

Das *Automated Similarity Judgment Program*

Sebastian Sauppe, MPI für Psycholinguistik





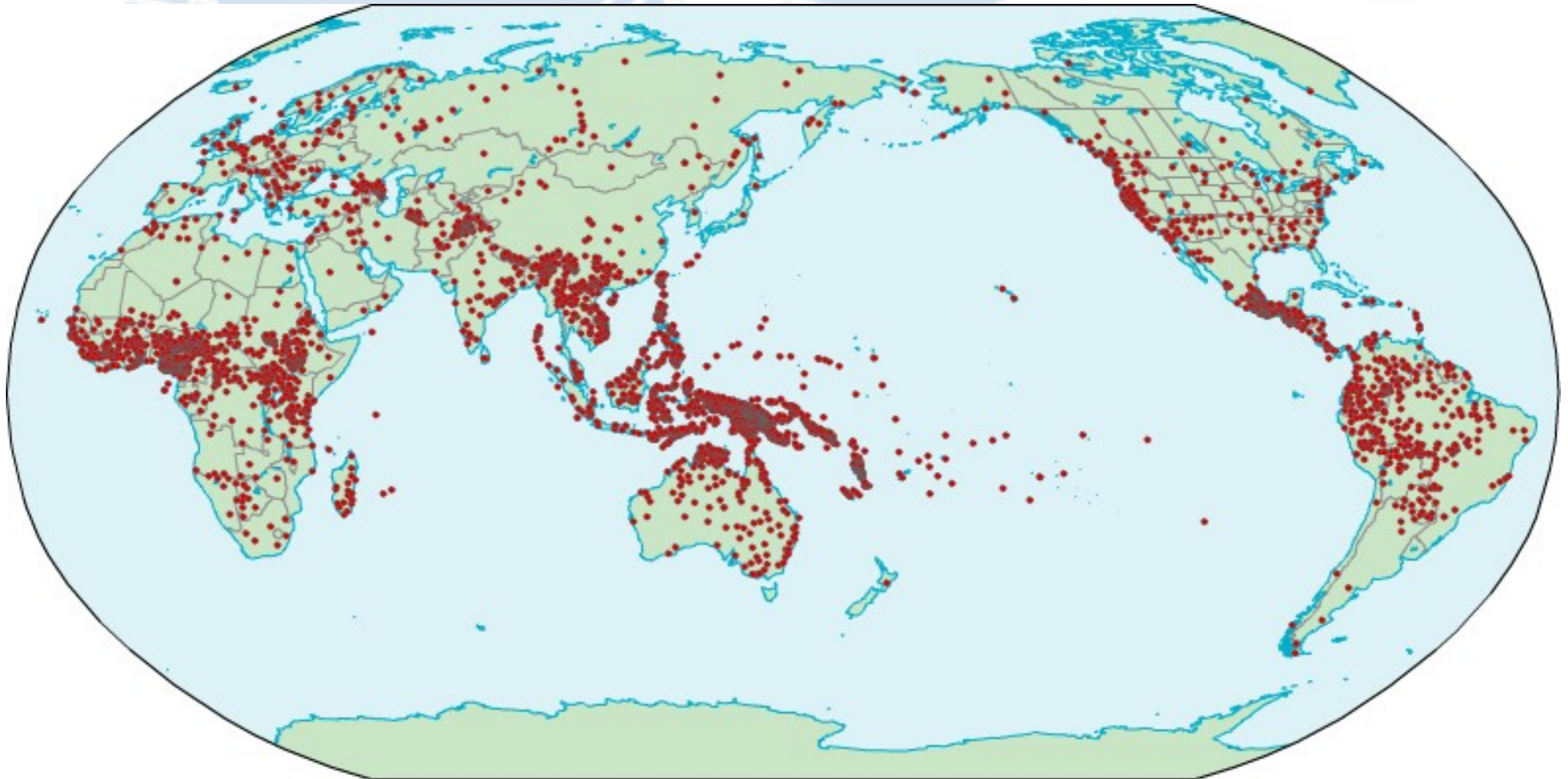
ASJP: Ziele und Methode

- großes Konsortium um Søren Wichmann (MPI EVA, Leipzig), Eric W. Holman (UCLA) und Cecil H. Brown (NIU)
- 2007 Arbeit am Projekt begonnen
- online unter <http://email.eva.mpg.de/~wichmann/ASJPHomePage.htm>
- Ziel: klassische Fragen der sprachlichen Vorgeschichte mit konsistentem, automatisiertem, objektivem und statistisch validierbarem Ansatz angehen



ASJP: Ziele und Methode

- ASJP-Datenbank besteht aus 40-Wort-Listen für derzeit 5732 Sprachen und Dialekte (ca. die Hälfte der Sprachen der Welt abgedeckt)





ASJP: Ziele und Methode

- ASJP-Code

ASJP symbol	Description
i	high front vowel, rounded and unrounded
e	mid front vowel, rounded and unrounded
E	low front vowel, rounded and unrounded
3	high and mid central vowel, rounded and unrounded
a	low central vowel, unrounded
u	high back vowel, rounded and unrounded
o	mid and low back vowel, rounded and unrounded

ASJP symbol	Description
p	voiceless bilabial stop and fricative
b	voiced bilabial stop and fricative
m	bilabial nasal
f	voiceless labiodental fricative
v	voiced labiodental fricative
8	voiceless and voiced dental fricative
4	dental nasal
t	voiceless alveolar stop
d	voiced alveolar stop
s	voiceless alveolar fricative
z	voiced alveolar fricative
c	voiceless and voiced alveolar affricate
n	voiceless and voiced alveolar nasal
S	voiceless postalveolar fricative
Z	voiced postalveolar fricative
C	voiceless palato-alveolar affricate
j	voiced palato-alveolar affricate
T	voiceless and voiced palatal stop
5	palatal nasal
k	voiceless velar stop
g	voiced velar stop
x	voiceless and voiced velar fricative
N	velar nasal
q	voiceless uvular stop
G	voiced uvular stop
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative
7	voiceless glottal stop
h	voiceless and voiced glottal fricative
l	voiced alveolar lateral approximant
L	all other laterals
w	voiced bilabial-velar approximant
y	palatal approximant
r	voiced apico-alveolar trill and all varieties of `r-sounds`
!	all varieties of `click-sounds`



ASJP: Ziele und Methode

- warum ein vereinfachtes Kodierungsschema?
- die meisten Quellen unterdifferenzieren Vokalqualität, Vokallänge oder Ton bzw. benutzen Nicht-IPA-Symbole
- die Benutzung von phonetisch sehr detaillierten Transkriptionen für einige Listen und sehr vereinfachte Transkriptionen für andere, würde die Ergebnisse unnötig verzerren



ASJP: Ziele und Methode

- denn die meisten Quellen zu außereuropäischen Sprachen sehen so aus:

Wörtersammlung
Brasilianischer Sprachen.

