Wie die Bioinformatik hilft, Sprachgeschichte zu rekonstruieren
Das Automated Similarity Judgment Program

Gerhard Jäger

8. Februar 2013
Forum Scientiarum

History of the ASJP-Project[1]

- Jan. 2007
  - Cecil Brown (US linguistic anthropologist) comes up with idea of comparing languages automatically and communicates this to
  - Eric Holman (US statistician) and Wichmann. 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 Wichmann what the two of them have come up with and I begin to take more active part, adding ideas.
- August 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.

[1] cf. Wichmann 2009

History of the ASJP-Project[2]

- September 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.
- October 2007
  - Hagen Jung (computer scientist, MPI, makes a preliminary online implementation).
  - Wichmann takes over the "administration" of the project.
  - A second paper is finished about stabilities of lexical items, defining a shorter Swadesh list, etc.
- November 2007
  - Robert Mailhammer (linguist, BRD) joins
- December 2007
  - Anthony Grant (linguist, GB) joins.
  - Dmitry Egorov (linguist, Kazan) joins.
  - Levenshtein distances are implemented instead of old "matching rules" identifying cognates.

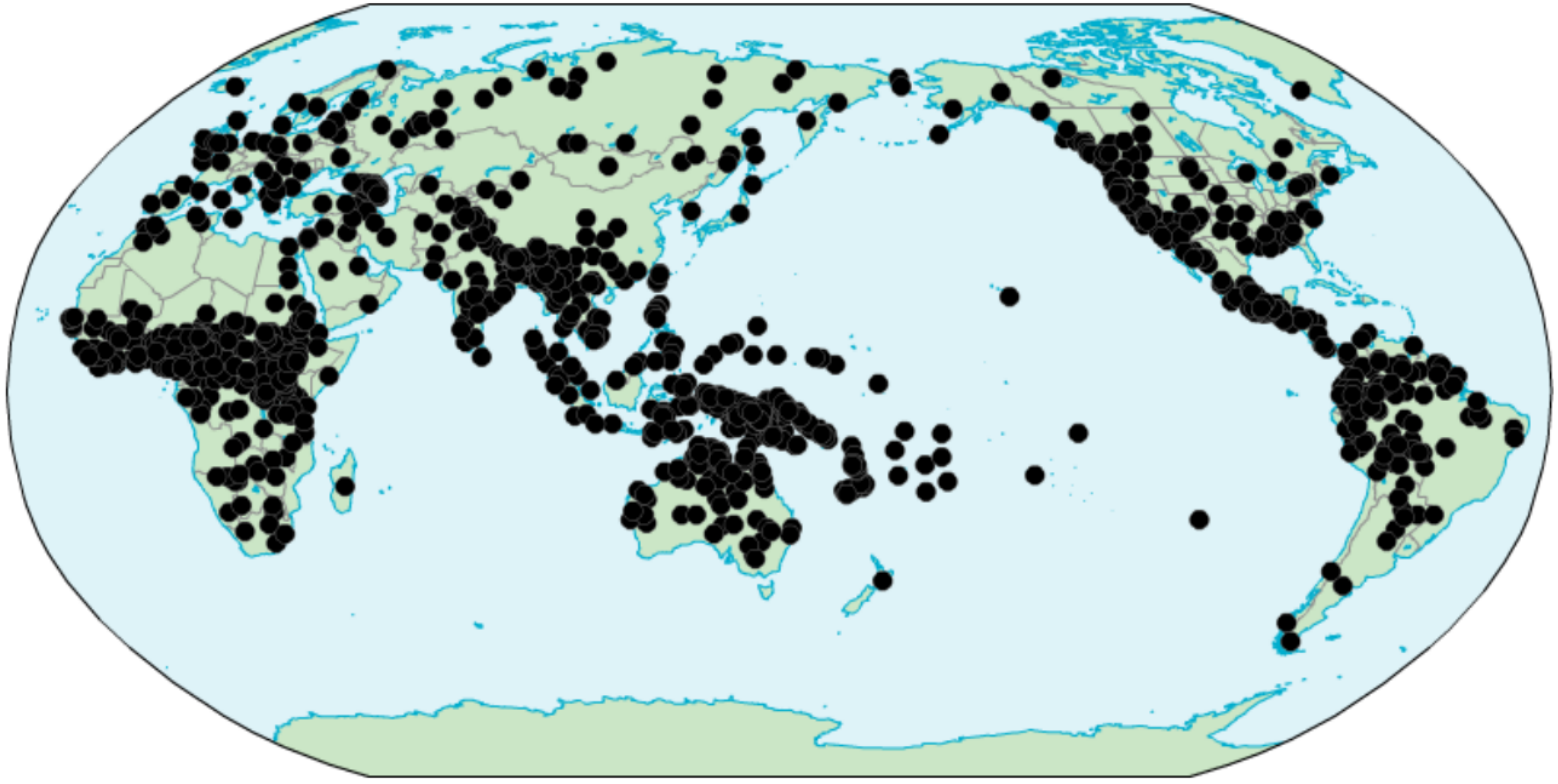
[2] cf. Wichmann 2009

History of the ASJP-Project[3]

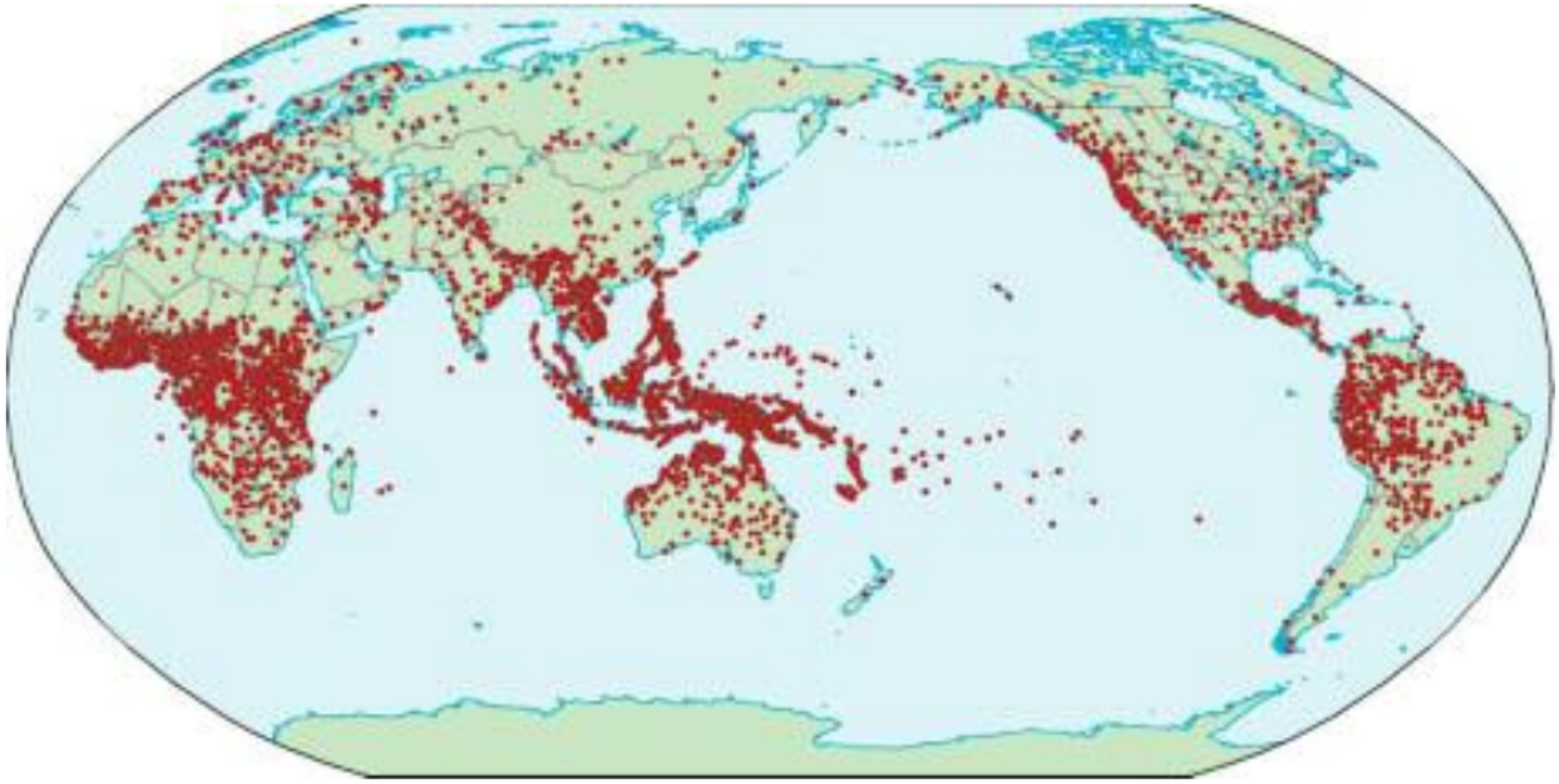
- January 2008
  - Kofi Yakpo (linguist) joins.
- February 2008
  - The two papers are accepted for publication without revision (in respectively Sprachtypologie und Universalienforschung and Folia Linguistica).
- April 2008
  - Oleg Belyaev (linguist, Moscow) joins.
- 2008
  - Papers presented at conferences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

