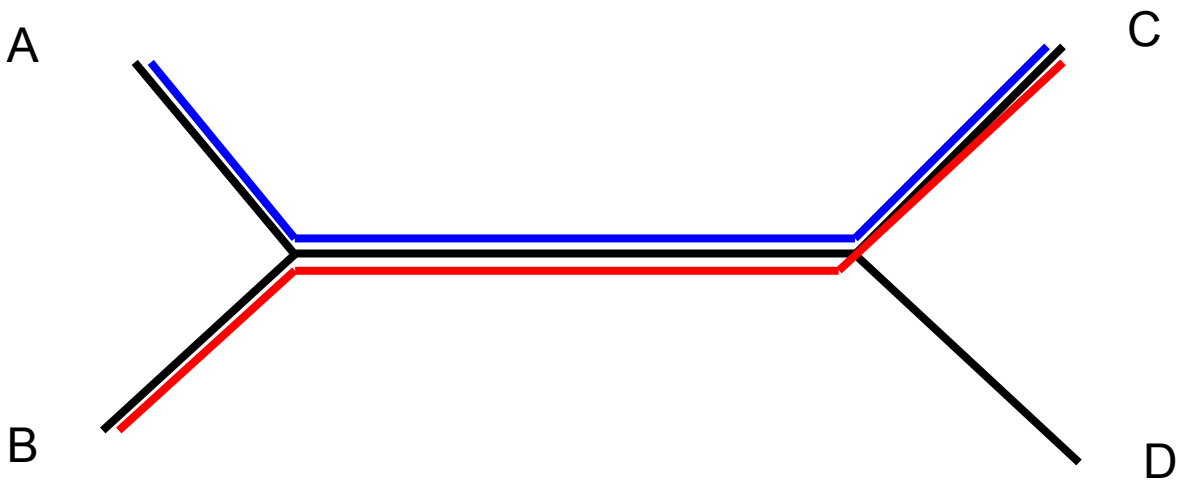



Distance C-A = Distance C-B

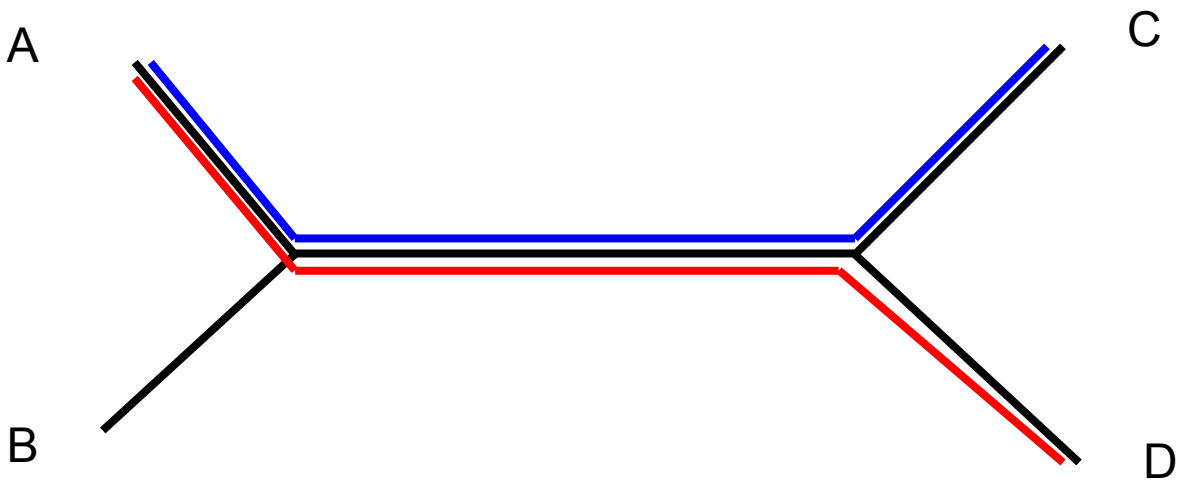
# Unrooted tree



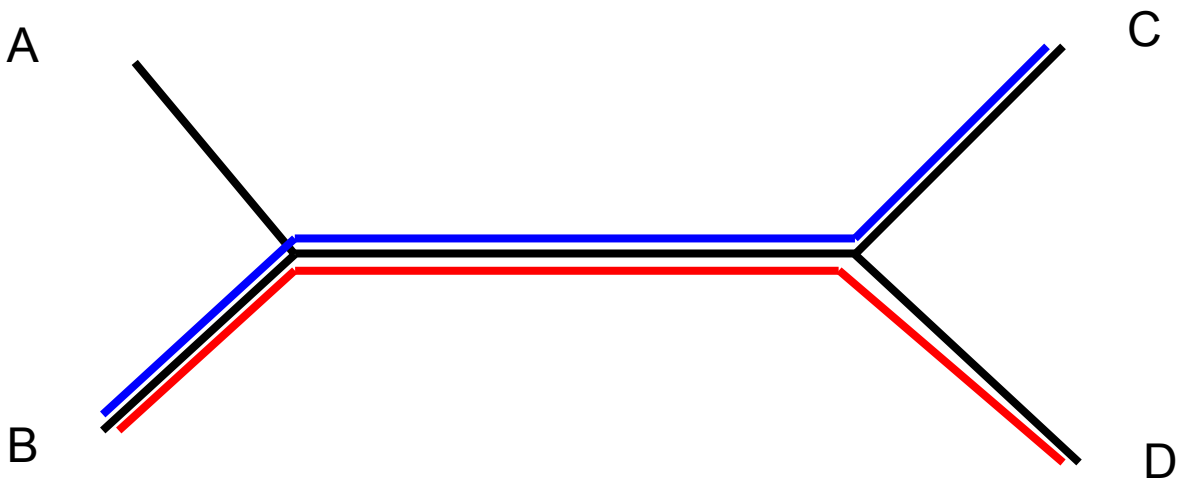

Distance A-D = Distance B-D




Distance A-C = Distance B-C

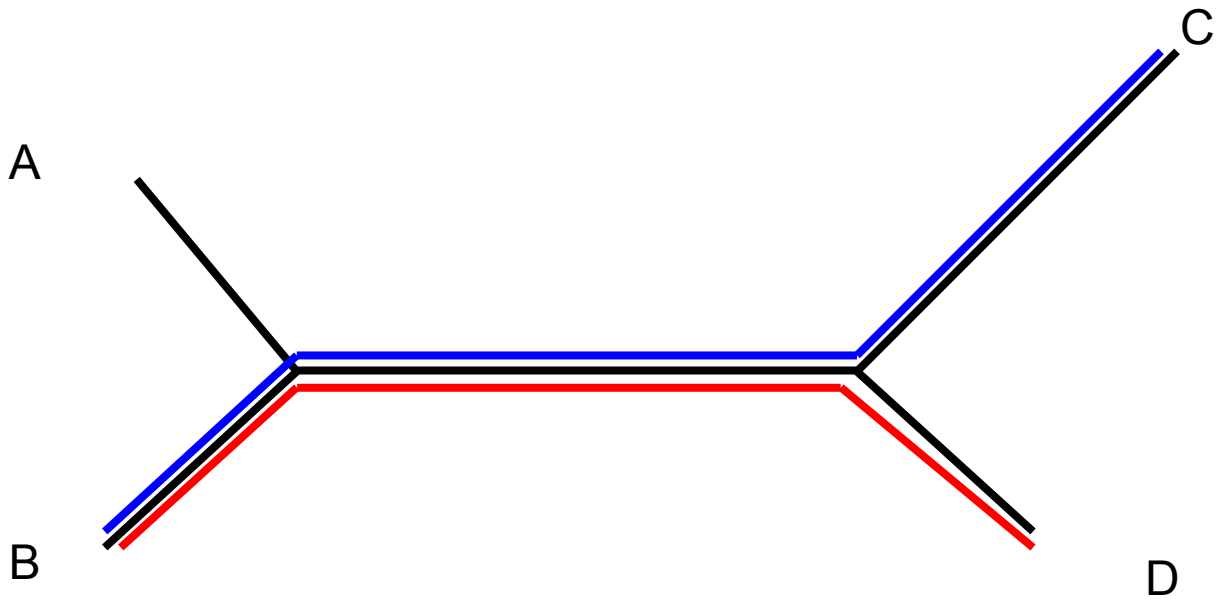



Distance A-C = Distance A-D




Distance B-C = Distance A-D

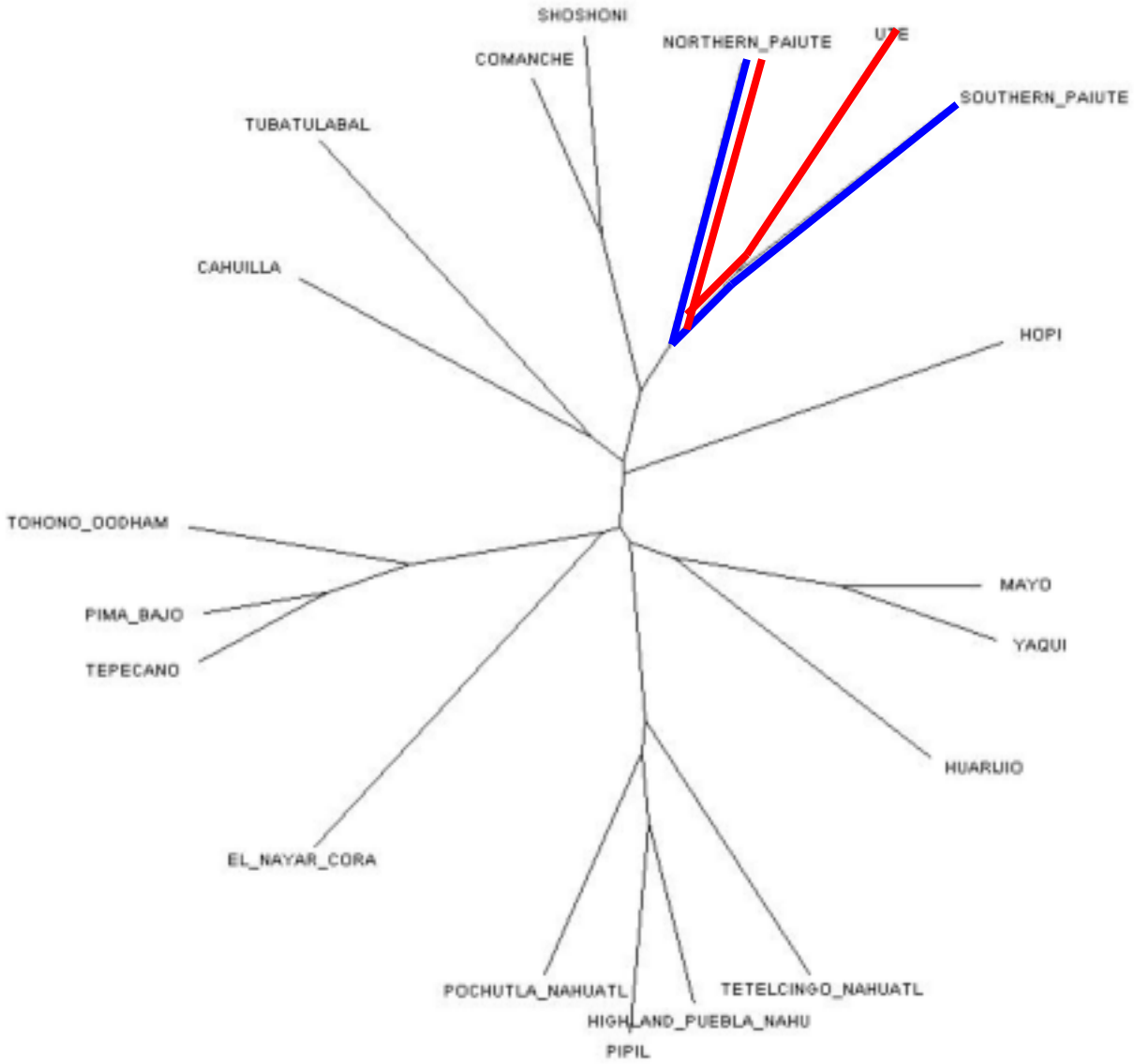