Glossaria linguarum Brasiliensium.

Glossarios
de diversas lingoas e dialectos, que fallao os Indios
no imperio do Brazil.

Von
Dr. Carl Friedrich Phil. v. Martius.

Linguae unitas et similitudo firmissimum est
vinculum societatis humanae et religionis.
S. August. de Civ. Dei c. 7.



Leipzig
Friedrich Fleischer
1867.

ygarapé jatimá timá — *rio de muitas voltas*, Fluss mit vielen Windungen.
— mirim — *riacho, regato, ri-beiro*, Bach, Canal.
— reapýra — *cabeceira ou origem do rio*, Quelle, Ursprung eines Flusses.
— remoçape — *boca ou foz do rio*, Mündung eines Flusses.
ygarité — *canoinha*, kleines Fahrzeug.
ygaropába — *porto*, Hafen.
ygatim — *proa da canoa*, Schiff-Schnabel.
ygatiýba — *proeiro da canoa*, Ruderknecht am Vordertheil.
yha — *especie de macaco*, Nyctipithecus.
yiçába — *palavra*, das Wort.
ymirá (imirá, ymyrá, moirá) — *arvore*, Baum, Holz.
ypó (ypú) — *por ventura, na verdade*, vielleicht, in Wahrheit *).
yque (adv. loci) — *aqui*, hier.
ýra — *mel*, Honig.
— máya — *abelha*, Biene (Honigmutter.)

yрати — *abelha cujo mel faz tetano*, Biene, deren Honig Starrkrampf macht.
yraitim — *cera*, Wachs.
— canéa (port. candeia) — *vela de cera*, Wachskerze.
— canéa rendába — *castiçal*, Leuchter.
yrób — *amargar*, bitter seyn.
— oaé marica póra — *colera*, Zorn.
yroiçang — *frescura, viração*, frisches Lüftchen.
yryri — *ostra*, Auster.
— çui † — *cal*, Kalk.
ytá (vide itá) — *pedra, ferro*, Stein, Eisen.
— beraba — *brilhante*, Diamant.
— cepú — *ouro*, Gold, i. e. lapis multi pretii (cepy).
— — mirim — *latão*, Messing.
— jinga (xinga) — *prata*, Silber.
— — cepu mirim — *estanho*, Zinn.
— membeca — *chumbo*, Blei (ferum molle.)
— una anga (unga) — *aço*, Stahl (anima ferri nigri). **)
ytan — *concha*, Muschel.



ASJP: Ziele und Methode

- Levenshtein-Distanz als automatisches Distanzmaß
- minimale Anzahl an Editionsschritten (Substitution, Einsetzung, Löschung), die benötigt wird, um ein Wort in ein anderes zu überführen

deu *Zunge* → eng *tongue*

cuN3

3 Schritte → LD = 3

tuN3 (Substitution)

toN3 (Substitution)

toN (Löschung)



ASJP: Ziele und Methode

- LDN: Levenshtein-Distanz geteilt durch die Anzahl der Segmente des längeren Wortes → korrigiert Unterschiede in der Wortlänge
- ASJP erlaubt bis zu zwei Synonyme pro Eintrag → Mittel zwischen beiden als LDN-Wert
- LDND: LDN-Mittel aller Wörter mit der selben Bedeutung geteilt durch Mittelwert aller Wörter mit unterschiedlicher Bedeutung → korrigiert Ähnlichkeiten aufgrund zufällig gleicher Phoneminventare und sorgt für lexikalisches (statt phonologisches) Maß
- ASJP-Distanz: $1 - \text{LDND}$



ASJP: Ziele und Methode

- Unterschied zu klassischen lexikostatistischen Ansätzen: keine Kognaten-Einschätzungen durch Experten mehr nötig (wie z.B. bei Nakhleh et al. 2005)



ASJP: Ziele und Methode

- ursprünglich wurde mit 100-Wort-Listen gearbeitet
- es hat sich jedoch herausgestellt, dass 40-Wort-Listen, die die historisch stabilsten Items enthalten, genauso gute Ergebnisse ergeben

Brown, C. H., Holman, E. W., Wichmann, S., and Velupillai, V. (2008). Automated classification of the world's languages: A description of the method and preliminary results. *Sprachtypologie und Universalienforschung*, 61(4):285–308.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., and Bakker, D. (2008). Explorations in automated language classification. *Folia Linguistica*, 42(2):331–354.