[3] cf. Wichmann 2009

The ASJP data



All languages of the world



The ASJP data

- current version (v. 15)
  - 5,844 languages
  - includes
    - artificial languages: *Esperanto*, *Klingon*, *Volapük*, ...
    - creoles
    - 72 reconstructed languages: *Proto-Indoeuropean*, *Proto-Austronesian*, *Proto-Mayan*, ...
    - extinct languages: *Latin*, *Hittite*, *Gothic*, *Sanskrit*, *Old Norse*, *Coptic*, ...
    - more than 5,500 living languages

The ASJP data

- started with 100 item Swadesh lists, later reduced to 40 items
- 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*

Transcription

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

Transcription

Another transcription example: Abaza (Northwest Caucasian)

18 person	ʕʷɨtʃʼʲʷʕʷɨs	Xw~3Cw"y$Xw~3s
19 fish	pslatʃʷa	pslaCw~a
21 dog	la	la
22 louse	ʦʼa	c"a
23 tree	ʦʼla	c"la
25 leaf	bɣʲɨ	bxy~3
28 skin	tʃʷazʲ	Cw~azy~
30 blood	ʃa	Sy~a
31 bone	bʕʷɨ	bXw~3
34 horn	tʃʼʷɨʕʷa	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	ʃamqa	Sy~amqa

Transcription

ASJPcode symbol	*Description*	*IPA symbols*
p	voiceless bilabial stop and fricative	p, ɸ
b	voiced bilabial stop and fricative	b, β
f	voiceless labiodental fricative	f
v	voiced labiodental fricative	v
m	bilabial nasal	m
w	voiced bilabial-velar approximant	w
8	voiceless and voiced dental fricative	θ, ð
4	dental nasal	n̪
t	voiceless alveolar stop	t̪
d	voiced alveolar stop	d
s	voiceless alveolar fricative	s
z	voiced alveolar fricative	z
c	voiceless and voiced alveolar affricate	ʦ, ʣ
n	alveolar nasal	n
r	voiced apico-alveolar flap and all other varieties of "r-sounds"	ɾ, r, ʀ, ɽ
l	voiced alveolar lateral approximant	l

Transcription

S	voiceless post-alveolar fricative	ʃ
Z	voiced post-alveolar fricative	ʒ
C	voiceless palato-alveolar affricate	ʧ
j	voiced palato-alveolar affricate	ʤ
T	voiceless and voiced palatal stop	c, ɟ
5	palatal nasal	ɲ
y	palatal approximant	j
k	voiceless velar stop	k
g	voiced velar stop	g
x	voiceless and voiced velar fricative	x, ɣ
N	velar nasal	ŋ
q	voiceless uvular stop	q
G	voiced uvular stop	ɢ
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative	χ, ʁ, ħ, ʕ
h	voiceless and voiced glottal fricative	h, ɦ
7	voiceless glottal stop	ʔ
L	all other laterals	ʟ, ɭ, ʎ
!	all varieties of “click-sounds”	!, ǀ, ǁ, ǂ

Transcription

Table 2. ASJPcode vowel symbols and their IPA values.