# A margin of error found by measuring the deviation from ultrametric inequality



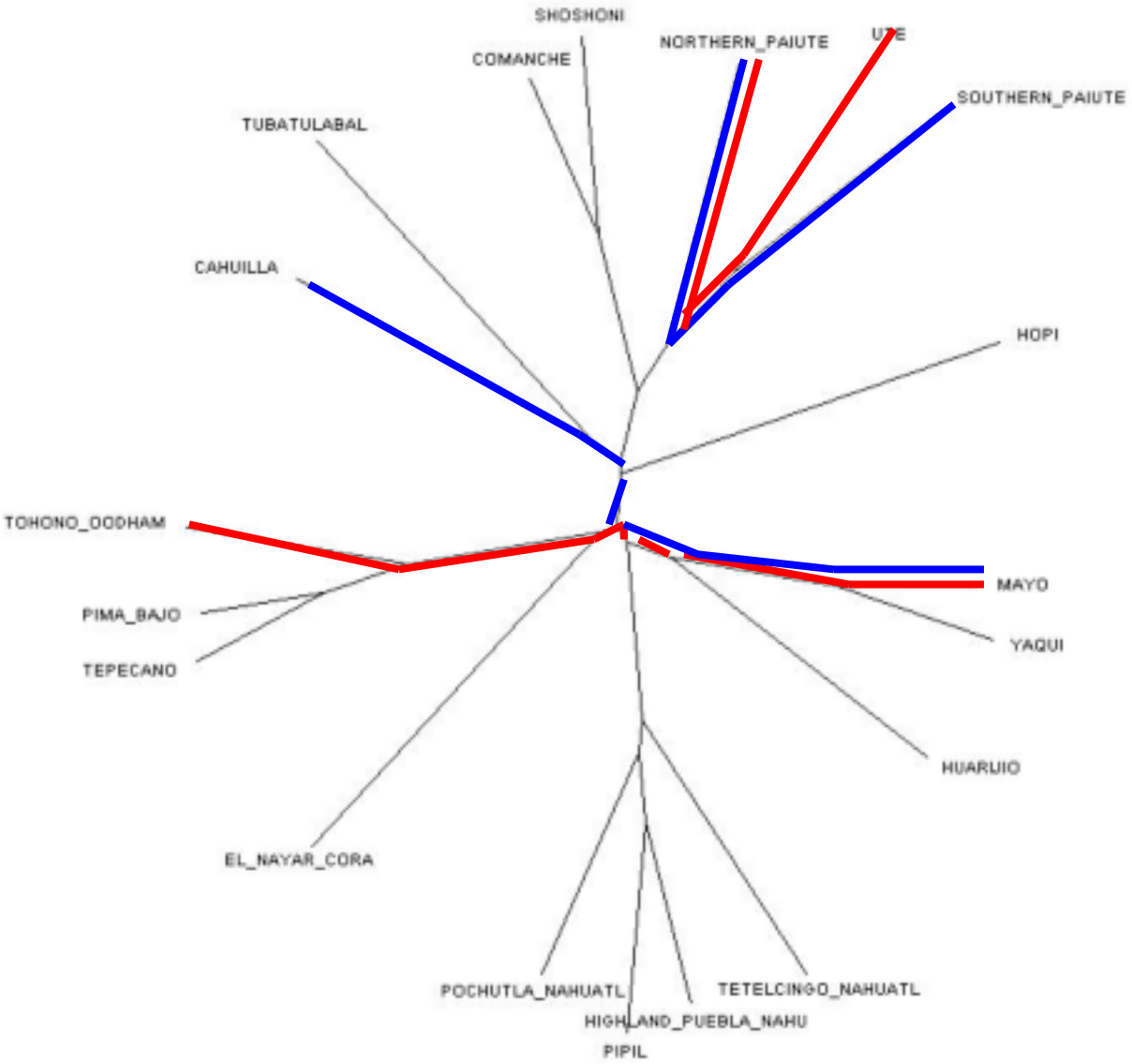

Margin of error = BC – BD/[(BC + BD)/2]


SHOSHONI
COMANCHE
NORTHERN_PAIUTE
UTE
SOUTHERN_PAIUTE
TUBATULABAL
CAHUILLA
HOPI
TOHONO_OODHAM
PIMA_BAJO
TEPECANO
MAYO
YAQUI
HUARUIO
EL_NAYAR_CORA
POCHUTLA_NAHUATL
TETELCINGO_NAHUATL
HIGHLAND_PUEBLA_NAHU
PIPIL