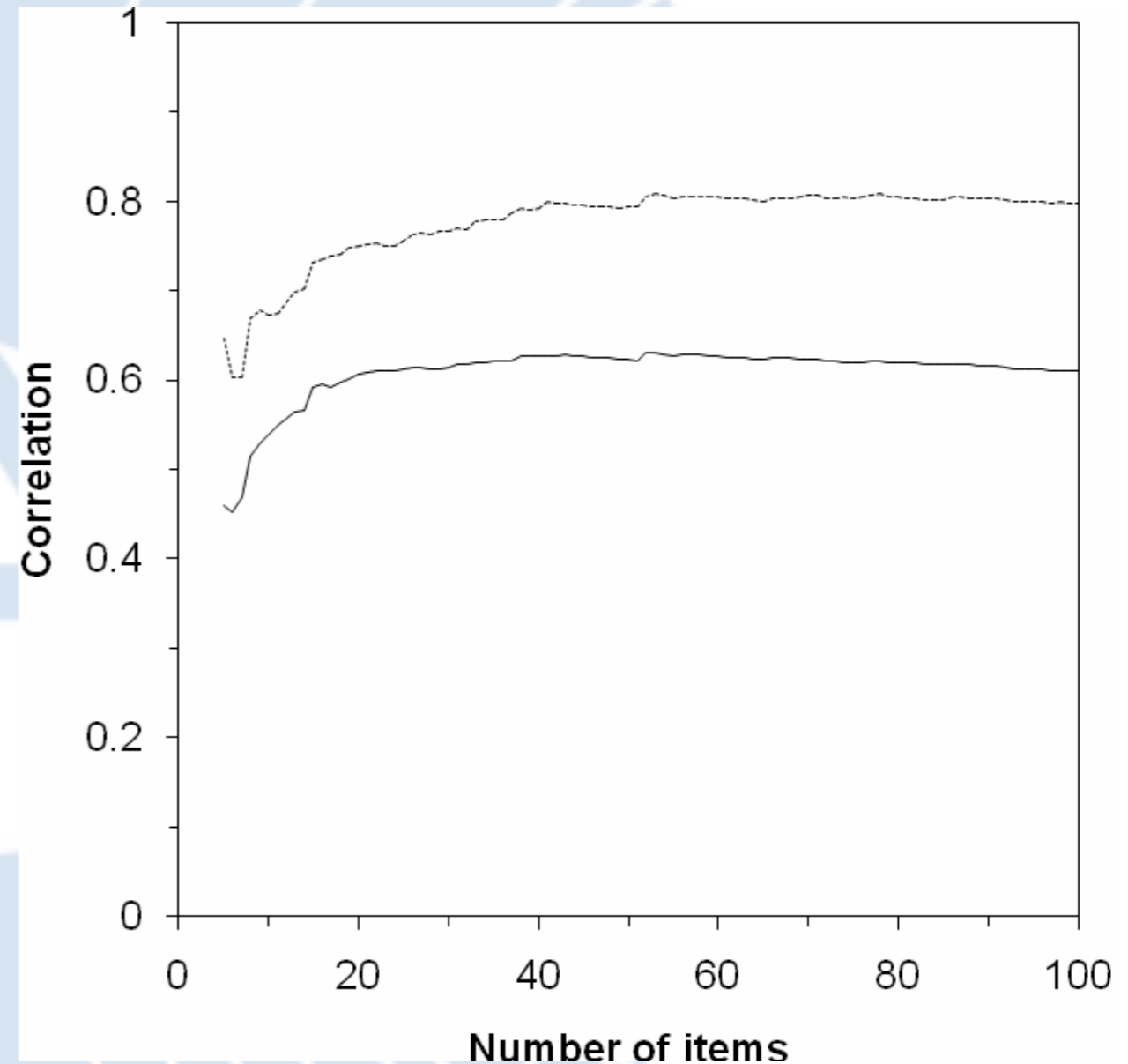
ASJP: Ziele und Methode

- stabile Items:
 - haben geringe Wahrscheinlichkeit, mit der Zeit durch andere Wörter derselben Sprache oder durch Entlehnungen ersetzt zu werden
 - haben größere Wahrscheinlichkeit, Kognaten in eng verwandten Sprachen/Dialekten zu haben
 - lexikalische Ähnlichkeit von Items innerhalb der Genera im WALS bestimmt (ähnliche phonologische Formen für gleiche Konzepte)
 - Daten aus Loanword Typology Project: Entlehnungsrate ca. 8,5%



ASJP: Ziele und Methode

- Größe der Wortlisten:
- Listen zwischen 5 und 100 Items untersucht
- Liste mit 99 Items beinhaltet die 99 stabilsten Items, Liste mit 98 Items die 98 stabilsten Items usw.
- Korrelation der Distanzmatrizen aus den Listen mit genealogischer Klassifikation im WALS und im Ethnologue
- ab Listen mit 40 Items keine signifikante Verbesserung mehr