ASJPcode symbol	*Description*	*IPA symbols*
i	high front vowel, rounded and unrounded	i, ɪ, y, ʏ
e	mid front vowel, rounded and unrounded	e, ø
E	low front vowel, rounded and unrounded	æ, ɛ, œ, ɶ
3	high and mid central vowel, rounded and unrounded	ɨ, ɘ, ə, ɜ, ʉ, ɵ, ɞ
a	low central vowel, unrounded	a, ɐ
u	high back vowel, rounded and unrounded	ɯ, u
o	mid and low back vowel, rounded and unrounded	ɤ, ʌ, ɑ, o, ɔ, ɒ

Transcription

ASJPcode also uses compound symbols, combining basic symbols. For example, it provides a means for transcribing consonant aspiration so that it is recognized as such in automation. In original sources, aspiration is often represented by a consonant followed by a raised *h*, as in t^h and p^h. In ASJPcode, these are transcribed by *th* and *ph*, respectively, followed by the symbol ~ to flag such clusters as being equivalent to single symbols. Typically, compound symbols involve labialization (e.g., k^w transcribed as *kw~*), palatalization (e.g., t^j transcribed as *ty~*), and prenasalization (e.g., *ŋg* transcribed as *Ng~*) in addition to aspiration. There are also other compound symbols involving less common modifications. In fact, ASJPcode provides for compound symbols consisting of as many as three symbols, e.g., *hkw*, which is a pre-aspirated, labialized voiceless velar stop. When followed by the modifier $, a sequence of three symbols is treated as a single symbol.

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)

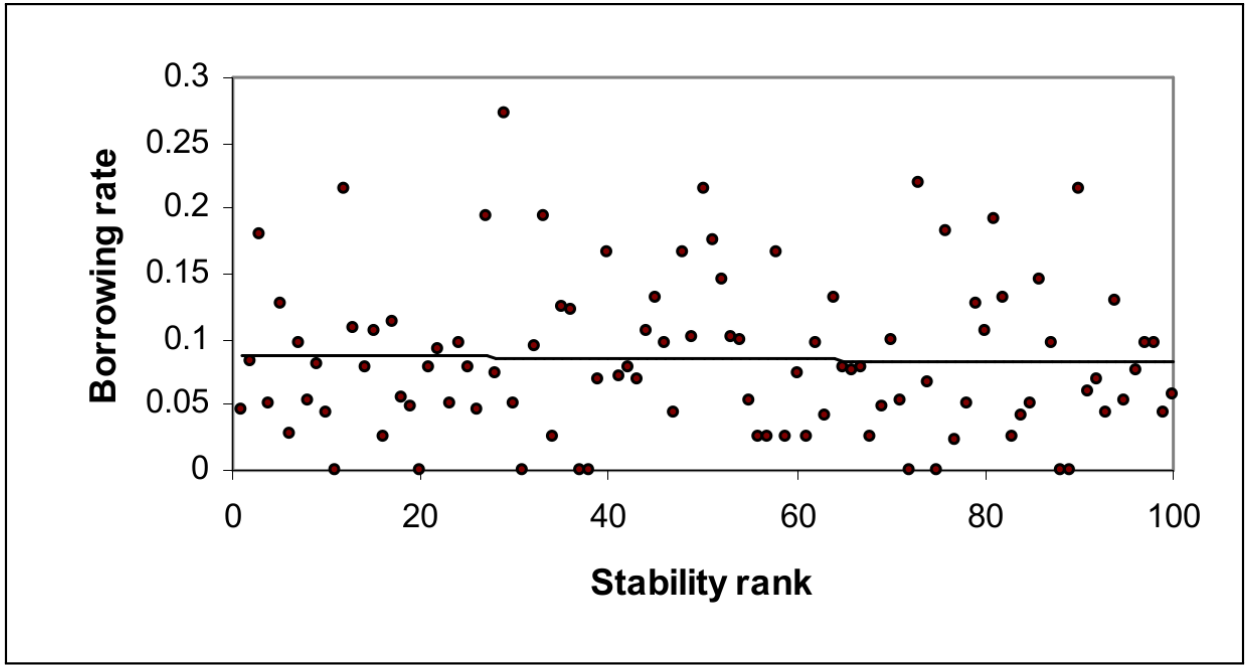
Items on the Swadesh 100-item list in order of descending stability