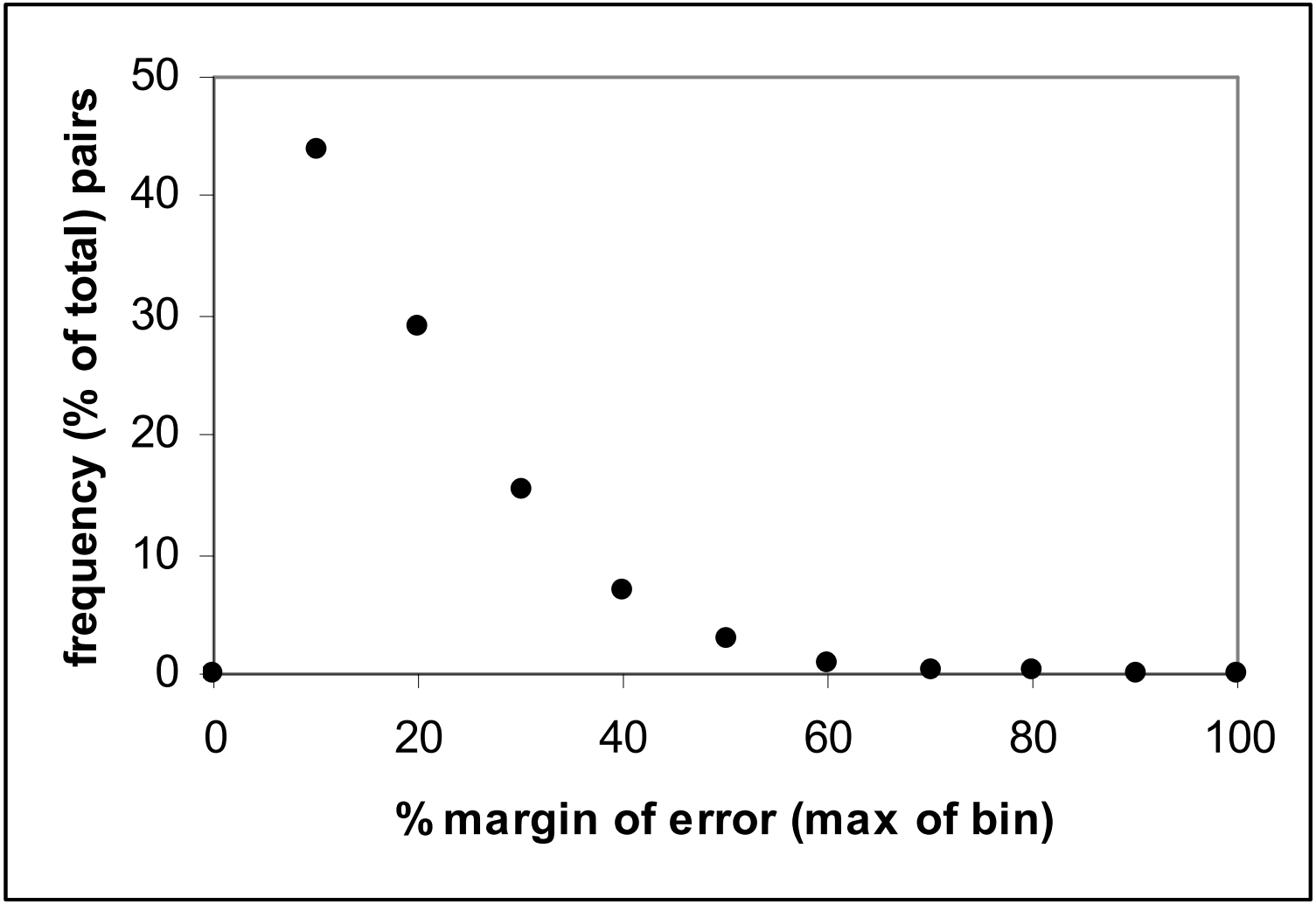
Uto-Aztecan


SHOSHONI
COMANCHE
NORTHERN_PAIUTE
UTE
SOUTHERN_PAIUTE
TUBATULABAL
CAHUILLA
HOPI
TOHONO_OODHAM
PIMA_BAJO
TEPECANO
MAYO
YAQUI
HUARUIO
EL_NAYAR_CORA
POCHUTLA_NAHUATL
TETELCINGO_NAHUATL
HIGHLAND_PUEBLA_NAHU
PIPIL

Uto-Aztecan

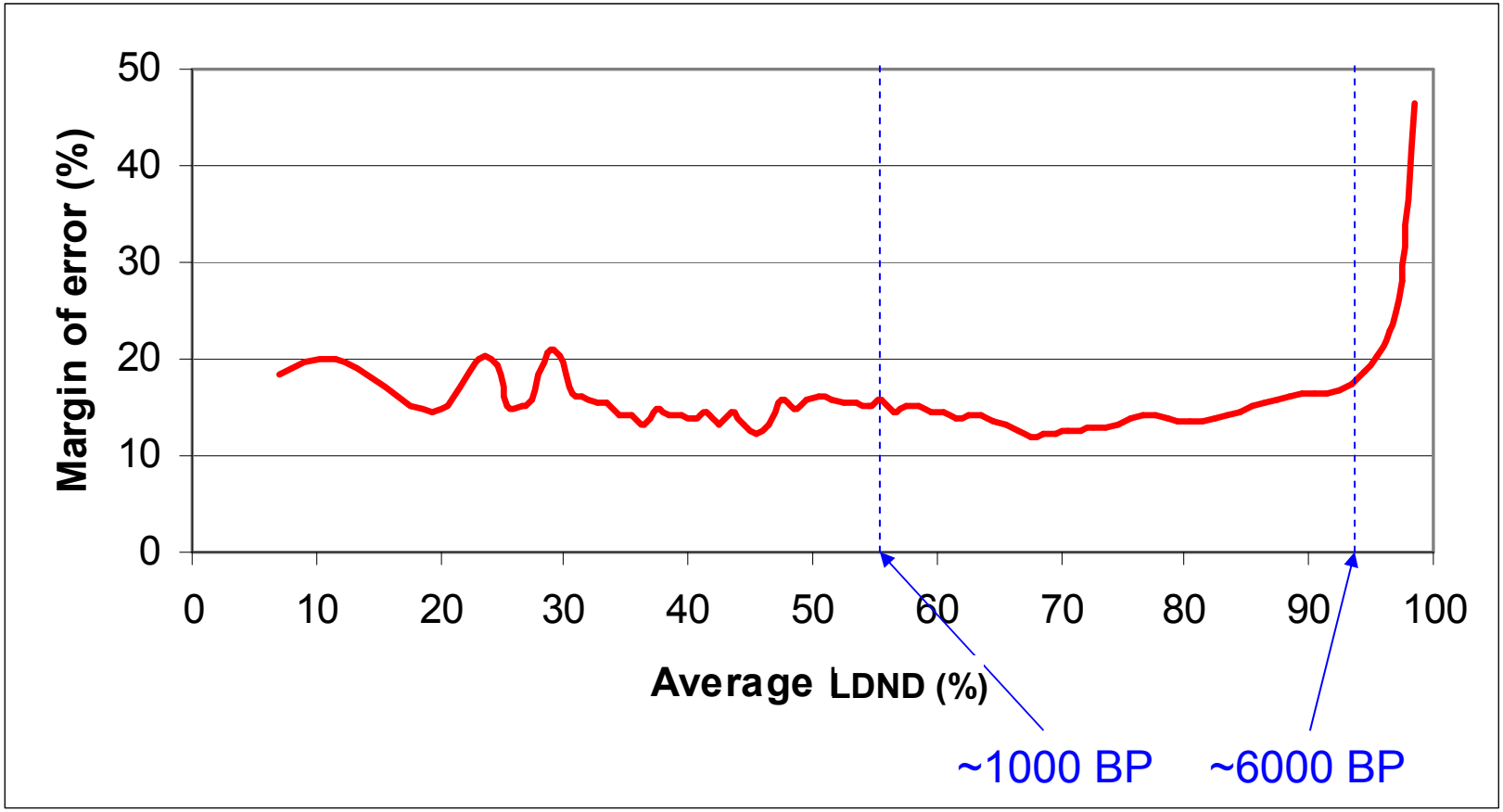

SHOSHONI
COMANCHE
NORTHERN_PAIUTE
UTE
SOUTHERN_PAIUTE
TUBATULABAL
CAHUILLA
HOPI
TOHONO_OODHAM
MAYO
PIMA_BAJO
YAQUI
TEPECANO
HUARUIO
EL_NAYAR_CORA
POCHUTLA_NAHUATL
TETELCINGO_NAHUATL
HIGHLAND_PUEBLA_NAHU
PIPIL

Uto-Aztecan




Binned frequencies of margins of errors for ages of single pairs (Indo-European)

# Margins of error for multiple language pairs as a function of LDND

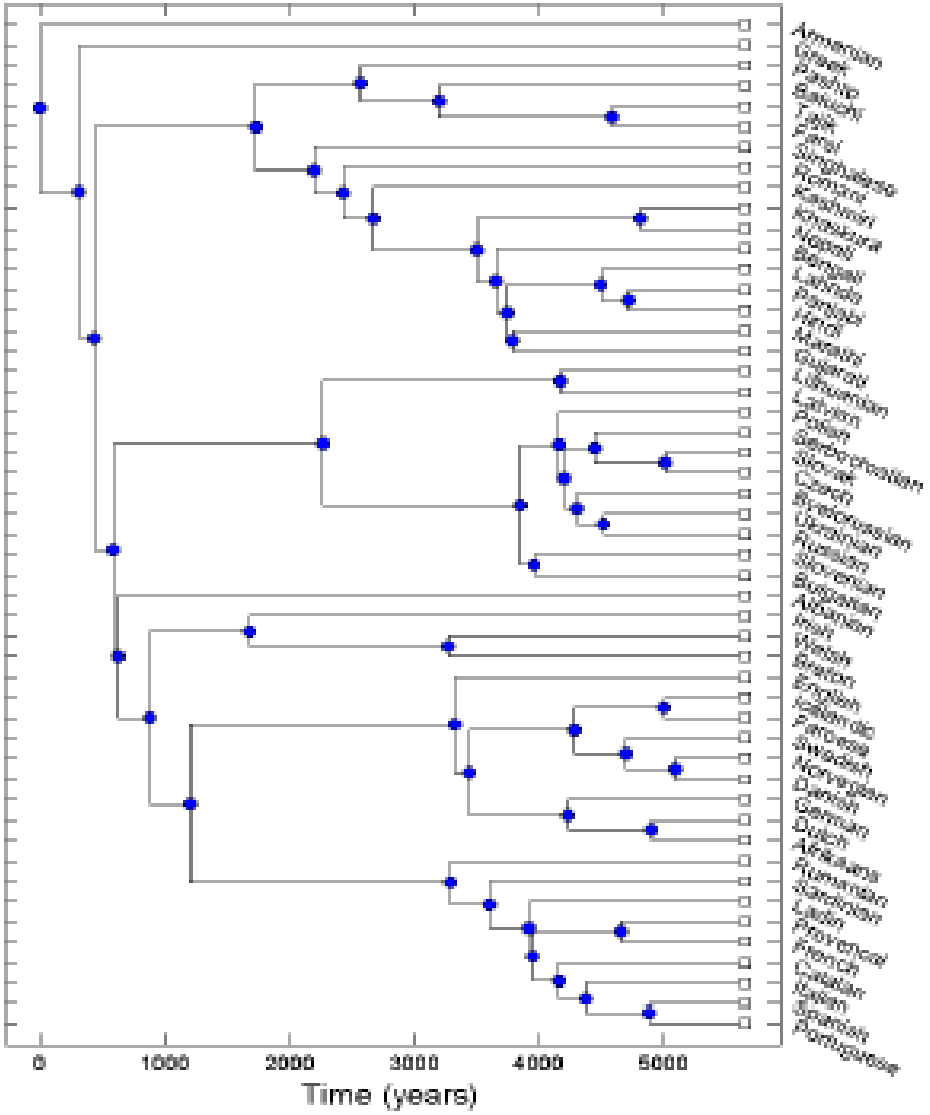



x-axis: average of the greatest LDNDs within all sets of three related languages that are within the same 1% interval.
y-axis: the margin of error estimated as the average of the differences between the (logarithms of) the two largest distances for the set of triplets in the interval divided by the (logarithm) of the average of these two largest distances.