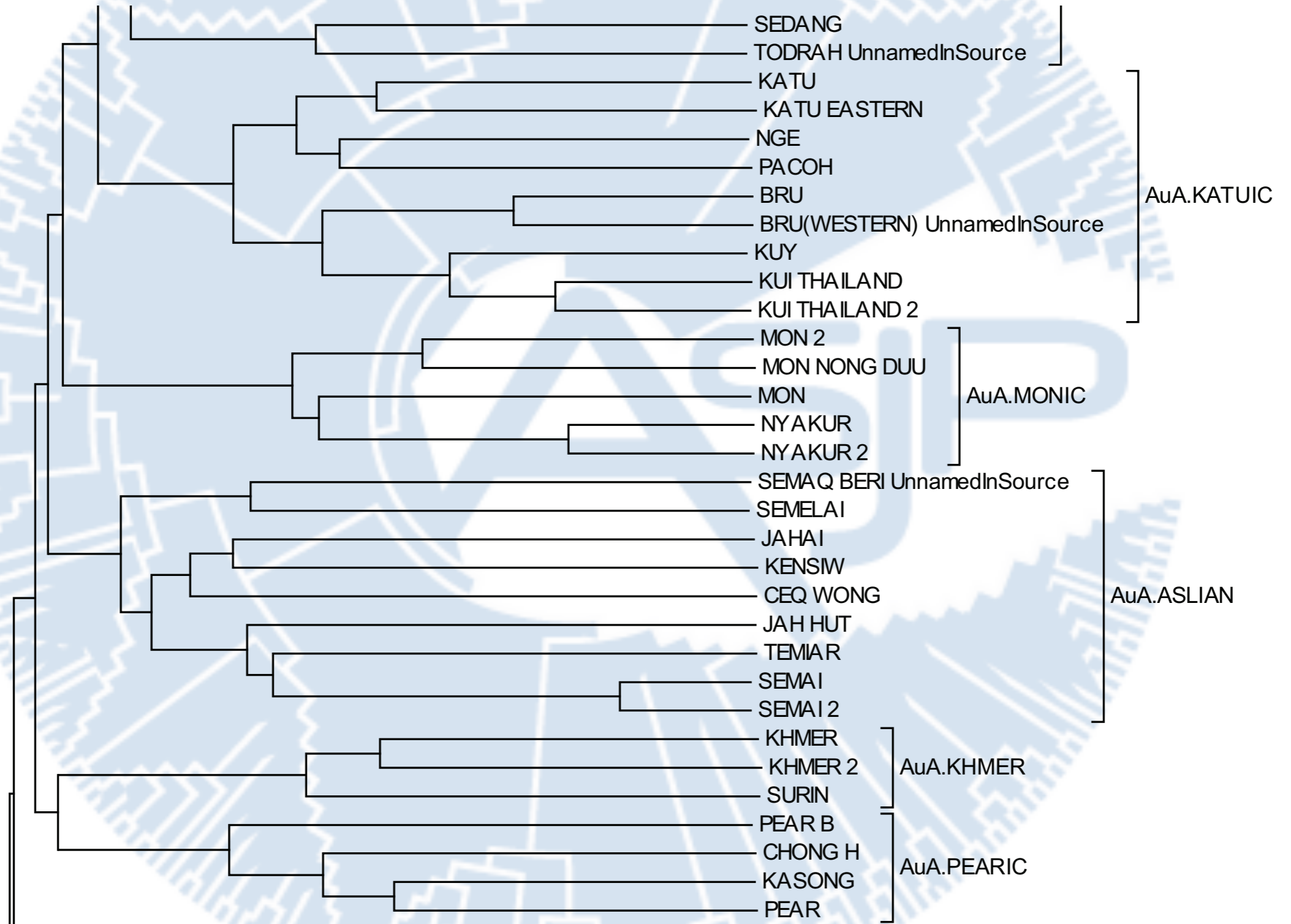


ASJP: Ziele und Methode

<i>Rank</i>	<i># in list</i>	<i>Meaning</i>	<i>Stability</i>	<i>Rank</i>	<i># in list</i>	<i>Meaning</i>	<i>Stability</i>
1	22	*louse	42.8	30	11	*one	27.4
2	12	*two	39.8	31	41	*nose	27.3
3	75	*water	37.4	32	95	*full	26.9
4	39	*ear	37.2	33	66	*come	26.8
5	61	*die	36.3	34	74	*star	26.6
6	1	*I	35.9	35	86	*mountain	26.2
7	53	*liver	35.7	36	82	*fire	25.7
8	40	*eye	35.4	37	3	*we	25.4
9	48	*hand	34.9	38	54	*drink	25.0
10	58	*hear	33.8	39	57	*see	24.7
11	23	*tree	33.6	40	27	bark	24.5
12	19	*fish	33.4	41	96	*new	24.3
13	100	*name	32.4	42	21	*dog	24.2
14	77	*stone	32.1	43	72	*sun	24.2
15	43	*tooth	30.7	44	64	fly	24.1
16	51	*breasts	30.7	45	32	grease	23.4
17	2	*you	30.6	46	73	moon	23.4
18	85	*path	30.2	47	70	give	23.3
19	31	*bone	30.1	48	52	heart	23.2
20	44	*tongue	30.1	49	36	feather	23.1
21	28	*skin	29.6	50	90	white	22.7
22	92	*night	29.6	51	89	yellow	22.5
23	25	*leaf	29.4	52	20	bird	21.8
24	76	rain	29.3	53	38	head	21.7
25	62	kill	29.2	54	79	earth	21.7
26	30	*blood	29.0	55	46	foot	21.6
27	34	*horn	28.8	56	91	black	21.6
28	18	*person	28.7	57	42	mouth	21.5
29	47	*knee	28.0	58	88	green	21.1



World Tree of Lexical Similarity





Fallstudie: Urheimaten

- Center of Gravity-Annahme: Region mit der größten Diversität ist wahrscheinlich Urheimat
- wenn Sprache sich in Tochtersprachen aufspaltet, bleiben sie wahrscheinlich näher zusammen
- setzt voraus, dass Ausbreitung von Sprachfamilien eher einem langsamen Spaziergang als einem zielgerichtetem Lauf ähnelt
- aber: meiste Sprachfamilien erstrecken sich über ein kontinuierliches Gebiet, Annahme scheint also gerechtfertigt



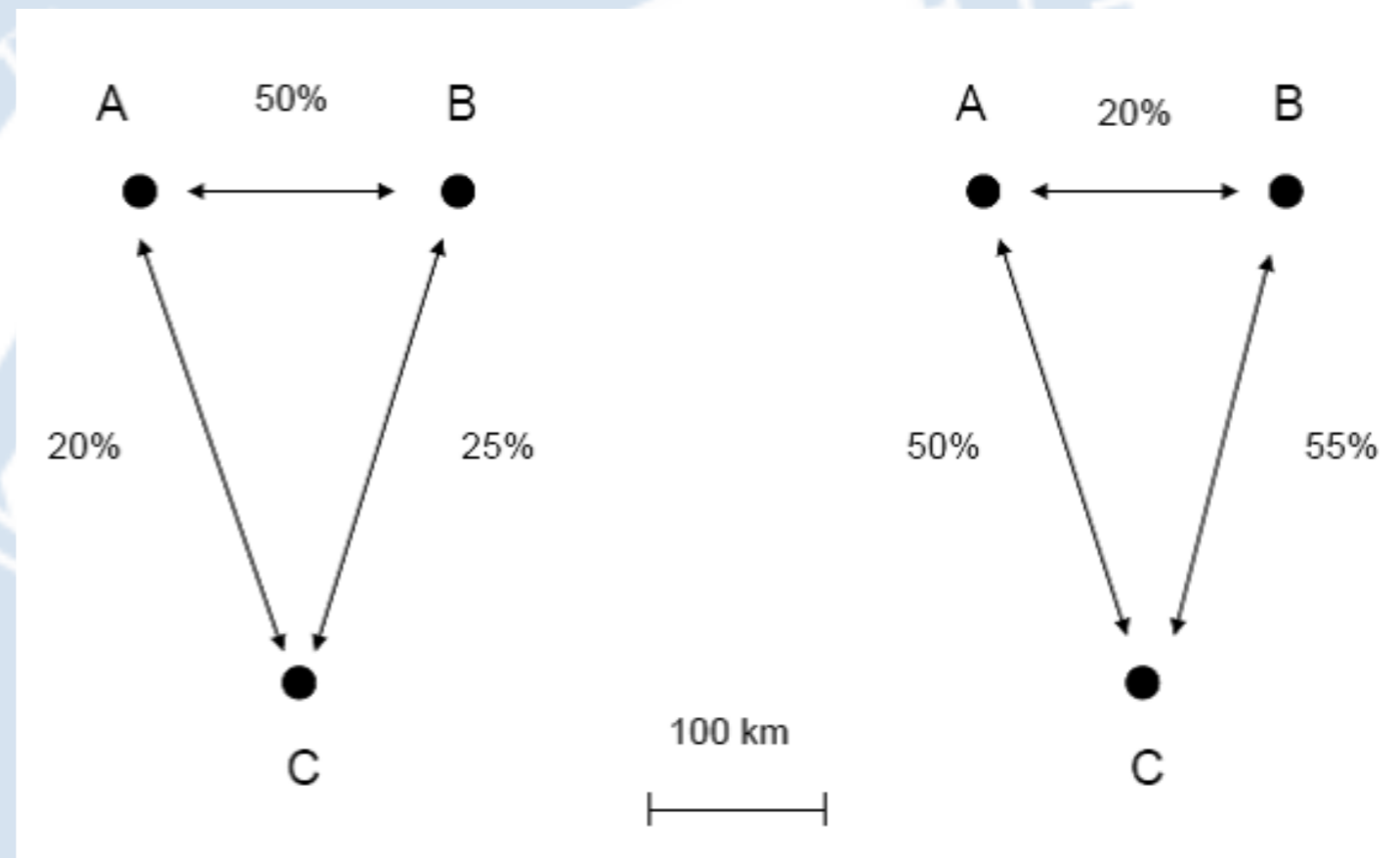
Fallstudie: Urheimaten

- **Areale Diversität:**
- Sprachen sind sehr ähnlich, liegen aber weit auseinander → geringe Diversität
- Sprachen sind sehr unterschiedlich und liegen nahe bei einander → hohe Diversität