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

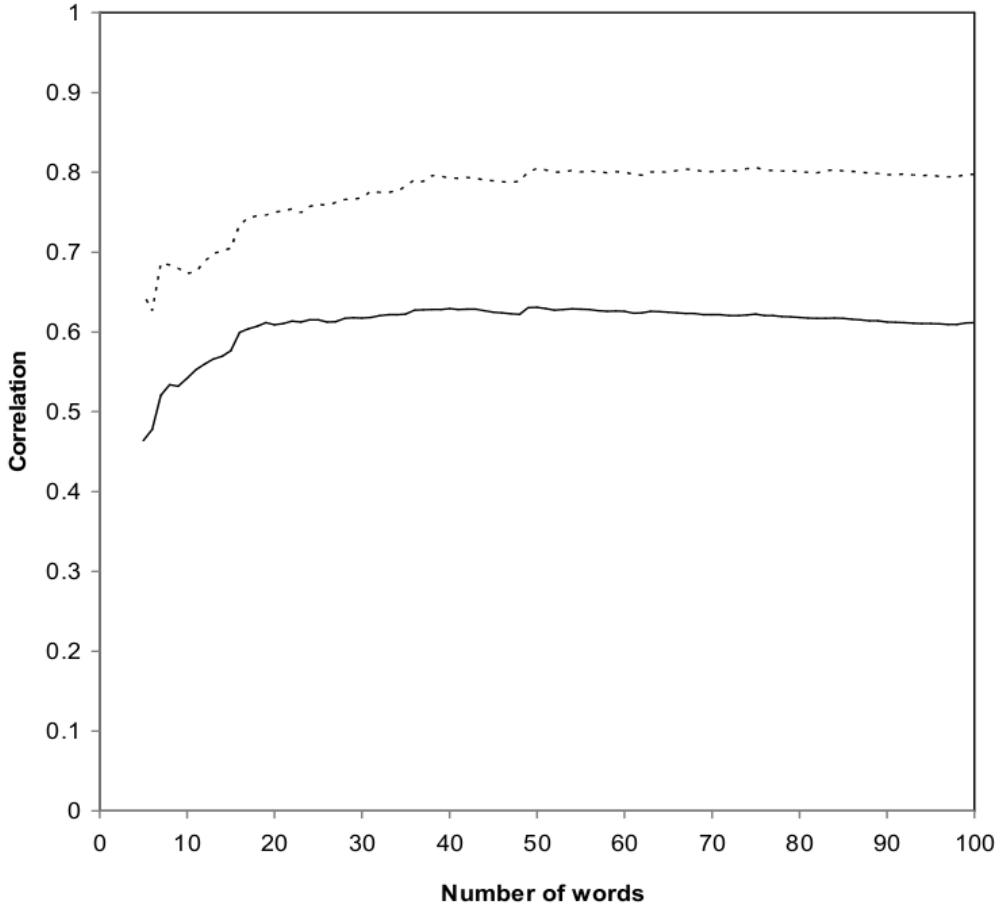


Potential explanations

- 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 the authors had only 36 language at their 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

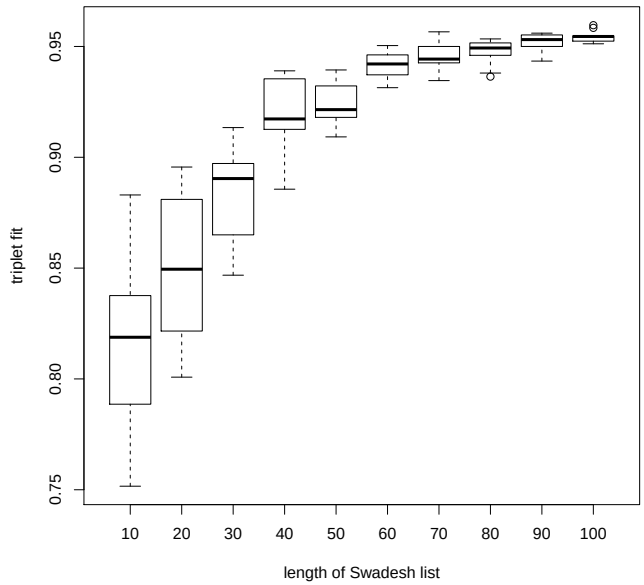


Ethnologue
(Goodman-Kruskal gamma)

WALS/Dryer
(Pearson product-moment correlation)

Selecting a shorter list

- GJ: phylogenetic inference **does** get better if you increase the number of Swadesh items



- Note: More data is better than less data.