# How to measure the age of a language group

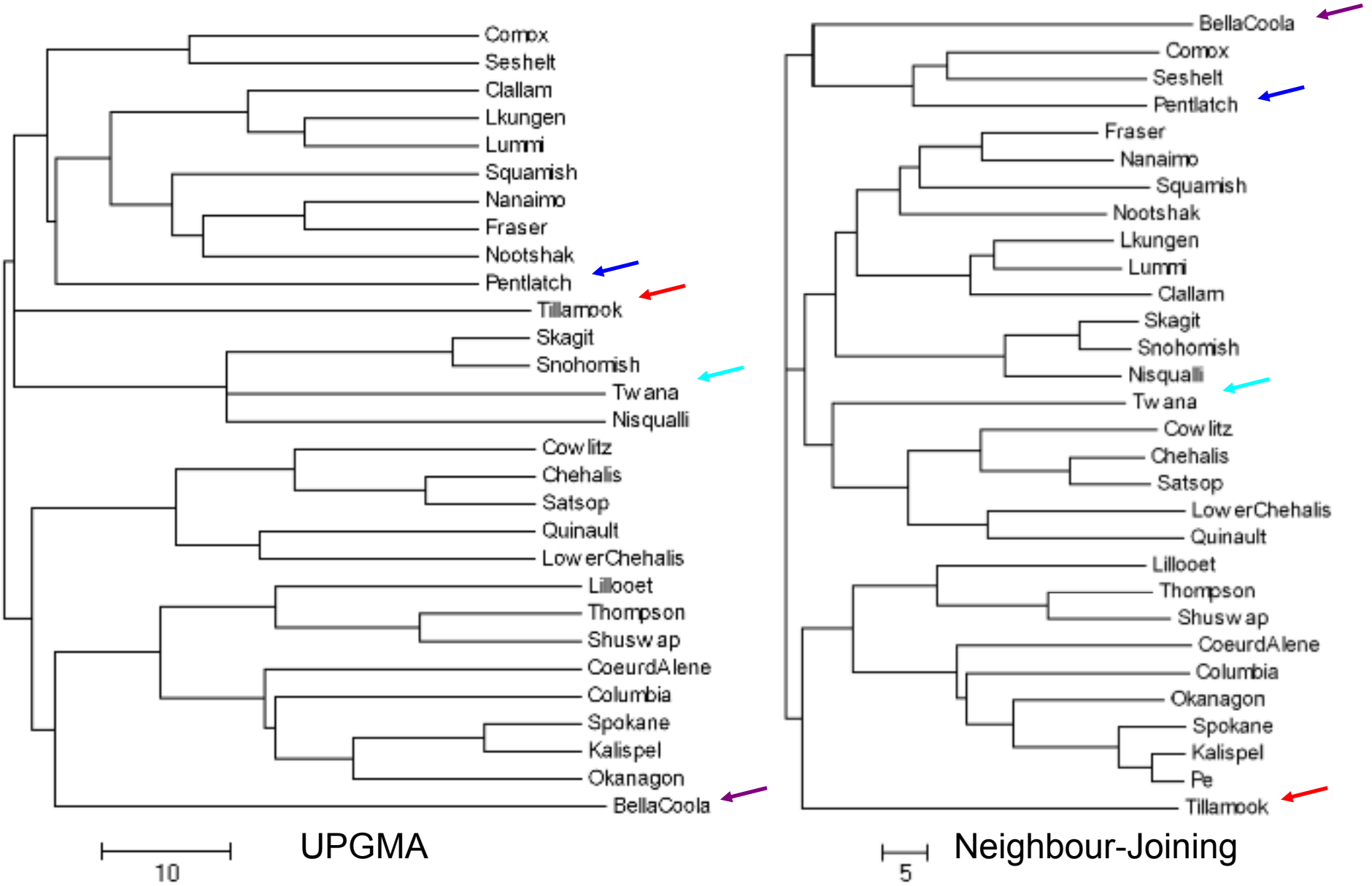
- Take the age of the two most divergent languages? No, this would bias the result high.
- Take the average age of all language pairs? No, this would bias the result low.
- Make the ages part of the lexicostatistical tree and measure lengths from root (midpoint) to tips? No, this is only doable for a UPGMA tree, which is far from an optimal phylogenetic algorithm.

# The last approach is taken by Serva and Petroni (2008)




Serva, Maurizio and Filippo Petroni. 2008. Indo-European languages by Levenshtein distances. Available at www.arXiv.org (and now published)

# Comparing two Salishan trees



# Our approach

- Find the midpoint in the tree of the language group and take the average modified Levenshtein distances of all pairs whose members are on either side of the midpoint.
- Calibrate with ages of known linguistic event.
- Find the LDND‘s at zero years = the LDND expected for dialects, and build that into the formula.

# The revised glottochronological formula

Standard formula: log(SIM) = [2log(R)]T

New formula taking into account inherent variability within languages
log(SIM) = [2log(R)] T + log(SIM')

SIM = observed similarity = 1-LDND
SIM' = baseline similarity at time 0
R = retention rate
T = time in millenia

R = .81 (slope of the line)
SIM' = .68 (the intercept). So

$$T = [\log(1-LDND)-\log(.68)]/2\log(.81)$$

# Some examples of results

| | | | |
|---|---|---|---|
| Arawakan | 5403 | Mixe-Zoque | 3672 |
| Austronesian | 5050 | Muskogean | 1812 |
| Cariban | 3511 | Nakh-Daghestanian | 5373 |
| Chibchan | 6146 | NW Caucasian | 5313 |
| Chukotko-Kamchatkan | 4312 | Pano-Tacanan | 5212 |
| Dravidian | 2959 | Romance | 2255 |
| Eskimo | 1749 | Salishan | 6097 |
| Germanic | 1506 | Semitic | 3274 |
| Hmong-Mien | 5384 | Slavic | 1187 |
| IndoEuropean | 5981 | TaiKadai | 3604 |
| Indo-Iranian | 4281 | Tupian | 4887 |
| Kartvelian | 4893 | Uralic | 4873 |
| Mayan | 2669 | Uto-Aztecan | 4629 |

# Outstanding problems

- Still not enough good calibration points, and they are hard to find.
- Ages greater than 6,000 BP cannot be trusted because randomness plays in (and ASJP classifications also typically break down beyond 6,000 years BP)
- Ages swallower than 1,000 show great variation from what‘s expected and cannot be trusted either.

# 4. Identifying homelands

The idea (going back to Vavilov 1926 in botany and Sapir‘s *Time Perspective in Aboriginal American Culture* of 1916) is that the area of highest diversity will tend to be the homeland.

Nikolai Vavilov (1887-1943)

Edward Sapir (1884-1939)

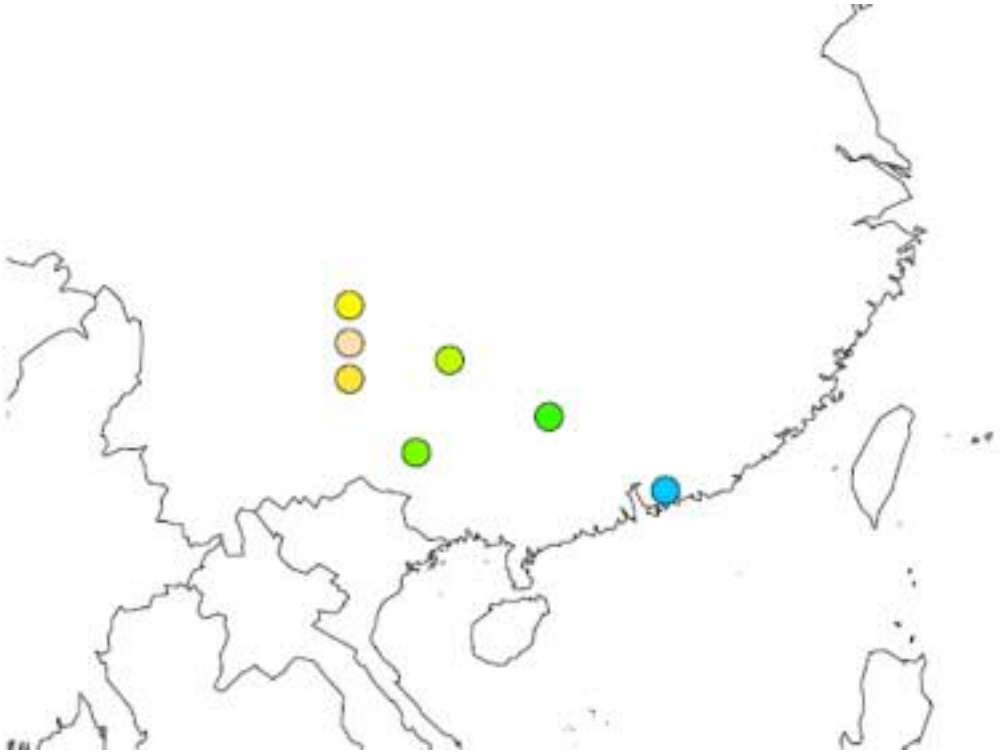
- A quantitative implementation:
  - For each language in a family, measure the proportion between the linguistic distance *L* and the geographical distance *G* to each of the other members of the family, and take the average. This produces a diversity measure *D* for the location where the given language is spoken.
  - The language with the highest *D* sits in the homeland.
  - Map the results by grouping *D*‘s into topographic color categories.