Fallstudie: Urheimaten

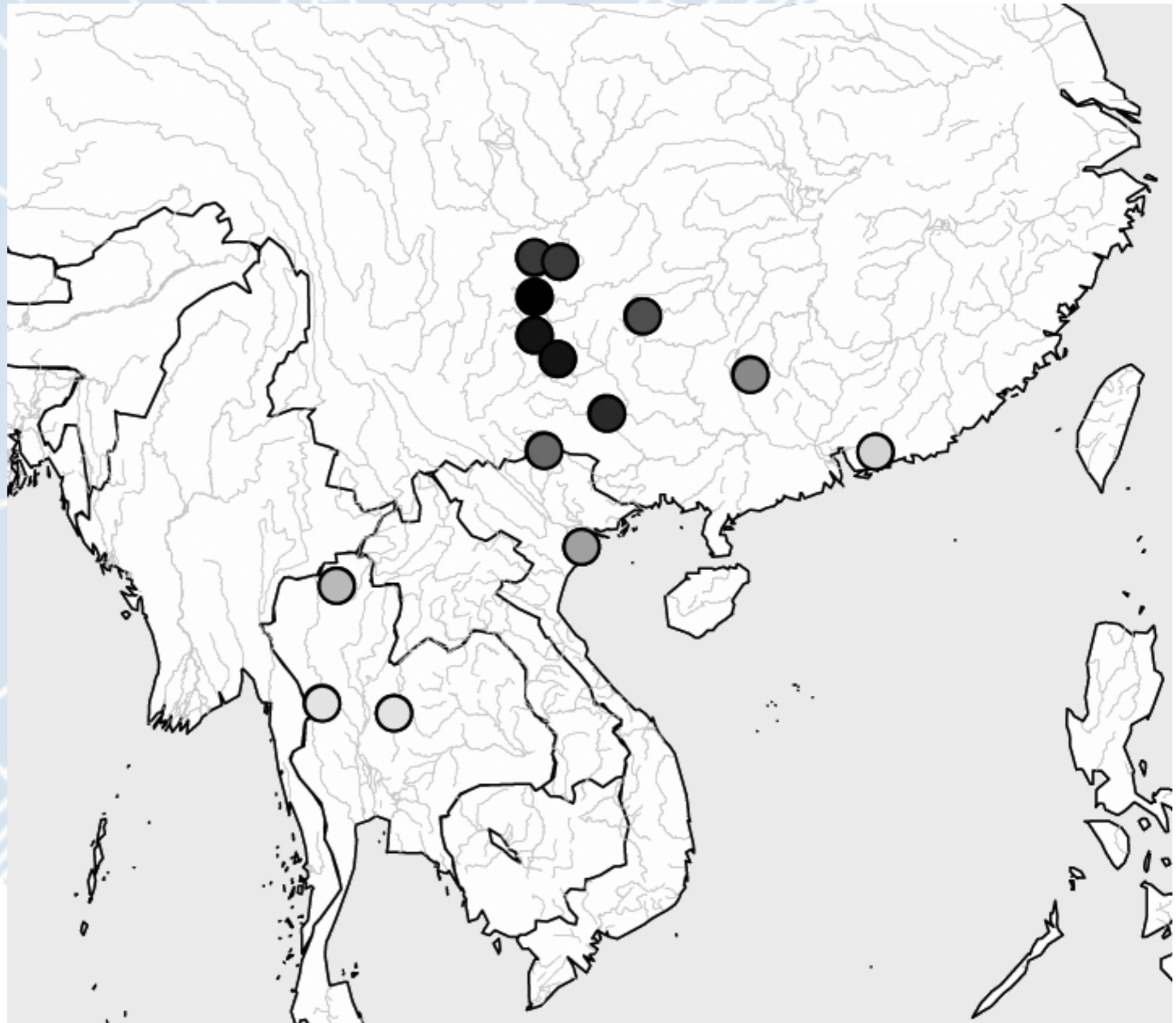
- areale Diversität =
lexikalische Distanz /
geographische Distanz
- um Urheimat
bestimmen zu können,
werden mindestens
drei Sprachen benötigt
- $D = \text{Mittelwert aus } l/g$
für jedes Sprachenpaar