Supplement with reconstruction of ecological vocabulary, known migration histories, archaeology, etc. when available.

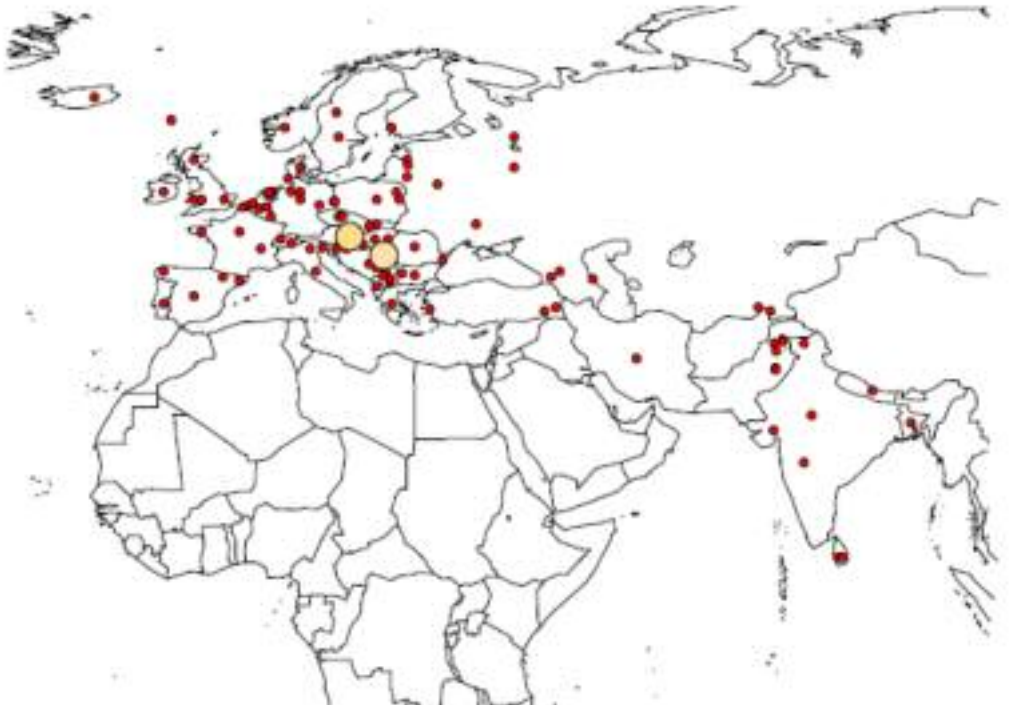
„*Any one criterion is never to be applied to the exclusion of or in opposition to all others. It is a comfortable procedure to attach oneself unreservedly or primarily to a single mode of historical inference and wilfully to neglect all others as of little moment, but the clean-cut constructions of the doctrinaire never coincide with the actualities of history* “ (Sapir 1916: 87).

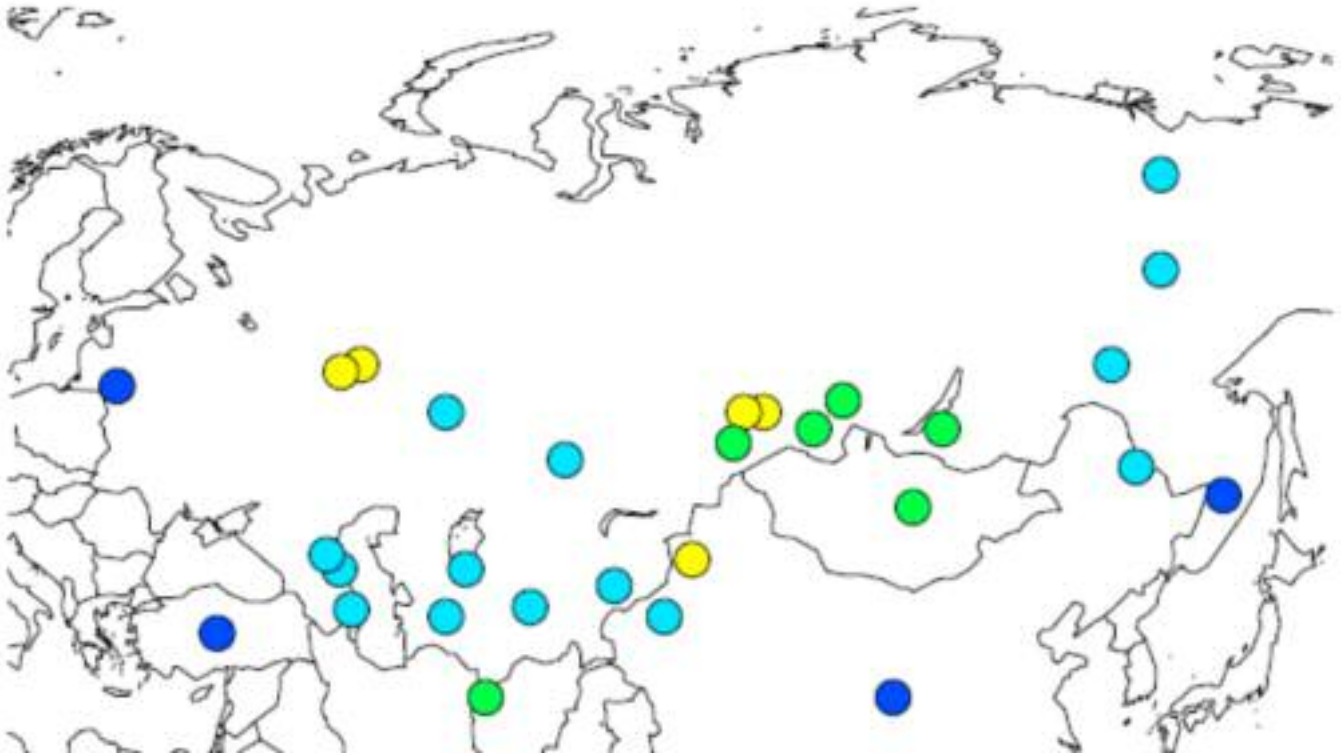
(cf. also critique of Vavilov by Harlan 1971)

HMONG-MIEN



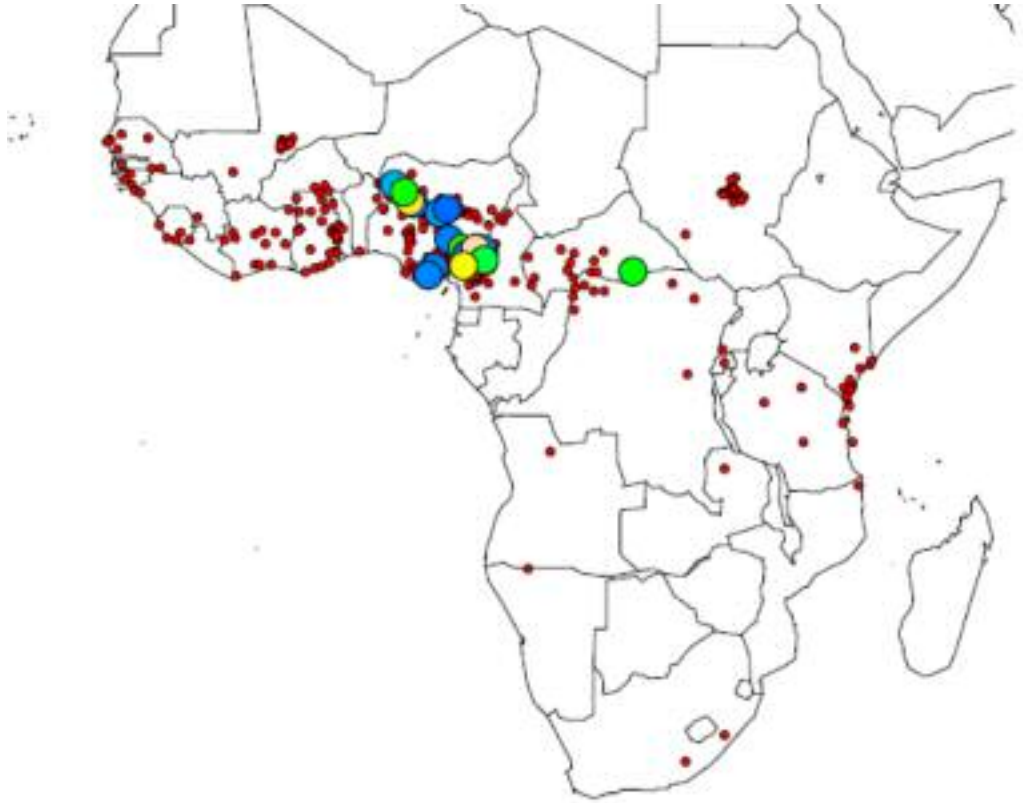
# CURRENTLY SPOKEN INDO-EUROPEAN LANGUAGES