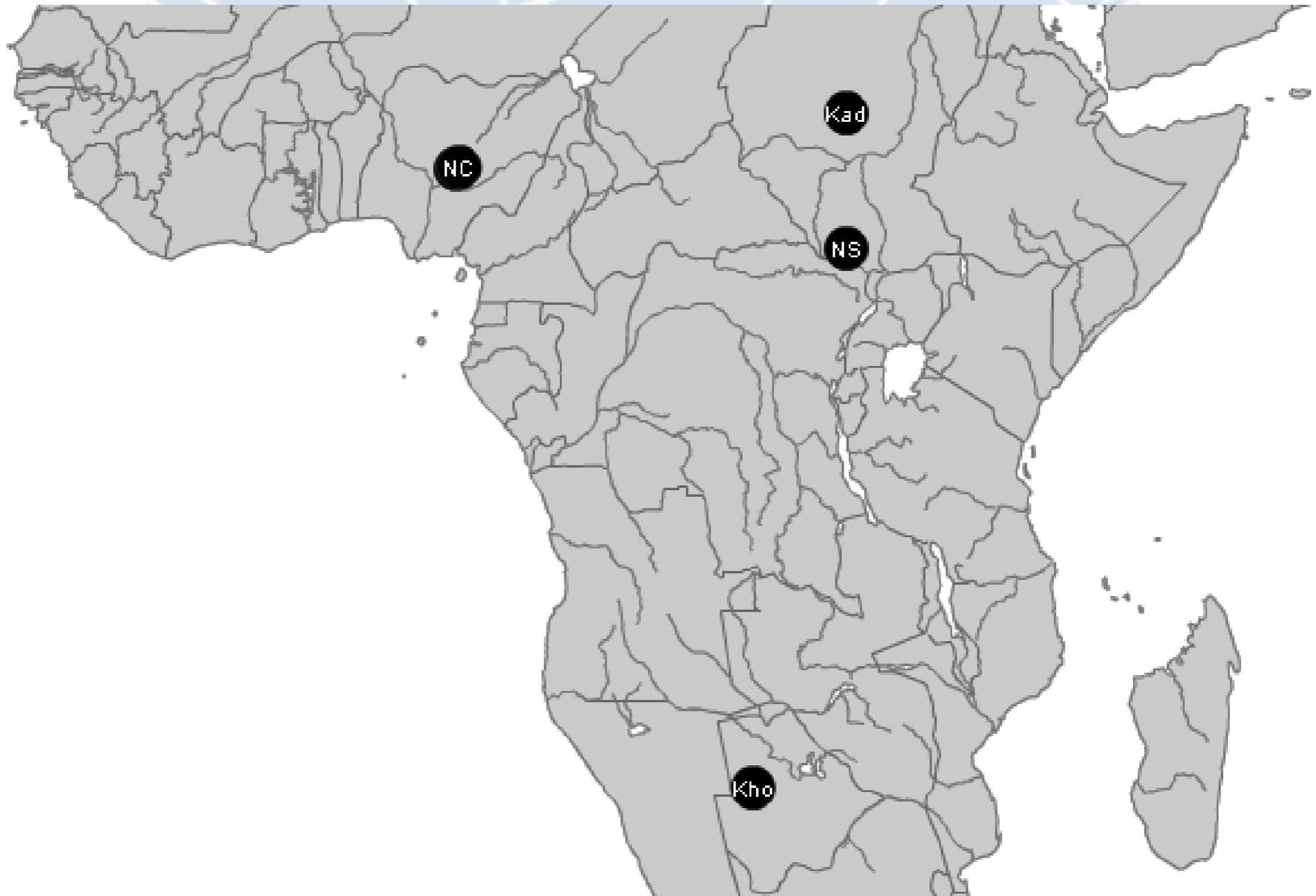
Fallstudie: Urheimaten

- Hmong-Mien



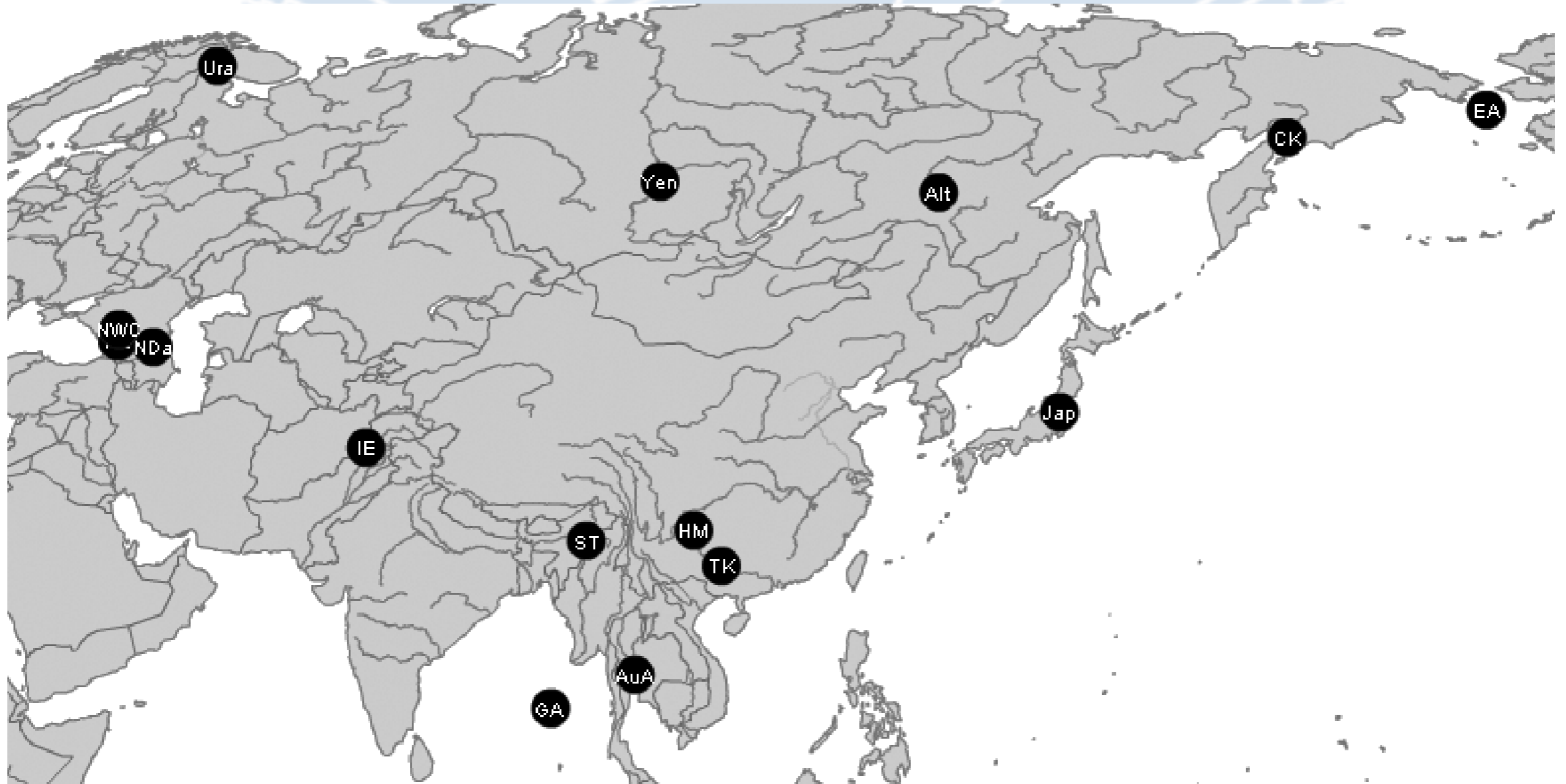


Fallstudie: Urheimaten





Fallstudie: Urheimaten





Fallstudie: Urheimaten

- Methode produziert Ergebnisse, die meistens plausibel sind
- automatische Natur der Methode umgeht ad-hoc-Einschätzungen und ermöglicht es, Urheimaten aller Sprachfamilien komparativ zu betrachten
- aber: konvergierende Evidenz lässt die besten Schlüsse zu





Fallstudie: Datierung von Ursprachen

- Grundannahme: Je unterschiedlicher zwei Sprachen einer Familie sind, desto mehr Zeit ist seit ihrer Aufspaltung vergangen (geht zurück auf Sapir und Swadesh)
- viele Diskussion um Vor- und Nachteile der Glottochronologie in den letzten Jahrzehnten → soll hier nicht fortgesetzt werden, sondern neue, objektivere Methode vorgestellt werden



Fallstudie: Datierung von Ursprachen

- Vorteil des ASJP-Ansatzes: keine historische Analyse für Kognaten-Einschätzungen nötig (einfach für gut untersuchte Familien, sehr schwierig für die meisten anderen)
- lexikalische Ähnlichkeiten/Unterschiede können entstehen durch
 - Lautwandel
 - Ersetzung
 - Entlehnung
- um Datierung zuverlässig zu machen, müssen sich verschiedene Prozesse ausgleichen um eine konstante Wandelrate zu ergeben



Fallstudie: Datierung von Ursprachen

- mit ASJP-Datenbank ist Test von Datierung mit konstanter Wandelrate möglich
- Ergebnisse mithilfe historischer, archäologischer und epigraphischer Evidenz kalibriert → gibt Maß für Genauigkeit der Ergebnisse und hilft einzuschätzen, wie sie verwendet werden können



Fallstudie: Datierung von Ursprachen

- Berechnung der Zeit seit der Aufspaltung an Formel von Swadesh angelehnt:

$$t = \frac{\log s - \log s_0}{2 \log r}$$

- t = Zeittiefe, s = ASJP-Distanz, r = Retentionsrate
- s_0 und r sind Konstanten → mithilfe von Kalibrierungsdaten ermittelt ($s_0 = 0.92$, $r = 0.72$)



Fallstudie: Datierung von Ursprachen

- Kalibrierung anhand von 52 archäologisch, historisch oder epigraphisch ermittelten Daten für Familien, deren Sprachen in der ASJP-Datenbank enthalten sind

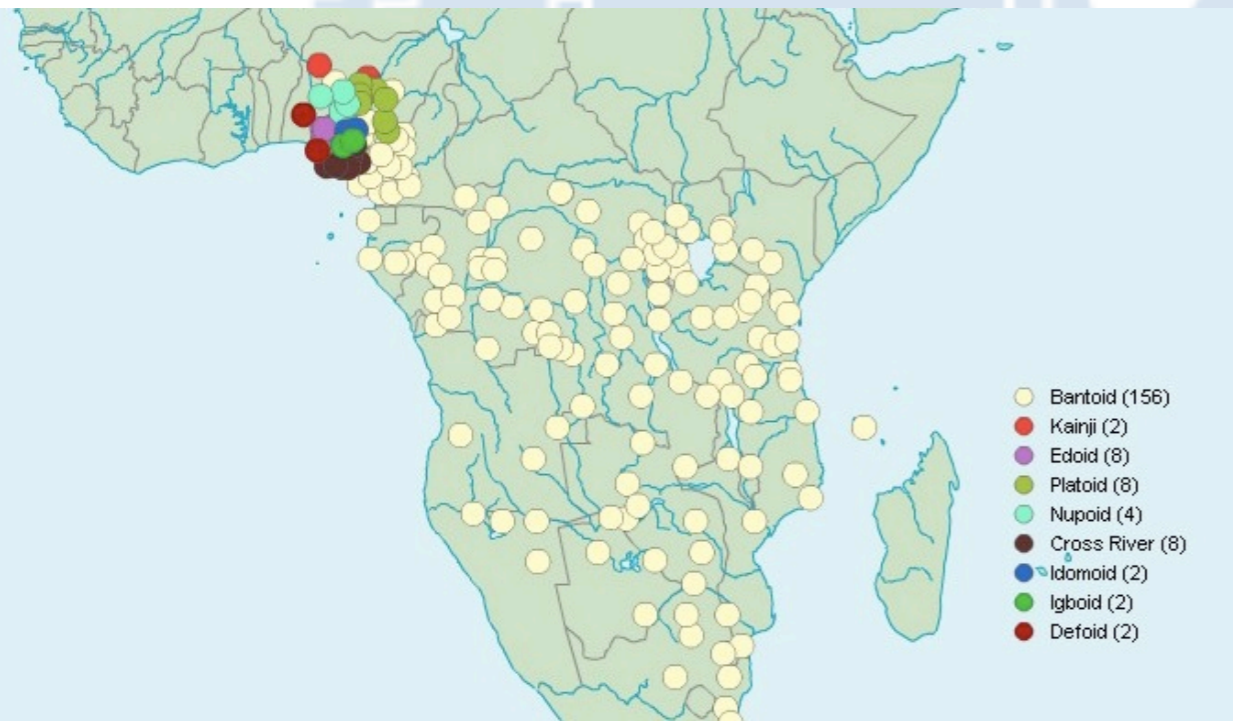
Benue-Congo

Date: 6500 (A).

Source: Bostoen and Grégoire (2007:77) link the introduction, during 7000–6000 BP, of new technologies such as macrolithic tools and pottery into the Grassfields region with the break off of Bantoid from Benue-Congo. They argue that this would fit the hypothesis that the center of dispersion of Benue-Congo is near the confluence of the Niger and Benue rivers, and they also mention that pottery-related terminology can be reconstructed to Proto-Benue-Congo.

Similarity: 3.58.