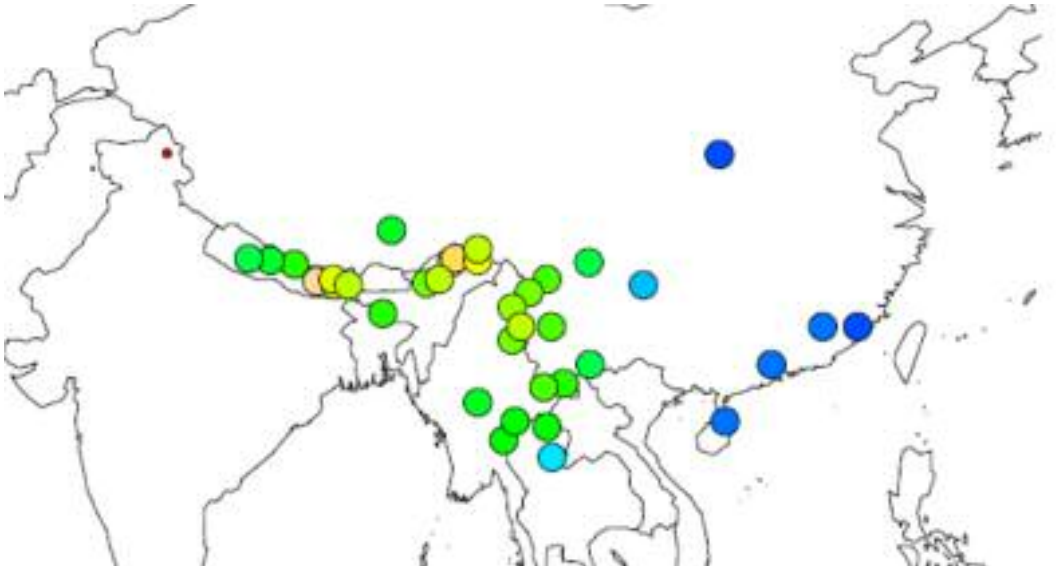




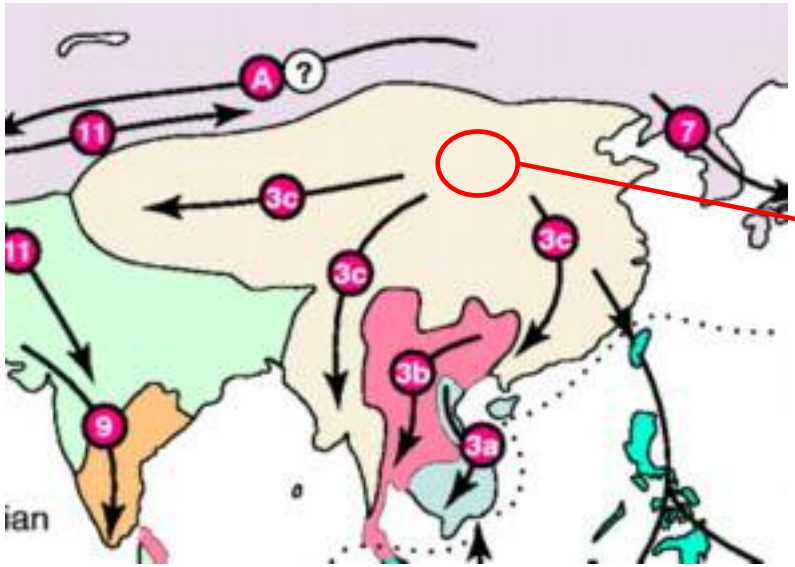
ALTAIC

NIGER-CONGO

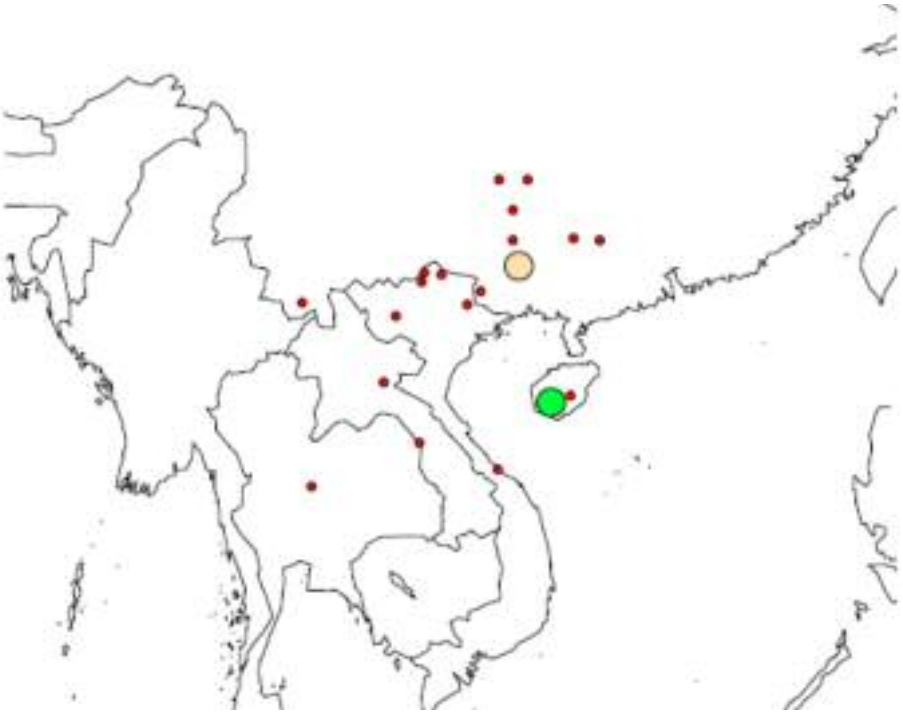




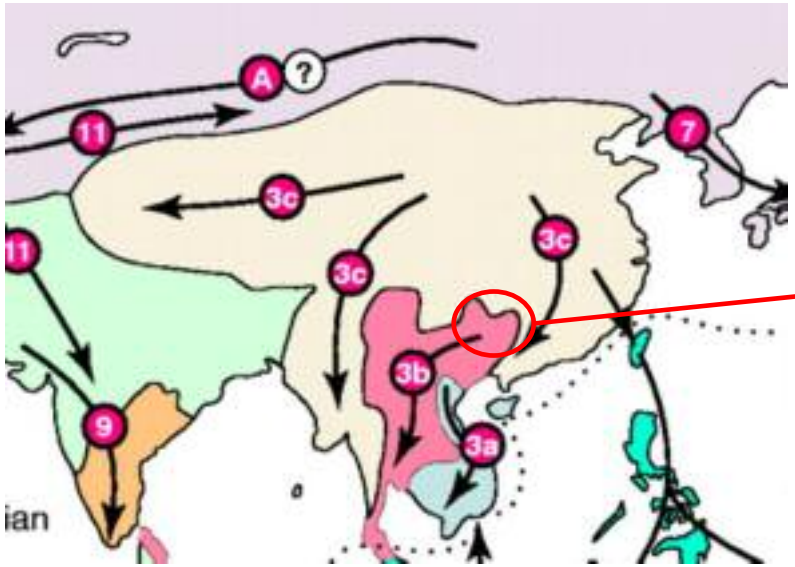
SINO-TIBETAN



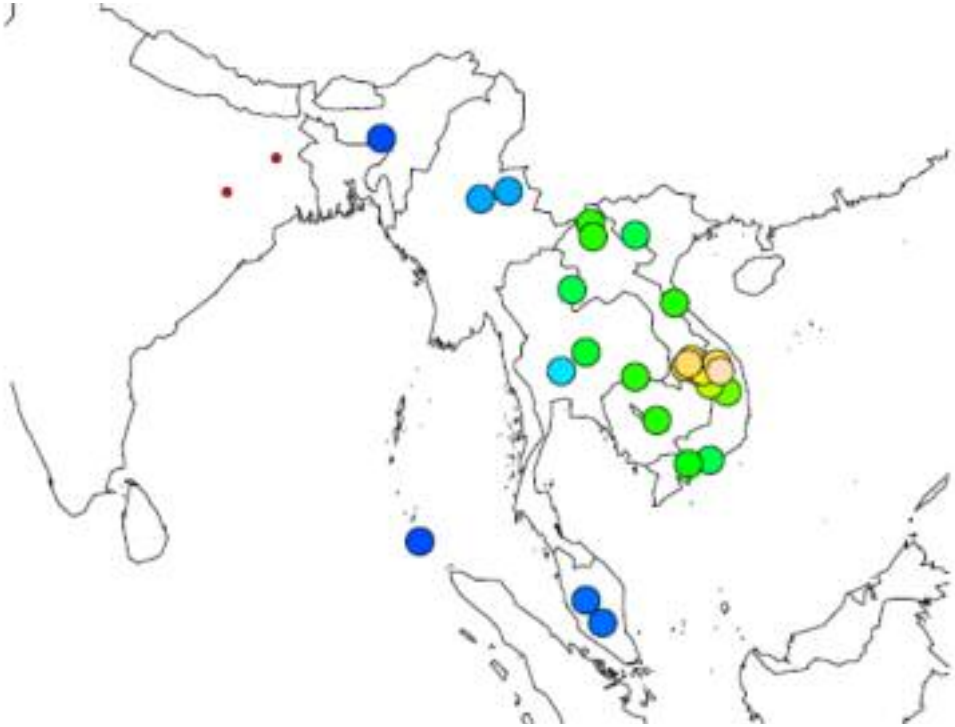

Sino-Tibetan homeland
According to Diamond &
Bellwood (2003)



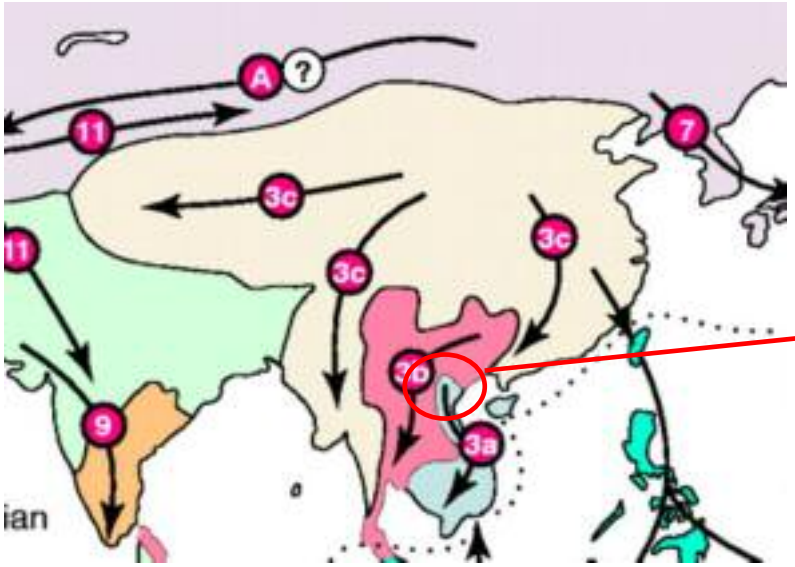
TAI-KADAI




Tai-Kadai homeland according to Diamond & Bellwood (2003)



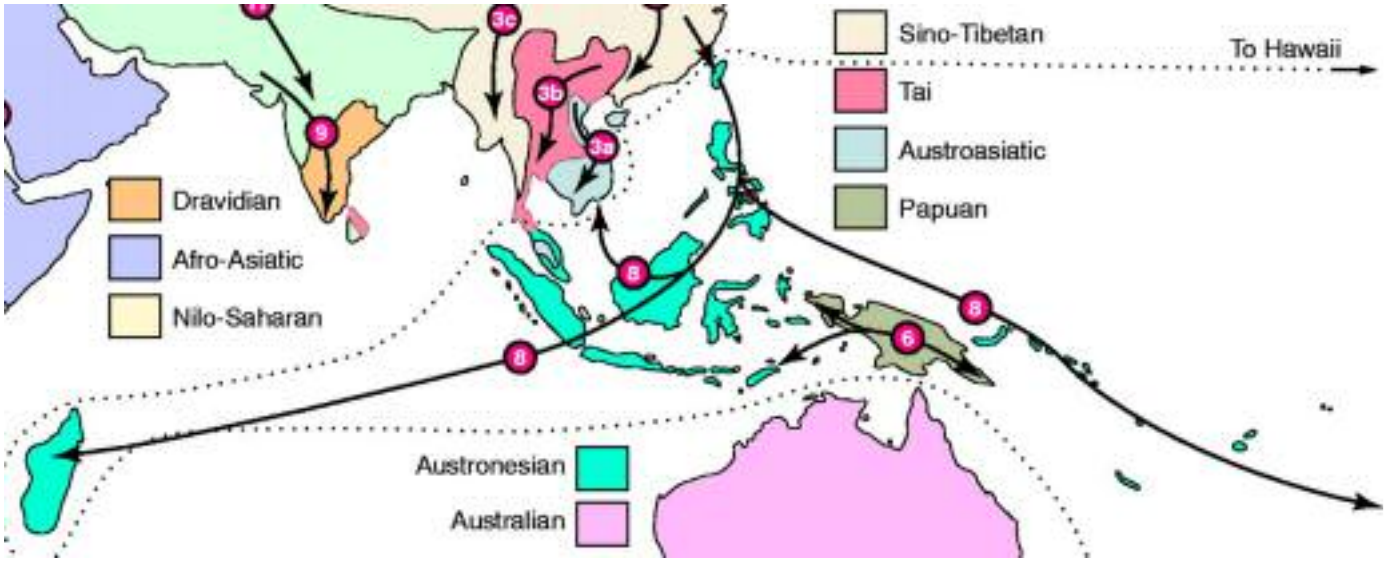
AUSTRO-ASIATIC

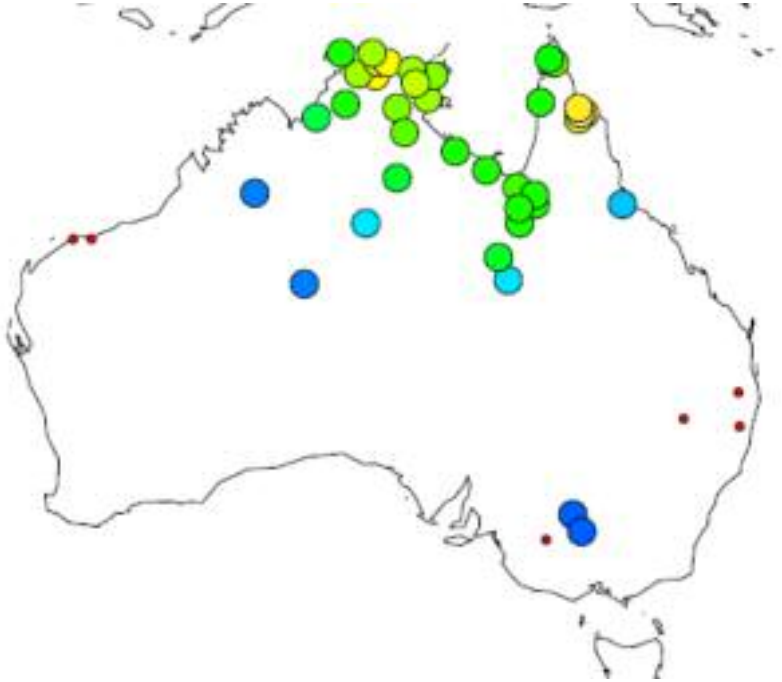



Austro-Asiatic homeland according to Diamond & Bellwood (2003)

AUSTRONESIAN

Austronesian dispersal according to Diamond & Bellwood (2003)

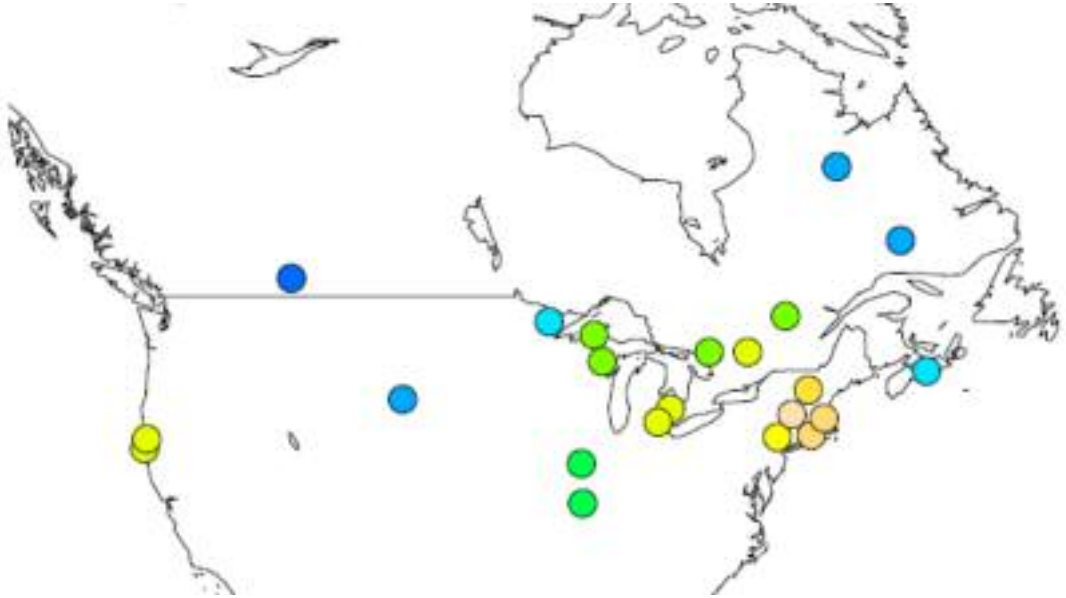




AUSTRALIAN

Nichols (1997: 377):
“Pama-Nyungan originated in the northeast of its range and spread by a combination of language shift and migration (…) (Evans & Jones 1997, McConvell 1996a,b). Northeastern Australia (southern Cape York), the likely Pama-Nyungan homeland, is a long-standing center of technological innovation (Morwood & Hobbs 1995), an area of deep divergence within Pama-Nyungan, and close to the Tangkic family, which represents a likely first sister to Pama-Nyungan (Evans 1995).”