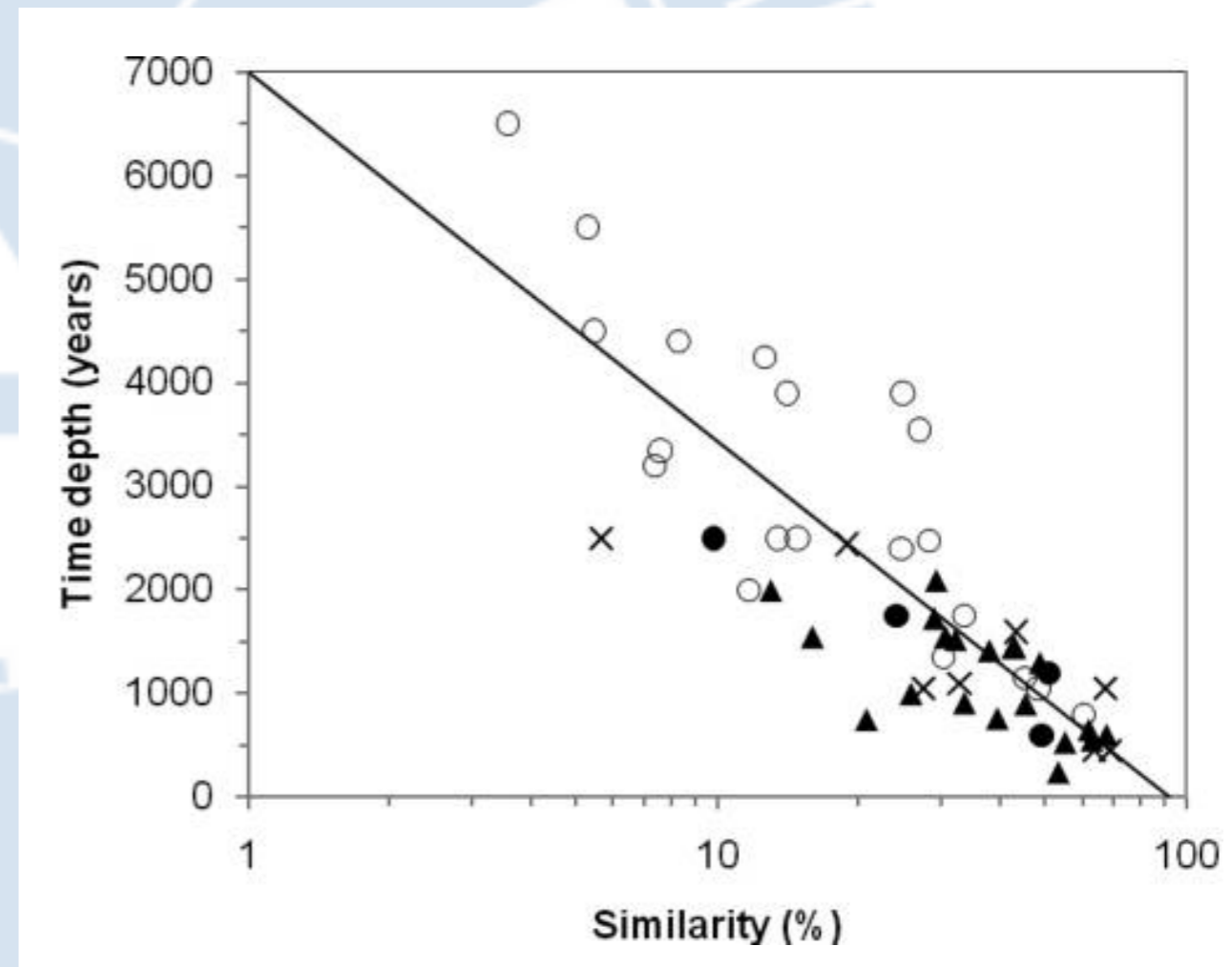
Comparisons: Akpes (1), Bantoid (258), Cross River (28), Defoid (4), Edoid (27), Idomoid (3), Igboid (5), Jukunoid (2), Kainji (20), Nupoid (4), Oko (1), Plateau (45), Ukaan (6); 46,303 pairs.





Fallstudie: Datierung von Ursprachen

- Ähnlichkeitsmaß:
- durchschnittliche lexikalische Distanz zwischen Sprachen aus verschiedenen Genera
- Klassifikation der Genera basierend auf Ethnologue
- nimmt mit zunehmender Zeittiefe linear ab





Fallstudie: Datierung von Ursprachen

- durchschnittliche Diskrepanz = 29% (größer für ältere Daten, kleiner für jüngere Daten)
- ergibt sich ev. aus unterschiedlichen Wandelraten und v.a. aus Unsicherheiten bei den Kalibrierungsdaten → eher obere Grenze

Language group	Calibration date	ASJP date	Difference (%)
Archaeological:			
Benue-Congo	6500	4940	-1,560 (-32)
Indo-European	5500	4348	-1,152 (-26)
Pama-Nyungan	4500	4295	-205 (-5)
Indo-Iranian	4400	3665	-735 (-20)
Malayo-Polynesian	4250	3024	-1,226 (-41)
Indo-Aryan (Indic)	3900	1996	-1,904 (-95)
Iranian	3900	2856	-1,044 (-37)
Dardic	3550	1868	-1,682 (-90)
Eastern Malayo-Polynesian	3350	3803	+453 (+12)
Temotu	3200	3844	+644 (+17)
Southern Nilotic	2500	2928	+428 (+15)
Wakashan	2500	2781	+281 (+10)
Mississippi Valley Siouan	2475	1798	-677 (-38)
Malayo-Chamic	2400	2003	-397 (-20)
Central Southern African Khoisan	2000	3143	+1,143 (+36)
Saami	1750	1532	-218 (-14)
Ma'anyan-Malagasy	1350	1690	+340 (+20)
Ongamo-Maa	1150	1083	-67 (-6)
East Polynesian	1050	979	-71 (-7)
Inuit	800	640	-160 (-25)



Fallstudie: Datierung von Ursprachen

- Methode kann nun auch für Sprachfamilien benutzt werden, für die es keine Kalibrierungsdaten gibt

Austronesian	19,212	10 (11)	8.46	3633
Atayalic	1	2 (2)	15.98	2664
East Formosan	5	3 (3)	19.11	2392
Malayo-Polynesian	268,972	16 (17)	12.62	3024
Celebic	814	4 (4)	28.27	1796
Eastern	43	2 (2)	29.92	1710
Kaili-Pamona	2	2 (2)	45.36	1076
Tomini-Tolitoli	20	2 (2)	35.07	1468
Central-Eastern	51,447	3 (4)	11.92	3111
Central Malayo-Polynesian	4,562	9 (10)	18.82	2415
Eastern Malayo-Polynesian	18,404	2 (2)	7.56	3803
Greater Barito	1,012	4 (4)	24.23	2031
East	156	3 (3)	26.73	1881
Sama-Bajaw	21	2 (2)	34.59	1489
West	15	2 (2)	45.03	1087
Javanese	2	2 (5)	63.45	566
Lampung	164	3 (3)	54.92	785
Land Dayak	3	3 (5)	34.11	1510
Malayo-Sumbawan	65	3 (3)	27.37	1845
North and East	221	3 (3)	26.44	1898
North Borneo	80	3 (5)	24.46	2016
Melanau-Kajang	1	2 (2)	37.34	1372
North Sarawakan	33	4 (5)	22.08	2172
Sabahan	7	3 (3)	38.32	1333
Northwest Sumatra–Barrier Islands	5	3 (5)	27.79	1822
Philippine	7,587	8 (10)	27.65	1830
Bashiic	9	2 (2)	57.43	717



<fin daNk>