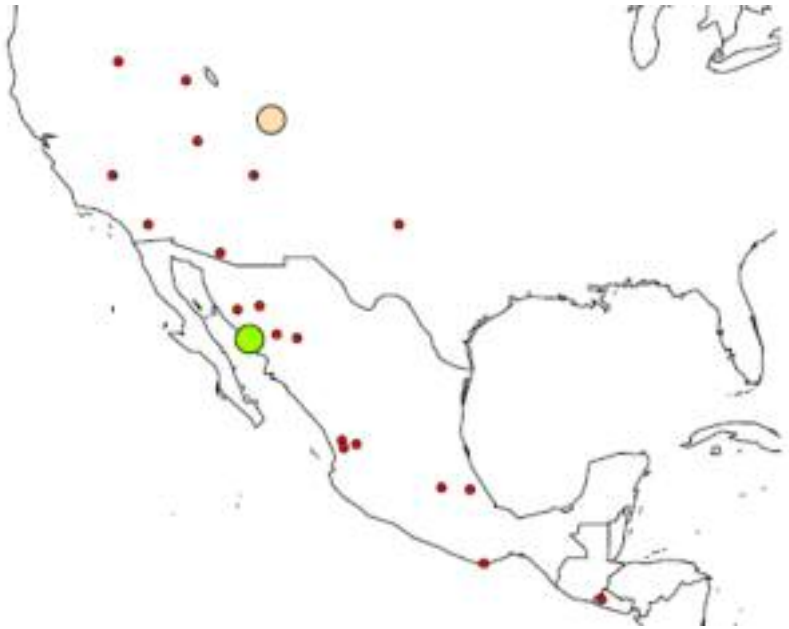
ALGIC



Ruhlen (1994): Proto-Algonkian in the southwest of the family's extent
F. Siebert: PA in the area of the eastern upper Great Lakes (cited without reference by Ruhlen)
Denny (1991): PA around Upper Columbia River in Oregon and Washington

# UTO-AZTECAN

Fowler (1983)

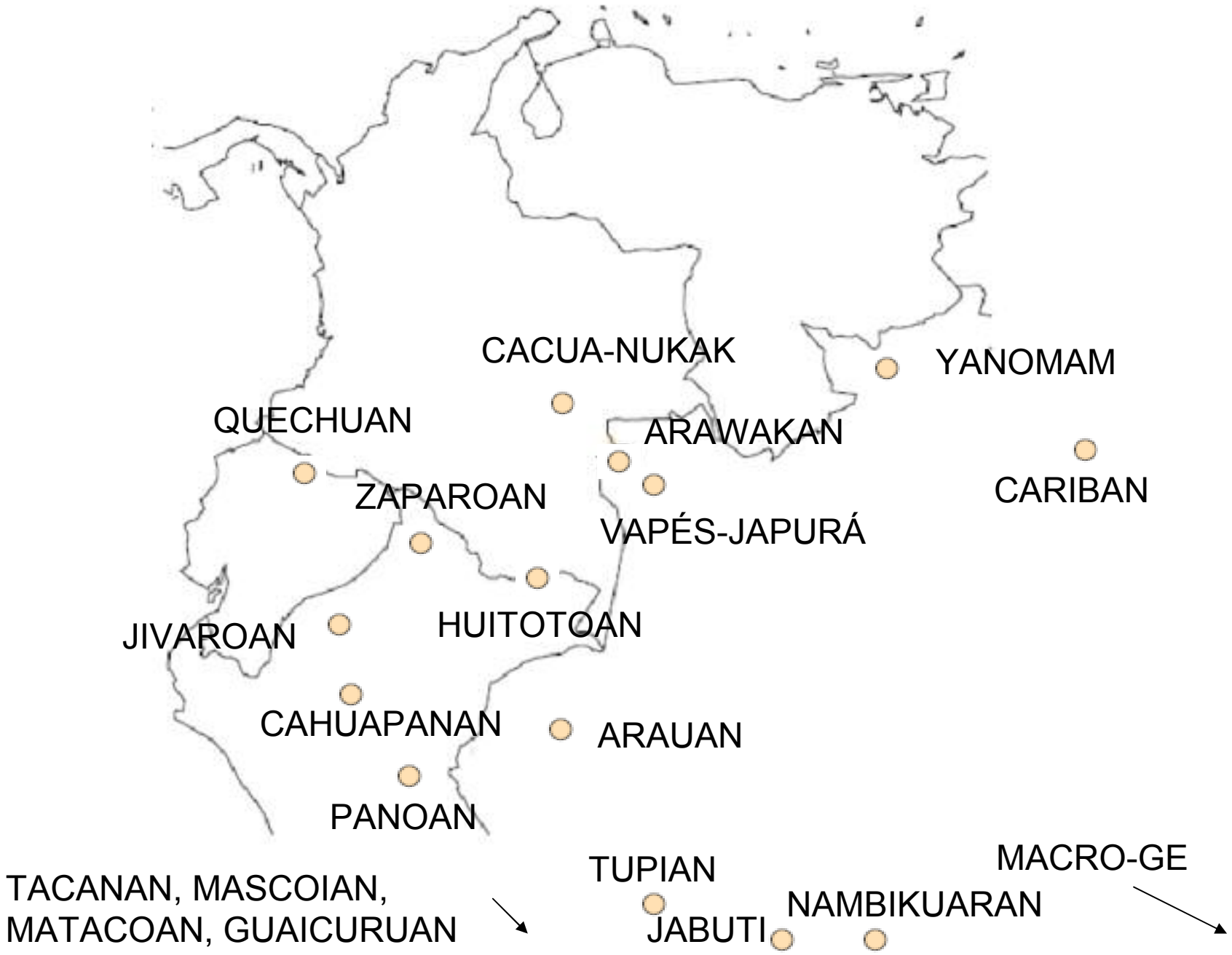


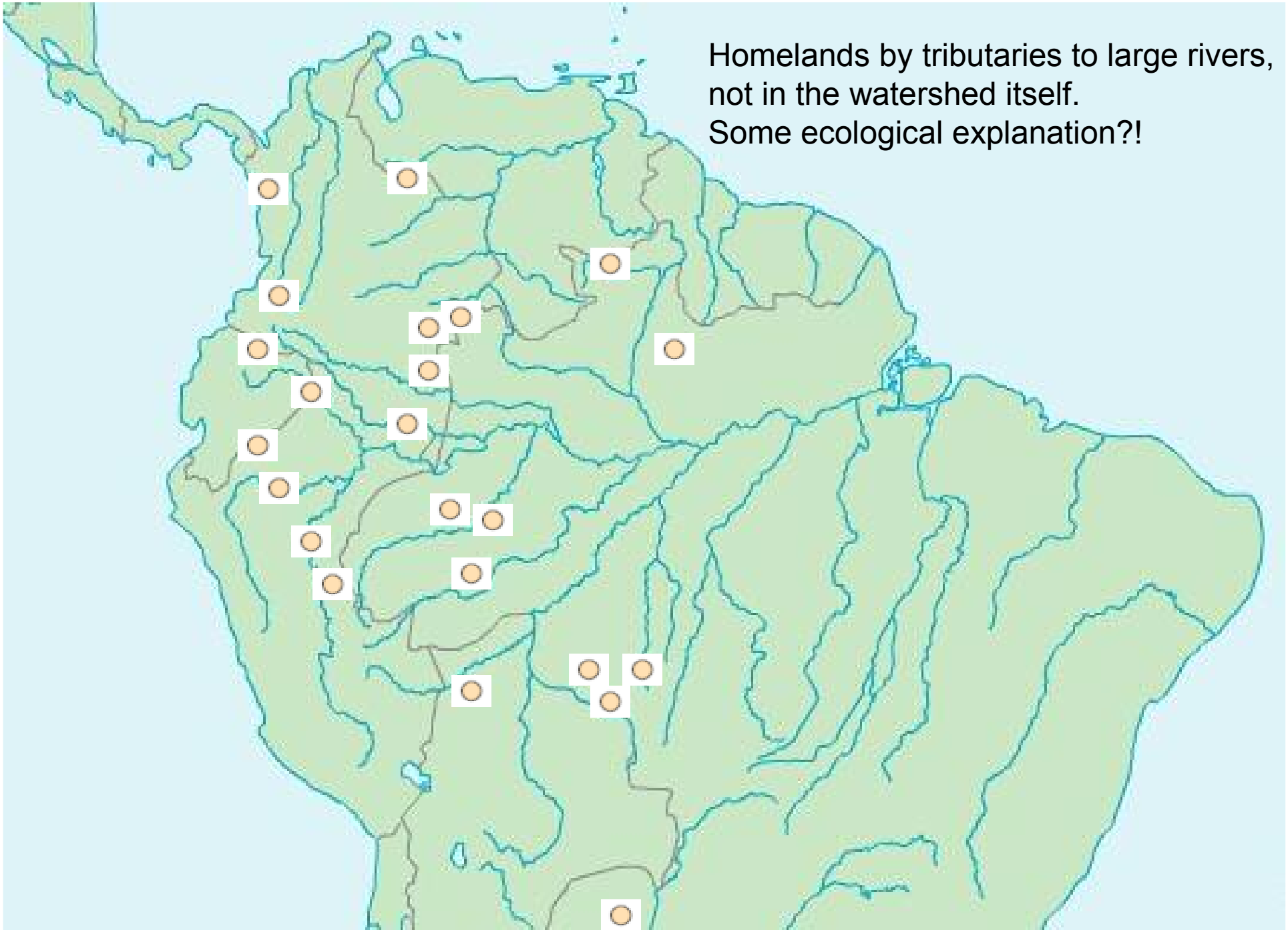
Hopkins (1965): Columbia Plateau
Fowler (1983: New Mexico
Hill (2001): Mesoamerica

CHIBCHAN

TUPIAN

Approximate homeland according to Dall‘Igna Rodrigues (1958), based on the presence
Of nearly all major subgroups of the family.


CACUA-NUKAK
YANOMAM
QUECHUAN
ARAWAKAN
CARIBAN
ZAPAROAN
VAPÉS-JAPURÁ
JIVAROAN
HUITOTOAN
CAHUAPANAN
ARAUAN
PANOAN
TUPIAN
MACRO-GE
TACANAN, MASCOIAN,
MATACOAN, GUAICURUAN
NAMBIKUARAN
JABUTI


Homelands by tributaries to large rivers,
not in the watershed itself.
Some ecological explanation?!


# Thank you for your attention!

Acknowledgment: thanks to Hans-Jörg Bibiko (the one to the right) for implementing the homeland identification procedure in R